Date: Friday, May 25, 2012

Instructions

- Answer only **two** questions out of Section L. If you answer more than two questions, then only the **FIRST TWO questions will be marked.**
- Answer only **two** questions out of Section G. If you answer more than two questions, then only the **FIRST TWO questions will be marked.**

<table>
<thead>
<tr>
<th>Questions</th>
<th>Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td></td>
</tr>
<tr>
<td>G2</td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td></td>
</tr>
</tbody>
</table>

This exam comprises the cover page and four pages of questions.
Section L (Linear Regression Models)
Answer only two questions out of L1–L3

L1.

The data in this question concerns the prices and other price-related variables for 24 houses sold in Erie, Pennsylvania. The variables in the data are:

\[ y: \text{Sale price of the house}/1000. \]
\[ x_1: \text{Taxes (local, school, county)}/1000. \]
\[ x_2: \text{Number of baths}. \]
\[ x_3: \text{Lot size (sq ft \times 1000)}. \]
\[ x_4: \text{Living space (sq ft \times 1000)}. \]
\[ x_5: \text{Number of garage stalls}. \]
\[ x_6: \text{Number of rooms}. \]
\[ x_7: \text{Number of bedrooms}. \]
\[ x_8: \text{Age of the house (years)}. \]
\[ x_9: \text{Number of fireplaces}. \]

The objective is to study the potential effect of the regressors \( x_1-x_9 \) on the sale house price variable, \( y \), through a multiple linear regression model.

(a) Figure 1 shows the scatter plots of the response variable \( y \) versus each regressor \( x_j \).

(i) Comment on the potential relationship between \( y \) and each \( x_j \).

\( (The \ correlation \ matrix \ of \ all \ the \ variables \ is \ given \ below \ Figure \ 1.) \)

2 MARKS

(ii) In general, do such marginal scatter plots provide insights about the potential relationship between the response variable \( y \) and the regressors in multiple linear regression context? Explain your answer.

2 MARKS

This question continues overleaf.
(b) The multiple linear regression model

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \varepsilon \]

is fitted to the data, with \( \varepsilon \sim N(0, \sigma^2) \). The R output for this model is \text{fit1}.

(i) Using the output \text{fit1}, comment on the overall significance of the fitted model.

1 MARK

(ii) Test for a statistically significant association of each of the covariates with the response variable \( y \) in the presence of all other covariates. Comment on your findings.

2 MARKS

(iii) Is multicollinearity a potential problem in this model?

1 MARK

(iv) Figure 2 shows the residual plots for \text{fit1}. Comment on the validity of the model assumptions. Suggest how you might resolve any problems that you observe with possible violations of the model assumptions.

2 MARKS

(c) After some further investigation, we have decided to fit the following model to the data:

\[ \frac{1}{y} = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_5 + \alpha_4 x_7 + \varepsilon. \]

The R output for this model is \text{fit2}.

(i) Explain how we may have come up with this model.

2 MARKS

(ii) Using the output \text{fit2}, comment on the overall significance of the fitted model.

1 MARK

(iii) Test for a statistically significant association of each of the covariates with the response variable \( 1/y \) in the presence of all other covariates in the model. Comment on your findings.

2 MARKS

(iv) Comment on the validity of the model assumptions for the above model using the residual plots in Figure 3 and also the results of the influence summary.

3 MARKS

(v) Provide careful interpretations of the estimated regression coefficients \( \hat{\alpha}_1 \) and \( \hat{\alpha}_2 \) in the context of the problem.

2 MARKS
Figure 1: Scatter plots of $y$ versus $x_1 \ldots x_9$.

Correlation matrix of all the variables:

```
> cor(Data[,1:10])
             y      x1     x2     x3     x4     x5     x6     x7     x8     x9
y  1.000  0.876  0.713  0.649  0.710  0.459  0.531  0.283 -0.399  0.262
x1 0.876  1.000  0.651  0.689  0.734  0.459  0.641  0.367 -0.437  0.147
x2 0.713  0.651  1.000  0.413  0.729  0.224  0.510  0.426 -0.101  0.204
x3 0.649  0.689  0.413  1.000  0.572  0.205  0.392  0.152 -0.353  0.306
x4 0.710  0.734  0.729  0.572  1.000  0.359  0.679  0.574 -0.139  0.107
x5 0.459  0.459  0.224  0.205  0.359  1.000  0.589  0.541 -0.020  0.102
x6 0.531  0.641  0.510  0.392  0.679  0.589  1.000  0.870  0.124  0.222
x7 0.283  0.367  0.426  0.152  0.574  0.541  0.870  1.000  0.314  0.000
x8 -0.399 -0.437 -0.101 -0.353 -0.139 -0.020  0.124  0.314  1.000  0.226
x9  0.262  0.147  0.204  0.306  0.107  0.102  0.222  0.000  0.226  1.000
```
Section L

> fit1<-lm(y~., data=Data)
> summary(fit1)
Call:
  lm(formula = y ~ ., data = Data)
Residuals:
     Min      1Q  Median      3Q     Max
-3.71960 -1.95575 -0.04503  1.62733  4.25259

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.92765    5.91285   2.525  0.0243 *
    x1       1.92472    1.02990   1.869  0.0827 .
    x2       7.00053    4.30037   1.628  0.1258
    x3       0.14918    0.49039   0.304  0.7654
    x4       2.72281    4.35955   0.625  0.5423
    x5       2.00668    1.37351   1.461  0.1661
    x6      -0.41012    2.37854  -0.172  0.8656
    x7      -1.40324    3.39554  -0.413  0.6857
    x8      -0.03715    0.06672  -0.557  0.5865
    x9       1.55945    1.93750   0.805  0.4343
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1

Residual standard error: 2.949 on 14 degrees of freedom
Multiple R-squared: 0.8531, Adjusted R-squared: 0.7587
F-statistic: 9.037 on 9 and 14 DF,  p-value: 0.0001850

> anova(fit1)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
    x1      1 636.16 636.160 73.1525 6.238e-07 ***
    x2      1  29.18  29.180  3.5511 0.08836
    x3      1  4.71  4.710  0.5416 0.47391
    x4      1  0.03  0.030  0.0032 0.95537
    x5      1  8.78  8.780  1.0091 0.33216
    x6      1 13.03 13.030  1.4982 0.24115
    x7      1  9.14  9.140  1.0515 0.32254
    x8      1  0.64  0.640  0.0741 0.78943
    x9      1  5.63  5.630  0.6478 0.43435
Residuals 14 121.75  8.700
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1  1
Figure 2: Residual plots for model `fit1`.

```r
> fit2 <- lm((1/y) ~ x1 + x2 + x5 + x7, data = Data)
> summary(fit2)

Call:
  lm(formula = (1/y) ~ x1 + x2 + x5 + x7, data = Data)

Residuals:  
            Min  1Q  Median  3Q  Max
-0.0030890 -0.0017039 -0.0003299  0.0014922  0.0037488

Coefficients:  
                         Estimate  Std. Error   t value Pr(>|t|)
(Intercept)             0.0456412  0.0029674  15.381  3.54e-12 ***
x1                     -0.0019024  0.0004222  -4.506  0.000242 ***
x2                     -0.0080470  0.0026907  -2.991  0.007514 **
x5                     -0.0023811  0.0009886  -2.409  0.026335 *
x7                      0.0027727  0.0010424   2.660  0.015467 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.00219 on 19 degrees of freedom  
Multiple R-squared: 0.8454, Adjusted R-squared: 0.8128  
F-statistic: 25.97 on 4 and 19 DF,  p-value: 1.792e-07
```
Section L

>anova(fit2)
Analysis of Variance Table

Response: (1/y)

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1</td>
<td>0.00043659</td>
<td>0.00043659</td>
<td>91.0184</td>
<td>1.117e-08 ***</td>
</tr>
<tr>
<td>x2</td>
<td>1</td>
<td>0.00002042</td>
<td>0.00002042</td>
<td>4.2568</td>
<td>0.05303</td>
</tr>
<tr>
<td>x5</td>
<td>1</td>
<td>0.00000740</td>
<td>0.00000740</td>
<td>1.5424</td>
<td>0.22937</td>
</tr>
<tr>
<td>x7</td>
<td>1</td>
<td>0.00003394</td>
<td>0.00003394</td>
<td>7.0754</td>
<td>0.01547 *</td>
</tr>
<tr>
<td>Residuals</td>
<td>19</td>
<td>0.00009114</td>
<td>0.00000480</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 . 0.1  1

>summary(influence.measures(fit2))

Potentially influential observations of

lm(formula = (1/y) ~ x1 + x2 + x5 + x7, data = Data):

dfb.1  dfb.x1  dfb.x2  dfb.x5  dfb.x7  dffit  cov.r  cook.d  hat
1   0.02  0.07  0.03 -0.05 -0.08 -0.16   1.81_  0.01  0.29
7  -0.01 -0.10  0.11  0.24 -0.12 -0.27   1.97_  0.02  0.36
12 -0.04 -0.01  0.00  0.03  0.03 -0.07   1.87_  0.00  0.30
16 -0.09 -1.15_  0.92  0.28 -0.01 -1.27   1.30  0.30  0.43

Figure 3: Residual plots for model fit2.
L2.

An engineer studied the effect of four variables on a dimensionless factor used to describe pressure drops in a screen-plate bubble column. The variables in the data are:

\[ y: \text{ Dimensionless factor for the pressure drop through a bubble cap.} \]
\[ x_1: \text{ Superficial fluid velocity of the gas (cm/s).} \]
\[ x_2: \text{ Kinematic viscosity.} \]
\[ x_3: \text{ Mesh opening (cm).} \]
\[ x_4: \text{ Dimensionless number relating the superficial fluid velocity of the gas to the superficial fluid velocity of the liquid.} \]

The objective is to study the potential relationship between the \( y \) and the regressors \( x_1-x_4 \) through a multiple linear regression model.

(a) Consider the multiple linear regression model

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon, \]

where we assume that \( \varepsilon \sim \mathcal{N}(0, \sigma^2) \). The R output for this model is \texttt{fit1}.

(i) Using the output \texttt{fit1}, comment on the overall significance of the fitted model. What conclusion can you draw? 1 MARK

(ii) Test for a statistically significant association of each of the covariates with the response variable \( y \) in the presence of all other covariates. Discuss your findings. 2 MARKS

(iii) Provide a careful interpretation of the value of \( R^2 \) in the context of the problem. Also comment on the value of \( R^2 \). 2 MARKS

(iv) In general, what are partial residual plots in multiple linear regression? Explain their use. Figure 4 gives the partial residual plots for the model \texttt{fit1}. Comment on these plots. 2 MARKS

(v) Figure 5 shows the residuals plots for \texttt{fit1}. Comment on the validity of the model assumptions. Suggest how you might resolve any problems that you observe with possible violations of the model assumptions. 2 MARKS

This question continues overleaf.
(b) Consider the multiple linear regression model

\[ y = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + \varepsilon, \]

where we assume that \( \varepsilon \sim N(0, \sigma^2) \). The R output for this model is \textit{fit2}.

(i) Using the output \textit{fit2}, comment on the overall significance of the fitted model, and also the individual covariates included in the model.  

2 MARKS

(ii) Using \textit{fit2}, provide careful interpretations of the estimated regression coefficients \( \hat{\alpha}_1 \) and \( \hat{\alpha}_2 \) in the context of the problem.  

2 MARKS

(iii) Comment on the validity of the model assumptions for the above model using the residual plots in Figure 6.  

2 MARKS

(iv) What conclusion do you draw from the partial residual plots in Figure 7?  

1 MARK

(c) Using the outputs \textit{fit1}, \textit{fit2} and the \textit{partial F-test}, perform the test

\[ H_0 : \beta_1 = \beta_4 = 0 \quad \text{vs} \quad H_1 : \text{at least one of them is non-zero} \]

for the parameters of the model in Part \textit{(a)}, at the significance level 0.05.  

2 MARKS

(d) Using the model in Part \textit{(b)} and the output \textit{fit2}, what is the prediction value of \( y \) when \( x_2 = 10.0 \) and \( x_3 = 0.34 \)? What is the estimated expected value (or mean) of \( y \) when \( x_2 = 10.0 \) and \( x_3 = 0.34 \)?

If you were to provide confidence intervals for these quantities, which one would have a wider confidence interval? Explain your answer.  

2 MARKS
Section L

> summary(fit1)

Call:
  lm(formula = y ~ ., data = Data)

Residuals:
     Min      1Q  Median      3Q     Max
-9.9958 -3.3092 -0.2419  3.3924  10.5668

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.89453    4.32508  1.363  0.17828
x1          -0.47790    0.34002 -1.406  0.16530
x2           0.18271    0.01718 10.633  3.78e-15 ***
x3          35.40284   11.09960  3.190  0.00232 **
x4           5.84391    2.90978  2.008  0.04935 *
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 . 0.1  1

Residual standard error: 5.014 on 57 degrees of freedom
Multiple R-squared:  0.6914, Adjusted R-squared:  0.6697
F-statistic: 31.92 on 4 and 57 DF,  p-value: 5.818e-14

> anova(fit1)
Analysis of Variance Table

Response: y
     Df  Sum Sq Mean Sq  F value    Pr(F)
x1    1   9.60   9.600   0.3818 0.539128
x2    1 12839.78 12839.78 3.835e-15 ***
x3    1  258.95  258.95 10.3018 0.002184 **
x4    1  101.39  101.39  4.0336 0.049350 *
Residuals 57 1432.79   25.14
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Figure 4: Partial residual plots for model fit1.

Figure 5: Residual plots for model fit1.
Section L

> fit2<-lm(y~x2+x3, data=Data)
> summary(fit2)

Call:
  lm(formula = y ~ x2+x3, data = Data)

Residuals:
  Min     1Q Median     3Q    Max
-10.0994 -3.6236 -0.6911  2.4722  14.1854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.18971    4.03031   1.784   0.07958 .
x2           0.18456    0.01755   10.518 3.74e-15 ***
x3          35.11616   11.33308    3.099   0.00298 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  

Residual standard error: 5.127 on 59 degrees of freedom
Multiple R-squared:  0.6659, Adjusted R-squared:  0.6546
F-statistic: 58.8 on 2 and 59 DF,  p-value: 9.007e-15

> anova(fit2)
Analysis of Variance Table

Response: y

            Df Sum Sq Mean Sq  F value Pr(>F)
  x2           1 2839.01 2839.01 5.975e-15 ***
  x3           1  252.41  252.41  9.601   0.002977 **
Residuals    59 1551.09   26.29
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Figure 6: Residual plots for model fit2.

Figure 7: Partial residual plots for model fit2.
(a) Consider the standard multiple linear regression model

\[ Y = X\beta + \varepsilon \]  
(1)

where \( \beta = (\beta_0, \ldots, \beta_k)^T \) is the vector of regression parameters, \( Y \) is the \( n \times 1 \) vector of observations of the response variable \( y \), \( X \) is the \( n \times p \) design matrix with \( p = k + 1 \) and \( k \) is the number of covariates included in the model; \( \varepsilon \) is the \( n \times 1 \) vector of errors \( \varepsilon_i \), and \( \varepsilon \sim N(0, \sigma^2 I_n) \).

(i) Explain in full detail how you will perform the general test

\[ H_0 : A\beta = a \quad \text{vs} \quad H_1 : A\beta \neq a \]

for model (1). You may assume that \( A_{m \times p} \) and \( a_{m \times 1} \) are known and that the rank of the matrix \( A \) is \( m \). (You need to derive an appropriate test statistic for the above test and explain how you will use it).

4 MARKS

(ii) Show that the least squares estimate of \( \beta \) subject to the set of equality constraints \( A\beta = a \), for known \( A_{m \times p} \) and \( a_{m \times 1} \), is given by

\[ \hat{\beta} = \hat{\beta} + (X^T X)^{-1} A^T \left[ A(X^T X)^{-1} A^T \right]^{-1} (a - A\hat{\beta}), \]

where \( \hat{\beta} = (X^T X)^{-1} X^T Y \) is the regular least squares estimate of \( \beta \). The estimator \( \hat{\beta} \) is called the constrained least squares estimator of \( \beta \).

4 MARKS

(iii) Discuss situations where you may use the constrained least squares estimator \( \hat{\beta} \) instead of \( \hat{\beta} \) to estimate \( \beta \) in model (1).

2 MARKS

This question continues overleaf.
(b) Denote $X_{(i)}$ as the design matrix $X$ with its $i$-th row $x_i^T = (1, x_{i1}, \ldots, x_{ik})$ removed. You may take for granted (and need not prove) that

$$
\left[ X_{(i)}^T X_{(i)} \right]^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}}
$$

where $h_{ii} = x_i^T (X^T X)^{-1} x_i$ is the $i$-th diagonal element of the hat matrix $H = X(X^T X)^{-1}X^T$.

Also, denote $Y_{(i)}$ as the vector $Y$ of observations of the response variable with the $i$-th observation $y_i$ removed.

(i) Let $\hat{y}_{(i)}$ be the fitted value based on the fitted regression model using all the data except the $i$-th observation $(X_i, y_i)$. Show the following relationship:

$$
e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1 - h_{ii}}, \quad i = 1, \ldots, n
$$

where $e_i = y_i - \hat{y}_i$ are the regular residuals based on the fitted model using all the data.

The $e_{(i)}$ are often called prediction, press or deleted residuals.

4 MARKS

(ii) Under the normality assumption $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, what are the distributional properties of the deleted residuals $e_{(i)}$?

4 MARKS

(iii) Discuss the main application(s) of the deleted residuals $e_{(i)}$ in multiple linear regression.

2 MARKS
Section G

Section G (Generalized Linear Models)
Answer only two questions out of G1–G3

G1.

(a) Consider a $2 \times 2 \times 2$ contingency table arising from the cross-classification by three binary factor variables, A, B and C. Explain how the fit of model

$$A \times B + A \times C + B \times C$$

can be used to assess the equality of odds ratios arising from two $2 \times 2$ sub-tables.  

4 MARKS

(b) The following data describe the numbers of different types of automobile accidents, and their outcome. These data constitute a $2 \times 2 \times 2 \times 2$ table, with the four binary factors being

- **Small** - whether a small car was involved
- **Eject** - whether the driver was ejected
- **Roll** - whether the crash was a rollover
- **Severe** - whether the injuries were severe

All variables are coded as Yes or No (Y or N).

<table>
<thead>
<tr>
<th></th>
<th>Small Car</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Rollover</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Ejected</td>
<td>Ejected</td>
<td>Ejected</td>
<td>Ejected</td>
</tr>
<tr>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Severe</td>
<td>N</td>
<td>350</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>150</td>
<td>23</td>
</tr>
<tr>
<td>Severe</td>
<td>N</td>
<td>1878</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>1022</td>
<td>161</td>
</tr>
<tr>
<td>Severe</td>
<td>N</td>
<td>148</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>404</td>
<td>265</td>
</tr>
</tbody>
</table>

Output related to the analysis of these data, stored in data frame AutoAcc, is included on pages 21–23. Use this output, or direct computations, to answer the following questions. Note that a table of the Chi-squared distribution is given on page 28.

This question continues overleaf.
Section G

(i) For small cars only, is there a statistically significant difference in the odds ratios relating severity of injury to ejection from the automobile in the two sub-tables defined by the nature of the accident (rollover or non-rollover)?

2 MARKS

(ii) Are outcomes significantly different (in any way) for small and standard (non-small) cars? Justify your answer.

4 MARKS

(iii) On the basis of the output given, establish the most appropriate model for these data. Comment on the global fit of this model, interpret the model, and write down any other models that you would consider fitting.

4 MARKS

(iv) Using your selected model, report the fitted proportion of severe injuries resulting from non-rollover accidents for small cars where the driver was not ejected.

2 MARKS

(c) Suppose that these data were recorded from road traffic accidents in a single province of Canada, but that data on all ten Canadian provinces were available separately in similar format. Explain how the modelling process should be extended to allow an analysis of all the data simultaneously.

4 MARKS
G2.

(a) Explain the concepts of overdispersion and underdispersion in the context of generalized linear modelling. Show how models for overdispersion and underdispersion can be constructed for non-negative count data based on the Poisson distribution. Give full details of the mathematical derivations. 

6 MARKS

(b) In the following ecology study, 96 agricultural plots (10 metres square) were sampled across ten regions. For each plot, the type of soil type was recorded, along with the soil nitrogen content. The number of beetles of a particular species found in the plots were then recorded as the response variable.

The predictors are denoted:

- **Type** - soil type (factor predictor, levels 1,2,3)
- **X** - nitrogen content (per cent)
- **Region** - region (factor predictor, levels 1 to 10)

Output related to the analysis of these data is included on pages 24–25. Use this output to select the best model to explain the observed data.

5 MARKS

(c) For the data in (b), carry out a formal test of the null hypothesis where the data are specified by a Poisson model, against a suitable alternative distributional assumption, where the linear predictor is specified by the main effects model

\[
\text{Type} + \text{Region}
\]

under both null and alternative. Give full details of the test you use.

4 MARKS

(d) Explain the estimation procedure behind the R command

```
glm(Y~Type+Region, family=quasipoisson)
```

and contrast the results with those that would be obtained if the argument

```
family=poisson
```

was used instead.

5 MARKS
G3.

(a) Suppose that $Y_1, \ldots, Y_K$ are independent Poisson random variables with means $\mu_1, \ldots, \mu_K$. Let

$$Y = \sum_{k=1}^{K} Y_k$$

Derive the conditional distribution of $Y_1, \ldots, Y_K$, given that $Y = m$. 4 MARKS

(b) Describe in detail two models that can be used for multinomial response data when the responses are ordered categorical data. Explain how such data might arise in the context of a clinical pain study. 6 MARKS

(c) The following table classifies 1681 residents in terms of the type of housing they had, their feeling of influence on apartment management, their degree of contact with other residents, and their satisfaction with housing conditions.

<table>
<thead>
<tr>
<th>Housing Type</th>
<th>Influence</th>
<th>Contact</th>
<th>Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>low</td>
<td>low</td>
</tr>
<tr>
<td>Tower block</td>
<td>low</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>medium</td>
<td>34</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>17</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Apartments</td>
<td>low</td>
<td>61</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>78</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>medium</td>
<td>43</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>48</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>26</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>Atrium</td>
<td>low</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>medium</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Terraced</td>
<td>low</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>57</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>medium</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>31</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

This question continues overleaf.
These data form a $3 \times 4 \times 3 \times 2$ table, with satisfaction the response variable. An analysis of these data was carried out in R, where the data are stored in the data frame `housing`, and the results presented on page 26. The three factor predictors are

- **Type**: the type of housing (four levels)
- **Infl**: the amount of influence (three levels)
- **Cont**: the contact level (two levels)

Using the output, decide upon the best model to describe the observed data.

5 MARKS

(d) The output on page 27 addresses a different aspect of model comparison. Describe exactly the two models fitted, and contrast the results obtained. How would you compute the estimated probability of a high satisfaction level for each model, given a specification for the three factor predictors?

Note that the AIC quantity quoted is computed as

$$-2 \times \text{maximized log-likelihood} + 2p,$$

where $p$ is the number of parameters fitted.

5 MARKS
Code and Output for Question G1.

```r
1 AutoAcc <- read.table("tab4-15.txt", header=F)
2 names(AutoAcc) <- c("Small", "Eject", "Severe", "Roll", "Count")
3 ####################################################################################################
4 #Part (b)(i)
5
6 msub1 <- glm(Count ~ Eject + Severe + Roll, data = AutoAcc, family = poisson, subset = (Small == "Y"))
7 msub2 <- glm(Count ~ Eject + Severe + Roll, data = AutoAcc, family = poisson, subset = (Small == "Y"))
8 msub3 <- glm(Count ~ Eject + Severe + Severe + Roll, data = AutoAcc, family = poisson,
9   subset = (Small == "Y"))
10 msub4 <- glm(Count ~ Eject + Severe + Eject + Roll + Severe + Roll, data = AutoAcc,
11   family = poisson, subset = (Small == "Y"))
12
13 ####################################################################################################
14 #Part (b)(ii-iv)
15
16 m1 <- glm(Count ~ Eject + Severe + Eject + Roll + Severe + Roll, data = AutoAcc, family = poisson)
17 m2 <- glm(Count ~ Eject + Severe + Roll, data = AutoAcc, family = poisson)
18 m3 <- glm(Count ~ Small + (Eject + Severe + Roll), data = AutoAcc, family = poisson)
19 m4 <- glm(Count ~ Small * (Eject + Severe + Eject + Roll + Severe + Roll), data = AutoAcc,
20   family = poisson)
21 m5 <- glm(Count ~ Small + Severe + Eject + Severe + Eject + Roll, data = AutoAcc,
22   family = poisson)
23 m6 <- glm(Count ~ Small + Severe + Eject + Severe + Eject + Roll + Severe + Roll + Small + Roll,
24   data = AutoAcc, family = poisson)
25
26 > anova(msub1, msub2, msub3, msub4)
27 # Analysis of Deviance Table
28
29 Model 1: Count ~ Eject + Severe + Roll
30 Model 2: Count ~ Eject + Severe + Roll
31 Model 3: Count ~ Eject + Severe + Severe + Roll
32 Model 4: Count ~ Eject + Severe + Eject + Roll + Severe + Roll
33 Resid. Df Resid. Dev Df Deviance
34 1 4 217.196
35 2 3 170.879 1 46.317
36 3 2 55.315 1 115.564
37 4 1 0.043 1 55.272
```
### Code and Output for Question G1.

```r
> anova(m1,m2,m3,m4,m5,m6)
Analysis of Deviance Table

| Model 1: Count ~ Eject * Severe + Eject + Roll + Severe + Roll |
| Model 2: Count ~ Eject * Severe + Roll |
| Model 3: Count ~ Small + (Eject * Severe + Roll) |
| Model 4: Count ~ Small * (Eject * Severe + Eject * Roll + Severe + Roll) |
| Model 5: Count ~ Small * Severe + Eject + Severe + Eject + Roll |
| Model 6: Count ~ Small * Severe + Eject * Severe + Eject + Roll + Severe + Roll + Small + Roll |

<table>
<thead>
<tr>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>2372.80</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>2371.09</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>74.64</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3.61</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>504.66</td>
<td>-6</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>9.02</td>
<td>2</td>
</tr>
</tbody>
</table>

> coef(summary(m3))

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|---------|
| (Intercept) | 7.5228467 | 0.02216193 | 339.449148 | 0.0000000e+00 |
| SmallY | -1.5874915 | 0.03832527 | -41.421532 | 0.0000000e+00 |
| EjectY | -2.7888787 | 0.08802331 | -31.683411 | 2.629748e-220 |
| SevereY | -0.6423926 | 0.03608428 | -17.802561 | 6.750941e-71 |
| RollY | -2.3713215 | 0.07250190 | -32.707026 | 1.240866e-234 |
| EjectY:SevereY | 0.9373475 | 0.11847417 | 7.911830 | 2.536326e-15 |
| EjectY:RollY | 1.1649127 | 0.19221352 | 6.060514 | 1.356876e-09 |
| SevereY:RollY | 1.5509613 | 0.08970926 | 17.288753 | 5.717580e-67 |
| EjectY:SevereY:RollY | 0.2840562 | 0.21925008 | 1.295581 | 1.951199e-01 |
```
### Code and Output for Question G1.

```r
> coef(summary(m4))
     Estimate Std. Error  z value         Pr(>|z|)
(Intercept)   7.5418240 0.02293643  328.814252 0.0000000e+00
SmallY      -1.6824952 0.05774285  -29.137724 1.195231e-186
EjectY      -2.8999903 0.09258160  -31.323512 2.226159e-215
SevereY     -0.6194418 0.03847575  -16.099337 2.570187e-58
Rolly      -2.5949496 0.08206460  -31.620317 1.910775e-219
EjectY:SevereY    1.1031518 0.11215619     9.835353 7.889647e-23
EjectY:Rolly     1.3295288 0.10261570     12.956388 2.161899e-38
SevereY:Rolly    1.6918067 0.09573065     17.672371 6.821559e-70
SmallY:EjectY    0.2797779 0.02010768      1.384301 1.662663e-01
SmallY:SevereY   -0.2325159 0.10256079    -2.267103 2.338392e-02
SmallY:Rolly     0.8217848 0.15574130      5.276602 1.316015e-07
SmallY:EjectY:SevereY -0.3337856 0.24041353    -1.388381 1.650211e-01
SmallY:EjectY:Rolly  0.1743610 0.23152589      0.753095 4.513928e-01
SmallY:SevereY:Rolly -0.2031591 0.19374265   -1.048603 2.943610e-01

> coef(summary(m5))
     Estimate Std. Error  z value         Pr(>|z|)
(Intercept)   7.41382803 0.02329733  318.2265136 0.0000000e+00
SmallY      -1.55710301 0.05158468  -30.1853753 3.685239e-200
SevereY      -0.35546528 0.03424098   -10.3812819 3.016965e-25
EjectY       -3.21287049 0.08820973   -36.4230852 1.835586e-290
Rolly       -1.54673932 0.04903083   -37.7890996 0.0000000e+00
SmallY:SevereY -0.06702105 0.07707569    -0.8695485 3.845472e-01
SmallY:EjectY  1.45601784 0.09225649    15.7822804 4.120291e-56
EjectY:Rolly  1.73113556 0.08591432     20.1495572 2.715316e-90

> coef(summary(m6))
     Estimate Std. Error  z value         Pr(>|z|)
(Intercept)   7.5352491 0.02275184  331.192893 0.0000000e+00
SmallY      -1.6458546 0.05337623  -30.834974 8.908270e-209
SevereY      -0.6051310 0.03744598   -16.160105 9.639748e-59
EjectY       -2.8371393 0.08149746   -34.812610 1.567630e-265
Rolly       -2.5724488 0.07369775   -34.905389 6.158794e-267
SmallY:SevereY -0.3194086 0.08485783    -3.764044 1.671876e-04
SmallY:EjectY  1.0234252 0.09872381    10.366549 3.520076e-25
EjectY:Rolly  1.3815725 0.09120438    15.148095 7.799523e-52
SevereY:Rolly  1.6383380 0.08284156    19.776764 4.720311e-87
```
Code and Output for Question G2.

```r
mod1<-glm(Y~X+Region+Type, family=poisson)
mod2<-glm(Y~X+Type, family=poisson)
mod3<-glm(Y~X+Region, family=poisson)
mod4<-glm(Y~Type+Region, family=poisson)
mod5<-glm(Y~Type, family=poisson)
mod6<-glm(Y~Region, family=poisson)

# library(MASS)
mod7<-glm.nb(Y~X+Region+Type)
mod8<-glm.nb(Y~Type+Region)
mod9<-glm.nb(Y~Region)
mod10<-glm.nb(Y~Type)

> anova(mod1,mod2,mod3,mod4,mod5,mod6)
Analysis of Deviance Table

Model 1: Y ~ X + Region + Type
Model 2: Y ~ X + Type
Model 3: Y ~ X + Region
Model 4: Y ~ Type + Region
Model 5: Y ~ Type
Model 6: Y ~ Region

Resid. Df Resid. Dev Df Deviance
  1  83   147.61
  2  92   564.19  -9  -416.58
  3  85   293.67   7  270.53
  4  84   152.90   1  140.77
  5  93   565.48  -9  -412.58
  6  86   313.08   7  252.40

> anova(mod7,mod8,mod9,mod10)
Likelihood ratio tests of Negative Binomial Models

Response: Y

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
<th>theta</th>
<th>rdf</th>
<th>2 log-lik.</th>
<th>Test df</th>
<th>LR stat.</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Type</td>
<td>1.238</td>
<td>93</td>
<td>-525.4176</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Region</td>
<td>2.758</td>
<td>86</td>
<td>-471.0324</td>
<td>1 vs 7</td>
<td>54.385</td>
<td>1.973292e-09</td>
</tr>
<tr>
<td>3</td>
<td>Type + Region</td>
<td>12.052</td>
<td>84</td>
<td>-416.9450</td>
<td>2 vs 3</td>
<td>54.087</td>
<td>1.799227e-12</td>
</tr>
<tr>
<td>4</td>
<td>X + Region + Type</td>
<td>13.051</td>
<td>83</td>
<td>-410.4605</td>
<td>3 vs 4</td>
<td>6.485</td>
<td>1.088147e-02</td>
</tr>
</tbody>
</table>
```
# Log likelihoods: df records the total numbers of parameters fitted.

> logLik(mod1); logLik(mod2); logLik(mod4); logLik(mod5); logLik(mod6)

'log Lik.' -213.8031 (df=13)
'log Lik.' -422.0931 (df=4)
'log Lik.' -216.4461 (df=12)
'log Lik.' -422.7371 (df=3)
'log Lik.' -296.5375 (df=10)

> logLik(mod7); logLik(mod8); logLik(mod9); logLik(mod10)

'log Lik.' -205.2302 (df=14)
'log Lik.' -208.4725 (df=13)
'log Lik.' -235.5162 (df=11)
'log Lik.' -262.7088 (df=4)
Section G

Code and Output for Question G3.

1 house.plr1<-polr(Sat\textasciicircum Infl\textasciicircum Type\textasciicircum Cont, weights=Freq, data=housing)  
2 house.plr2<-polr(Sat\textasciicircum Infl\textasciicircum Type+Infl\textasciicircum Cont+Type\textasciicircum Cont, weights=Freq, data=housing)  
3 house.plr3<-polr(Sat\textasciicircum Infl+Infl\textasciicircum Type+Type\textasciicircum Cont, weights=Freq, data=housing)  
4 house.plr4<-polr(Sat\textasciicircum Infl+Infl\textasciicircum Type+Infl\textasciicircum Type, weights=Freq, data=housing)  
5 house.plr5<-polr(Sat\textasciicircum Infl+Infl\textasciicircum Cont+Type\textasciicircum Infl\textasciicircum Cont, weights=Freq, data=housing)  
6 house.plr6<-polr(Sat\textasciicircum Infl+Infl\textasciicircum Type+Type\textasciicircum Cont, weights=Freq, data=housing)  
7 house.plr7<-polr(Sat\textasciicircum Infl+Type\textasciicircum Cont, weights=Freq, data=housing)  

8 > anova(house.plr1,house.plr2,house.plr3,house.plr4,house.plr5,house.plr6,house.plr7)  
9 Likelihood ratio tests of ordinal regression models

11 Response: Sat

<table>
<thead>
<tr>
<th>Model</th>
<th>Res.df</th>
<th>Res. Dev</th>
<th>Test</th>
<th>Df</th>
<th>LRstat.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Infl+Type+Cont</td>
<td>1673</td>
<td>3479.149</td>
<td>1 2</td>
<td>2</td>
<td>0.209</td>
<td>0.9008</td>
</tr>
<tr>
<td>2 Infl+Cont+Type+Infl+Cont</td>
<td>1671</td>
<td>3478.940</td>
<td>1 2</td>
<td>2</td>
<td>0.209</td>
<td>0.9008</td>
</tr>
<tr>
<td>3 Infl+Cont+Type+Infl+Type</td>
<td>1670</td>
<td>3470.483</td>
<td>2 3</td>
<td>3</td>
<td>8.457</td>
<td>0.0036</td>
</tr>
<tr>
<td>4 Infl+Cont+Type+Infl+Type</td>
<td>1667</td>
<td>3456.640</td>
<td>3 4</td>
<td>3</td>
<td>13.843</td>
<td>0.0031</td>
</tr>
<tr>
<td>5 Infl+Cont+Infl+Type+Type+Cont</td>
<td>1664</td>
<td>3448.695</td>
<td>4 5</td>
<td>3</td>
<td>7.945</td>
<td>0.0472</td>
</tr>
<tr>
<td>6 Infl+Type+Infl+Cont+Type+Cont</td>
<td>1662</td>
<td>3448.583</td>
<td>5 6</td>
<td>2</td>
<td>0.112</td>
<td>0.9457</td>
</tr>
<tr>
<td>7 Infl+Type+Cont</td>
<td>1656</td>
<td>3446.458</td>
<td>6 7</td>
<td>6</td>
<td>2.125</td>
<td>0.9079</td>
</tr>
</tbody>
</table>

21

22 > logLik(house.plr1);logLik(house.plr2);logLik(house.plr3);  
23 'log Lik.' -1723.229 (df=25)  
24 'log Lik.' -1724.292 (df=19)  
25 'log Lik.' -1724.347 (df=17)  
26

28 > logLik(house.plr4);logLik(house.plr5);logLik(house.plr6);  
29 'log Lik.' -1728.32 (df=14)  
30 'log Lik.' -1739.47 (df=10)  
31 'log Lik.' -1735.242 (df=11)  
32

33 > logLik(house.plr7)  
34 'log Lik.' -1739.575 (df=8)
Code and Output for Question G3.

```r
> summary(house.plr7, digits=4)

Call: polr(formula = Sat ~ Infl + Type + Cont, data = housing, weights = Freq)

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Err.</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>InflMedium</td>
<td>0.5664</td>
<td>0.10465</td>
<td>5.412</td>
</tr>
<tr>
<td>InflHigh</td>
<td>1.2888</td>
<td>0.12716</td>
<td>10.136</td>
</tr>
<tr>
<td>TypeApartment</td>
<td>-0.5724</td>
<td>0.11924</td>
<td>-4.800</td>
</tr>
<tr>
<td>TypeAtrium</td>
<td>-0.3662</td>
<td>0.15517</td>
<td>-2.360</td>
</tr>
<tr>
<td>TypeTerrace</td>
<td>-1.0910</td>
<td>0.15149</td>
<td>-7.202</td>
</tr>
<tr>
<td>ContHigh</td>
<td>0.3603</td>
<td>0.09554</td>
<td>3.771</td>
</tr>
</tbody>
</table>

Intercepts:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Medium</td>
<td>-0.4961</td>
<td>0.1248</td>
</tr>
<tr>
<td>Medium</td>
<td>High</td>
<td>0.6907</td>
<td>0.1255</td>
</tr>
</tbody>
</table>

Residual Deviance: 3479.149
AIC: 3495.149

> house.mlr <- multinom(Sat ~ Infl+Type+Cont, weights=Freq, data=housing)

> summary(house.mlr, digits=4)

Call: multinom(formula = Sat ~ Infl + Type + Cont, data = housing, weights = Freq)

Coefficients:

(Intercept) InflMedium InflHigh TypeApart TypeAtrium TypeTerr ContHigh

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Err.</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td>-0.4192</td>
<td>0.4464</td>
<td>0.665</td>
</tr>
<tr>
<td>High</td>
<td>-0.1387</td>
<td>0.7349</td>
<td>1.613</td>
</tr>
</tbody>
</table>

Std. Errors:

(Intercept) InflMedium InflHigh TypeApart TypeAtrium TypeTerr ContHigh

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Err.</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td>0.1729</td>
<td>0.1416</td>
<td>0.1863</td>
</tr>
<tr>
<td>High</td>
<td>0.1592</td>
<td>0.1369</td>
<td>0.1671</td>
</tr>
</tbody>
</table>

Residual Deviance: 3470.084
AIC: 3498.084
### Table of the Chi-squared distribution

Entries in table $\chi^2_\nu (\nu)$: the α tail quantile of Chi-squared(\nu) distribution α given in columns, \nu given in rows.

<table>
<thead>
<tr>
<th>\nu</th>
<th>Left-tail</th>
<th>Right-tail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.99500</td>
<td>0.99000</td>
</tr>
<tr>
<td>1</td>
<td>3.8416</td>
<td>2.7005</td>
</tr>
<tr>
<td>2</td>
<td>5.9914</td>
<td>4.6051</td>
</tr>
<tr>
<td>3</td>
<td>7.8788</td>
<td>5.9914</td>
</tr>
<tr>
<td>5</td>
<td>14.0626</td>
<td>12.8316</td>
</tr>
<tr>
<td>7</td>
<td>15.6465</td>
<td>15.0862</td>
</tr>
<tr>
<td>8</td>
<td>16.6604</td>
<td>15.9871</td>
</tr>
<tr>
<td>9</td>
<td>17.6745</td>
<td>17.0260</td>
</tr>
<tr>
<td>10</td>
<td>18.6880</td>
<td>18.1493</td>
</tr>
<tr>
<td>14</td>
<td>22.7356</td>
<td>22.6710</td>
</tr>
<tr>
<td>15</td>
<td>23.7474</td>
<td>23.8069</td>
</tr>
<tr>
<td>18</td>
<td>26.7824</td>
<td>27.1135</td>
</tr>
<tr>
<td>20</td>
<td>28.8055</td>
<td>29.1835</td>
</tr>
<tr>
<td>21</td>
<td>29.8170</td>
<td>30.2186</td>
</tr>
<tr>
<td>22</td>
<td>30.8284</td>
<td>31.2537</td>
</tr>
<tr>
<td>23</td>
<td>31.8399</td>
<td>32.2888</td>
</tr>
<tr>
<td>24</td>
<td>32.8513</td>
<td>33.3240</td>
</tr>
<tr>
<td>25</td>
<td>33.8628</td>
<td>34.3591</td>
</tr>
<tr>
<td>26</td>
<td>34.8742</td>
<td>35.3943</td>
</tr>
<tr>
<td>27</td>
<td>35.8857</td>
<td>36.4295</td>
</tr>
<tr>
<td>28</td>
<td>36.9971</td>
<td>37.4647</td>
</tr>
<tr>
<td>29</td>
<td>38.0086</td>
<td>38.5000</td>
</tr>
</tbody>
</table>

*Note: The table continues with similar entries for \nu values from 31 to 100.*