The University of Melbourne

Semester 2 Assessment —November, 2002

Department of Mathematics and Statistics

620-372  Applied Statistical Analysis

Exam duration: three hours
Reading time: fifteen minutes
This paper has sixteen (16) pages, including 8 pages of formulae and tables.

Authorised materials:
Hand-held electronic calculators may be used.

Instructions to invigilators:
Statistical tables will be supplied.

Instructions to students:
There are eight (8) questions. All questions may be attempted.
Some notes and tables are attached.
The number of marks for each question is indicated.
The total number of marks available is 120.
1. A random sample of size $n$ is obtained on $X$, where $X$ has probability density function

$$f(x; \theta) = \frac{4\theta^4}{x^5} e^{-\theta/x^4}, \quad x > 0, \ \theta > 0.$$  

The results $\mathbb{E}(X) = \theta \Gamma(3/4)$, $\mathbb{E}(X^2) = \theta^2 \Gamma(1/2)$ and $\mathbb{E}(X^{-4}) = \theta^{-4}$ may be used without proof.

(a) Find the method of moments estimator of $\theta$.
(b) Find the maximum likelihood estimator of $\theta$.
(c) Show that the asymptotic efficiency of the method of moments estimator of $\theta$ is

$$\frac{\Gamma(3/4)^2}{16[\Gamma(1/2)-\Gamma(3/4)^2]}.$$  

$[2 + 3 + 5 = 10 \text{ marks}]$

2. Let $x_1, x_2, \ldots, x_n$ be a random sample from a distribution with probability density function

$$f(x; \theta, \alpha) = \frac{\alpha \theta^\alpha}{(x + \theta)^{\alpha+1}}, \quad x \geq 0, \ \theta > 0, \ \alpha > 2.$$  

The result $\mathbb{E}[(X + \theta)^r] = \frac{\alpha^r}{\alpha - r}$, for $r < \alpha$, may be used without proof.

(a) Find, if one exists,

(i) the (single) sufficient statistic for $\alpha$ when $\theta$ is known;
(ii) the (single) sufficient statistic for $\theta$ when $\alpha$ is known.
(b) Show that the (expected) information matrix for $(\alpha, \theta)$ is

$$\begin{bmatrix}
\frac{n}{\alpha^2} & -\frac{n}{\theta(\alpha+1)} \\
-\frac{n}{\theta(\alpha+1)} & \frac{n\theta}{\theta^2(\alpha+2)}
\end{bmatrix}.$$  

(c) Find the lower bounds for the variance of unbiased estimators of $\theta$:

(i) when $\alpha = 10$;
(ii) when $\alpha$ is unknown.
(d) Explain how the method of scoring could be used to obtain the maximum likelihood estimates of $\alpha$ and $\theta$.  

$[2 + 5 + 3 + 3 = 13 \text{ marks}]$
3. Let $x_1, x_2, \ldots, x_n$ be a random sample from a distribution with probability density function

$$f(x; \theta) = \frac{2e^{-(x-\theta)^4}}{\Gamma(1/4)}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$ 

(a) Find:

(i) the form of the most powerful test for testing $H_0 : \theta = 0$ versus $H_1 : \theta = 1$

(ii) the form of the locally most powerful test for testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$

(iii) the form of the likelihood ratio test for testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$

(b) Give the asymptotic forms of the tests in (ii) and (iii) of part (a) for a test of size (level of significance) 0.05. [You may use the result that $I(\theta) = \frac{12n\Gamma(3/4)}{\Gamma(1/4)}$ without proof.]

[8 + 4 = 12 marks]

4. A study was conducted to investigate the toxicity to the tobacco budworm *Heliothis virescens* of doses of the pyrethroid *trans*-cypermethrin to which the moths were beginning to show resistance. Batches of 20 moths of each sex were exposed for three days to the pyrethroid and the number in each batch that were killed was recorded. The results were

<table>
<thead>
<tr>
<th>Dose (µg)</th>
<th>Dosage (=log₂(dose))</th>
<th>Number killed (out of 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
<td>20</td>
</tr>
</tbody>
</table>

Nine (logistic regression) models (bud.1 to bud.9) were fitted to these data and resulted in the residual deviances given in the table below, together with the Splus model specifications. For these models, sex is a factor with 2 levels (1 = male; 2 = female), dose.f refers to dose treated as a factor with 6 levels, dose is the dose treated as a variable and dosage is log₂(dose) treated as a variable (logs to the base 2 have been used for numerical convenience). Also, dose² = dose², etc.
(a) For each of the following pairs of models explain what hypothesis is being tested if the difference between the residual deviances is compared with the relevant $\chi^2$ distribution and give the result of the test, or explain why a test of the difference is not valid:

(i) model bud.1 and model bud.2;
(ii) model bud.1 and model bud.3;
(iii) model bud.1 and model bud.7.

(b) Which of the models listed in the table above is most appropriate for these data? Give details of any tests that you carry out and clearly state the conclusions drawn from each test.

(c) For the model found in (b), use the parameter estimates from the relevant summaries given below to answer the following questions.

(i) Is the effect of sex significant in the model? Justify your answer.
(ii) Quantify, in terms of odds ratios, the effects of sex and dose or dosage on toxicity.
(iii) Find an estimate of the dose, or dosage, that would kill 50% of male budworm moths (the LD(50)).
(iv) Obtain an estimate of the probability that a female moth will die within three days if subjected to a dose of 16$\mu$g of pyrethroid (or a dosage of 4) for three days.

(d) A total of 240 budworms were used in this study. Describe what would have been the same and what would have been different in the output (residual deviances and their degrees of freedom, differences between residual deviances and their degrees of freedom, parameter estimates and their standard errors) had the data been treated as 240 ungrouped, 0–1 observations, rather than as 12 groups of 20 moths.

\[3 + 5 + 8 + 4 = 20 \text{ marks}\]
Splus output for question 4 (using contrasts = contr.treatment)

```r
> summary(bud.1)$coef
   Value Std. Error t value
(Intercept) -3.254121 1.0197836 -3.190992
sex -1.089652 0.3529994 -3.086838
dose.f2 1.961727 1.1111702 1.765461
dose.f3 2.502078 1.0729401 2.346500
dose.f4 4.124006 1.0729401 3.843650
dose.f5 4.972161 1.0906944 4.558711
dose.f6 6.112219 1.1578986 5.278717

> summary(bud.2)$coef
   Value Std. Error t value
(Intercept) -1.9277146 0.40188764 -4.7966505
sex 0.2119349 0.51518049 0.4113799
dose 0.2972343 0.06252197 4.7540777
sex:dose -0.1815578 0.06689466 -2.7140853

> summary(bud.3)$coef
   Value Std. Error t value
(Intercept) -1.1660678 0.26154205 -4.458433
sex -0.9685483 0.32953474 -2.939139
dose 0.1599557 0.02341137 6.832392

> summary(bud.4)$coef
   Value Std. Error t value
(Intercept) -1.932167799 0.359495120 -5.374670
sex -1.071214890 0.351186208 -3.050276
I(dose^2) -0.007159676 0.001991659 -3.594830

> summary(bud.5)$coef
   Value Std. Error t value
(Intercept) -2.9893581726 0.5939929023 -5.032650
sex -1.0973248809 0.3557157461 -3.084836
I(dose^2) -0.0494398191 0.0169558730 -2.915793
I(dose^3) 0.0008911737 0.0003504321 2.543071

> summary(bud.6)$coef
   Value Std. Error t value
(Intercept) -2.8185549 0.5479524 -5.1437955
sex -0.1749868 0.7782843 -0.2248366
dosage 1.2589494 0.2120484 5.9370858
sex:dosage -0.3529130 0.2699765 -1.3071990

> summary(bud.7)$coef
   Value Std. Error t value
(Intercept) -2.372412 0.3855067 -6.154009
sex -1.100743 0.3558238 -3.093507
I(dosage^2) -0.07940384 0.08440062 -0.9407969
I(dosage^3) 0.05643178 0.06318746 0.8933679

> summary(bud.8)$coef
   Value Std. Error t value
(Intercept) -3.26959244 0.90621320 -3.6079726
sex -1.08956802 0.35287167 -3.0877175
dosage 2.46160217 1.24299869 1.9803882
I(dosage^2) -0.52941369 0.51654746 -1.0249081
I(dosage^3) 0.05643178 0.06318746 0.8933679
```
5. The following data refer to the effect of academic achievement on self-esteem among black and white female college students.

<table>
<thead>
<tr>
<th>Academic Achievement</th>
<th>Self – Esteem</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Race</td>
</tr>
<tr>
<td>High</td>
<td>Black</td>
</tr>
<tr>
<td></td>
<td>White</td>
</tr>
<tr>
<td>Low</td>
<td>Black</td>
</tr>
<tr>
<td></td>
<td>White</td>
</tr>
</tbody>
</table>

From these data the following residual deviances were obtained using log-linear models where A, R and E are factors which denote academic achievement, race and self-esteem, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>residual deviance</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>A + R + E</td>
<td>17.32</td>
<td>4</td>
</tr>
<tr>
<td>A + R*E</td>
<td>11.53</td>
<td>3</td>
</tr>
<tr>
<td>A*R + E</td>
<td>8.64</td>
<td>3</td>
</tr>
<tr>
<td>A*E + R</td>
<td>16.71</td>
<td>3</td>
</tr>
<tr>
<td>A<em>R + R</em>E</td>
<td>2.85</td>
<td>2</td>
</tr>
<tr>
<td>A<em>E + R</em>E</td>
<td>10.91</td>
<td>2</td>
</tr>
<tr>
<td>A<em>R + A</em>E</td>
<td>8.02</td>
<td>2</td>
</tr>
<tr>
<td>A<em>R + A</em>E + R*E</td>
<td>0.95</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) State which model corresponds to each of the following interpretations and state, with reasons, which model you consider to be the most appropriate for these data:

(i) Self-esteem is independent of both academic achievement and race.
(ii) Given academic achievement, self-esteem is independent of race.
(iii) Given race, self-esteem is independent of academic achievement.
(iv) The association between self-esteem and academic achievement, as measured by the odds ratio, is the same for blacks and whites.

(b) For each of the factors A and R state, with reasons, whether collapsing the three-way table to form a two-way table is reasonable, and determine whether or not a test of association in the collapsed table would be statistically significant, at the 5% level.

(c) Taking self-esteem to be a response factor with two levels, the above three-way contingency table can be analysed using logistic regression models, though not all of the log-linear models can be expressed as a logistic regression model. List the log-linear models from the above table for which there is an equivalent logistic regression model, together with the corresponding Splus logistic regression model specification [eg, \( y/n \sim A + R \) where \( y \) denotes the number with high self-esteem and \( n \) denotes the group size].

[A log-linear and logistic regression model are equivalent if they give the same residual deviance, and degrees of freedom.]

\[ 5 + 6 + 4 = 15 \text{ marks} \]
6. (a) Explain what an improper prior is and why it is used.

(b) Suppose data \( x \sim (x_1, x_2, \ldots, x_n) \) are drawn from the distribution \( N(\mu, \phi) \) where \( \mu \) is known and \( \phi \) is unknown. What is the commonly used reference prior for \( \phi \) and why is it preferred?

(c) Suppose a random sample on \( X \sim N(1, \phi) \) gives

\[
1.4 \quad 1.1 \quad 0.8 \quad 1.5 \quad 1.0 \quad 0.9 \quad 1.6 \quad 1.5
\]

Assuming the reference prior for \( \phi \) specified in (b):
(i) find the posterior distribution of \( \phi \);
(ii) find a 95% highest density region (HDR) for \( \phi \), based on a log chi-squared distribution;
(iii) give a reason why a log chi-squared distribution is preferred to an inverse chi-squared distribution.

(3 + 2 + 6 = 11 marks)

7. (a) (i) State Jeffreys’ rule and give a reason why Jeffreys’ prior is desirable.
(ii) Suppose a random sample is available from a Poisson distribution with mean \( \lambda \). Derive Jeffreys’ prior for \( \lambda \).

(b) Explain what a conjugate prior is and show that the family of beta distributions is the conjugate family for the binomial distribution.

(c) In a sequence of 10 Bernoulli trials with \( \pi = \Pr(\text{success}) \), 4 successes were observed. Assume the reference prior \( \text{Be}(1/2, 1/2) \).
(i) Find the Bayes estimate and the generalized maximum-likelihood estimate (GMLE) of \( \pi \).
(ii) Show that this prior is uniform in \( \sin^{-1} \sqrt{\pi} \).
(iii) It can be shown that, for \( X \sim \text{Bi}(n, \pi) \), the transformation

\[
Z = \sin^{-1} \sqrt{\frac{X}{n}}
\]

is approximately normally distributed. Specifically,

\[
Z \approx N(\theta, \frac{1}{4n})
\]

where \( \theta = \sin^{-1} \sqrt{\pi} \). (Do not show this.) Find the posterior distribution of \( \theta \).

(7 + 5 + 10 = 22 marks)
8. (a) Explain carefully why the plug-in estimate for the mean \( \mu = E(X) \), of a random variable \( X \) is the sample mean \( \bar{x} \). Hence derive the plug-in estimate of the variance \( \sigma^2 \) of \( X \).

(b) The following data were obtained on 6 oranges randomly selected from a truckload of oranges

\[
\begin{array}{cccccc}
    x (\text{gm}) & 9.5 & 11.3 & 15.0 & 8.6 & 10.4 & 11.2 \\
    y (\text{gm}) & 181 & 194 & 227 & 177 & 191 & 200 \\
\end{array}
\]

where \( x = \) sugar content and \( y = \) weight. It is desired to estimate the ratio \( \theta = \frac{E(Y)}{E(X)} \), using the estimator \( T = \frac{\bar{Y}}{\bar{X}} \). (For these data \( \sum x_i = 66.0 \) and \( \sum y_i = 1170 \).)

(i) Describe in detail how the non-parametric bootstrap method can be used to obtain an estimate of the standard error of \( T \).

(ii) Compute the jackknife estimate of the standard error of \( T \).

(7 + 6 + 4 = 17 marks)

2001

1. (a) (i) Give an example of an estimator which is consistent but not unbiased.

(ii) Give an example of an estimator which is unbiased but not consistent.

(b) (i) Suppose that \( T_1 \) and \( T_2 \) are independent unbiased estimators of \( \theta \).

Let \( T = a_1 T_1 + a_2 T_2 \). What determines your choice of \( a_1 \) and \( a_2 \)?

(ii) The following estimates are from independent sources:

\[
\begin{align*}
    \hat{\theta}_1 &= 4.32, \text{se}(\hat{\theta}_1) = 0.47; \quad \hat{\theta}_2 = 4.92, \text{se}(\hat{\theta}_2) = 0.83.
\end{align*}
\]

Derive an estimate and an approximate 95% confidence interval for \( \theta \).

[4 + 6 = 10 marks]

2. (a) A random sample of \( n \) observations is obtained on the random variable \( Y \) which has a lognormal distribution, \( \ell N(\theta, 1) \); the pdf of \( Y \) is given by

\[
f(y | \theta) = \frac{1}{y\sqrt{2\pi}} e^{-\frac{1}{2}(\ln y - \theta)^2} \quad (y > 0).
\]

(i) Find the method of moments estimator of \( \theta \), and derive an expression for its variance.

(ii) Find the maximum likelihood estimator of \( \theta \), and derive an expression for its variance.
(iii) Comment on the consistency, unbiasedness and relative efficiency of these two estimators.

Note: if \( X \overset{d}{=} t \mathcal{N}(\alpha, \beta^2) \) then \( \mu = e^{\alpha + 0.5 \beta^2} \), \( \sigma^2 = e^{2\alpha + \beta^2}(e^{\beta^2} - 1) \); \( \ln X \overset{d}{=} t \mathcal{N}(\alpha, \beta^2) \).

(b) In a two parameter model, with parameters \( \theta \) and \( \phi \), the expected information matrix based on a random sample of \( n \) is given by:

\[
H(\theta, \phi) = n \begin{bmatrix}
\theta + \phi & \frac{\phi}{\theta} \\
\frac{\phi}{\theta} & \frac{\phi}{\theta^2}
\end{bmatrix}.
\]

Specify the minimum variance bound for unbiased estimators of \( \theta \) (i) when \( \phi = \phi_0 \); (ii) when \( \phi \) is unknown. \[8+4=12 \text{ marks}\]

3. (a) Define observed information \( (V) \) and expected information \( (I) \).

(b) Consider a random sample \( Y = (Y_1, Y_2, \ldots, Y_n) \) on \( Y \) for which the distribution depends on the parameter \( \theta \). Let \( T = \psi(Y) \). Show that \( I_T(\theta) \leq I_Y(\theta) \). How does this relate to sufficiency?

(c) A random sample of \( n \) observations is obtained on \( U \overset{d}{=} \exp(\theta) \). Quantify the expected loss of information if all observations greater than \( c \) are censored. Give an expression in terms of \( \theta \) and \( c \).

(d) A random sample is obtained from a population with pdf

\[
f(x; \theta) = \frac{g(x)}{a(\theta)} \quad (0 < x < \theta), \quad [\theta > 0].
\]

(i) Show that \( X_{(n)} \) is sufficient for \( \theta \).

(ii) Show that a 95% confidence interval for \( \theta \) is given by

\[
0.975^{-1/n} G(X_{(n)}) < a(\theta) < 0.025^{-1/n} G(X_{(n)})
\]

where \( G(x) = \int_0^x g(u) du \). \[2+3+3+6=14 \text{ marks}\]

4. (a) Define the score test.

(b) Explain why, in testing \( H_0: \theta = \theta_0 \) against \( H_1: \theta > \theta_0 \), if a uniformly most powerful test exists, then the score test must be the uniformly most powerful test.

(c) For an exponential family model in which the link function is the natural parameter, it is found that

\[
\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^n (y_i - \mu_i)x_{ij},
\]

\[
\frac{\partial^2 \ln L}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^n x_{ij}x_{ik}\sigma_i^2.
\]

Explain how the method of scoring leads to the method of iteratively re-weighted least squares in this case. \[1+3+6=10 \text{ marks}\]
5. Suppose that $Y_i \overset{d}{=} P_n(\lambda_i)$, where $\ln \lambda_i = \alpha + \beta x_i \quad (i = 1, 2, \ldots, n)$.

Note: if $Y \overset{d}{=} P_n(\lambda)$ then $\Pr(Y = k) = e^{-\lambda} \lambda^k/k! \quad (k = 0, 1, 2, \ldots); \quad \mu = \lambda, \sigma^2 = \lambda.$

(a) (i) What motivates this form of model for the mean? Give at least two reasons.

(ii) Show that $(\sum y_i, \sum x_i y_i)$ is sufficient for $(\alpha, \beta)$.

(iii) Derive the score equations to be solved for the MLEs of $\alpha$ and $\beta$; and find the expected information matrix.

(b) The following data are available:

\[
\begin{array}{cccccccc}
x & -1 & -1 & 0 & 0 & 1 & 1 \\
y & 2 & 3 & 4 & 8 & 7 & 11 \\
\ln y & 0.7 & 1.1 & 1.4 & 2.1 & 1.9 & 2.4
\end{array}
\]

(i) Draw a rough scatter plot of $\ln y$ against $x$ and hence obtain reasonable initial estimates of $\alpha$ and $\beta$.

(ii) Use these reasonable initial estimates to perform one iteration of the Newton-Raphson procedure to obtain improved estimates and their standard errors. Comment on how close you think your estimates are to the MLEs.

(iii) The following GLIM output was obtained for this data set

```
[1] ? $units 6$
[1] ? $read x$
[1] ? -1 -1 0 0 1 1
[1] ? $read y$
[1] ? 2 3 4 8 7 11
[1] ? $yvar y$
[1] ? $fit x$
[0] scaled deviance = 2.829 at cycle 3
[0] residual df = 4
[1] ? $fit -x$
[0] scaled deviance = 10.271 (change = +7.442) at cycle 4
[0] residual df = 5 (change = +1 )
```

Comment on the goodness of fit of the model and the significance of the effect of $x$ in the model.

6. Answer any four (4) of the following short questions.

(a) For the log-linear model specified by $[ABC][CD][DE][DF]$, list four (4) of the implied independence relationships.

(b) Explain, briefly, what is meant by overdispersion in the context of logistic regression. Describe one mechanism that may give rise to overdispersion and describe, briefly, what effect overdispersion has on the “usual” tests of

(i) adequacy of fit and

(ii) model comparisons.

(c) Explain what is meant by “quasi-independence”, and use an example to illustrate when, and how, you might fit a quasi-independence model.

(d) The following GLIM output was obtained when fitting a logistic regression model to data on the mortality of male and female tobacco budworms 72 hours after exposure to a range of doses of cypermethrin. In the output G refers to gender (1 = male; 2 = female) and x refers to dose. Give estimates of the “LD(50)” dose (ie the dose that will kill 50% of the budworms within 72 hours) for both male and female tobacco budworms.
620-372  Applied Statistical Analysis

[i] ? $\text{fit G+X}$
[o] scaled deviance = 6.7571 at cycle 3
[o] residual df = 9

[i] ? $\text{dis e}$
[o] estimate s.e. parameter
[o] 1 -2.372 0.3854 1
[o] 2 -1.101 0.3557 G(2)
[o] 3 1.535 0.1890 X

(e) Describe circumstances under which you would consider fitting a uniform association model to a two-way contingency table. The following output was obtained for such a model, give an interpretation of the model in terms of odds ratios.

[i] ? $\text{calc x=s*b}$
[i] ? $\text{fit s+b:+x}$
[o] scaled deviance = 49.494 at cycle 4
[o] residual df = 4

[o] scaled deviance = 5.6389 (change = -43.86) at cycle 4
[o] residual df = 3 (change = -1 )

[i] ? $\text{dis e}$
[o] estimate s.e. parameter
[o] 1 4.299 0.1529 1
[o] 2 -0.8495 0.1798 S(2)
[o] 3 -1.148 0.3266 S(3)
[o] 4 -3.999 0.3890 B(2)
[o] 5 -6.801 0.8209 B(3)
[o] 6 0.7811 0.1430 X

[4+4+4+4=16 marks]

7. The following data were obtained from a study of coronary heart disease, where $N$ is the total number of subjects in each group and $Y$ is the number diagnosed with coronary heart disease. The factor CHOL refers to serum cholesterol in mg/100cc where:

\[
1 = < 200, \ 2 = 200 - 219, \ 3 = 220 - 259, \ 4 = 260+ \]

while the factor BP refers to blood pressure in mm of mercury where:

\[
1 = < 127, \ 2 = 127 - 146, \ 3 = 147 - 166, \ 4 = 167+ \]
Four models have been fitted to these data, GLIM output for which is given below.

```
[i] ? $UNITS 16
[i] ? $DATA Y N$READ 2 119 3 124 etc
[i] ? $CALC CHOL=%GL(4,4) : BP=%GL(4,1)$
[i] ? $FACTOR CHOL 4 BP 4$
[i] ? $yvar y$error b n$
[i] ? $fit$dis e$
[o] scaled deviance = 58.726 at cycle 4
[o] residual df = 15
[o] estimate s.e. parameter
[o] 1  -2.599 0.1081 1
[o] scale parameter 1.000

[i] ? $fit chol$dis e$
[o] scaled deviance = 26.805 at cycle 4
[o] residual df = 12
[o] estimate s.e. parameter
[o] 1  -3.242 0.2942 1
[o] 2  -0.1839 0.4643 CHOL(2)
[o] 3  0.5914 0.3480 CHOL(3)
[o] 4  1.454 0.3391 CHOL(4)
[o] scale parameter 1.000

[i] ? $fit bp$dis e$
[o] scaled deviance = 35.163 at cycle 4
[o] residual df = 12
[o] estimate s.e. parameter
[o] 1  -2.965 0.2293 1
[o] 2  0.03028 0.3003 BP(2)
[o] 3  0.6429 0.3278 BP(3)
[o] 4  1.373 0.3205 BP(4)
[o] scale parameter 1.000
```
620-372 Applied Statistical Analysis

(?fit chol+bp$dis e$)

-8.0762 at cycle 4

-9

estimate    s.e.  parameter
1  -3.482  0.3486  1
2  -0.2080  0.4664  CHOL(2)
3  0.5622  0.3508  CHOL(3)
4  1.344  0.3430  CHOL(4)
5  -0.04146  0.3037  BP(2)
6  0.5324  0.3324  BP(3)
7  1.200  0.3269  BP(4)

scale parameter 1.000

(a) Which of the four models is “best”? Give details of any formal tests that you use in reaching your decision.

(b) Describe briefly (no calculations required) what your chosen model says, if anything, about the relationships between:

(i) coronary heart disease and serum cholesterol levels;
(ii) coronary heart disease and blood pressure;
(iii) serum cholesterol levels and blood pressure.

c) The model with CHOL and BP included as variables, rather than as factors, was fitted to the data and resulted in a scaled deviance of 14.847. What conclusions do you draw from this? [Give details of any formal tests that you use.]

d) The data have been analysed using logistic regression models. An alternative would have been to use log-linear models in a 3-way contingency table with factors CHOL, BP and CHD, where CHD is a factor with 2 levels indicating whether or not subjects have coronary heart disease. For each of the four (logistic regression) models given in the GLIM output, specify the equivalent log-linear model (eg CHOL + BP + CHD).

[2+3+4+4=13 marks]

8. The GLIM output below refers to a study of the relationship between chronic bronchitis, cigarette consumption (CIG) and the smoke level of the locality of the respondent’s home (POLL). A total of 212 subjects participated in the study.

(a) (i) What, if anything, can you say about the adequacy of fit of each of the four models. Give reasons for your answer(s).

(ii) Carry out tests to determine whether the two covariates, cigarette consumption (CIG) and smoke level (POLL) are significantly related to chronic bronchitis. Show all your workings.

(b) Using the model specified by ($fit CIG + POLL$):

(i) Estimate the odds ratios associated with “CIG” and “POLL”, and explain exactly what is meant by each of the odds ratios.

(ii) Find a 95% confidence interval for the odds ratio for “CIG”.

(iii) Estimate the probability that an individual who smokes 10 cigarettes per day (CIG = 10) and lives in a locality for which POLL = 50 suffers from chronic bronchitis.

[6+8=14 marks]

- 13 -
The data below come from a survey of opinion on the Vietnam war conducted among 1st – 3rd year undergraduate students at the University of North Carolina in 1967.

The policies listed were:
A – defeat power of Vietnam by widespread bombing and land invasion
B – follow the present policy
C – withdraw troops to strong points and open negotiations on elections involving the Vietcong
D – immediate withdrawal of all U.S. troops.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Year</th>
<th>Opinion (op)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Male</td>
<td>175</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>120</td>
</tr>
<tr>
<td>Female</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>29</td>
</tr>
</tbody>
</table>
(a) Of the seven models that were fitted, four of them are probably not worth considering (even if they did provide an adequate fit to the data). Which ones, and why?

(b) For each of the following “conclusions” state, with reasons, whether or not it is reasonable for these data.

(i) Opinion is independent of year level and gender.
(ii) Given the year level, opinion is independent of gender.
(iii) Given gender, opinion is independent of year level.
(iv) The ratio of the odds between any two opinions (B and C say) for males and females is the same for each year level.

Which, if any, of these conclusions do you consider to be the most appropriate based on the GLIM output provided? Justify your answer.

(c) It has been suggested that it might be useful to include year as a variable (as yr) rather than as a factor in some way. Write down the form of such a model which is likely to provide an adequate fit to the data, and give an interpretation of the model.

(d) Write down the scaled deviance and the degrees of freedom that would be obtained by fitting the model year+op to the 2–way table obtained by collapsing over gender. State, with reasons, whether collapsing over gender would be a reasonable thing to do.

[4+5+4+4=17 marks]

---

2000

1. The cosine of the angle at which electrons are emitted in muon decay has a distribution with pdf given by

\[ f(x; \theta) = \frac{1}{2}(1 + \theta x) \quad (-1 < x < 1); \quad [-1 \leq \theta \leq 1] \]

The parameter \( \theta \) is related to polarisation. You may use without proof the results that

\[ \mathbb{E}(X) = \frac{1}{3} \theta, \ \text{var}(X) = \frac{1}{5}(3-\theta^2) \text{ and } I(\theta) \approx (\frac{1}{3} + \frac{1}{5}\theta^2)n \text{ (which is a satisfactory approximation for } |\theta| < \frac{3}{4}). \]

(a) Suppose that we have a random sample of \( n \) observations on \( X \).

(i) Find the form of the most powerful test for testing \( H_0 : \theta = 0 \) against \( H_1 : \theta = 1 \); and explain how the required constant could be determined.

(ii) Is there a uniformly most powerful test for testing \( H_0 : \theta = 0 \) against \( H_1 : \theta > 0 \)? Either specify it, or explain why such a test does not exist.
(iii) Specify the form of the locally most powerful test for testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$ and give an approximate value for the constant required for a test of size 0.05.

(b) Suppose that we have a random sample of three observations on $X$:

$$x_1 = 0.1, \quad x_2 = 0.6, \quad x_3 = -0.4.$$ 

(i) Find the method of moments estimate (MME) and its standard error for these data.

(ii) Find the equation satisfied by the maximum likelihood estimate (MLE) of $\theta$. Using the MME as initial value, perform one iteration of the method of scoring. Give an approximate value for the MLE and its standard error for these data. [15 marks]

2. (a) In a two parameter model, with parameters $\gamma$ and $\beta$, it is found that the expected information matrix is given by

$$H(\gamma, \beta) = n \begin{bmatrix} \gamma^2 + \gamma \beta \beta^2 & \beta^2 \beta / \gamma \end{bmatrix}.$$ 

Specify the minimum variance bound for unbiased estimators of $\gamma$

(i) when $\beta = 1$;

(ii) when $\beta$ is unknown.

(b) The Weibull distribution, $W(\alpha, \theta)$ has pdf and cdf given by:

$$f(x; \alpha, \theta) = \alpha \theta x^{\theta - 1} e^{-\alpha x^\theta}, \quad F(x; \alpha, \theta) = 1 - e^{-\alpha x^\theta} \quad (x > 0)$$

The following three independent observations are obtained:

$$x_1 = e^{-1}, \quad x_2 = 1, \quad x_3 > e^{0.5}$$

Show that for these data

$$\ln L = 2 \ln \alpha + 2 \ln \theta - (\theta - 1) - \alpha (e^{-\theta} + 1 + e^{0.5 \theta}),$$

and find

$$\frac{\partial \ln L}{\partial \alpha}, \quad \frac{\partial \ln L}{\partial \theta}, \quad \frac{\partial^2 \ln L}{\partial \alpha^2}, \quad \frac{\partial^2 \ln L}{\partial \alpha \partial \theta} \quad \text{and} \quad \frac{\partial^2 \ln L}{\partial \theta^2}.$$ 

Starting with $\hat{\alpha}_0 = 0.6, \hat{\theta}_0 = 1.4$ perform one iteration of the method of scoring to obtain $\hat{\alpha}_1$ and $\hat{\theta}_1$.

Assuming that $\hat{\alpha}_1$ and $\hat{\theta}_1$ give reasonable approximations for $\hat{\alpha}$ and $\hat{\theta}$, obtain approximate values for $\text{se}(\hat{\alpha})$ and $\text{se}(\hat{\theta})$.

(c) For an exponential family model in which the link function is the natural parameter, it is found that

$$\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^n (y_i - \mu_i) x_{ij}$$

$$\frac{\partial^2 \ln L}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^n x_{ij} x_{ik} \sigma_i^2$$

Explain how the method of scoring leads to the method of iteratively re-weighted least squares in this case. [15 marks]
3. The following data were obtained in a survey of attitudes to women staying at home with respect to sex and educational level. The education variable has 12 levels corresponding to years of study.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>27</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>75</td>
<td>49</td>
</tr>
<tr>
<td>9</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>10</td>
<td>32</td>
<td>45</td>
</tr>
<tr>
<td>11</td>
<td>36</td>
<td>59</td>
</tr>
<tr>
<td>12</td>
<td>115</td>
<td>245</td>
</tr>
<tr>
<td>13</td>
<td>31</td>
<td>70</td>
</tr>
<tr>
<td>14</td>
<td>28</td>
<td>79</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>23</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>110</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>29</td>
</tr>
</tbody>
</table>

Using, \texttt{SEX} as a factor with 2 levels (1 = male; 2 = female), \texttt{EDUC} as a factor with 12 levels and \texttt{E} as a variable (taking values 6 to 17), various logistic regression models were fitted as indicated in the following table.

[Note: “Agree” was used as the response category for the logistic regression.]

<table>
<thead>
<tr>
<th>Model</th>
<th>Scaled Deviance</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. –</td>
<td>337.9</td>
<td>23</td>
</tr>
<tr>
<td>2. \texttt{SEX}</td>
<td>337.5</td>
<td>22</td>
</tr>
<tr>
<td>3. \texttt{EDUC}</td>
<td>15.3</td>
<td>12</td>
</tr>
<tr>
<td>4. \texttt{SEX + EDUC}</td>
<td>15.2</td>
<td>11</td>
</tr>
<tr>
<td>5. \texttt{E}</td>
<td>25.2</td>
<td>22</td>
</tr>
<tr>
<td>6. \texttt{SEX + E}</td>
<td>25.1</td>
<td>21</td>
</tr>
<tr>
<td>7. \texttt{SEX*E}</td>
<td>20.2</td>
<td>20</td>
</tr>
<tr>
<td>8. \texttt{EDUC + SEX + SEX.E}</td>
<td>9.9</td>
<td>10</td>
</tr>
</tbody>
</table>

(a) With 8 models there are \(8 \times 7 = 28\) pairs of models [(1) and (2), (1) and (3), \ldots, (7) and (8)]. Of these 28 pairs, there are five (5) that cannot legitimately be compared. List any two (2) of the five pairs; there is no need to justify your answer.

(b) The only interaction term that has been considered is that between \texttt{SEX} and \texttt{E} (education level treated as a variable). Explain why it would not be helpful to include the \texttt{SEX.EDUC} interaction in any of the models.

(c) Which of the above models do you consider to be the most appropriate for these data? Show enough of your working to make clear how you have reached your decision.

(d) The following output is from the model \texttt{SEX + E}, which may or may not be the “best” model.

(i) Explain, in terms of odds ratios, what this model implies about the effects of sex and education on attitudes to women staying at home.

(ii) Find (also) a 95% confidence interval for the odds ratio for \texttt{SEX}. 

\[ \]
(iii) Use the model to find an estimate of the probability that a female with 10 years of education agrees that women should stay at home.

\[
\text{[i]} \quad \text{estimate} \quad \text{s.e.} \quad \text{parameter}
\]

\[
\begin{array}{ccc}
1 & 2.900 & 0.2192 \\
2 & -0.03344 & 0.08662 \\
3 & -0.3028 & 0.01855
\end{array}
\]

\[2 + 2 + 4 + 6 = 14 \text{ marks}\]

4. Data were collected by the National Opinion Research Centre on attitudes towards abortion during each of the years 1972, 1973 and 1974. Respondents were identified by their years of education and their religious group. The data presented below were from Southern Protestants, that is, those that live in or south of Texas, Oklahoma, Arkansas, Kentucky, West Virginia, Maryland, and Delaware. Attitudes towards abortion were determined by whether the respondent thought that legal abortions should be available under each of the following three sets of circumstances:

(a) a strong chance of a serious birth defect,
(b) the woman’s health is threatened,
(c) the pregnancy was the result of rape.

A negative response in the table consists of negative responses to all circumstances. A positive response is three positives. A mixed response is any other pattern.

<table>
<thead>
<tr>
<th>Year</th>
<th>Years of Education</th>
<th>Negative</th>
<th>Mixed</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>1974</td>
<td>0–8</td>
<td>1</td>
<td>19</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>9–12</td>
<td>5</td>
<td>21</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>12+</td>
<td>2</td>
<td>11</td>
<td>87</td>
</tr>
<tr>
<td>1973</td>
<td>0–8</td>
<td>4</td>
<td>16</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>9–12</td>
<td>6</td>
<td>29</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>12+</td>
<td>1</td>
<td>4</td>
<td>82</td>
</tr>
<tr>
<td>1972</td>
<td>0–8</td>
<td>9</td>
<td>17</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>9–12</td>
<td>6</td>
<td>10</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>12+</td>
<td>1</td>
<td>8</td>
<td>68</td>
</tr>
</tbody>
</table>

The following scaled deviances and degrees of freedom were obtained from these data where YR denotes year, ED denotes education level and ATT denotes attitude to abortion.

<table>
<thead>
<tr>
<th>Model</th>
<th>Scaled Deviance</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>YR+ED+ATT</td>
<td>78.32</td>
<td>20</td>
</tr>
<tr>
<td>YR+ED+ATT</td>
<td>73.54</td>
<td>16</td>
</tr>
<tr>
<td>YR*ATT+ED</td>
<td>72.88</td>
<td>16</td>
</tr>
<tr>
<td>YR+ED+ATT</td>
<td>21.81</td>
<td>16</td>
</tr>
<tr>
<td>YR+ED+YR*ATT</td>
<td>68.09</td>
<td>12</td>
</tr>
<tr>
<td>YR+ED+ED*ATT</td>
<td>17.02</td>
<td>12</td>
</tr>
<tr>
<td>YR+ATT+ED*ATT</td>
<td>16.36</td>
<td>12</td>
</tr>
<tr>
<td>YR+ATT+ED*ATT+YR+ED</td>
<td>11.39</td>
<td>8</td>
</tr>
</tbody>
</table>

(a) Indicate which model corresponds to each of the following interpretations and state, with reasons, whether the interpretation is reasonable (at the 5% level).
(i) Given the year, attitude is independent of level of education.

(ii) Given the level of education, attitude is independent of year.

(b) It is suggested that the interpretation of these data would be simpler if the (three-way) table were collapsed to form two $3 \times 3$ tables; one for attitude versus education, and one for attitude versus year. For both of these $3 \times 3$ tables, give the scaled deviance and degrees of freedom that would be obtained if the “no association” model is fitted, and state, with reasons, whether collapsing would be a reasonable thing to do.

[4 + 5 = 9 marks]

5. Data were obtained on the education levels of married couples who participated in the 1972 General Social Survey (in the USA).

<table>
<thead>
<tr>
<th>Husband</th>
<th>Wife</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–11</td>
<td>175</td>
</tr>
<tr>
<td>12</td>
<td>180</td>
</tr>
<tr>
<td>13–15</td>
<td>104</td>
</tr>
<tr>
<td>16+</td>
<td>52</td>
</tr>
</tbody>
</table>

The following GLIM output was obtained from these data, where $H$ and $W$ are factors denoting the husband’s and wife’s education, respectively, with levels ($1–4$), and $X$ is the product of $H$ and $W$.

```
? $units 16
? $data y$read 253 101 22 4 175 180 104 52 25 43 41 4 14 20 69
? $calc H=%gl(4,1):W=%gl(4,4)$y$var y$fact H 4 W 4$error p$
? $calc X=H*W$fit H+W:+X$

scaled deviance = 415.90 at cycle 4
residual df = 9

scaled deviance = 11.472 (change = -404.4) at cycle 3
residual df = 8 (change = -1 )
```

```
? $dis e$
<table>
<thead>
<tr>
<th>estimate</th>
<th>s.e.</th>
<th>parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.734</td>
<td>0.06527</td>
</tr>
<tr>
<td>2</td>
<td>-1.576</td>
<td>0.1029</td>
</tr>
<tr>
<td>3</td>
<td>-3.754</td>
<td>0.1970</td>
</tr>
<tr>
<td>4</td>
<td>-5.927</td>
<td>0.3383</td>
</tr>
<tr>
<td>5</td>
<td>-0.9849</td>
<td>0.09896</td>
</tr>
<tr>
<td>6</td>
<td>-3.992</td>
<td>0.2198</td>
</tr>
<tr>
<td>7</td>
<td>-6.712</td>
<td>0.3934</td>
</tr>
<tr>
<td>8</td>
<td>0.7545</td>
<td>0.04724</td>
</tr>
</tbody>
</table>

scale parameter 1.000
```

```
? $tprint(s=1) %fv H;W$
```

– 19–
(a) From the output given above, it is possible to carry out three formal tests of hypotheses, two of which lead to the conclusion that there is significant association between the education levels of husbands and wives, without reference to either of the other tests. Specify which two of the tests lead to this conclusion. There is no need to justify your answer.

(b) Explain, in terms of odds ratios, the meaning of the model “H + W + X”.

(c) Suggest another (non-saturated) model that might be expected to provide a good fit to these data, and give an interpretation of your suggested model.

[2 + 2 + 3 = 7 marks]

1999

1. Kihlberg, Narragon and Campbell (1964) [Automobile crash injury in relation to car size. Cornell Aero. Lab. Report No. VJ–1823–R11.] report on severity of drivers’ injuries in car accidents along with the type of accident and whether or not the driver was ejected from the vehicle during the accident. We consider the results for small cars only.

<table>
<thead>
<tr>
<th>Driver Ejected</th>
<th>Accident Type</th>
<th>Injury Not Severe</th>
<th>Injury Severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Collision</td>
<td>350</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>Rollover</td>
<td>60</td>
<td>112</td>
</tr>
<tr>
<td>Yes</td>
<td>Collision</td>
<td>26</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Rollover</td>
<td>19</td>
<td>80</td>
</tr>
</tbody>
</table>

Using GLIM, logistic regression models, with severe injuries as response, were fitted and the following output obtained.

[1] ? $units 4
[1] ? $data y n $read 150 500 23 49 112 172 80 99$
[1] ? $calc eject=%gl(2,1):type=%gl(2,2)$
[1] ? $yvar y$error b n$
[1] ? $factor eject 2 type 2$
[1] ?
(a) (i) Which model or models provide an adequate fit to the data, and what is the interpretation of this model or models?

(ii) For the model you consider to be most appropriate, interpret the parameter estimates (other than that for the parameter “1”) in terms of odds ratios, and give 95% confidence intervals for the odds ratios. What are the implications of these odds ratios, if any, for the wearing of seat belts?

(iii) Using the model you consider to be most appropriate, obtain an estimate of the probability that a driver ejected from a (small) car in a rollover accident will suffer a serious injury.

(b) For each of the four (4) logistic regression models fitted above, it is possible to fit an equivalent log-linear model. Specify (all four of) these equivalent models using the factors “EJECT”, “TYPE” and “SEVERITY”.

[1] fit $dis e$
  [o] scaled deviance = 128.96 at cycle 3
  [o] residual df = 3
  [o] estimate   s.e.   parameter
  [o]  1  -0.2204  0.07027  1

[1] fit eject$dis e$
  [o] scaled deviance = 82.644 at cycle 3
  [o] residual df = 2
  [o] estimate   s.e.   parameter
  [o]  1  -0.4478  0.07909  1
  [o]  2   1.276   0.1954   EJECT(2)

[1] fit type$dis e$
  [o] scaled deviance = 13.397 at cycle 3
  [o] residual df = 2
  [o] estimate   s.e.   parameter
  [o]  1  -0.7763  0.09187  1
  [o]  2   1.664  0.1622   TYPE(2)

[1] fit eject+type$dis e$
  [o] scaled deviance = 0.043326 at cycle 2
  [o] residual df = 1
  [o] estimate   s.e.   parameter
  [o]  1  -0.8520  0.09507  1
  [o]  2   0.7694  0.2126   EJECT(2)
  [o]  3   1.489   0.1684   TYPE(2)

Scale parameter 1.000
2. The following data refer to graduate admissions at the University of California, Berkeley, for the six largest departments.

<table>
<thead>
<tr>
<th>Dept</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admitted</td>
<td>Rejected</td>
</tr>
<tr>
<td>A</td>
<td>512</td>
<td>313</td>
</tr>
<tr>
<td>B</td>
<td>353</td>
<td>207</td>
</tr>
<tr>
<td>C</td>
<td>120</td>
<td>205</td>
</tr>
<tr>
<td>D</td>
<td>138</td>
<td>279</td>
</tr>
<tr>
<td>E</td>
<td>53</td>
<td>138</td>
</tr>
<tr>
<td>F</td>
<td>22</td>
<td>351</td>
</tr>
</tbody>
</table>

(a) From these data the following scaled deviances and residual df were obtained, where $D$ refers to the department, $G$ to gender and $A$ to admission.

<table>
<thead>
<tr>
<th>Model</th>
<th>scaled deviance</th>
<th>residual df</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D+G+A$</td>
<td>2097.7</td>
<td>16</td>
</tr>
<tr>
<td>$D+G*A$</td>
<td>2004.2</td>
<td>15</td>
</tr>
<tr>
<td>$D*G+A$</td>
<td>877.1</td>
<td>11</td>
</tr>
<tr>
<td>$D*A+G$</td>
<td>1242.4</td>
<td>11</td>
</tr>
<tr>
<td>$D<em>A+G</em>A$</td>
<td>1148.9</td>
<td>10</td>
</tr>
<tr>
<td>$D<em>G+G</em>A$</td>
<td>783.6</td>
<td>10</td>
</tr>
<tr>
<td>$D<em>G+D</em>A$</td>
<td>21.7</td>
<td>6</td>
</tr>
<tr>
<td>$D<em>G+D</em>A+G*A$</td>
<td>20.2</td>
<td>5</td>
</tr>
</tbody>
</table>

State which model corresponds to each of the following interpretations and state, with reasons, whether the hypothesis implied by the interpretation would be accepted or rejected (at the 5% level).

(i) Given the department, admission is independent of gender.
(ii) The ratio of the odds for the admission of males and females is the same for each department.

(b) The same set of models was also fitted to the data from departments B–F (only) with the following results.

<table>
<thead>
<tr>
<th>Model</th>
<th>scaled deviance</th>
<th>residual df</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D+G+A$</td>
<td>1292.9</td>
<td>13</td>
</tr>
<tr>
<td>$D+G*A$</td>
<td>1254.3</td>
<td>12</td>
</tr>
<tr>
<td>$D*G+A$</td>
<td>539.5</td>
<td>9</td>
</tr>
<tr>
<td>$D*A+G$</td>
<td>756.1</td>
<td>9</td>
</tr>
<tr>
<td>$D<em>A+G</em>A$</td>
<td>717.5</td>
<td>8</td>
</tr>
<tr>
<td>$D<em>G+G</em>A$</td>
<td>500.9</td>
<td>8</td>
</tr>
<tr>
<td>$D<em>G+D</em>A$</td>
<td>2.68</td>
<td>5</td>
</tr>
<tr>
<td>$D<em>G+D</em>A+G*A$</td>
<td>2.56</td>
<td>4</td>
</tr>
</tbody>
</table>

(i) What GLIM output, in addition to that given in (a) above, may have been used in making the decision to omit department A from the analysis, and what might the output have shown?
(ii) What conclusions(s) do you draw about departments B–F?
(iii) What would be the scaled deviance and residual df if the “no association” model were to be fitted to the data (for departments B–F, only) collapsed over the five departments? State, with reasons, whether collapsing over (the five) departments is reasonable.
3. A sample of men between the ages of 40 and 59 were taken from the city of Framingham, Massachusetts. The men were cross-classified by their serum cholesterol and systolic blood pressure. The data below were obtained from those men who did not develop coronary heart disease during a 6-year follow-up period.

<table>
<thead>
<tr>
<th>Cholesterol (in mg/100 cc)</th>
<th>Blood Pressure (in mm Hg)</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 200</td>
<td>&lt; 127</td>
<td>47</td>
</tr>
<tr>
<td>200–219</td>
<td>127–146</td>
<td>43</td>
</tr>
<tr>
<td>200–259</td>
<td>147–166</td>
<td>68</td>
</tr>
<tr>
<td>≥ 260</td>
<td>167+</td>
<td>46</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>388</td>
</tr>
</tbody>
</table>

The following GLIM output was obtained where $B$ and $C$ are factors which refer to Blood Pressure and Cholesterol, respectively, and $x$ is a variable, the product of $B$ and $C$.

```
[1] ? $units 16
[1] ? $calc x=B*C$
[1] ?
[1] ? $yvar y$error p$
[1] ?
[1] ? $fit B+C$dis e$
[o] scaled deviance = 20.378 at cycle 3
[o] residual df = 9

[o] estimate s.e. parameter
[o] 1 4.567 0.07089 1
[o] 2 0.3062 0.06689 B(2)
[o] 3 -0.6429 0.08648 B(3)
[o] 4 -1.190 0.1051 B(4)
[o] 5 -0.2215 0.08557 C(2)
[o] 6 0.3577 0.07440 C(3)
[o] 7 -0.2256 0.08566 C(4)
[o] scale parameter 1.000

[1] ? $tprint(s=1) %rs C;B$
[0] +------------------------------------+
[0] | B | 1 | 2 | 3 | 4 |+
[0] |---|---|---|---|---|
[0] | 1 | 2.1100 -0.8562 -0.5100 -1.3463 |
[0] | 2 | 0.8924 -0.6646 0.3817 -0.7156 |
[0] | 3 | -1.5934 1.6067 -0.5169 0.1735 |
[0] | 4 | -1.1233 -0.5264 0.8803 1.9918 |
[0] +------------------------------------+

[1] ?
[1] ? $fit +x$dis e$
[o] scaled deviance = 7.4291 (change = -12.95) at cycle 3
```
residual df = 8  (change = -1 )

<table>
<thead>
<tr>
<th>estimate</th>
<th>s.e.</th>
<th>parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.614</td>
<td>0.06988</td>
<td>1</td>
</tr>
<tr>
<td>0.05164</td>
<td>0.09654</td>
<td>B(2)</td>
</tr>
<tr>
<td>-1.164</td>
<td>0.1698</td>
<td>B(3)</td>
</tr>
<tr>
<td>-1.991</td>
<td>0.2522</td>
<td>B(4)</td>
</tr>
<tr>
<td>-0.4253</td>
<td>0.1015</td>
<td>C(2)</td>
</tr>
<tr>
<td>-0.05894</td>
<td>0.1363</td>
<td>C(3)</td>
</tr>
<tr>
<td>-0.8645</td>
<td>0.1985</td>
<td>C(4)</td>
</tr>
<tr>
<td>0.1044</td>
<td>0.02925</td>
<td>X</td>
</tr>
</tbody>
</table>

scale parameter 1.000

---

What do you conclude about the association between Cholesterol and Blood Pressure, for these men?

Explain, in terms of odds ratios, the meaning of the (GLIM) model “B + C + x”.