

McGill University

Part A Examination in Statistics

Methodology Paper

Department of Mathematics & Statistics

Date: 22 August 2003

Instructions

- Calculators are allowed; tables are supplied at the end of this exam.
- Answer two questions out of Section L. Only two questions will be marked.
- Answer two questions out of Section G. Only two questions will be marked.
- If you do not indicate which questions you wish to have marked, the questions will be marked in the order in which they appear in the answer book until the quota has been reached.
- All questions are weighted equally (20 marks each).
- Each question will be assessed independently by at least two members of the statistics group, and the final result determined after discussion within the Part A Exam Subcommittee.
- Good luck!

This exam comprises the cover and 6 questions on 32 pages, plus tables.

V3 20030821
Section L

Answer two questions out of questions L1 (page 1), L2 (page 12) and L3 (page 20).

L1. The age of abalone (a snail-like mollusc) is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Therefore, data were collected for 4175 abalone for purposes of modelling the age using other covariates. The ranges of the continuous values have been scaled by dividing by 200. The data are described in the table below:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>categorical</td>
<td>M, F, and I (infant)</td>
</tr>
<tr>
<td>Length</td>
<td>continuous, (\geq 0)</td>
<td>Longest shell measurement (in mm)</td>
</tr>
<tr>
<td>Diameter</td>
<td>continuous, (\geq 0)</td>
<td>Measured perpendicular to length (in mm)</td>
</tr>
<tr>
<td>Height</td>
<td>continuous, (\geq 0)</td>
<td>With meat in shell (in mm)</td>
</tr>
<tr>
<td>Wholeweight</td>
<td>continuous, (\geq 0)</td>
<td>Whole Abalone (in g)</td>
</tr>
<tr>
<td>ShuckedWeight</td>
<td>continuous, (\geq 0)</td>
<td>Weight of meat (in g)</td>
</tr>
<tr>
<td>VisceraWeight</td>
<td>continuous, (\geq 0)</td>
<td>Gutted weight after bleeding (in g)</td>
</tr>
<tr>
<td>ShellWeight</td>
<td>continuous, (\geq 0)</td>
<td>After being dried (in g)</td>
</tr>
<tr>
<td>Rings</td>
<td>integer, (\geq 0)</td>
<td>(+1.5) gives the age in years</td>
</tr>
</tbody>
</table>

We consider linear models for \(\text{Rings}\) (and functions of \(\text{Rings}\)) on the other covariates or on transformations of them. We assume independence of observations, a common variance, and Normality of errors. **Wherever testing is required, be absolutely clear about your hypotheses, your test statistic and its distribution. So long as these aspects are clear, you can use any reasonable shorthand to answer the questions.**

(a) [5 marks]

Figure 1 shows scatterplots of the response variable, \(\text{Rings}\), and all of the weight covariates. Figure 2 shows scatterplots of the response variable, \(\text{Rings}\), and all of the weight covariates after a log transformation. **In seven sentences or fewer**, explain at least 2 potential problems that might result from fitting a linear model for \(\text{Rings}\) containing all of the weight covariates based on the information in the pairwise scatterplots. Note whether any of these problems would be solved by transforming the response (to \(\log(\text{Rings})\)) and/or the covariates.

*(Question L1 continues on next page.)*
L1. - continued

(b) [5 marks]
Refer to the code and output for part (b) below:

i. Test for an effect of each covariate on $\log(\text{Rings})$, by itself.

ii. Test for an effect of each covariate on $\log(\text{Rings})$, except for Sex, allowing for all the others.

iii. Test for an effect of all the covariates simultaneously on $\log(\text{Rings})$.

(c) [5 marks]
Consider Figures 3 and 4. Evaluate the assumptions of the linear model and suggest (but do not perform) additional analyses to further test and possibly correct any suspect assumptions.

(d) [2 marks]
Refer to the backward step-wise regression results below for part (d). In five lines or fewer, explain the discrepancy between the results from the tests in part (b) and the results of the backward step-wise regression with respect to the covariates deleted from the model.

(e) [3 marks]
Now use the reduced model selected by the backward stepwise regression in part (d) as a starting point and also refer to the output below for part (e) and to Figure 5.

i. Give interpretations for the coefficients of the interaction term in the new model in part (e) below.

ii. Is there any evidence for the addition of the interaction terms to the model?

(R code and output for Question L1 start on next page.)
R code and output for Question L1:

```r
# ********** Code for part (b) **********
> deviance(glm(log(Rings)~1))
[1] 426.1015

> deviance(glm(log(Rings)~log(Wholeweight)))
[1] 219.7579

> deviance(glm(log(Rings)~log(ShuckedWeight)))
[1] 257.8

> deviance(glm(log(Rings)~log(VisceraWeight)))
[1] 229.1688

> deviance(glm(log(Rings)~log(ShellWeight)))
[1] 192.7262

> deviance(glm(log(Rings)~log(Length)))
[1] 230.5402

> deviance(glm(log(Rings)~log(Diameter)))
[1] 222.5500

> deviance(glm(log(Rings)~log(Height)))
[1] 217.9372

> deviance(glm(log(Rings)~Sex))
[1] 323.0519

> deviance(glm(log(Rings)~factor(Sex)))
[1] 323.0519
```

(R code and output for Question L1 continue on next page.)
> summary(glm(log(Rings)~log(Wholeweight)+log(ShuckedWeight) + log(VisceraWeight)+log(ShellWeight)+log(Length)+log(Diameter)+ log(Height)+Sex))

               Estimate Std. Error t value Pr(>|t|)  
(Intercept)   2.436963   0.076355  31.916  < 2e-16 ***  
log(Wholeweight) 0.539691   0.049604  10.880  < 2e-16 ***  
log(ShuckedWeight) -0.566420   0.025045 -22.616  < 2e-16 ***  
log(VisceraWeight)  -0.077381   0.019059  -4.060 5.00e-05 ***  
log(ShellWeight)   0.373229   0.025592  14.584  < 2e-16 ***  
log(Length)      -0.301817   0.083556  -3.612  0.000307 ***  
log(Diameter)    0.170492   0.076928   2.216  0.026728 *  
log(Height)      0.120443   0.025759   4.676  3.02e-06 ***  
SexI             -0.051299   0.008994  -5.703  1.26e-08 ***  
SexM             0.001091   0.007246   0.151  0.880338  

(Dispersion parameter for gaussian family taken to be 0.03658178)

Null deviance: 426.10  on 4174  degrees of freedom  
Residual deviance: 152.36  on 4165  degrees of freedom

# ********** Code for parts (d) and (e)**********
> stepAIC(large.mod, scope=list(lower=glm(log(Rings)~1), + direction=c("backward")))

Start:  AIC= -1888.26
log(Rings) ~ log(Wholeweight) + log(ShuckedWeight) + log(VisceraWeight) + log(ShellWeight) + log(Length) + log(Diameter) + log(Height) + Sex

            Df Deviance    AIC
<none>                1 152.54  -1891.68
- log(Diameter)    1 152.54  -1891.68
- log(Length)      1 152.84  -1883.54
- log(VisceraWeight) 1 152.97  -1880.11
- log(Height)      1 153.16  -1874.74
- Sex              2 153.94  -1861.91
- log(Wholeweight) 1 156.69  -1779.60
- log(ShellWeight)  1 160.14  -1688.67
- log(ShuckedWeight) 1 171.07  -1413.00

(R code and output for Question L1 continue on next page.)
Step: AIC= -1891.68
\[
\text{log(Rings)} \sim \text{log(Wholeweight)} + \text{log(ShuckedWeight)} + \text{log(VisceraWeight)} + \text{log(ShellWeight)} + \text{log(Length)} + \text{log(Height)} + \text{Sex}
\]

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>152.84</td>
<td>-1891.86</td>
</tr>
<tr>
<td>- \text{log(Length)}</td>
<td>1</td>
<td>152.54</td>
</tr>
<tr>
<td>- \text{log(VisceraWeight)}</td>
<td>1</td>
<td>153.15</td>
</tr>
<tr>
<td>- \text{log(Height)}</td>
<td>1</td>
<td>153.36</td>
</tr>
<tr>
<td>- Sex</td>
<td>2</td>
<td>154.15</td>
</tr>
<tr>
<td>- \text{log(Wholeweight)}</td>
<td>1</td>
<td>156.98</td>
</tr>
<tr>
<td>- \text{log(ShellWeight)}</td>
<td>1</td>
<td>161.20</td>
</tr>
<tr>
<td>- \text{log(ShuckedWeight)}</td>
<td>1</td>
<td>171.12</td>
</tr>
</tbody>
</table>

Step: AIC= -1891.86
\[
\text{log(Rings)} \sim \text{log(Wholeweight)} + \text{log(ShuckedWeight)} + \text{log(VisceraWeight)} + \text{log(ShellWeight)} + \text{log(Height)} + \text{Sex}
\]

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>152.84</td>
<td>-1891.86</td>
</tr>
<tr>
<td>- \text{log(VisceraWeight)}</td>
<td>1</td>
<td>153.58</td>
</tr>
<tr>
<td>- \text{log(Height)}</td>
<td>1</td>
<td>153.73</td>
</tr>
<tr>
<td>- Sex</td>
<td>2</td>
<td>154.71</td>
</tr>
<tr>
<td>- \text{log(Wholeweight)}</td>
<td>1</td>
<td>157.02</td>
</tr>
<tr>
<td>- \text{log(ShellWeight)}</td>
<td>1</td>
<td>161.23</td>
</tr>
<tr>
<td>- \text{log(ShuckedWeight)}</td>
<td>1</td>
<td>173.03</td>
</tr>
</tbody>
</table>

Call: \text{glm(formula = log(Rings) \sim log(Wholeweight) + log(ShuckedWeight) + log(VisceraWeight) + log(ShellWeight) + log(Height) + Sex)}

Coefficients:

\[
\begin{align*}
\text{(Intercept)} & : 2.450191 \\
\text{log(Wholeweight)} & : 0.522231 \\
\text{log(ShellWeight)} & : 0.367196 \\
\text{log(VisceraWeight)} & : -0.577179 \\
\text{log(Height)} & : 0.126745 \\
\text{Sex}\text{I} & : -0.084843 \\
\text{Sex}\text{M} & : 0.055245 \\
\end{align*}
\]

Degrees of Freedom: 4174 Total (i.e. Null); 4167 Residual
Null Deviance: 426.1 Residual Deviance: 152.8 AIC: -1943

\text{(R code and output for Question \textbf{L1} continue on next page.)}
# ********** Code for part (e)**********

> summary(glm(formula = log(Rings) ~ Sex + Sex*log(Wholeweight) + log(ShuckedWeight) + log(VisceraWeight) + log(ShellWeight) + log(Height)))

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 2.415670 | 0.069450   | 34.783  | < 2e-16  ***|
| SexI       | -0.022401| 0.010188   | -2.199  | 0.02796  * |
| SexM       | 0.001542 | 0.007326   | 0.210   | 0.83329  |
| log(Wholeweight) | 0.484046 | 0.049451   | 9.788   | < 2e-16  ***|
| log(ShuckedWeight) | -0.580027| 0.024468   | -23.705 | < 2e-16  ***|
| log(VisceraWeight) | -0.084551| 0.018803   | -4.497  | 7.09e-06  ***|
| log(ShellWeight) | 0.353986 | 0.024233   | 14.607  | < 2e-16  ***|
| log(Height)  | 0.119704 | 0.025602   | 4.676   | 3.02e-06  ***|
| SexI:log(Wholeweight) | 0.082634 | 0.012521   | 6.600   | 4.64e-11  ***|
| SexM:log(Wholeweight) | 0.041281 | 0.013172   | 3.134   | 0.00174  ** |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.03626683)

Null deviance: 426.10 on 4174 degrees of freedom
Residual deviance: 151.05 on 4165 degrees of freedom

(Figures for Question L1 start on next page.)
Figure 1: Pairwise scatterplots of weight variables and Rings variable for Part (a)
Figure 2: Pairwise scatterplots of log(weight) variables and log(Rings) variable for Part (b)
Figure 3: Residual diagnostic plots for Part (c)
Figure 4: Residual diagnostic plots for Part (c)
Figure 5: Scatterplots of log(Rings) by log(Whole Weight) and Gender for Part (e)
L2. A total of 4000 adults were surveyed for television advertisements efficiency. Regular product users were asked to cite a commercial they had seen for that product category in the past week. Data include the amount of money spent in advertising (variable SPEND, in millions of US$) and the estimated number of retained impressions in the U.S.A. population (variable MILIMP, in millions of impressions per week). The data are shown for 21 brands in the R output below.

We wish to model MILIMP on SPEND. We assume $N(0, \sigma^2)$ independent errors on outcome variable MILIMP throughout the models below. Refer to the R code and output to answer questions (a) to (f).

(a) [2 marks]
Refer to model $M_A : E[MILIMP] = \beta_0 + \beta_1 \times SPEND$ in the R output for part (a) below. Is there an effect of SPEND on MILIMP?

(b) [2 marks]
Interpret Figure L012-1. Use no more than three lines.

(c) [2 marks]
We create variable SPEND2 = SPEND$^2$, and fit model $M_B : E[MILIMP] = \beta_0 + \beta_2 \times SPEND2$. The residuals of this model are plotted against SPEND in Figure L012-2. Explain the similarity of shape between Figures L012-1 and L012-2. Use no more than three lines.

(d) i. [2 marks]
Model $M_C : E[MILIMP] = \beta_0 + \beta_1 \times SPEND + \beta_2 \times SPEND2$ is fitted in the R output for part (d). Test for the simultaneous effect of SPEND and SPEND2 on $E[MILIMP]$.

ii. [2 marks]
Use statistical notions and the nature of the data to argue that the intercept $\beta_0$ can be removed from model $M_C$. Use no more than three lines.

(e) [2 marks]
Model $M_D : E[MILIMP] = \beta_1 \times SPEND + \beta_2 \times SPEND2$ is fitted. According to this model, the number of impressions retained decreases with increased spending once this spending exceeds a certain amount. Use the R output for part (f) to compute the maximum likelihood estimate of this amount.

(Question L2 continues on next page.)
L2. - continued

An alternative to quadratic models is a change-point model, where the relationship between MILIMP and SPEND is assumed linear but with possibly different slopes before and after a given value of SPEND.

Suppose that the slope change occurs at a value $s$ of SPEND. We create two new covariates, SPENDBEF and SPENDAFT, as follows:

$$
\text{SPENDBEF}_i = \begin{cases} 
\text{SPEND}_i & \text{if } \text{SPEND}_i \leq s \\
s & \text{otherwise}
\end{cases}
$$

and

$$
\text{SPENDAFT}_i = \begin{cases} 
0 & \text{if } \text{SPEND}_i \leq s \\
\text{SPEND}_i - s & \text{otherwise}
\end{cases}
$$

so that $\text{SPEND} = \text{SPENDBEF} + \text{SPENDAFT}$.

The following provides an example for $s = 40$.

\[
\begin{align*}
\text{SPEND} & = \begin{bmatrix} 5.0 & 5.7 & \ldots & 27.0 & 32.4 & 40.1 & 45.6 & \ldots & 185.9 \end{bmatrix}' \\
\text{SPENDBEF} & = \begin{bmatrix} 5.0 & 5.7 & \ldots & 27.0 & 32.4 & 40 & 40 & \ldots & 40 \end{bmatrix}' \\
\text{SPENDAFT} & = \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 & 0 & 0.1 & 5.6 & \ldots & 145.9 \end{bmatrix}'
\end{align*}
\]

We call the value $s$ a change-point.

(f) [2 marks] Express model $M_A$ as an appropriate null hypothesis for model $M_s : \mathbb{E} [\text{MILIMP}] = \beta_0 + \beta_3 \times \text{SPENDBEF} + \beta_4 \times \text{SPENDAFT}$.

(g) Assume for now that $s = 75$. Change-point model $M_{75}$ is fitted in the R output for part (g).

i. [2 marks] Sketch the curve representing the fitted values.

ii. [2 marks] Test $M_A$ against $M_{75}$.

(h) [2 marks] We now treat $s$ as a (nonlinear) parameter. Assume that $\sigma^2$ is known and equal to 456.6, the estimated $\sigma^2$ for model $M_{75}$. We consider every integral value for the change-point $s$ between 50 and 150, fit model $M_s$ and record the residual sum of squares (RSS) of the model.

The graph of RSS/$\sigma^2$ against the change-point $s$ is presented in Figure L012-3, reaching a minimum value at $s = 75$. Inverting an appropriate likelihood ratio test, use this graph to produce an approximate 95% confidence interval for the change-point.

\[(R \text{ code and output for Question L2 start on next page.)}\]
R code and output for Question L2:

> tvads

<table>
<thead>
<tr>
<th>SPEND</th>
<th>MILIMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MILLER LITE</td>
<td>50.1</td>
</tr>
<tr>
<td>PEPSI</td>
<td>74.1</td>
</tr>
<tr>
<td>STROH’S</td>
<td>19.3</td>
</tr>
<tr>
<td>FED’L EXPRESS</td>
<td>22.9</td>
</tr>
<tr>
<td>BURGER KING</td>
<td>82.4</td>
</tr>
<tr>
<td>COCA-COLA</td>
<td>40.1</td>
</tr>
<tr>
<td>MC DONALD’S</td>
<td>185.9</td>
</tr>
<tr>
<td>MCI</td>
<td>26.9</td>
</tr>
<tr>
<td>DIET COLA</td>
<td>20.4</td>
</tr>
<tr>
<td>FORD</td>
<td>166.2</td>
</tr>
<tr>
<td>LEVI’S</td>
<td>27.0</td>
</tr>
<tr>
<td>BUD LITE</td>
<td>45.6</td>
</tr>
<tr>
<td>ATT/BELL</td>
<td>154.9</td>
</tr>
<tr>
<td>CALVIN KLEIN</td>
<td>5.0</td>
</tr>
<tr>
<td>WENDY’S</td>
<td>49.7</td>
</tr>
<tr>
<td>POLAROID</td>
<td>26.9</td>
</tr>
<tr>
<td>SHASTA</td>
<td>5.7</td>
</tr>
<tr>
<td>MEOW MIX</td>
<td>7.6</td>
</tr>
<tr>
<td>OSCAR MEYER</td>
<td>9.2</td>
</tr>
<tr>
<td>CREST</td>
<td>32.4</td>
</tr>
<tr>
<td>KIBBLES ’N BITS</td>
<td>6.1</td>
</tr>
</tbody>
</table>

> attach(tvads)

(R code and output for Question L2 continue on next page.)
> # R code and output for questions (a) to (e)#
> # Question (a)##########################################################
> # (also required for (f) and (g)ii.) ###########################

> MA<-glm(MILIMP~SPEND); summary(MA)
> Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.16269    7.08948  3.126 0.005566
SPEND        0.36317    0.09712  3.739 0.001389

(Dispersion parameter for gaussian family taken to be 552.3215)

  Null deviance: 18217 on 20 degrees of freedom
  Residual deviance: 10494 on 19 degrees of freedom

> # Question (b): Plot of Figure L012-1 ######
> plot(SPEND,resid(MA),xlab="Money SPEND on TV ads",
+ ylab=expression(paste("Residual of model ",M[A]))
+ main=expression(paste("Fig. L012-1: Residuals of model ",M[A],
+ " against SPEND")))
> abline(h=0,lty=2)

> # Question (c)##########################################################
> SPEND2<-SPEND^2
> MB<-glm(MILIMP~SPEND2)
> summary(MB)
> Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.166e+01  6.405e+00  4.943 9.02e-05
SPEND2      1.652e-03  5.722e-04  2.887 0.009444

(Dispersion parameter for gaussian family taken to be 666.3947)

  Null deviance: 18217 on 20 degrees of freedom
  Residual deviance: 12661 on 19 degrees of freedom

(R code and output for Question L2 continue on next page.)
> # Plot of Figure L012-2

```r
plot(SPEND,resid(MB),xlab="Money spend on TV ads",
+ ylab=expression(paste("Residual of model ",M[B])),
+ main=expression(paste("Fig. L012-2: Residuals of model ",M[B],
+ " against SPEND")))
> abline(h=0,lty=2)
```

> # Question (d)

```r
MC<-glm(MILIMP~SPEND+SPEND2)
> summary(MC)
Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 7.059322 | 9.986180 | 0.707 | 0.4887 |
| SPEND | 1.084708 | 0.369944 | 2.932 | 0.0089 |
| SPEND2 | -0.003990 | 0.001984 | -2.011 | 0.0595 |
```

(Dispersion parameter for gaussian family taken to be 476.0474)

Null deviance: 18217.4 on 20 degrees of freedom
Residual deviance: 8568.9 on 18 degrees of freedom

> # Question (e)

```r
MD<-glm(MILIMP~SPEND+SPEND2-1)
> summary(MD)
Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| SPEND | 1.304460 | 0.197899 | 6.592 | 2.62e-06 |
| SPEND2 | -0.005045 | 0.001290 | -3.910 | 0.000942 |
```

(Dispersion parameter for gaussian family taken to be 463.5128)

Null deviance: 52606.0 on 21 degrees of freedom
Residual deviance: 8806.7 on 19 degrees of freedom

(R code and output for Question L2 continue on next page.)
> # R code and output for questions (f) to (h)#
> # R code and output for question (g) #######
> SPENDBEF<-SPEND
> SPENDAFT<-SPEND
> SPENDBEF[SPEND>75]<-75
> SPENDAFT<-SPENDAFT-75
> SPENDAFT[SPENDAFT<0]<-0

> M75<-glm(MILIMP~SPENDBEF+SPENDAFT)
> summary(M75)
Coefficients:

Estimate  Std. Error   t value Pr(>|t|)
(Intercept)  8.818800   8.791935  1.003  0.3291
SPENDBEF    0.861323   0.240029  3.588  0.0021
SPENDAFT    0.007767   0.182086  0.043  0.9664

(Dispersion parameter for gaussian family taken to be 456.6347)

Null deviance: 18217.4 on 20 degrees of freedom
Residual deviance: 8219.4 on 18 degrees of freedom

# Question (h) #################################
# Code to produce Figure L012-3 ################
> sigma2<-456.6
> chpt<-seq(50,150,1)
> RSS<-chpt
> for (i in 1:length(chpt)) {
+   SPENDBEF<-SPEND
+   SPENDAFT<-SPEND
+   SPENDBEF[SPEND>chpt[i]]<-chpt[i]
+   SPENDAFT<-SPENDAFT-chpt[i]
+   SPENDAFT[SPENDAFT<0]<-0
+   Mtemp<-glm(milimp~SPENDBEF+SPENDAFT)
+   RSS[i]<-Mtemp$deviance
}
Fig. L012–1: Residuals of model $M_A$ against SPEND

Fig. L012–2: Residuals of model $M_B$ against SPEND
Fig. L-012-3: Residual sum of squares over variance against change-point
L3. Discuss model assessment using the Mean Squared Error of Prediction (MSEP) or the equivalent Sum of Squared Error of Prediction (SSEP) to assess linear models with Gaussian error and identity link. Make sure to discuss the following elements:

- compare the squared error of prediction to the residual sum of squares as a tool for model assessment;
- provide three estimators of MSEP or SSEP and discuss their bias (with proofs).
Section G

Answer two questions out of questions G1 (page 21), G2 (page 28) and G3 (page 31).

G1. A survey of student opinion on the Vietnam War was taken at the University of North Carolina in Chapel Hill in May 1967. Students were asked to fill in questionnaires, stating which policy out of A, B, C or D they supported. Responses were cross-classified by sex and by undergraduate year or graduate status:

Policy A: The USA should increase its military activity.
B: The USA should follow the present policy.
C: The USA should de-escalate its military activity.
D: The USA should withdraw its military forces immediately.

The data are reproduced below:

<table>
<thead>
<tr>
<th>Sex:</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy:</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>175</td>
<td>116</td>
</tr>
<tr>
<td>2</td>
<td>160</td>
<td>126</td>
</tr>
<tr>
<td>3</td>
<td>132</td>
<td>120</td>
</tr>
<tr>
<td>4</td>
<td>145</td>
<td>95</td>
</tr>
<tr>
<td>Grad</td>
<td>118</td>
<td>176</td>
</tr>
</tbody>
</table>

We treat both Year and Policy as ordered factors; in the latter case, Policy levels are ordered from the most to the least bellicose (pro-war). We treat the counts as Poisson data. Refer to the accompanying R output to answer the questions.

(a) [2 marks] Select appropriate hypotheses and test the independence of Sex and Year.

(b) [2 marks] Can you test the goodness of fit of model count~Sex+Year+Sex:Year? If so, do so; if not, explain why not.

(c) Carry out suitable tests of hypotheses to answer the following questions. Does the proportion of students favouring each policy differ according to

i. [1 mark] the Sex of the respondent;

ii. [1 mark] the Year of the respondent;

iii. [1 mark] both Sex and Year of the respondents?

(Question G1 continued on next page.)
G1. - continued

(d) [2 marks] Explain why the models \( \text{count} \sim \text{Sex} + \text{Year} + \text{Policy} + \text{Sex:Year} \) and \( \text{count} \sim \text{Policy} + \text{Sex:Year} \) have exactly the same deviance.

Consider, in the output, model \( M_0: \text{count} \sim \text{Sex:Year} + \text{Sex:Policy} + \text{Year:Policy} - 1 \), without an intercept.

(e) [2 marks] Under \( M_0 \), what is the maximum likelihood estimate of the number of second-year respondents (either male or female) who support Policy B?

(f) [2 marks] The estimated correlation coefficient between the parameter estimates for \( \text{SexM:PolicyB} \) and \( \text{SexF:PolicyB} \) is 0.4862 (not shown in output). Use this figure and the summary output for \( M_0 \) to create an approximate 95% confidence interval for the difference between these coefficients.

The parameter estimates for all interaction terms in model \( M_0 \) are displayed graphically (the code to produce the graph is included for completeness).

(g) [1 mark] Consider Figure G019-1. Bearing in mind that there is no intercept in model \( M_0 \), provide a simple interpretation of the fact that the curve corresponding to Females lies below the curve corresponding to Males.

**Answer in no more than one line.**

(h) [1 mark] In Figure G019-2, the parameter estimates for both \( \text{SexM:PolicyA} \) and \( \text{SexF:PolicyA} \) were artificially set to zero for the purpose of plotting. Justify this practice in **no more than three lines**.

(i) [1 mark] Refer to Figure G019-2 again. Provide an interpretation of the fact that the curve corresponding to Males lies below the curve corresponding to Females. **Answer in no more than one line.**

(j) [1 mark] Refer to Figure G019-3. The curve for Policy A is absent from the graph. What should it be? **Answer in no more than one line.**

(k) [1 mark] Interpret Figure G019-3. **Use no more than three lines.**

(l) [2 marks] In light of Figure G019-3, suggest an alternative model to \( M_0 \) which might capture the main characteristics of the \( \text{Year:Policy} \) interaction with fewer parameters.

*(R code and output for Question G1 follow on next page.)*
R code and output for question G1:

```r
> viet
count Sex Year Policy
1 175 M 1 A
2 116 M 1 B
[... some output skipped...]
39 128 F Grad C
40 13 F Grad D
> attach(viet)
> deviance(glm(count~Sex+Year,family=poisson))
[1] 1142.417
> glm(count~Sex+Year+Sex:Year,family=poisson)

Call: glm(formula = count ~ Sex + Year + Sex:Year, family = poisson)

Coefficients:
(Intercept)      SexF   Year2   Year3     Year4
   4.698205  -1.740694  0.006810 -0.009153  0.066103
YearGrad SexF:Year2 SexF:Year3 SexF:Year4 SexF:YearGrad
  0.747947  -0.438590  0.783346  0.205212  0.312513

Degrees of Freedom: 39 Total (i.e. Null); 30 Residual
Null Deviance: 2708 Residual Deviance: 1080
```

```r
> deviance(glm(count~Policy,family=poisson))

> deviance(glm(count~Sex+Year+Policy,family=poisson))
[1] 423.8291

> deviance(glm(count~Sex+Year+Policy+Sex:Year,family=poisson))
[1] 361.7166

> deviance(glm(count~Policy+Sex:Year,family=poisson))
[1] 361.7166

> deviance(glm(count~Sex+Year+Policy+Sex:Year+Sex:Policy,family=poisson))
[1] 216.3122 # Note: the above model has 24 residual degrees of freedom

> deviance(glm(count~Sex+Year+Policy+Sex:Year+Year:Policy,family=poisson))
[1] 153.9352 # Note: the above model has 15 residual degrees of freedom
```

(R code and output for Question G1 continue on next page.)
R code and output for Question G1 - continued:

> M0<-glm(count~Sex:Year+Sex:Policy+Year:Policy-1,family=poisson)
> summary(M0)

Coefficients:

|                | Estimate | Std. Error | z value | Std. Error | Pr(>|z|) |
|----------------|----------|------------|---------|------------|----------|
| SexM:Year1     | 5.15414  | 0.07409    | 69.562  | 9.2e-16    | ***      |
| SexF:Year1     | 2.69824  | 0.16891    | 15.975  | 2e-16      | ***      |
| SexM:Year2     | 5.05397  | 0.07843    | 64.440  | 9.2e-16    | ***      |
| SexF:Year2     | 2.12310  | 0.19488    | 10.894  | 2e-16      | ***      |
| SexM:Year3     | 4.88041  | 0.08336    | 58.543  | 2e-16      | ***      |
| SexF:Year3     | 3.10527  | 0.14957    | 20.761  | 2e-16      | ***      |
| SexM:Year4     | 4.96618  | 0.08100    | 61.309  | 2e-16      | ***      |
| SexF:Year4     | 2.60429  | 0.16733    | 15.564  | 2e-16      | ***      |
| SexM:YearGrad  | 4.82869  | 0.08645    | 55.857  | 2e-16      | ***      |
| SexF:YearGrad  | 2.48094  | 0.16255    | 15.263  | 2e-16      | ***      |
| SexM:PolicyB   | -0.38340 | 0.11418    | -3.358  | 0.000786   | ***      |
| SexF:PolicyB   | 0.13458  | 0.18622    | 0.723   | 0.469862   |          |
| SexM:PolicyC   | -0.29954 | 0.10941    | -2.738  | 0.006184   | **       |
| SexF:PolicyC   | 1.05527  | 0.16396    | 6.436   | 1.23e-10   | ***      |
| SexM:PolicyD   | -2.18280 | 0.22640    | -9.641  | 2e-16      | ***      |
| SexF:PolicyD   | -1.78913 | 0.29926    | -5.979  | 2.25e-09   | ***      |
| Year2:PolicyB  | 0.14893  | 0.16213    | 0.919   | 0.358306   |          |
| Year3:PolicyB  | 0.25659  | 0.16181    | 1.586   | 0.112792   |          |
| Year4:PolicyB  | 0.02394  | 0.16669    | 0.144   | 0.885807   |          |
| YearGrad:PolicyB | 0.71910  | 0.15814    | 4.547   | 5.43e-06   | ***      |
| Year2:PolicyC  | 0.18157  | 0.15499    | 1.172   | 0.241388   |          |
| Year3:PolicyC  | 0.49017  | 0.15029    | 3.262   | 0.001108   | **       |
| Year4:PolicyC  | 0.51505  | 0.15010    | 3.431   | 0.000601   | ***      |
| YearGrad:PolicyC | 1.31478  | 0.14590    | 9.012   | 2e-16      | ***      |
| Year2:PolicyD  | 0.23077  | 0.31387    | 0.735   | 0.462190   |          |
| Year3:PolicyD  | 0.63363  | 0.29347    | 2.159   | 0.030845   | *        |
| Year4:PolicyD  | 1.07484  | 0.27501    | 3.908   | 9.29e-05   | ***      |
| YearGrad:PolicyD | 2.25855  | 0.25403    | 8.891   | 2e-16      | ***      |

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 23969.438 on 40 degrees of freedom
Residual deviance: 19.194 on 12 degrees of freedom

(Figures for Question G1 start on next page.)
Fig. G019–1: Sex:Year interaction parameter estimates

- Male
- Female
Fig. G019–2: Sex:Policy interaction parameter estimates
Fig. G019–3: Year:Policy interaction parameter estimates
G2. Data were collected to determine the relationship of time in use to the number of fissures that develop in turbine wheels. The data are presented below:

<table>
<thead>
<tr>
<th>Variable name:</th>
<th>hours</th>
<th>n</th>
<th>count</th>
<th>prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description:</td>
<td>Hours in use</td>
<td>Total turbines studied</td>
<td>Number with fissures</td>
<td>Proportion</td>
</tr>
<tr>
<td>400</td>
<td>39</td>
<td>0</td>
<td>0.00000000</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>53</td>
<td>4</td>
<td>0.07547170</td>
<td></td>
</tr>
<tr>
<td>1400</td>
<td>33</td>
<td>2</td>
<td>0.06060606</td>
<td></td>
</tr>
<tr>
<td>1800</td>
<td>73</td>
<td>7</td>
<td>0.09589041</td>
<td></td>
</tr>
<tr>
<td>2200</td>
<td>30</td>
<td>5</td>
<td>0.16666667</td>
<td></td>
</tr>
<tr>
<td>2600</td>
<td>39</td>
<td>9</td>
<td>0.23076923</td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>42</td>
<td>9</td>
<td>0.21428571</td>
<td></td>
</tr>
<tr>
<td>3400</td>
<td>13</td>
<td>6</td>
<td>0.46153846</td>
<td></td>
</tr>
<tr>
<td>3800</td>
<td>34</td>
<td>22</td>
<td>0.64705882</td>
<td></td>
</tr>
<tr>
<td>4200</td>
<td>40</td>
<td>21</td>
<td>0.52500000</td>
<td></td>
</tr>
<tr>
<td>4600</td>
<td>36</td>
<td>21</td>
<td>0.58333333</td>
<td></td>
</tr>
</tbody>
</table>


We assume that the count data have Binomial distributions.

(a) [2 marks]
Suppose we counted not the number of turbines with fissures but the total number of fissures on all turbines. Discuss the consequences on modelling options. **Use no more than five lines.**

(b) [3 marks]
We assume a logit link and regress the proportion of fissured turbines on hours. By writing out the binomial deviance (for the general case), explain why we must use n as a weight to maximize the likelihood at hand.

(c) [2 marks]
Refer to the R output below. Under the setting of (b), test that hours has an effect on prop.

(d) [3 marks]
Refer to the R output below. A quadratic term in hours, hours$^2$, is introduced in the model. Test for a quadratic effect using two different statistics. Briefly explain the discrepancy in p-values.

(Question G2 continues on next page.)
G2. -continued-

(e) [3 marks]
Suppose that we use quasi-likelihood with logit link and variance function
\[ V(\mu) = \mu(1 - \mu), \quad 0 < \mu < 1 \]
to fit the model in (c) (prop\_hours). Write down the parameter estimate and standard error for covariate hours.

(f) [2 marks]
A suitable Poisson model can sometimes be used to approximate logistic regression, under some condition on the expectations. Write down this model and the condition on the expectations, and indicate why this approximation might fail in the present context. (Hint: The model sought contains an offset.)

Questions (g), (h) and (i) refer to a probit-link model, which does not appear in the output.

(g) [2 marks]
Suppose that the first fissure in a turbine is caused by a latent (unobserved) random variable \( Z \sim N(0, \tau^2) \) for an unknown \( \tau^2 \), and that a fissure is observed if \( Z \leq \alpha + \beta \times \text{hours} \). Show that this model can be fitted using the probit link \( \Phi^{-1}(\cdot) \), where \( \Phi(\cdot) \) is the Standard Normal CDF. Can \( \alpha, \beta \) and \( \tau^2 \) all be estimated simultaneously?

(h) [1 mark]
Let \( \Phi(\cdot) \) be the Standard Normal CDF. Show that
\[
\frac{d}{dp} \Phi^{-1}(p) = \sqrt{2\pi} \exp \left( \frac{z_p^2}{2} \right)
\]
where \( z_p \) is such that \( \text{Prob}[Z \leq z_p] = p \) for \( Z \sim N(0,1) \).

(i) [2 marks]
Use a Taylor series about \( p = 1/2 \) to approximate the probit link, and propose a Gaussian model to approximate the Binomial model with probit link. Indicate whether you think this model would be appropriate for the data at hand and why.

(R code and output for Question G2 starts on next page.)
\textbf{R code and output for question G2:}

```r
> turb$prop <- turb$count/turb$n
> turb
     hours  n count  prop
   1   400  39  0.00000000
   2  1000  53  0.07547170
   3  1400  33  0.06060606
   4  1800  73  0.09589041
   5  2200  30  0.16666667
   6  2600  39  0.23076923
   7  3000  42  0.21428571
   8  3400  13  0.46153846
   9  3800  34  0.64705882
  10  4200  40  0.52500000
  11  4600  36  0.58333333
> attach(turb)
> summary(glm(prop~hours, weights=n, family=binomial))

Coefficients:

\begin{verbatim}
            Estimate Std. Error     z value  Pr(>|z|)  
(Intercept) -3.9236  0.37783  -10.384   <2e-16 ***
   hours      0.0009  0.00011   8.756    <2e-16 ***
\end{verbatim}

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 112.670 on 10 degrees of freedom
Residual deviance: 10.331 on 9 degrees of freedom

Number of Fisher Scoring iterations: 3

> hours2 <- hours^2
> summary(glm(prop~hours+hours2, weights=n, family=binomial))

Coefficients:

\begin{verbatim}
            Estimate Std. Error     z value  Pr(>|z|)  
(Intercept) -4.728e+00  9.165e-01  -5.159   2.48e-07 ***
   hours      1.645e-03  6.590e-04   2.496  0.0126   *
  hours2    -1.104e-07  1.096e-07  -1.008  0.3136
\end{verbatim}

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 112.6700 on 10 degrees of freedom
Residual deviance: 9.2677 on 8 degrees of freedom

```
G3. Suppose we are analyzing data regarding the effect of smoking and/or sex on the number of colds in a five year period (COLDS). Suppose that the true five-year rate of colds within each group is as follows:

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsmokers</td>
<td>8.20</td>
<td>6.15</td>
</tr>
<tr>
<td>Smokers</td>
<td>16.40</td>
<td>12.30</td>
</tr>
</tbody>
</table>

Suppose that the percentage of smokers among both males and females is 30%. Let MALE be an indicator variable that is 1 for males and 0 for females, SMOKE be an indicator variable that is 1 for smokers and 0 for nonsmokers, and MALE : SMOKE = MALE \times SMOKE.

(a) [5 marks]
Suppose we fit the Poisson regression model:
\[
\log(\mathbb{E}[\text{COLDS}]) = \beta_0 + \beta_1 \times \text{MALE}
\]
What are the interpretations and asymptotic tendencies of the Poisson regression model coefficients \( \hat{\beta}_0, \hat{\beta}_1 \) (i.e., as the sample size increases, toward what values will the estimates tend and how do you interpret the values?)

(b) [5 marks]
Suppose we fit the Poisson regression model:
\[
\log(\mathbb{E}[\text{COLDS}]) = \beta_0 + \beta_1 \times \text{SMOKE} + \beta_2 \times \text{MALE}
\]
What are the interpretations and asymptotic tendencies of the Poisson regression model estimates \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \) ?

(c) [5 marks]
Suppose we fit the Poisson regression model:
\[
\log(\mathbb{E}[\text{COLDS}]) = \beta_0 + \beta_1 \times \text{SMOKE} + \beta_2 \times \text{MALE} + \beta_3 \times \text{MALE : SMOKE}
\]
What are the interpretations and asymptotic tendencies of the Poisson regression estimates \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3 \) ?

(d) [3 marks]
For the model in question (c), let \( \text{se}_n(\hat{\beta}_0) \) be the standard error of the maximum likelihood estimate \( \hat{\beta}_0 \) for a certain sample of size \( n \). The quantity \( \sqrt{n}\text{se}_n(\hat{\beta}_0) \) converges in probability to a certain value. What is this value?

(e) [2 marks]
Based on the table of true five-year rates and the Poisson assumption, what is the expected duration between colds among female smokers?

End of Methodology Paper. Tables follow.