UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday, 3 June, 2011 1:30 pm to 4:30 pm

PAPER 31

APPLIED STATISTICS

*Attempt no more than **FOUR** questions, with at most **THREE** from Section A.*

*There are **SIX** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

SECTION A

1

An investigation into car noise with two different types of oil filters (the standard type and a new type) was carried out with three different sizes of car. For each type of filter there were six noise-tests for each of the three car sizes. The edited R output on the following page shows part of a statistical analysis of the results of the investigation. The object `noise` contains the measured noise levels in decibels, the object `size` is a factor with three levels (`1`, `2` and `3` denoting small, medium and large car sizes respectively), and `type` is a factor with two levels (`1` and `2` denoting the standard and the new type of oil filter respectively). Corner point constraints have been used.

(a) Write down the algebraic form of the model fitted in `model1`, defining your notation carefully and stating the constraints explicitly. Write down the parameter estimates and standard errors.

(b) The degrees of freedom in the analysis of variance table shown in the output to the command `anova(model1)` have been replaced by four asterisks. Write down what the degrees of freedom should be. Without carrying out any calculations, explain how you would use the parameter estimates and the data to find the residuals. Explain how the value `1962.5` in the analysis of variance table may be obtained from the residuals. One of the entries in the analysis of variance table has been replaced by `?`. Explain how this value may be calculated and give its value to one significant figure. What hypothesis is being tested by this test statistic? Carry out this hypothesis test in detail and give your conclusion.

(c) Give a detailed summary of the results of the analysis. If lower noise levels are preferable, what advice would you give about which filter to use?

```
> carnoise
   noise size type
1    810    1    1
2    820    1    1
3    820    1    1

<output omitted>

34   770    3    2
35   760    3    2
36   765    3    2
> model1 <- lm(noise~size*type)
> anova(model1)
Analysis of Variance Table
Response: noise
          Df  Sum Sq Mean Sq  F value    Pr(>F)
size       * 26051.4 13025.7 199.1189 < 2.2e-16
type       *  1056.2  1056.2  16.1465 0.0003631
size:type  *   804.2   402.1     ?    0.0057915
Residuals  *  1962.5    65.4

> summary(model1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  825.833      3.302 250.106  < 2e-16
size2         20.000      4.670   4.283 0.000175
size3        -50.833      4.670 -10.886 6.11e-12
type2         -3.333      4.670  -0.714 0.480849
size2:type2  -20.833      6.604  -3.155 0.003638
size3:type2   -1.667      6.604  -0.252 0.802471

Residual standard error: 8.088 on 30 degrees of freedom
Multiple R-squared: 0.9343
F-statistic: 85.34 on 5 and 30 DF,  p-value: < 2.2e-16
```

2

In an investigation into the effectiveness of three training programmes, A, B and C, designed to reduce a person's fear of snakes, forty subjects were initially given an approach test to see how close they could walk to a snake without feeling uncomfortable. The subjects were then randomly assigned to four groups so that there were 10 subjects in each group. Group 1 (the control group) received no training, groups 2, 3 and 4 received training programmes A, B and C respectively. At the end of the experiment, all subjects were again given the approach test. The (edited) R output on the following page shows part of an analysis of the data by a statistician. The R objects `Initial` and `Final` contain the outcome of the initial and final distance tests. These show the closest distance (in feet) between the subject and the snake without the subject experiencing discomfort. The R object `Group` shows the group for each subject. Let Y_{ij} be the distance in the final test for the jth subject in group i, $i = 1, 2, 3, 4$, $j = 1, \ldots, 10$.

Write down the algebraic forms of the models for Y_{ij} in `snake1` and `snake2`, defining any notation carefully and giving any constraints explicitly. Explain what these models mean. Explain why the statistician decided to fit `snake2`, giving the details of any hypothesis tests.

Consider the 40×1 vector $Y = (Y_{11} \ldots, Y_{1,10}, Y_{21}, \ldots, Y_{4,10})^T$, where ^{T} denotes transpose. Show that the model in `snake2` can be written in the form

$$Y = X\beta + \varepsilon,$$

where ε is a 40×1 vector of independent normally distributed random variables each with mean zero and variance σ^2, β is a vector of unknown parameters whose estimates are given in the output to `summary(snake2)`, and where you should give β and X explicitly. Find $X^T X$ and write down equations satisfied by the elements of the maximum likelihood estimator $\hat{\beta}$ of β.

Give a detailed discussion of your conclusions (together with your reasons) about the effectiveness or otherwise of the various training programmes. Comment on the adequacy of `snake2`. What further model checking would you carry out?

```
> snakedata
   Initial Final Group
1       25    25     1
2       13    25     1
3       10    12     1

<output omitted>

40      13     9     4
> Group <- factor(Group)
> snake1 <- lm(Final~Initial*Group)
> anova(snake1)
Analysis of Variance Table
              Df  Sum Sq Mean Sq F value    Pr(>F)
Initial        1 1388.07 1388.07 41.0030 3.394e-07
Group          3 1861.23  620.41 18.3267 4.206e-07
Initial:Group  3  237.31   79.10  2.3366   0.09227
Residuals     32 1083.29   33.85
> snake2 <- lm(Final~Initial+Group)
> anova(snake2)
Analysis of Variance Table
          Df Sum Sq Mean Sq F value    Pr(>F)
Initial    1 1388.1 1388.07  36.788 6.334e-07
Group      3 1861.2  620.41  16.443 7.788e-07
Residuals 35 1320.6   37.73
> summary(snake2)

Residuals:
    Min      1Q  Median      3Q     Max
-10.237  -3.506  -1.639   1.742  20.320

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.8101     2.6931   5.871 1.14e-06
Initial       0.6427     0.1091   5.891 1.07e-06
Group2      -14.1930     2.7497  -5.162 9.84e-06
Group3      -18.4072     2.7472  -6.700 9.34e-08
Group4      -10.3432     2.7541  -3.756 0.000629

Residual standard error: 6.143 on 35 degrees of freedom
Multiple R-squared: 0.711
F-statistic: 21.53 on 4 and 35 DF,  p-value: 4.939e-09
```

3

The (edited) R output below refers to data on the number of applications to UK universities from non-UK regions of the world for entry in 2008 and 2009. The R objects `app` and `succ` denote the number of applicants and the number of successful applicants respectively. The factor `Gender` is `F` (`M`) for male (female) applicants. The factor `Year` is `2008` (`2009`) for applications for 2008 (2009) entry. The factor `Region` has 10 levels, each representing a different Region. The levels are called:
`Africa`, `Americas`, `Australasia`, `Europe(EU)`, `Europe(nonEU)`, `FarEast1`, `HongKong`, `Malaysia`, `MiddleEast` and `Other`.

(a) Write a short paragraph summarising in words what you learn from the output for part (a). For hypothesis tests, state the null hypothesis, the result of the test and your conclusion.

(b) Write down the algebraic form of the model fitted in `applications1`, defining your notation carefully and giving any constraints explicitly. Use the output to determine whether or not there is a significant difference in success rates in 2009 for male and female applicants, in a model that inlcudes `Region`. Carry out any hypothesis tests in detail, giving the null hypothesis, the test statistic, its null distribution, the result and your conclusion in words. Comment on the fit of the model.

(c) Use relevant hypothesis testing to determine which of the models `applications2` and `applications3` is preferred. Using `applications3`, find the odds in favour of a successful application for a female applicant from Hong Kong for entry in 2009.

```
> admissionsdat
          Region Gender   app  succ Year
1         Africa      M  4370  2265 2009
2         Africa      F  3043  1551 2009
3       Americas      M  2654  1263 2009
4       Americas      F  3873  1816 2009

 <output omitted>

39         Other      M   144   105 2008
40         Other      F   165   109 2008
# Output for part (a)
> sum(succ)/sum(app)
[1] 0.5963653
> sum(app[Year=="2008"])
[1] 86228
> sum(succ[Year=="2008"])/sum(app[Year=="2008"])
[1] 0.5987614
> sum(app[Year=="2009"])
[1] 95575
> sum(succ[Year=="2009"])/sum(app[Year=="2009"])
[1] 0.5942035
> succM8 <- sum(succ[(Gender=="M")&(Year=="2008")])
> appM8 <- sum(app[(Gender=="M")&(Year=="2008")])
> succM9 <- sum(succ[(Gender=="M")&(Year=="2009")])
> appM9 <- sum(app[(Gender=="M")&(Year=="2009")])
> succM8/appM8
[1] 0.6038099
```

```
> succM9/appM9
[1] 0.5998165
> chisq.test( cbind(c(succM8,succM9),c(appM8-succM8,appM9-succM9)))
data:  cbind(c(succM8, succM9), c(appM8 - succM8, appM9 - succM9))
X-squared = 1.5035, df = 1, p-value = 0.2201

# Output for parts (b) and (c)
> psucc <- succ/app
> applications1 <- glm(psucc~Region+Gender,binomial,weights=app,subset=(Year=="2009"))
> anova(applications1,test="Chisq")
Analysis of Deviance Table
Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      19    2038.02
Region  9  1944.09        10      93.93 < 2.2e-16
Gender  1    21.75         9      72.18 3.110e-06
> applications2 <- glm(psucc~(Region+Year+Gender)^2,binomial,weights=app)
> deviance(applications2)
[1] 19.83222
> applications3 <- update(applications2,.~.-Year:Gender)
> deviance(applications3)
[1] 19.87395
> summary(applications3)
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 0.1584034  0.0298875   5.300 1.16e-07
RegionAmericas             -0.3199475  0.0423629  -7.553 4.27e-14
RegionAustralasia          -0.3214867  0.1226289  -2.622 0.008751
RegionEurope(EU)            0.2371956  0.0325562   7.286 3.20e-13
RegionEurope(nonEU)         0.3760754  0.0493698   7.618 2.59e-14
RegionFarEast1              0.3783132  0.0348244  10.863  < 2e-16
RegionHongKong              0.8278952  0.0542301  15.266  < 2e-16
RegionMalaysia              0.5479052  0.0565147   9.695  < 2e-16
RegionMiddleEast           -0.2762561  0.0502328  -5.500 3.81e-08
RegionOther                 0.4946775  0.1431422   3.456 0.000549
Year2009                   -0.1004949  0.0323631  -3.105 0.001901
GenderM                     0.0020256  0.0329246   0.062 0.950943
RegionAmericas:Year2009     0.1483974  0.0483213   3.071 0.002133
RegionAustralasia:Year2009 -0.2896508  0.1420282  -2.039 0.041411
RegionEurope(EU):Year2009   0.0311396  0.0357268   0.872 0.383425
RegionEurope(nonEU):Year2009 0.0927646 0.0560996   1.654 0.098215
RegionFarEast1:Year2009     0.0835878  0.0377268   2.216 0.026718
RegionHongKong:Year2009    -0.0945854  0.0609007  -1.553 0.120397
RegionMalaysia:Year2009     0.2957536  0.0628595   4.705 2.54e-06
RegionMiddleEast:Year2009   0.1063519  0.0508873   2.090 0.036622
RegionOther:Year2009        1.5536293  0.1591791   9.760  < 2e-16
RegionAmericas:GenderM     -0.0009137  0.0491059  -0.019 0.985155
RegionAustralasia:GenderM   0.1152904  0.1421459   0.811 0.417325
RegionEurope(EU):GenderM    0.1985802  0.0362646   5.476 4.35e-08
RegionEurope(nonEU):GenderM -0.0235954 0.0565494  -0.417 0.676493
RegionFarEast1:GenderM     -0.1143450  0.0382877  -2.986 0.002822
RegionHongKong:GenderM      0.0069423  0.0611210   0.114 0.909569
RegionMalaysia:GenderM     -0.0473346  0.0631941  -0.749 0.453835
RegionMiddleEast:GenderM    0.1236702  0.0533503   2.318 0.020445
RegionOther:GenderM         0.3521938  0.1553807   2.267 0.023412
    Null deviance: 3459.334  on 39  degrees of freedom
Residual deviance:   19.874  on 10  degrees of freedom
> 1-pchisq(19.874,10)
[1] 0.03046721
```

4

The random variable Y has density $f_\mu(y) = \frac{4}{\mu^2} y e^{-2y/\mu}$, $y > 0$, where $\mu > 0$. Show that $f_\mu(y)$ can be written in the form

$$\exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right),$$

with $\theta = 1/\mu$. Identify $b(\theta)$ and ϕ. Verify that, for this density, $b'(\theta) = \mathbb{E}(Y)$ and $\phi b''(\theta) = \mathrm{var}(Y)$.

Now suppose that $Y_1, \ldots, Y_n$ are independent random variables and that Y_i has density $f_{\mu_i}(y)$, $i = 1, \ldots, n$. Suppose that

$$\log(\mu_i) = x_i^T \beta,$$

where β is a p-dimensional vector of unknown parameters, and x_i is a p-dimensional vector of known covariates. Find the log-likelihood $l(\beta)$ and find equations satisfied by the maximum likelihood estimator $\hat{\beta}$ of β. Find the asymptotic distribution of $\hat{\beta}$ as n tends to infinity. Derive the deviance for this model.

There are alternative production lines, A and B, where a particular procedure can be carried out in a manufacturing process. In order to investigate whether or not the mean time to carry out the procedure is the same for both production lines, observations $Y_1, \ldots, Y_m$ are made of the times taken to carry out the procedure for production line A, and observations $Y_{m+1}, \ldots, Y_{2m}$ are made of the corresponding times for production line B. It is assumed that Y_i has density f_{μ_i} and

$$\log(\mu_i) = \alpha + \beta v_i,$$

where $v_1 = \ldots = v_m = -1$ and $v_{m+1}, \ldots v_{2m} = +1$. Find $\hat{\alpha}$, $\hat{\beta}$, the asymptotic variances of $\hat{\alpha}$ and $\hat{\beta}$, and their asymptotic covariance. Explain how you would test whether or not the mean procedure times are the same for both production lines.

[Hint: If a random variable U has density $f(u) = \lambda^2 u e^{-\lambda u}$ then $\mathbb{E}(U) = 2/\lambda$ and $\mathrm{var}(U) = 2/(\lambda^2)$.]

SECTION B

5

A researcher has collected data for the calendar year 2010 on the bonuses paid out (in millions of pounds) to the Chief Executive Officers (CEOs) of 100 leading banks in the world. Also acquired was information on the profit made in 2010 by each bank (in billions of pounds) and the years of service (up to the end of 2010) as current CEO of the bank.

The researcher wishes to determine the relationship of the CEO's bonus on the bank's profit and years of service as CEO. She approaches two statisticians, Statistician A and Statistician B, for help with analysing the data. The data-set contains the variables `bonus`, `profit` and `CEOyrs` corresponding to the bonus paid out to the CEO, the profit made by the bank and the years of service as CEO respectively.

Statistician A decides to fit an *additive* model in R to the data and produces the (edited) R output below.

```
> library(mgcv)
> logbonus <- log(bonus)
> bonus.am <- gam(logbonus~s(profit,bs="cr",k=10)+s(CEOyrs,bs="cr",k=10),
+ family=gaussian(link=identity))
> summary(bonus.am)

Family: gaussian
Link function: identity

Formula:
logbonus ~ s(profit, bs = "cr", k = 10) + s(CEOyrs, bs = "cr",
    k = 10)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.250801   0.008893  -253.1   <2e-16
---

Approximate significance of smooth terms:
            edf  p-value
s(profit) 3.345   <2e-16
s(CEOyrs) 6.061   <2e-16
---

R-sq.(adj) =  0.993   Deviance explained = 99.4%
GCV score = 0.0088264  Scale est. = 0.0079079  n = 100
```

Statistician B, on the other hand, decides to fit a *generalized additive* model in R to the data and produces the following edited R output:

```
> library(mgcv)
> bonus.gam <- gam(bonus~s(profit,bs="cr",k=10)+s(CEOyrs,bs="cr",k=10),
+ family=Gamma(link=log))
> summary(bonus.gam)

Family: Gamma
Link function: log

Formula:
bonus ~ s(profit, bs = "cr", k = 10) + s(CEOyrs, bs = "cr", k = 10)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.247264   0.008888  -252.8   <2e-16
---

Approximate significance of smooth terms:
            edf  p-value
s(profit) 3.347   <2e-16
s(CEOyrs) 6.099   <2e-16
---

R-sq.(adj) =  0.997   Deviance explained = 99.5%
GCV score = 0.0088206  Scale est. = 0.0078992  n = 100
```

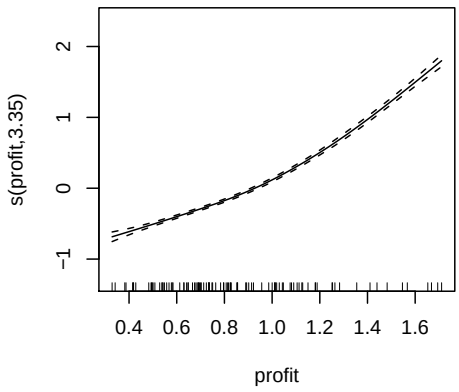
The plots (using `plot(bonus.am)` and `plot(bonus.gam)`) corresponding to Statistician A and Statistician B models are shown in the accompanying figure. The upper two plots correspond to Statistician A's model and the lower two plots correspond to Statistician B's model.

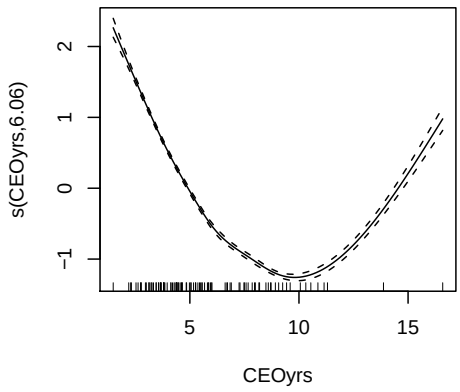
(i) Write down the algebraic forms of the two models fitted by the statisticians, making sure to define all notation used and stating all assumptions made.

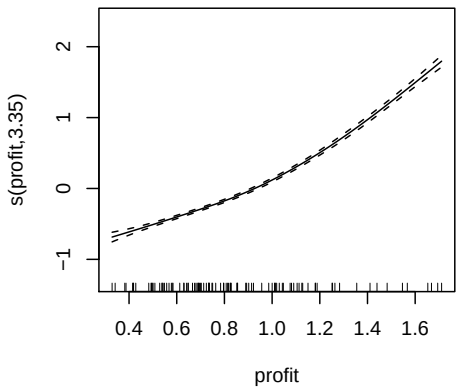
(ii) Explain briefly and interpret the (edited) output from the R command `summary(bonus.gam)`. What is the total effective degrees of freedom for Statistician B's model?

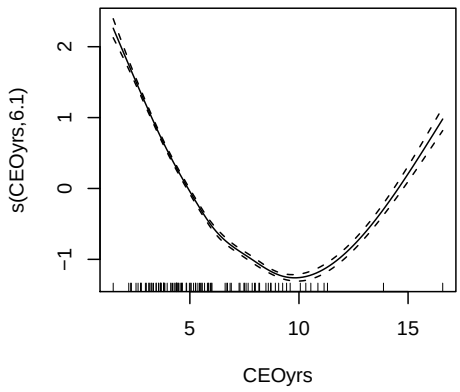
(iii) Why are the results and plots from the two statisticians' analyses so similar?

(iv) From the figure suggest and justify an alternative more parsimonious model that either Statistician A or Statistician B could fit next. How many total degrees of freedom does this model have?


s(profit,3.35)
−1
0
1
2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
profit


s(CEOyrs,6.06)
−1
0
1
2
5
10
15
CEOyrs


s(profit,3.35)
−1
0
1
2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
profit


s(CEOyrs,6.1)
−1
0
1
2
5
10
15
CEOyrs

6

A (i) Write down the probability mass functions for the counts from a zero-inflated Poisson (ZIP) model and a two-part (hurdle) model with non-zero Poisson counts, where no explanatory variables are included in the models.

(ii) Are the two models referred to in (i) equivalent? You must justify your answer.

B Below is the (edited) R output from regression analyses of recurrent episodes of self-harm (`count`) on treatment (`trt`: 0 = standard treatment; 1 = experimental treatment), controlling for age (`age`), sex (`sex`: 0 = female; 1 = male) and type of personality disorder (`bpd`: 1 = no personality disorder; 2 = borderline personality disorder; 3 = other personality disorder).

```
> library(pscl)
> library(lmtest)
> slfhrm.pois <- glm(count~trt+factor(bpd)+sex+age,family="poisson",
+ data=slfhrm.dat)
> summary(slfhrm.pois)

Call:
glm(formula = count ~ trt + factor(bpd) + sex + age, family = "poisson",
    data = slfhrm.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8652  -1.9522  -1.3300  -0.1328  11.9150

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.569157   0.189970   8.260  < 2e-16
trt          -0.379682   0.098403  -3.858 0.000114
factor(bpd)2  0.012501   0.127745   0.098 0.922044
factor(bpd)3  0.040630   0.125698   0.323 0.746518
sex           0.533676   0.119477   4.467 7.94e-06
age          -0.043950   0.005028  -8.741  < 2e-16
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1560.0  on 254  degrees of freedom
Residual deviance: 1416.6  on 249  degrees of freedom
AIC: 1681.6

Number of Fisher Scoring iterations: 7
```

```
> slfhrm.hurdle <- hurdle(count~trt+factor(bpd)+sex+age | trt+factor(bpd)+sex+age,
+ dist="poisson",zero.dist="binomial",link="cloglog",data=slfhrm.dat)
> summary(slfhrm.hurdle)

Call:
hurdle(formula = count ~ trt + factor(bpd) + sex + age | trt + factor(bpd) +
    sex + age, data = slfhrm.dat, dist = "poisson", zero.dist = "binomial",
    link = "cloglog")

Pearson residuals:
     Min       1Q   Median       3Q      Max
-0.91032 -0.65550 -0.55861 -0.04763 11.73068

Count model coefficients (truncated poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.846309   0.195913  14.528  < 2e-16
trt          -0.305595   0.103135  -2.963  0.00305
factor(bpd)2 -0.181050   0.133466  -1.357  0.17493
factor(bpd)3 -0.200671   0.130431  -1.539  0.12392
sex           0.642190   0.127675   5.030 4.91e-07
age          -0.054866   0.005514  -9.950  < 2e-16
Zero hurdle model coefficients (binomial with cloglog link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.10343    0.39008  -2.829  0.00467
trt          -0.25566    0.21603  -1.183  0.23664
factor(bpd)2  0.55326    0.28165   1.964  0.04949
factor(bpd)3  0.15608    0.27386   0.570  0.56872
sex           0.04375    0.22896   0.191  0.84845
age           0.00740    0.00909   0.814  0.41560
---

Number of iterations in BFGS optimization: 17
Log-likelihood: -530.2 on 12 Df

> lrtest(slfhrm.pois,slfhrm.hurdle)
Likelihood ratio test

Model 1: count ~ trt + factor(bpd) + sex + age
Model 2: count ~ trt + factor(bpd) + sex + age | trt + factor(bpd) + sex +
    age
  #Df  LogLik Df Chisq Pr(>Chisq)
1   ?       ?
2   ?       ?  6     ?  < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Warning message:
In modelUpdate(objects[[i - 1]], objects[[i]]) :
  original model was of class "glm", updated model is of class "hurdle"
```

(i) Write down the algebraic forms of the regression equations corresponding to `slfhrm.hurdle`. You must define all notation used.

(ii) Interpret the treatment effects and the type of personality disorder effects in the model corresponding to `slfhrm.hurdle`.

(iii) Determine the values of the five numbers that have been replaced by question marks (`?`) in the table corresponding to the likelihood ratio test of the Poisson model `slfhrm.pois` versus the hurdle model `slfhrm.hurdle`, produced from the R command `lrtest(slfhrm.pois,slfhrm.hurdle)`. Was it appropriate to do this standard likelihood ratio test here? You need to justify your answer.

END OF PAPER