UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Tuesday, 3 June, 2014 1:30 pm to 4:30 pm

PAPER 33

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.

1

Suppose that $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ (^{T} denotes transpose), X is a known $n \times p$ matrix with rank p ($< n$), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown parameters, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ where $\varepsilon_1, \ldots, \varepsilon_n$ are independent normally distributed random variables with mean 0 and variance σ^2. Derive the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ of $\boldsymbol{\beta}$ and σ^2 respectively. Write down the distribution of $\hat{\boldsymbol{\beta}}$. Define the fitted values, the residuals and the residual sum of squares (RSS). Write down the distribution of RSS/σ^2. Show that $\hat{\sigma}^2$ is a biased estimator of σ^2 and find an unbiased estimator $\tilde{\sigma}^2$ of σ^2.

A paper manufacturer is investigating how the tensile strength of paper is related to the percentage of hardwood in the pulp from which the paper is made. There are 19 samples of paper, each produced from a different batch of pulp. For each sample of paper, the manufacturer measures its tensile strength and the percentage of hardwood in the corresponding batch of pulp. The (edited) R output below refers to statistical analysis of the resulting data. The R objects `strength` and `hardwood` contain the tensile strengths of the paper samples and the percentages of hardwood respectively.

Write down the algebraic forms of the models fitted in `lm1` and `lm2`. For each model, write down the value of $\tilde{\sigma}$. For the model `lm1`, explain how to find an estimate of the expected tensile strength of a sample of paper produced from a new batch of pulp containing x% harwood. How would you estimate the variance of your estimate using values in the output? Justify your answer.

Determine which of `lm1` and `lm2` is your preferred model, giving full details for any hypothesis tests that you carry out (ie state the null and alternative hypotheses, the test statistic, the null distribution of the test statistic, and your conclusion). What model checking would you carry out?

```
> hardwoodnew <- hardwood - mean(hardwood)

> lm1 <- lm(strength~hardwoodnew)
> summary(lm1)

Residuals:
    Min      1Q  Median      3Q     Max
-25.986  -3.749   2.938   7.675  15.840

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.1842     2.7108  12.611 4.69e-10
hardwoodnew   1.7710     0.6478   2.734   0.0141

Residual standard error: 11.82 on 17 degrees of freedom
Multiple R-squared: 0.3054,
F-statistic: 7.474 on 1 and 17 DF,  p-value: 0.01414

> hardwoodnew2<- hardwoodnew*hardwoodnew
> lm2 <- lm(strength~hardwoodnew+hardwoodnew2)
> summary(lm2)
Residuals:
    Min      1Q  Median      3Q     Max
-5.8503 -3.2482 -0.7267  4.1350  6.5506
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.29497    1.48287   30.55 1.29e-15
hardwoodnew   2.54634    0.25384   10.03 2.63e-08
hardwoodnew2 -0.63455    0.06179  -10.27 1.89e-08

Residual standard error: 4.42 on 16 degrees of freedom
Multiple R-squared: 0.9085,
F-statistic: 79.43 on 2 and 16 DF,  p-value: 4.912e-09
```

2

An engineer is investigating the relationship between the speed of a lathe and the lifetimes of four types of cutting tool used on the lathe. The lifetimes (in hours) of 20 cutting tools (five of each type) are measured at various lathe speeds (in revolutions per minute). The engineer has asked a statistician to analyse the results, and the statistician has produced the (edited) R output below. Suppose that the R objects `lifetime`, `speed` and `type` contain the lifetimes, the lathe speeds and the type of cutting tool, respectively, and that `type` has been declared a factor.

Write down the algebraic forms for each of the three models fitted, including any constraints.

In line `(A)` four values have been replaced by asterisks. Find the missing values, giving reasons for your answers. Give details of the hypothesis test that is carried out in line `(A)`, stating the null and alternative hypotheses, the test statistic, the null distribution of the test statistic, and state whether or not the null hypothesis should be rejected.

State with reasons which of the three models you would recommend to the engineer. For your chosen model, give a detailed interpretation of each line of the output to the relevant `summary` command. Draw a sketch graph to illustrate the relationship between the lifetimes of the types of cutting tools and the lathe speed.

```
> type
 [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4

> toollife.lm1 <- lm(lifetime ~ speed)
> anova(toollife.lm1)
Analysis of Variance Table

Response: lifetime
          Df  Sum Sq Mean Sq F value Pr(>F)
speed      1  293.01 293.005  4.1137 0.0576
Residuals 18 1282.08  71.227
> summary(toollife.lm1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.61672    9.60323   4.542 0.000253
speed       -0.02545    0.01255  -2.028 0.057600

Residual standard error: 8.44 on 18 degrees of freedom
Multiple R-squared: 0.186,
F-statistic: 4.114 on 1 and 18 DF,  p-value: 0.0576

> toollife.lm2 <- lm(lifetime ~ type + speed)
> anova(toollife.lm2)
Analysis of Variance Table

Response: lifetime
          Df  Sum Sq Mean Sq F value    Pr(>F)
type      3 1238.66  412.89  52.252 3.566e-08
speed     1  217.91  217.91  27.577 9.769e-05
Residuals 15  118.53    7.90
```

```
> summary(toollife.lm2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.891690   4.039533   8.885 2.31e-07
type2       -0.833408   1.904707  -0.438    0.668
type3       12.607660   1.777937   7.091 3.68e-06
type4       16.539121   1.876116   8.816 2.55e-07
speed       -0.024585   0.004682  -5.251 9.77e-05

Residual standard error: 2.811 on 15 degrees of freedom
Multiple R-squared: 0.9247,
F-statistic: 46.08 on 4 and 15 DF,  p-value: 2.974e-08


> toollife.lm3 <- lm(lifetime ~ type*speed)
> anova(toollife.lm3)
Analysis of Variance Table

Response: lifetime
          Df  Sum Sq Mean Sq F value    Pr(>F)
type       3 1238.66  412.89 54.1453 3.015e-07
speed      1  217.91  217.91 28.5760 0.0001748
type:speed *     *       *     *     0.3579213 (A)
Residuals 12   91.51    7.63
> summary(toollife.lm3)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.123237   7.374601   3.678  0.00316
type2        11.576083   9.749008   1.187  0.25804
type3        15.304382  11.601988   1.319  0.21175
type4        31.816249   9.431954   3.373  0.00554
speed        -0.013892   0.008866  -1.567  0.14314
type2:speed  -0.016095   0.012836  -1.254  0.23374
type3:speed  -0.003252   0.014029  -0.232  0.82057
type4:speed  -0.020099   0.012151  -1.654  0.12400

Residual standard error: 2.761 on 12 degrees of freedom
Multiple R-squared: 0.9419,
F-statistic: 27.79 on 7 and 12 DF,  p-value: 1.687e-06
```

3

(a) Show that the gamma distribution, with probability density function

$$f(y;\alpha,\gamma) = \frac{\gamma^\alpha}{\Gamma(\alpha)} e^{-\gamma y} y^{\alpha-1}, \quad y>0, \alpha>0, \gamma>0,$$

is an exponential dispersion family over the unknown parameters α, γ. Identify the variance function, canonical link function, and the dispersion parameter.

(b) Define what is meant by the term *generalized linear model* (GLM).

(c) The times to failure of 61 components on a ship were recorded, along with the type of component (labelled `type1`, `type2`, `type3`) and the position of the component (`pos`) on the inside (`in`) or outside (`out`) of the ship. In the R code below, `mod1` fits an exponential, and `mod2` fits a gamma generalized linear model to these data.

For a generalized linear model, it can be shown that the asymptotic covariance matrix of the parameter estimators $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ is $\phi(X^T W X)^{-1}$, with X the design matrix, ϕ the dispersion parameter and W the diagonal matrix with ith diagonal entry

$$[a_i V(\mu_i) g'(\mu_i)^2]^{-1}.$$

Here V is the variance function, g the link function, and $\mathrm{Var}(Y_i) = a_i \phi V(\mu_i)$. Use this information to derive the relationship between the standard errors from `mod1` with those from `mod2`, which both use the canonical link. Your answer should include a numerical value that you are able to determine from the output below.

(d) Explain what hypotheses are being tested with the test statistics `test1` and `test2` defined in the R code below.

(e) Use your interpretation of the code given for part (d) to determine which of the two analysis of deviance tables is appropriate. What do you conclude about which variables affect the time to failure?

```
> summary(mod1,dispersion=1)

Call:
glm(formula = times ~ type + pos, family = Gamma, data = ship)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1493  -0.5085  -0.2127   0.3178   1.1005

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.15032    0.03607   4.168 3.08e-05 ***
type2       -0.01228    0.04776  -0.257    0.797
type3        0.08323    0.06417   1.297    0.195
posout       0.02368    0.04556   0.520    0.603
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Gamma family taken to be 1)

    Null deviance: 21.061  on 60  degrees of freedom
Residual deviance: 18.185  on 57  degrees of freedom
AIC: 304.48

Number of Fisher Scoring iterations: 5


 > summary(mod2)

Call:
glm(formula = times ~ type + pos, family = Gamma, data = ship)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1493  -0.5085  -0.2127   0.3178   1.1005

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.15032    0.02009   7.481 5.02e-10 ***
type2       -0.01228    0.02661  -0.461   0.6463
type3        0.08323    0.03575   2.328   0.0235 *
posout       0.02368    0.02538   0.933   0.3548
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Gamma family taken to be 0.3103711)

    Null deviance: 21.061  on 60  degrees of freedom
Residual deviance: 18.185  on 57  degrees of freedom
AIC: 304.48

Number of Fisher Scoring iterations: 5


> test1<-57*summary(mod2)$dispersion/1
> test1
[1] 17.69115
> test2<-57*summary(mod2)$dispersion/(1/3)
> test2
[1] 53.07346
> pchisq(test1, df=57, lower=TRUE)
[1] 1.199504e-07
> pchisq(test2, df=57, lower=TRUE)
[1] 0.3768748
```

```
> anova(mod1,dispersion=1,test="Chisq")
Analysis of Deviance Table

Model: Gamma, link: inverse

Response: times

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    60     21.061
type  2  2.60027        58     18.461   0.2725
pos   1  0.27603        57     18.185   0.5993


 > anova(mod2,test="F")
Analysis of Deviance Table

Model: Gamma, link: inverse

Response: times

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev      F  Pr(>F)
NULL                    60     21.061
type  2  2.60027        58     18.461 4.1890 0.02007 *
pos   1  0.27603        57     18.185 0.8894 0.34963
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

4

The total number of floods (`nf`) at a randomly-selected gauging station on 100 different rivers was recorded over a 25-year period. Also recorded were:

1. The proportion of the river catchment area which is covered with buildings (`pc`) in the final year of the dataset;
2. A classification of whether the land is deemed "Urban" (`urban` = 1) or not (`urban` = 0) in the final year of the dataset.

A sample of the data is provided in the Table below.

Floods (`nf`)	Proportion Building Cover (`pc`)	Urban (`urban`)
3	0.05	0
4	0.31	0
5	0.42	0
18	0.62	0
23	0.77	1

The following 0.95 quantiles of chi-squared distributions may be useful when answering the questions below.

Degrees of freedom	0.95 quantile
1	3.84
2	5.99
96	119.9
97	121.0
98	122.1
99	123.2
100	124.3

(a) Three Poisson generalized linear models (GLMs), `m1`, `m2`, and `m3` were fitted to these data; the R output is given below for each fit. State, with justification, which of these three models is preferable for these data.

(b) For your chosen model, interpret the fitted model coefficients.

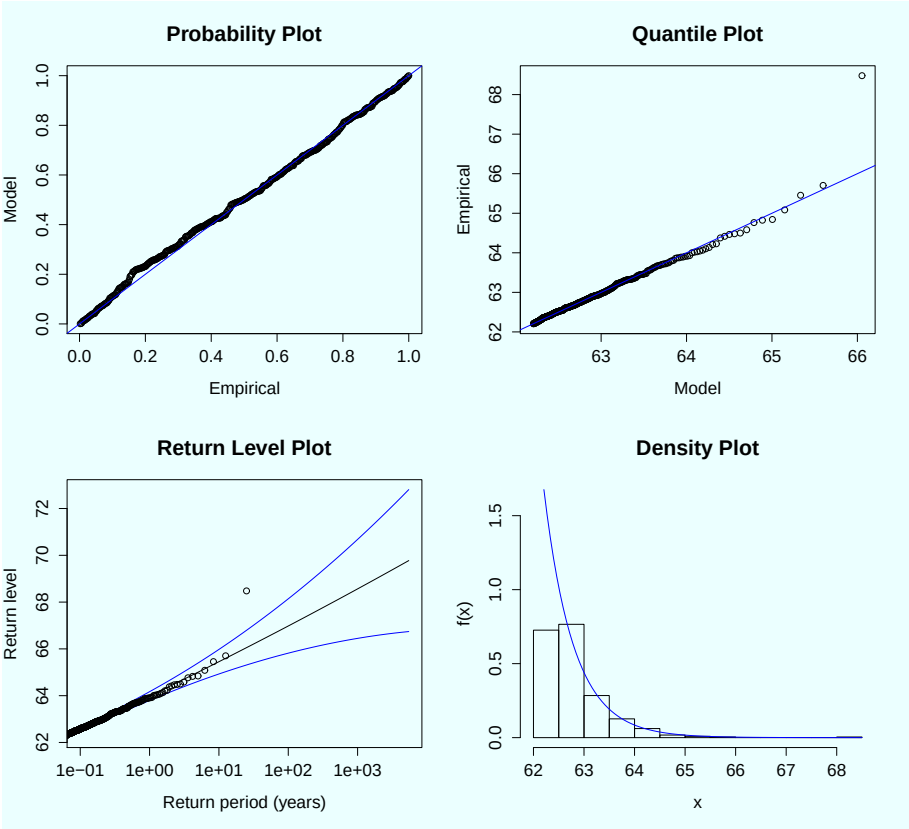
(c) State, with justification, whether your chosen model provides a good fit.

(d) The daily maximum flow rates of river water ($m^3\ s^{-1}$), over the 25 years, were recorded at one of the stations. Hydrologists are interested in estimating the 100-year return level for the flow rate. Define what is meant by the term "100-year return level". You do not need to include mathematical expressions in your answer.

(e) The distribution below was used to model all excesses above a high threshold, u, making a working assumption of independence between the observations. Name this distribution and give a reason why it is the natural choice. State what the possible values of the parameter ξ mean for the heaviness of the tail.

$$\Pr(X \leqslant x | X > u) = 1 - [1 + \xi(x-u)/\sigma]_+^{-1/\xi}, \quad x > u.$$

(f) The model in part (e) was fitted to excesses of the 95% quantile; diagnostic plots for the fit are given in the Figure below. Comment on the fit of the model. Use the diagnostic plots to estimate approximately the 100-year return level with 95% confidence interval.



(g) The waiting time, T, (in years) between years containing a flow rate in excess of a particular level follows a geometric distribution, with probability mass function

$$\Pr(T = t) = (1-p)^{t-1}p, \quad p \in (0,1), \quad t \in \{1, 2, 3, \ldots\}.$$

Derive the expected waiting time between exceedances of this level. For $p = 0.02$, what return level does this correspond to?

```
> summary(m1)

Call:
glm(formula = nf ~ pc + urban, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8059  -1.2783  -0.0995   1.0497   3.8102

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99734    0.09589  10.400   <2e-16 ***
```

```
pc           2.93702    0.18613  15.780   <2e-16 ***
urbanTRUE    0.04831    0.08929   0.541    0.588
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1258.10  on 99  degrees of freedom
Residual deviance:  290.64  on 97  degrees of freedom
AIC: 712.65

Number of Fisher Scoring iterations: 4

> summary(m2)

Call:
glm(formula = nf ~ urban, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3353  -2.2117  -0.8194   1.4691   5.7632

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.24042    0.03818   58.68   <2e-16 ***
urbanTRUE    1.32336    0.05007   26.43   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1258.10  on 99  degrees of freedom
Residual deviance:  559.17  on 98  degrees of freedom
AIC: 979.19

Number of Fisher Scoring iterations: 5

> summary(m3)

Call:
glm(formula = nf ~ pc, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8182  -1.2903  -0.0858   1.0953   3.8528

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)  0.96657    0.07768   12.44   <2e-16 ***
pc           3.02055    0.10468   28.86   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1258.10  on 99  degrees of freedom
Residual deviance:  290.93  on 98  degrees of freedom
AIC: 710.94

Number of Fisher Scoring iterations: 4
```

5

(a) Suppose that $n \geqslant 2$ and consider a set of points $\{(x_i, y_i) : i = 1, \ldots, n\}$, where $x_1 < \cdots < x_n$. Let $g(x)$ be a natural cubic spline *interpolating* these points. That is, $g(x)$ is a function made up of separate cubic polynomials on each interval $[x_1, x_2], (x_2, x_3], \ldots [x_{n-1}, x_n]$ and at its knots is continuous, as well as having its first two derivatives continuous. In addition, $g(x_i) = y_i$ and $g(.)$ is linear beyond the boundary knots (i.e. $g''(x_1) = g''(x_n) = 0$).

Show that of all functions, f, that are continuous on $[x_1, x_n]$, having absolutely continuous first derivatives and *interpolating* $\{(x_i, y_i) : i = 1, \ldots, n\}$, $g(x)$ is the one that minimises

$$\alpha \int_{x_1}^{x_n} f''(x)^2 dx$$

over f with α a fixed known positive quantity. (Hint: Define $\tilde{g}(x)$ to be an interpolant of $\{(x_i, y_i) : i = 1, \ldots, n\}$ other than $g(x)$ and let $h(x) = \tilde{g}(x) - g(x)$.)

(b) After the discovery of a severe leakage of major chemical contaminants into soil and groundwater supplies, an environmental agency undertook a 10-year follow-up study of women of child-bearing age, who at the time of this environmental disaster (baseline) were aged between 18 and 30 years (inclusive). In addition to the recording of the ages of the women at exposure (`age`), baseline overall exposure levels (`exposure`) to these contaminants were measured from blood samples taken.

One of the aims of the study was to determine if this disaster could be linked to any birth defects seen subsequently. Two hundred of the women followed-up had children in the intervening 10-year period. Data on the total number of children born in this period (`nbirths`) and the number with birth defects (`nbrthdfcts`) were collected for these women.

The statistician analysing the data (`disaster.dat`), fits a number of models to investigate the link between the disaster and children with birth defects. The (edited) R output of some of the statistician's performed analyses is shown below.

```
> disaster.glm <- glm(nbrthdfcts/nbirths ~ age + exposure,
family = binomial, weights = nbirths, data=disaster.dat)
> summary(disaster.glm)

Call:
glm(formula = nbrthdfcts/nbirths ~ age + exposure, family = binomial,
    data = disaster.dat, weights = nbirths)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.74272    0.75224  -2.317   0.0205 *
age          0.02030    0.02897   0.701   0.4834
exposure     2.33028    0.28581   8.153 3.54e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 744.61  on 199  degrees of freedom
Residual deviance: 671.27  on 197  degrees of freedom
AIC: 755.37


> library(mgcv)
> disaster.gam <- gam(nbrthdfcts/nbirths ~ s(age,bs="cr") +
s(exposure,bs="cr"), family = binomial, weights = nbirths,
data=disaster.dat, scale=-1) # scale=-1 means estimate scale
> summary(disaster.gam)

Family: binomial
Link function: logit

Formula:
nbrthdfcts/nbirths ~ s(age, bs = "cr") + s(exposure, bs = "cr")

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2031     0.1474  -1.378     0.17

Approximate significance of smooth terms:
              edf Ref.df     F p-value
s(age)      1.257   1.47 0.087   0.859
s(exposure) 1.895   2.36 9.071 8.3e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

R-sq.(adj) =  0.144   Deviance explained = 11.9%
GCV score = 3.4217  Scale est. = 3.3506    n = 200
```

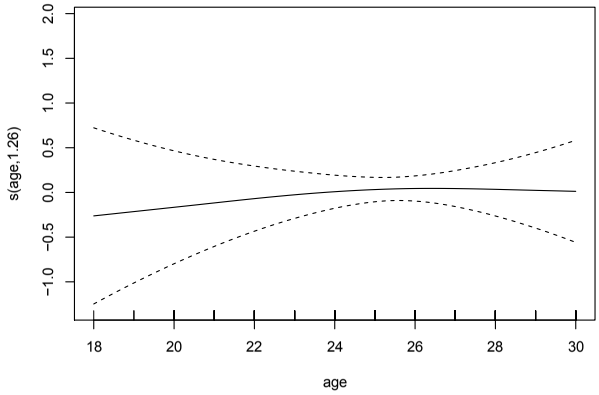
The plots (using `plot(disaster.gam)`) corresponding to the second model fitted by the statistician are shown in the accompanying figure.

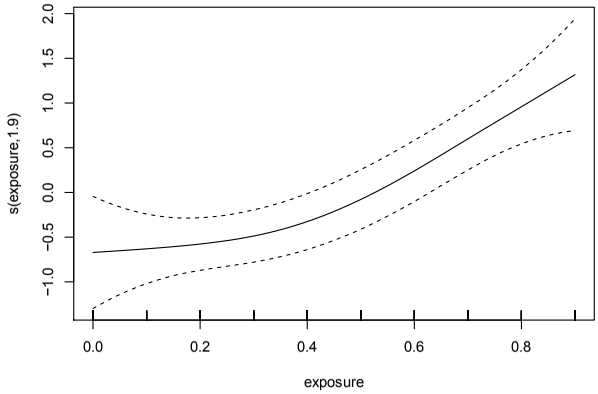
(i) Write down the algebraic forms of the two models fitted by the statistician, making sure to define all notation used and stating all assumptions made.

(ii) Explain the statistical reasons why the statistician would proceed to fit the model, `disaster.gam`, after seeing the results from `disaster.glm`.

(iii) What is the trace of the influence matrix used in the calculation of the GCV score shown in the output of `summary(disaster.gam)`?

(iv) From the figure, suggest an alternative simpler model to fit which would still capture adequately the relationships shown in the two plots. You need to justify your answer.


s(age,1.26)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
18
20
22
24
26
28
30
age


s(exposure,1.9)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
exposure

6

(a) Let $y_1, \ldots y_n$ be realisations of independent zero-inflated Poisson random variables $Y_1, \ldots, Y_n$ with $Y_i \sim ZIP(\pi_i, \mu_i)$, where π_i is the probability of the ith individual having a structural zero and μ_i is the mean of a Poisson random variable corresponding to the count component of the ith individual. Now suppose that we wish to model the dependence of Y_i on a binary treatment variable x_i, where x_i enters into the count component through the Poisson mean parameter, μ_i, on the log-scale, and enters into the structural zero component through a logistic regression for π_i.

Apply the E-M algorithm to this estimation problem, providing explicit expressions for the parameter updates obtained in the M-step of the algorithm to compute the maximum likelihood estimates. You are not required to find the standard errors corresponding to the maximum likelihood estimates.

(b) Below is the (edited) R output from a zero-inflated Poisson model analysis of recurrent episodes of self-harm (`count`) over a six-month period on treatment (`trt`: 0 = standard treatment; 1 = Cognitive Behavioural Therapy (CBT)), age (`age`), sex (`sex`: 0 = female; 1 = male), type of personality disorder (`bpd`: 1 = no personality disorder; 2 = borderline personality disorder; 3 = other personality disorder) and centre (`centre`: 0 = Centre A; 1 = Centre B).

```
> slfhrm.zip <- zeroinfl(count ~ trt + age + centre + factor(bpd)*sex
| centre+factor(bpd), dist="poisson", data=slfhrm.dat, EM=TRUE)
> summary(slfhrm.zip)

Call:
zeroinfl(formula = count ~ trt + age + centre + factor(bpd)*sex
 | centre + factor(bpd), data = slfhrm.dat, dist = "poisson", EM = TRUE)

Count model coefficients (poisson with log link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.470469   0.228756  10.800  < 2e-16 ***
trt              -0.322566   0.103304  -3.122 0.001793 **
age              -0.046394   0.005252  -8.833  < 2e-16 ***
centre            0.434635   0.111825   3.887 0.000102 ***
factor(bpd)2     -0.224673   0.240602  -0.934 0.350409
factor(bpd)3     -1.564576   0.522565  -2.994 0.002753 **
sex               0.630047   0.157306   4.005  6.2e-05 ***
factor(bpd)2:sex -0.240213   0.286260  -0.839 0.401390
factor(bpd)3:sex  1.462261   0.539107   2.712 0.006680 **

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.3757     0.2108   1.782   0.0748 .
centre         0.6715     0.3006   2.234   0.0255 *
factor(bpd)2  -0.8636     0.3943  -2.190   0.0285 *
factor(bpd)3  -0.5971     0.4031  -1.481   0.1385
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -515.6 on 13 Df

> exp(unlist(slfhrm.zip$coefficients))
  count.(Intercept)              count.trt              count.age
         11.8279887              0.7242881              0.9546655
       count.centre     count.factor(bpd)2     count.factor(bpd)3
          1.5443995              0.7987771              0.2091766
          count.sex count.factor(bpd)2:sex count.factor(bpd)3:sex
          1.8776983              0.7864606              4.3157053
   zero.(Intercept)            zero.centre      zero.factor(bpd)2
          1.4559782              1.9571323              0.4216574
  zero.factor(bpd)3
          0.5504194

> newindividual <- data.frame(trt=1,age=45,centre=0,bpd=3,sex=0)

> pi.new <- predict(slfhrm.zip,newindividual,type="zero")
> pi.new
        1
0.4448758

> mu.new <- predict(slfhrm.zip,newindividual,type="count")
> mu.new
       1
0.222147

> exp(-1*mu.new)
        1
0.8007976
```

(i) Interpret *carefully* the centre effects and the type of personality disorder effects in `slfhrm.zip`, providing the effect estimates on the *relevant scale*. Confidence intervals are not required.

(ii) What is the probability that a female patient, aged 45, with a personality disorder (other than borderline) who was prescribed CBT in Centre A and followed up for six months, is never at risk of self-harming, if she was observed to have not self-harmed during the six months after receiving CBT?

END OF PAPER