UNIVERSITY OF
CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Wednesday, 5 June, 2013 9:00 am to 12:00 pm

PAPER 30

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

(a) Write down the general form of an ordinary linear model on n observations $\boldsymbol{Y}$ with an $n \times p$ covariate matrix X. Make sure you define all the parameters you mention, and state any conditions on X.

(b) Show that $P = X(X^TX)^{-1}X^T$ is a symmetric projection matrix; that is, show that $P^T = P$ and $P^2 = P$.

(c) Define the *fitted values* and the *residuals* in terms of P and $\boldsymbol{Y}$. Derive their distributions. [*Hint: For random vectors* $\boldsymbol{Y}$, $\boldsymbol{Z}$ *and matrices of appropriate dimensions* A, B, *we have* $\operatorname{Cov}(A\boldsymbol{Y}, B\boldsymbol{Z}) = A\operatorname{Cov}(\boldsymbol{Y}, \boldsymbol{Z})B^T$. *You may assume any other standard results about multivariate normal distributions, but these must be stated clearly.*]

(d) Show that the fitted values and residuals are independent.

A statistician is considering the relationship between two variables `x` and `y`. She fits a linear model using the `R` command `lm1 = lm(y ~ x)`. Figure 1 shows the residuals from the model `lm1` plotted against the fitted values.

(e) The plot suggests a violation of the modelling assumptions made by the linear model. Explain what the assumption is, and how the plot diagnoses the violation.

(f) Can you suggest another model that the statistician might try? Write down the algebraic form of this model in full.

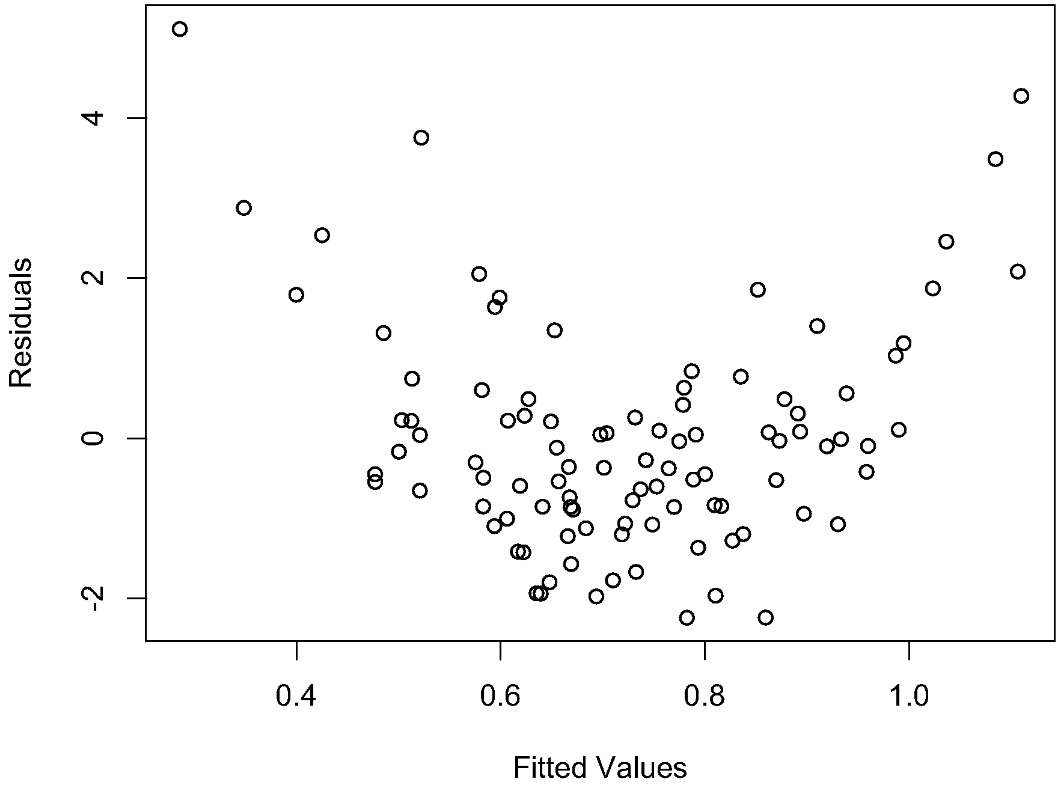

Figure 1: Scatter plot of residuals against fitted values for the model `lm1`.

2

Let $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ for each $i = 1, \ldots, I$ and $j = 1, \ldots, J$, where the ε_{ij}'s are independent normal random variables with mean 0 and variance σ^2, and subject to the restriction that $\alpha_1 = 0$.

(a) Derive the maximum likelihood estimates of μ and $\alpha_2, \ldots, \alpha_I$. Call them $\hat{\mu}$ and $\hat{\alpha}_2, \ldots, \hat{\alpha}_I$ respectively.

(b) Find the univariate distributions of $\hat{\mu}$ and $\hat{\alpha}_2$, and hence show that

$$\frac{\mathrm{se}(\hat{\alpha}_i)}{\mathrm{se}(\hat{\mu})} = \sqrt{2}$$

for each $i = 2, \ldots, I$, where $\mathrm{se}(\hat{\beta})$ is the standard error of $\hat{\beta}$.

Professor Eccentric wishes to test the effect of various types of chocolate bar upon his lecturing speed. Over a 24 lecture course he records:

`day`	the day of the week on which the lecture was given (Tuesday, Thursday or Saturday);
`choc`	the type of chocolate bar he ate (A, B, C or D);
`boards`	the number of chalkboards he gets through during the lecture.

Each combination of `choc` and `day` is observed exactly twice. Professor Eccentric fits two models using `R`, `lm.cd` and `lm.c`; an edited version of his output appears below. The professor uses `R`'s default corner-point constraints.

(c) Write down algebraic form of the model being fitted as `lm.cd`, defining all symbols and making clear any distributional assumptions or identifiability constraints.

(d) Carry out the hypothesis test being performed on the line of output marked with `(++)` (some of the values have been deleted and replaced with `x`'s). You should be sure to state the null hypothesis, test statistic, null distribution of the test statistic, your conclusion, and what it tells Professor Eccentric about his model.

[*Hint: The 0.95-quantile of an $F_{a,b}$-distribution for various a, b are given in the table below to assist you.*]

$a \backslash b$	17	18	19	20
1	4.451	4.414	4.381	4.351
2	3.592	3.555	3.522	3.493
3	3.197	3.160	3.127	3.098

(e) Look at all the `R` output. What would you report to Professor Eccentric about the effect of his choice of chocolate bar and the day of the week on his lecturing?

```
> profecc
   day choc boards
1 Thur    A     11
2  Sat    B     17
```

```
3  Tues    C     10
4  Thur    D      9
...
24 Tues    D     13

> lm.cd = lm(boards ~ choc + day, data=profecc)
> anova(lm.cd)
Analysis of Variance Table

Response: boards
          Df  Sum Sq Mean Sq F value  Pr(>F)
choc       3  70.792 23.5972  3.9665 0.02473
day       xx   9.750    xxxx    xxxx    xxxx    (++)
Residuals 18 107.083  5.9491
---

>
> lm.c = lm(boards ~ choc, data=profecc)
> summary(lm.c)

Call:
lm(formula = boards ~ choc, data = profecc)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.1667     0.9867  14.357 5.38e-12
chocB        -2.5000     1.3954  -1.792  0.08835
chocC        -4.8333     1.3954  -3.464  0.00245
chocD        -2.8333     1.3954  -2.030  0.05582
---

Residual standard error: 2.417 on 20 degrees of freedom
Multiple R-squared: 0.3773,Adjusted R-squared: 0.2839
F-statistic: 4.039 on 3 and 20 DF,  p-value: 0.02133
```

3

(a) Define an exponential dispersion family with natural parameter θ. Define the mean parameter μ and variance function V.

(b) Let Y be a Poisson random variable with mean λ. Show that the collection of distributions of Poisson random variables for varying λ is an exponential dispersion family with dispersion parameter 1, and find the natural parameter, canonical link function, and variance function.

A council officer measures the number of car crashes reported to the police (`nacc`) in Cambridgeshire each day for two years. In the second year, the police begin a road safety campaign to reduce accidents. The council officer wishes to determine whether or not the campaign has been effective in reducing accidents. Study the edited R output below.

(c) Carefully write down the model being fitted as `glm1`. Give estimates and approximate 95% confidence intervals each of the parameters, and interpretations of them with regard to the original problem.

(d) Describe the model being fitted as `glm2`; explain why this model may be more appropriate than `glm1`, in the context of the council officer's data.

(e) What conclusion should the council officer draw about the road safety campaign's effect on accident rates?

```
> dat
    nacc yr
1      5  1
2      8  1
3      3  1
4      4  1
5      7  1
...
726    5  2
727    1  2
728    3  2
729    5  2
730    2  2
>
> glm1 = glm(nacc ~ yr, family=poisson, data=dat)
> summary(glm1)

Call:
glm(formula = nacc ~ yr, family = poisson, data = dat)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.78810    0.02141  83.526  < 2e-16
```

```
yr2         -0.09816    0.03105  -3.162  0.00157
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2587.8  on 729  degrees of freedom
Residual deviance: 2577.8  on 728  degrees of freedom

>
> glm2 = glm(nacc ~ yr, family=quasipoisson, data=dat)
> summary(glm2)

Call:
glm(formula = nacc ~ yr, family = quasipoisson, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.78810    0.04014  44.549   <2e-16
yr2         -0.09816    0.05821  -1.686   0.0922
---

(Dispersion parameter for quasipoisson family taken to be 3.515351)

    Null deviance: 2587.8  on 729  degrees of freedom
Residual deviance: 2577.8  on 728  degrees of freedom
```

4

There are 100 people taking part in Mr Mussel's 10 week weight training programme; they each measure the amount of weight in kilograms that they can lift in weeks 0, 2, 4, 6, 8 and 10. The data look like this:

Subject	**Week**					
	0	2	4	6	8	10
1	66	84	87	120	121	128
2	48	63	59	73	76	88
3	42	42	54	56	55	38
⋮		⋮			⋮	
100	56	79	87	84	117	107

Mr Mussel wants to know how much extra weight his trainees can expect to be able to lift after completing his programme. He stores the data in an `R` data frame called `dat`.

```
> dat
subject wks weight
      1   0     66
      1   2     84
      1   4     87
      1   6    120
      1   8    121
      1  10    128
      2   0     48
      2   2     63
      2   4     59
......
    100   6     84
    100   8    117
    100  10    107
```

(a) Explain why an ordinary linear model with outcome `weight` and covariate `wks` might not be appropriate for these data.

(b) Consider the `R` output below. Write down, in algebraic form, the model fitted as `lmem1` and explain how it overcomes the difficulty mentioned in part (a).

(c) Write a short report to answer Mr Mussel's question, using the analyses in the `R` output. Your answer should include:

- a full algebraic description of the model `lmem2`;
- which of `lmem1` and `lmem2` you prefer and why;
- interpretations of `all` parameters in your preferred model;
- how you estimate an average person might expect to perform over the 10 week programme;

- an estimate of how the most successful individuals perform during the programme.

```
> lmem1 = lme(weight ~ wks, data=dat, random = ~ 1 | subject)
> summary(lmem1)
Linear mixed-effects model fit by REML

Random effects:
 Formula: ~1 | subject
        (Intercept) Residual
StdDev:    21.67593 14.23496

Fixed effects: weight ~ wks
               Value Std.Error  DF  t-value p-value
(Intercept) 61.09286 2.3999739 499 25.45563       0
wks          2.83143 0.1701403 499 16.64173       0
 Correlation:
    (Intr)
wks -0.354

Number of Observations: 600
Number of Groups: 100

> lmem2 = lme(weight ~ wks, data=dat, random = ~ wks | subject)
> summary(lmem2)
Linear mixed-effects model fit by REML

Random effects:
 Formula: ~wks | subject
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 15.785582 (Intr)
wks          2.738666 0.117
Residual     9.923280

Fixed effects: weight ~ wks
               Value Std.Error  DF  t-value p-value
(Intercept) 61.09286 1.7342575 499 35.22710       0
wks          2.83143 0.2984464 499  9.48723       0
 Correlation:
    (Intr)
wks -0.038

Number of Observations: 600
Number of Groups: 100

> lmem1a = lme(weight ~ wks, data=dat, random = ~ 1 | subject, method="ML")
> lmem2a = lme(weight ~ wks, data=dat, random = ~ wks | subject, method="ML")
> anova(lmem1a, lmem2a)
       Model df      AIC      BIC    logLik   Test
lmem1a     1  4 5165.779 5183.367 -2578.890
lmem2a     2  6 4941.819 4968.200 -2464.909 1 vs 2
        L.Ratio p-value
lmem1a
lmem2a 227.9604  <.0001
```

SECTION B

5

(a) Briefly describe what is meant by

(i) observed heterogeneity

(ii) unobserved heterogeneity

(iii) (true) contagion

(b) A recent randomised controlled smoking cessation trial compared the effectiveness of a combined intervention of cognitive behavioural therapy and nicotine replacement therapy (CBT & NRT) with monotherapy of nicotine replacement (NRT only) in improving the likelihood for smokers to quit smoking. The trial recruited 100 smokers who, after being advised by their General Practitioners (GPs) to give up smoking four weeks prior to randomisation, failed to spend at least one full week without smoking over that one month period immediately before the start of the study.

The study was conducted over a ten-week period where at the end of each week the smokers were asked to record in their diary whether or not they had a smoking-free week (coded 1 for a smoking-free week; 0 for a smoking week). Participants were randomised with probability of a half to either receipt of CBT & NRT or NRT only (coded 1 for CBT & NRT; 0 for NRT only) and, at the start of the study, explanatory variable information on gender of participant (1 for male; 0 for female) and age was recorded. The data collected are stored in the R data-frame, `quitsmoke.dat`, and the first few lines of the data-set are presented below using the R command:

```
> head(quitsmoke.dat)
  id sex age trt y1 y2 y3 y4 y5 y6 y7 y8 y9 y10
1  1   0  38   0  0  0  0  0  1  0  0  0  0   0
2  2   1  28   0  0  0  0  1  0  0  0  0  1   0
3  3   0  35   1  0  1  1  0  1  0  0  1  1   0
4  4   0  23   1  0  0  1  0  0  1  0  1  1   1
5  5   1  25   1  1  0  0  0  0  0  0  0  0   0
6  6   1  30   1  0  0  0  0  1  0  0  0  1   0
```

The reported variables are

`id` = unique smoker identification number,
`sex` = gender of smoker,
`age` = smoker's age at start (in years),
`trt` = intervention type, and
`y1`,..., `y10` = binary week 1 to week 10 smoking cessation outcome variables.

Examine carefully the following edited R code and output.

```
> n <- 100
> id <- rep(quitsmoke.dat$id, rep(10,n))
> sex <- rep(quitsmoke.dat$sex, rep(10,n))
> age <- rep(quitsmoke.dat$age, rep(10,n))
> trt <- rep(quitsmoke.dat$trt, rep(10,n))
> y <- as.vector(t(quitsmoke.dat[,5:14]))
> ylag.one <- as.vector(t(cbind(0,quitsmoke.dat[,5:13])))
> ylag.two <- as.vector(t(cbind(0,0,quitsmoke.dat[,5:12])))
> ycumprev <- as.vector(apply(cbind(0,quitsmoke.dat[,5:13]),MARGIN=1,FUN=cumsum))
```

```
> id[1:20]
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
> y[1:20]
 [1] 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0
> ylag.one[1:20]
 [1] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1
> ylag.two[1:20]
 [1] 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0
> ycumprev[1:20]
 [1] 0 0 0 0 0 1 1 1 1 1 0 0 0 0 1 1 1 1 1 2
```

```
> # First Model
> model1.glm <- glm(y~sex+age+trt+ylag.one, family=binomial)
> summary(model1.glm)

Call:
glm(formula = y ~ sex + age + trt + ylag.one, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1074  -0.7798  -0.7551   1.2683   1.8568

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.121966   0.413227  -2.715  0.00662
sex         -0.458154   0.146050  -3.137  0.00171
age          0.002295   0.011661   0.197  0.84399
trt          0.397702   0.145662   2.730  0.00633
ylag.one     0.449495   0.160742   2.796  0.00517
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1189.7  on 999  degrees of freedom
Residual deviance: 1164.4  on 995  degrees of freedom

Number of Fisher Scoring iterations: 4
```

```
> # Second Model
> model2.glm <- glm(y~sex+age+trt+ylag.one+ylag.two, family=binomial)
> summary(model2.glm)

Call:
glm(formula = y ~ sex + age + trt + ylag.one + ylag.two, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2621  -0.8648  -0.7278   1.2761   1.8872

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.227399   0.417261  -2.942  0.00327
sex         -0.426611   0.147121  -2.900  0.00373
age          0.002509   0.011718   0.214  0.83042
trt          0.370279   0.146682   2.524  0.01159
ylag.one     0.399898   0.162741   2.457  0.01400
ylag.two     0.538786   0.173447   3.106  0.00189
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1189.7  on 999  degrees of freedom
Residual deviance: 1155.0  on 994  degrees of freedom

Number of Fisher Scoring iterations: 4

> # Third Model
> model3.glm <- glm(y~sex+age+trt+ycumprev, family=binomial)
> summary(model3.glm)

Call:
glm(formula = y ~ sex + age + trt + ycumprev, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7749  -0.7961  -0.6689   1.1100   1.9562

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.417630   0.423986  -3.344 0.000827
sex         -0.364092   0.149944  -2.428 0.015175
age          0.001224   0.011880   0.103 0.917950
trt          0.327118   0.149391   2.190 0.028548
ycumprev     0.399296   0.058739   6.798 1.06e-11
---
```

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

    Null deviance: 1189.7  on 999  degrees of freedom
Residual deviance: 1122.0  on 995  degrees of freedom

Number of Fisher Scoring iterations: 4
```

(i) Write out mathematically the model being fitted in `model1.glm`, remembering to define all notation used and stating any assumptions. Additionally, write out the likelihood contribution for the ith individual in the study.

(ii) Carry out a likelihood ratio test to determine whether `model2.glm` is an improvement over `model1.glm`.

(iii) Determine whether `model3.glm` would be preferred over the better fitting model in b(ii).

(iv) Interpret the output from the best of the three models presented. This should be done in the context of the study.

(v) Based on the estimates from the best model, give an expression for the probability that a 20-year old female smoker (who satisfies the conditions for entering the study) would abstain from smoking (i.e. stop smoking) during her first three weeks on NRT.

6

Parkinson's disease is a degenerative disorder of the central nervous system that is more common in the elderly with mean age of onset around 60 years. Mild cognitive impairment may be a troublesome symptom of the disease and can progress onto dementia, a more severe loss of intellectual abilities that interferes with activities of daily living so profoundly that it may not be possible for a person to live independently.

Researchers into ageing and neurological diseases have decided to investigate the progression of cognitive functioning over time (from diagnosis) in an observational cohort of Parkinson's patients who were diagnosed in the last ten years and referred to a specialist neurological clinic. In particular, they are interested in modelling the transitions between three states – mild cognitive impairment (state 1), dementia (state 2) and death (state 3), and investigating the impact of time-independent explanatory variables, age at diagnosis of Parkinson's (in years), educational attainment (coded 0 for $\leqslant$12 years of full time education; and 1 for $>$12 years of full time education), and gender (coded 0 for female; and 1 for male), on the rates of transition between the various states. Patients in this clinic are followed up intermittently, but if they had died the exact death time was recorded.

The data were read into R and named `parkinson.dat`. The data for patients 7, 8 and 9 are shown below.

```
> parkinson.dat[(parkinson.dat$subject >=7) & (parkinson.dat$subject <=9),]
   subject     time state sex edu diagage
35       7 0.000000     1   0   1      69
36       7 2.473380     2   0   1      69
37       7 6.181523     2   0   1      69
38       7 6.295681     2   0   1      69
39       7 7.107472     2   0   1      69
40       7 7.466763     2   0   1      69
41       7 7.924641     2   0   1      69
42       7 7.990554     3   0   1      69
43       8 0.000000     1   0   0      64
44       8 3.261286     1   0   0      64
45       8 4.492359     1   0   0      64
46       9 0.000000     1   0   1      51
47       9 1.289561     1   0   1      51
48       9 3.984003     1   0   1      51
49       9 4.434085     1   0   1      51
50       9 8.772825     2   0   1      51
51       9 9.046087     2   0   1      51
```

The variables are

`subject` = unique patient number,
`time` = time (in years) from diagnosis of Parkinson's,
`state` = cognition or death states,
`sex` = gender of patient,
`edu` = educational attainment,
`diagage` = age (in years) at diagnosis of Parkinson's.

(a) Construct the likelihood contributions of these three patients (7, 8 and 9) for time-homogeneous Markov multi-state processes given by the `state` and `time` variables in `parkinson.dat`. You need to define any terms used. (Approximation of the `time` variable to two decimal places is permitted.)

(b) The following is the edited R output from a multi-state modelling analysis using the `msm` package in R.

```
> parkinson.msm0

Call:
msm(formula = state ~ time, subject = subject, data = parkinson.dat, qmatrix = Qmat,
death = TRUE, method = "Nelder-Mead")

Maximum likelihood estimates:
Transition intensity matrix

        State 1                    State 2                     State 3
State 1 -0.2043 (-0.2557,-0.1632) 0.1849 (0.1435,0.2383)      0.01935 (0.006759,0.05541)
State 2 0                         -0.06143 (-0.1063,-0.03552) 0.06143 (0.03552,0.1063)
State 3 0                         0                           0

-2 * log-likelihood:  462.7531


> pmatrix.msm(parkinson.msm0,t=1)
          State 1   State 2    State 3
State 1 0.8152585 0.1620358 0.02270565
State 2 0.0000000 0.9404163 0.05958370
State 3 0.0000000 0.0000000 1.00000000


> parkinson.msm1

Call:
msm(formula = state ~ time, subject = subject, data = parkinson.dat, qmatrix = Qmat,
covariates = ~sex + edu + diagage, death = TRUE, fixedpars = c(5, 6),
method = "Nelder-Mead")

Maximum likelihood estimates:
Transition intensity matrix with covariates set to  their means
```

```
        State 1                   State 2                   State 3
State 1 -0.1947 (-0.2478,-0.1529) 0.1821 (0.1418,0.2338)    0.0126 (0.001996,0.07954)
State 2 0                         -0.045 (-0.0918,-0.02206) 0.045 (0.02206,0.0918)
State 3 0                         0                         0

Log-linear effects of sex

        State 1 State 2                  State 3
State 1 0       0.09534 (-0.3968,0.5875) 0
State 2 0       0                        0
State 3 0       0                        0

Log-linear effects of edu

        State 1 State 2                   State 3
State 1 0       -0.4306 (-0.9363,0.07503) 1.223 (-2.507,4.953)
State 2 0       0                         -1.49 (-2.547,-0.4327)
State 3 0       0                         0

Log-linear effects of diagage

        State 1 State 2                     State 3
State 1 0       0.007627 (-0.03452,0.04978) 0.1262 (-0.06247,0.3148)
State 2 0       0                           0.07984 (-0.007765,0.1674)
State 3 0       0                           0

-2 * log-likelihood:  442.0713

> qchisq(0.95,1:12)
 [1]  3.841459  5.991465  7.814728  9.487729 11.070498 12.591587 14.067140
 [8] 15.507313 16.918978 18.307038 19.675138 21.026070
```

(i) This part of the question applies to the model as specified in fitting `parkinson.msm0`. Draw the transition diagram, including on it the estimated transition intensities corresponding to each type of transition. What is the estimated mean time spent (with 95% confidence interval) in the dementia state before dying? Given that a Parkinson's disease patient currently has dementia, what is this patient's estimated probability of being dead 2 years later based on this model?

(ii) Write out mathematically the multi-state model corresponding to `parkinson.msm1`, making sure to define all notation used and stating all assumptions being made.

(iii) Formally compare `parkinson.msm1` to `parkinson.msm0` to decide if there is an improvement from fitting this more complicated model. Interpret the effects of *all* explanatory variables in `parkinson.msm1` on the dementia to death transition.

END OF PAPER