S9209/9210/9211/9212/102C

DUBLIN INSTITUTE OF TECHNOLOGY

School of Mathematical Sciences

DT9209 MSc Applied Mathematics
DT9210 MSc Applied Mathematics
DT9211 MSc Applied Mathematics
DT9212 MSc Applied Mathematics

SUMMER EXAMINATIONS 2015/2016

MATH 9952: Modern Applied Statistical Modelling

Dr J Condon

Dr C Hills

Professor E O'Riordan

9.30am-12.30pm, Friday, 13 May 2016

Duration: 3 hours

Attempt four questions only

All questions carry equal marks

Approved calculators may be used

Mathematical tables are provided

New Cambridge Statistical Tables are provided

1. a) Consider the multiple regression model model:

$$Y = X\beta + \varepsilon \qquad E[\varepsilon] = 0 \qquad Var[\varepsilon] = \sigma^2 I_n$$

where ε is an $n \times 1$ vector of independent normally distributed random variables, β is a $p \times 1$ vector of unknown parameters ($1 < p < n$), and X is an $n \times p$ matrix of known constants.

Give the formula for the SS error and hence show that the expected value of MS error is σ^2. (9)

b) Research is conducted on the effect of two types of antibiotics in treating leprosy. For each patient, the volume of leprosy bacillus are recorded from six sites on their body both pre-treatment and again post-treatment. Five patients were assigned to one of two antibiotic treatments (drugs A and D) and five were assigned to a standard drug as a control (drug F). The data are shown in table 1.1 below. A model is fitted to these data with post-treatment bacillus volume as the response, drug as a categorical predictor, pre-treatment bacilli volume as a continuous predictor **and** the interaction of drug and pre-treatment bacillus volume.

Table 1.1: Leprosy Data.

Drug	PreTreatment Volume	PostTreatment Volume
A	11	6
A	8	0
A	5	2
A	14	8
A	19	11
D	6	0
D	6	2
D	7	3
D	8	1
D	18	18
F	16	13
F	13	10
F	11	18
F	9	5
F	21	23

i) Give the X matrix for this model using the constraint that the coefficient for the drug=A group is 'set to zero'. (2)

ii) Explain for the context of this experiment (perhaps using a suitable sketch), what the interaction term is modelling. (5)

iii) Figure 1.1 shows some output from the R console.

```
1  > fit=lm(post_treat~factor(drug)*pre_treat,data=leprosy)
2  > drop1(fit,test='F')
3  Single term deletions
4
5  Model:
6  post_treat ~ factor(drug) * pre_treat
7                        Df Sum of Sq    RSS    AIC F value Pr(>F)
8  <none>                              106.30 41.374
9  factor(drug):pre_treat  2    24.939 131.24 40.535  1.0557 0.3874
10 > drop1(update(fit,.~.-factor(drug):pre_treat),test='F')
11 Single term deletions
12
13 Model:
14 post_treat ~ factor(drug) + pre_treat
15              Df Sum of Sq    RSS    AIC F value    Pr(>F)
16 <none>                     131.24 40.535
17 factor(drug)  2     74.01 205.25 43.243  3.1017 0.0854662
18 pre_treat     1    365.56 496.80 58.502 30.6390 0.0001766
```

Figure 1.1: Results for Leprosy data.

There are outputs from three hypothesis tests in Figure 1.1.

- Give the formula of the tests statistic for general linear hypothesis being conducted in line 9 of the output explaining all terms. State the distribution of the test statistic under the null hypothesis. (5)
- With reference to the three hypothesis tests shown, state the conclusions for the model being fitted. (4)

[25]

2. a) Let $\eta_i = x_i'\beta$ be the linear predictor for a regression model. Using η_i, give the likelihood and log likelihood for the logistic regression model fitted to binary data. (6)

b) Describe the likelihood ratio test and its uses in model building when fitting generalised linear models. (8)

c) Customer 'churn' is sometimes defined as the loss of a customer to a competitor. A telecommunications supplier is conducting an analysis of their customer churn using historic data. All customer accounts are examined and any customers that have 'churned' in the last month are identified. They analyse the data using logistic regression. The response being modelled is the probability that a customer is a 'churn' with the following predictors:

MonthlyCharges: Average monthly charges ($) levied over the last 4 months.

OnlineBackup: Whether an online backup service was used by the customer with three possible values: Yes; No; No internet service

MultipleLines: Whether multiple phone lines are in use by the customer with three possible values: No (i.e. just one phone line); Yes; No phone service.

A portion of the output from R is given in Figure 2.1 below.

```
Call:
glm(formula = churn~monthlycharges + factor(onlinebackup) +
    factor(multiplelines), family = binomial(), data = churn)

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -1.026489   0.212949  -4.820 1.43e-06
monthlycharges                          0.008011   0.002829   2.831  0.00463
factor(onlinebackup)No internet service -1.735128   0.214896  -8.074 6.79e-16
factor(onlinebackup)Yes                 -0.919506   0.092537  -9.937  < 2e-16
factor(multiplelines)No phone service   -0.122050   0.171242  -0.713  0.47601
factor(multiplelines)Yes                -0.014634   0.096450  -0.152  0.87940

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4019.1  on 3526  degrees of freedom
Residual deviance: 3683.3  on 3521  degrees of freedom
AIC: 3695.3

Number of Fisher Scoring iterations: 5
```

Figure 2.1: Partial R output for logistic regression model.

i) Find the estimated odds ratio that a customer with no internet service will churn over a customer that uses the online backup service - all other variables being equal. (2)

ii) Discuss the evidence for the predictor 'MonthlyCharges' being related to the response based on the output given. (3)

iii) Predict the probability that a customer with the following values of the predictors will be a churn:

MonthlyCharges=$50,
OnlineBackup=Yes,
MultipleLines=No phone service. (2)

iv) Describe how a 95% confidence interval for the fitted probability predicted in part (iii) could be calculated. (4)

[25]

3. An experiment is conducted to assess three different machine designs in a manufacturing setting. Each of six employees operates each machine once. The response is an overall rating, which takes into account the number and quality of components produced by the employee over a set period. Table 3.1 shows the data.

The following model is proposed for the data.

$$Y_{ij} = \beta_0 + \beta_1 d_2 + \beta_2 d_3 + \gamma_i + \epsilon_{ik}$$

where Y_{ij} is the rating for employee i ($i = 1, \ldots, 6$) using machine j ($j = 1, \ldots, 3$)); β_0, β_1, β_2 are unknown fixed effect parameters; d_k is the dummy variable taking value 1 where $j = k$ and zero otherwise; $\gamma_i \overset{iid}{\sim} N(0, \sigma_b^2)$; $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, and γ and ϵ are mutually independent.

Table 3.1: Machine data

machine	person	rating	machine	person	rating
1	1	52.00	1	4	51.10
1	2	51.8	1	5	50.9
1	3	60.0	1	6	46.4
2	1	64.0	2	4	63.2
2	2	59.7	2	5	64.8
2	3	68.6	2	6	43.7
3	1	67.5	3	4	64.1
3	2	61.5	3	5	72.1
3	3	70.8	3	6	62.0

a) State the following: (1) the variance of Y_{ij}; (2) the correlation of Y_{ij} with $Y_{ij'}$ where $j \neq j'$; (3) the correlation of Y_{ij} with $Y_{i'j}$ where $i \neq i'$. (5)

b) Give the log likelihood for this model, explaining all terms. (5)

c) Derive the profile log likelihood for the model in terms of the covariance parameters. Outline the Newton-Raphson method for fitting this model (there is no need to give detailed analytic derivatives). (7)

d) The resulting output from fitting the model using R is shown in Figure 3.1.

i) Give the R code for fitting this random effects model. (2)

ii) The log-likelihood for a model excluding the random person effect is -56.88. Conduct an appropriate test to determine the statistical significance of the random person effect and state your conclusions. (3)

iii) Give the estimated variance-covariance matrix for the responses in this model. (3)

```
Linear mixed model fit by maximum likelihood

     AIC      BIC   logLik deviance df.resid
   117.8    122.2    -53.9    107.8       13

Random effects:
 Groups   Name        Variance Std.Dev.
 person   (Intercept) 19.12    4.373
 Residual             13.40    3.661
Number of obs: 18, groups:  person, 6

Fixed effects:
                  Estimate Std. Error t value
(Intercept)         52.033      2.328  22.348
factor(machine)2     8.633      2.114   4.084
factor(machine)3    14.300      2.114   6.765
```

Figure 3.1 Output from the machine data analysis.

[25]

4. a) The Rpart algorithm (recursive partitioning for classification) is applied to a set of data concerning the probability that a customer of a mobile phone company will churn (i.e. be lost as a customer) over the next 12 months. The modelled classification is the value of churn (yes or no) with two predictors: age of the customer and type of data allowance they have, one of: A all-you-can-eat (i.e. unlimited) data allowance (coded A) ; limited data allowance with a monthly data cap (coded D).

The model is fitted to the small data fragment given in Table 4.1.

Table 4.1: Churn data fragment

Observation	1	2	3	4	5	6	7	8	9
Churn	yes	yes	yes	yes	yes	no	no	no	no
Age	28	28	24	24	26	28	26	26	24
Allowance	A	D	A	D	D	A	D	D	A

The R command used to fit this model is given below.

```
rpart(Churn~Age+Allowance,data=churn,minsplit=1,maxdepth=2,cp=-1,minbucket=1)
```

Fit this model and draw the resulting partition tree, labelling each node appropriately.

[Hint: An outline of the resulting tree is shown in Figure 4.1]

(10 marks)

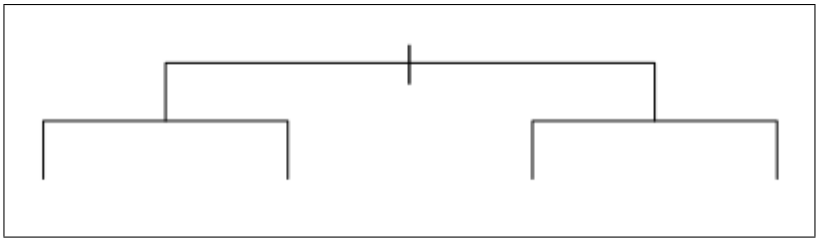

Figure 4.1: Tree outline.

b) The kyphosis data represents data on children who have had corrective spinal surgery. The dataset consists of the following variables: Kyphosis with levels 'absent' or 'present' indicating if a kyphosis (a type of deformation) was present after the operation; Age = age in months; Number = the number of vertebrae involved; Start = the number of the first vertebra operated on. A rpart model was fitted to these data and edited output from the R console is shown in Figure 4.2 below.

```
> fit=rpart(Kyphosis~Age+Number+Start,data=kyphosis)
> fit
n= 81

 1) root 81 17 absent
   2) Start>=8.5 62  6 absent
     4) Start>=14.5 29  0 absent
     5) Start< 14.5 33  6 absent
      10) Age< 55 12  0 absent
      11) Age>=55 21  6 absent
        22) Age>=111 14  2 absent
        23) Age< 111 7  3 present
   3) Start< 8.5 19  8 present
```

Figure 4.2: Rpart fit to the Kyphosis data.

i) Draw the resulting tree, labelling each node appropriately.

ii) Predict the classification for a child with the following predictors: (a) showing how the tree is used to determine this classification and (b) give the predicted probability for that class membership.
Predictors: Age=120, start=12, number=5.

(10 marks)

c) Discuss the role of the complexity parameter (cp) is the rpart function.

(5 marks)

[25]

5. a) A density that is sometimes used in modelling survival times is the log-logistic distribution. It has the following density function:

$$f(t) = \frac{e^{\theta}\kappa t^{\kappa-1}}{(1+e^{\theta}t^{\kappa})^2}$$

Given that κ is known to be equal to 1, find expressions for the following:

i) the cumulative distribution function,

ii) the survivor function,

iii) the hazard function,

iv) an expression for the median survival time.

(8)

b) The following data are survival times in months following heart surgery. Any time followed by + indicates that that time was a censored survival time.

$$2,\ 14^{+},\ 22,\ 22,\ 22,\ 22^{+},\ 39^{+},\ 50,\ 55^{+},\ 50^{+},\ 78^{+},\ 119,\ 121^{+},\ 125$$

i) Calculate the K-M of the survivor function.

ii) Sketch the plot of the K-M estimate of the survivor function.

iii) Estimate the median survival time for these data.

(10)

c) Imagine a Cox PH model being fitted to data with a single treatment covariate x, where x is coded 0=placebo, 1=treatment. Assume that there are n observations which may be assumed to be independent. Give the partial log likelihood for this model assuming no tied death times and explaining all terms used. (7)

[25]