RS9209/9210/9211/9212/102C

DUBLIN INSTITUTE OF TECHNOLOGY

School of Mathematical Sciences

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

AUTUMN EXAMINATIONS 2015/2016

MATH 9952: Modern Applied Statistical Modelling

Dr J Condon

Dr C Hills

Professor E O'Riordan

sometime , someday, August 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. Consider the model

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

where Y is an $n \times 1$ vector of 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 with rank p.

a) i) Derive the least squares estimator $\hat{\beta}$ of β. Show that $\hat{\beta}$ is an unbiased estimator of β and determine its variance-covariance matrix. (9 marks)

ii) Derive an expression for the variance-covariance matrix of fitted values $\hat{Y}$ from a multiple regression model. Hence, give the expression for calculating a $(1-\alpha)\%$ confidence interval for a fitted value with a vector of predictors x_i'. (5 marks)

b) A group of medical researchers conducts an experiment to investigate the relationship between the level of a particular steroid in the human body and heart rate. They record the standardised blood steroid level in the blood of each of seven human subjects. They also record each subject's average resting diastolic heart rate (mmHg). These data are shown in the table below.

Standardised blood steroid level	-2	-2	0	0	1	2	4
Average resting heart rate	67	70	71	72	75	75	74

Medical theory suggests that a quadratic model in terms of steroid level would be a good predictor of resting heart rate, i.e. the model is $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, where y = heart rate and x = blood steroid level.

The $(X'X)^{-1}$ matrix for this model is,

$$(X'X)^{-1} = \frac{1}{21,956}\begin{pmatrix} 5596 & 738 & -670 \\ 738 & 1294 & -312 \\ -670 & -312 & 194 \end{pmatrix}$$

i) Give the matrix X for this model. (2 marks)

ii) Give the formula for the test statistic for a general linear hypothesis of the form $L\beta = 0$ and state the distribution of the test statistics under the null hypothesis. (3 marks)

iii) The mean square error (s^2) for this model is found to be 2.23. The least squares parameter estimates are:

$$\hat{\beta} = \begin{pmatrix} 72.46 \\ 1.56 \\ -0.27 \end{pmatrix}$$

Construct a 95% confidence interval for the mean response when the standardised blood steroid level is 3. (6 marks)

[25]

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

(7 marks)

b) Define the score vector and the observed Fisher information matrix. Hence, outline the Newton-Raphson method for finding the maximum likelihood estimates of the parameters of a Poisson regression model. [NB. there is no need to give detailed derivatives].

(8 marks)

c) A Poisson regression model was used to analyse data relating the rate of caesarean sections used in 20 Irish maternity units to the following predictors: mean age of mothers in the unit (mage); the proportion of private to public patients (pprivate); the proportion of multiple births i.e. twins, triplets etc. (pmultiple). The response variable was the number of deliveries by caesarean section in 2011. The natural logarithm of the total number of births in each unit over that year (ln) was used as an offset term in the model.

Figure 1.1 shows a portion of the output from fitting the model using R's GLM function.

```
Call:
glm(formula = cs ~ pmultiple + mage + pprivate + offset(ln),
    family = poisson(), data = cs)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.819035   0.120470 -15.099  < 2e-16
pmultiple   11.178327   0.283216  39.469  < 2e-16
mage        -0.015414   0.003664  -4.206 2.59e-05
pprivate     1.224237   0.038749  31.594  < 2e-16
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3072.64  on 19  degrees of freedom
Residual deviance:  545.88  on 16  degrees of freedom
AIC: 718.37
```

Figure 1.1 Partial R console output for Poisson regression model.

i) Discuss why the offset term is used in this model. (3 marks)

ii) Estimate the rate of caesarean sections per 100 deliveries for a hospital with mage=29, pprivate=0.22 and pmultiple=0.08. (3 marks)

iii) Estimate the effect that a 0.1 increase in the proportion of private patients would be expected to have on a unit's ceasarean section rate and give a 95% confidence interval for this estimate. (4 marks)

[25]

3. a) Give the general log likelihood for a linear mixed model and show how the regression parameters may be profiled out of this likelihood to give a reduced log likelihood in terms of the variance components only.

(6 marks)

b) The data shown in Table 3.1 is a portion of dataset recording the a particular variable of dental growth measured from 11 boys and 16 girls. The measurement were repeated at ages 8, 10, 12 and 14.

[The variable was the distance in mm from the centre of the pituitary to the pteryo-maximillary fissure. This distance is the relative distance of the two points and may occasionally decrease between ages.]

Table 3.1 Dental growth data.

person	gender	distance	age
1	F	21	8
1	F	20	10
1	F	21.5	12
1	F	23	14
⋮	⋮	⋮	⋮
13	M	21.5	8
13	M	22.5	10
13	M	23	12
13	M	26.5	14
⋮	⋮	⋮	⋮

A model is fitted to these data using the following R code.

```
1 fit=gls(distance~age+factor(gender)+age:factor(gender),correlation = corAR1(
      form = ~1|person),data=dental,method="ML")
2 summary(fit)
```

i) Describe both the mean and variance structures of the vector of responses being fitted by this model.

(7 marks)

ii) A general linear mixed model formulation is as follows:

$$Y = X\beta + Zb + \epsilon$$

where X is an $n \times p$ design matrix for the fixed effects, β is a $p \times 1$ vector of unknown fixed effect parameters, Z is an $n \times q$ design matrix for the random effects $b \sim MVN(0, G)$ and $\epsilon \sim MVN(0, R)$.
[NB: MVN = multivariate normal]
Give a full specification of X, Z, G, and R for the model fitted to the dental data.

(6 marks)

iii) Figure 3.1 shows the (edited) output from the R code given above. Construct the estimated covariance matrix for distance measurement within a person.

(6 marks)

```
Generalized least squares fit by maximum likelihood
  Model: distance ~ age + factor(gender) + age:factor(gender)
  Data: dental
       AIC      BIC    logLik
  452.681 468.7738 -220.3405

Correlation Structure: AR(1)
 Formula: ~1 | person
 Parameter estimate(s):
      Phi
0.6071166

Coefficients:
                        Value Std.Error   t-value p-value
(Intercept)         17.321720 1.6345089 10.597507  0.0000
age                  0.483732 0.1409898  3.430973  0.0009
factor(gender)M     -0.729724 2.1232893 -0.343676  0.7318
age:factor(gender)M  0.285840 0.1831511  1.560676  0.1216

Residual standard error: 2.211512
Degrees of freedom: 108 total; 104 residual
```

Figure 3.1 R output for dental growth data.

[25]

4. a) Define the GINI index of impurity and describe how it is used in the rpart algorithm to build recursive partitioning for classification trees. (You may assume a binary response and an identity loss function matrix.)

(6 marks)

b) The Titanic data represents data on the survivors of the Titanic disaster of 1912. The dataset consists of the following variables: survived with levels 'died' or 'suvived' indicating the survival for the passenger; age = age in years; sex = male or female; embarked = one of S, C or Q indicating the embarkation port of the passenger. An rpart model was fitted to these data and edited output from the R console is shown in Figure 4.1 below.

```
> tfit1=rpart(factor(survived,levels=0:1,labels=c('died','survived'))~age+
    factor(sex)+factor(embarked),cp=-0,minsplit=1,minbucket=1,data=train,method
    ='class',maxdepth=3)
> tfit1
n= 575
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 575 214 died (0.6278261 0.3721739)
   2) factor(sex)=male 378  69 died (0.8174603 0.1825397)
     4) age>=1.5 370  62 died (0.8324324 0.1675676) *
     5) age< 1.5 8   1 survived (0.1250000 0.8750000) *
   3) factor(sex)=female 197  52 survived (0.2639594 0.7360406)
     6) factor(embarked)=S 132  43 survived (0.3257576 0.6742424)
      12) age< 3 4   1 died (0.7500000 0.2500000) *
      13) age>=3 128  40 survived (0.3125000 0.6875000) *
     7) factor(embarked)=,C,Q 65   9 survived (0.1384615 0.8615385) *
```

Figure 4.1 R output for rpart fit to Titanic Data.

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

ii) Predict the classification for a passenger: (a) showing how the tree is used to determine this classification and (b) give the predicted probability for that passenger surviving.
Predictors: sex=female, age=7, embarked=S.

iii) Calculate both the specificity and sensitivity of this tree for classifying passengers that survived.

(13 marks)

c) Write descriptive notes on the following and how they relate to classification:

i) The Receiver Operating Characteristic curve (ROC)

ii) The area under the curve (AUC)

(6 marks)

[25]

5. a) The following data are the times in days to failure of pieces of mechanical equipment following the expiration of their design lifetime. There are two types of equipment, A and B and the event variable is coded as: 0=censored observation; 1=failure time..

Time	1	21	23	42	46	55	83	2	7	57	71	154	361
Event	0	0	0	1	0	1	0	0	1	0	1	0	0
Type	A	A	A	A	A	A	A	B	B	B	B	B	B

i) Construct the Kaplan-Meier estimates of the survivor functions for both groups separately and sketch a plot of both functions on the same chart.

(10 marks)

ii) Conduct a log-rank test to compare the survival times of types A and B. State the null and alternative hypotheses, compute an approximate p-value for the test and state your conclusions.

(7 marks)

b) Customer 'churn' is sometimes defined as the loss of a customer to a competitor. A telecommunications supplier is conducting an analysis of the times to customer churn using historic data. They analyse the data using a Cox proportional hazards (Cox PH) model. The response variables are tenure (time in months with the company) and status (0=has not churned, 1=has churned). The following predictors are included in the model:

MonthlyCharges:	Average monthly charges ($) levied over the last 4 months.
Gender:	Male or Female
Paperlessbilling:	Whether the customer paid using paperless billing or not (levels Yes or No).

Figure 5.1 shows the ouput from the fitted model using the R coxph function.

```
coxph(formula = Surv(tenure, status) ~ factor(gender) + monthlycharges +
    factor(paperlessbilling), data = churn)

  n= 7032, number of events= 1869

                                  coef  exp(coef)   se(coef)      z Pr(>|z|)
factor(gender)Male          -0.0223701  0.9778783  0.0462703 -0.483 0.628765
monthlycharges               0.0027708  1.0027747  0.0008342  3.322 0.000895
factor(paperlessbilling)Yes  0.6566247  1.9282728  0.0562386 11.676  < 2e-16
```

Figure 5.1 R output for Cox PH fit to Churn Data.

i) Give the Cox PH hazard function as specified by this model, explaining all terms.

ii) Give the hazard ratio for a male customer over a female customer and calculate a 95% confidence interval for this hazard ratio.

iii) Discuss the evidence from the output shown in Figure 5.1 that 'paperless-billing' is related to the response.

(8 marks)

[25]