S9002/228B/642C

DUBLIN INSTITUTE OF TECHNOLOGY

School of Mathematical Sciences

DT9002 PG Cert Applied Statistics

SUMMER EXAMINATIONS 2016/2017

MATH 9903: LINEAR & GENERALISED LINEAR REGRESSION MODELS

DR J CONDON

DR C HILLS

DR K HAYES

16.00 – 18.00pm, Tuesday, 09 May 2017

Duration: 2 hours

Answer three questions

All questions carry equal marks

Approved calculators may be used

Mathematical tables are NOT permitted

New Cambridge Statistical Tables are provided

1. Intraocular pressure (IOP) is the fluid pressure inside the eye and can be a risk factor for the disease glaucoma. A multiple regression model is fitted to a dataset consisting of observations made on 262 elderly subjects. The model consists of: intraocular pressure in mmHG (`iop`) as the response variable; age in years (`age`), central corneal thickness in μm (`cct`) and systolic blood pressure in mmHg (`sbp`) as three continuous predictor variables.

   The data is read in to an R data.frame called `IOP`. Part of the output obtained from fitting this model in R is shown here:

```
> summary(fit_iop)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1509 -1.3481  0.1358  1.3869  5.6070

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.20309   41.80696   1.081   0.2806
age         -0.10073    0.02241  -4.496 1.05e-05
cct         -0.06671    0.07661  -0.871   0.3847
sbp          0.07905    0.03545   2.230   0.0266
---

Residual standard error: 2.015 on 258 degrees of freedom
Multiple R-squared:  0.09018, Adjusted R-squared:  0.07961
F-statistic: 8.525 on 3 and 258 DF,  p-value: 2.03e-05
```

   a) Mathematically state the *least squares criterion* for fitting a multiple regression model and illustrate the criterion with reference to the IOP model used here. (6)

   b) Give the R code for fitting the model shown in Figure 1. (4)

   c) Give the complete ANOVA table for the model fitted. State the default null and alternative hypotheses for this table and state your conclusions. (6)

   d) Calculate the estimated IOP for an individual with the following characteristics: age=70, cct=530, sbp=160. (7)

   e) Give the R code that would be used to calculate a 95% confidence interval for the estimate calculated in part d. (5)

   f) Discuss the evidence from the output given that central corneal thickness is related to intraocular pressure. Explain your conclusion by specifying the relevant null and alternative hypotheses, the test statistic and associated p-value. (5)

   [33]

2. A study was conducted to compare a new method of teaching children to learn the manipulation of fractions to a standard teaching method. Children were randomly assigned to receive either the new method or the standard method. The children had no prior exposure to fractions. At the end of 10 hours of instruction, all children are scored on their ability to use fractions using a standardised test. Before the study started, the children were also scored on their general mathematical ability using a standardised test. A portion of the data are shown in the table below.

Score for Fractions	Score for General Maths	Method
103	97	standard
111	104	new
99	94	standard
⋮		

a) Describe how dummy variables are used in multiple regression to model categorical predictors. (6)

b) A model is fitted to these data with the following R code.

```
fit_fractions=lm(fractions~general_maths+factor(method))
```

Describe in detail the model being fitted by this code, including the assumed relationship between the predictors and the response. (9)

c) A further analysis are carried out on these data. The code and edited output from this analysis are shown here:

```
> fit_new=update(fit,.~.+general_maths:factor(method))
> summary(fit_new)

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         98.61928   17.82127   5.534 0.000551
general_maths                        0.09398    0.18822   0.499 0.631017
factor(method)standard              -1.88660   21.81349  -0.086 0.933204
general_maths:factor(method)standard -0.06639    0.22588  -0.294 0.776291
```

i) For the `fit_new` model, predict the fractions score for a student taught under the new method who has a general maths score of 90. (4)

ii) For the `fit_new` model, predict the fractions score for a student taught under the standard method who has a general maths score of 102. (4)

iii) Explain the difference between the `fit` and `fit_new` models. Discuss the evidence for/against the `fit_new` model. (10)

[33]

3. a) Define what is meant by underfitting and overfitting in the context of multiple regression and outline the consequences of both for model fitting and statistical inference.

(8)

b) Researchers were interested in learning how the composition of the cement affected the heat evolved during the hardening of the cement. Therefore, they measured and recorded the following data on 13 samples of cement:

Response	Heat: heat evolved in calories during hardening of cement on a per gram basis
Predictor 1	aluminate: % of tricalcium aluminate
Predictor 2	silicate: % of tricalcium silicate
Predictor 3	ferrite: % of tetracalcium alumino ferrite
Predictor 4	dicalcium: % of dicalcium silicate

Table 3 shows the log likelihood values for a series of linear regression model fits to these data. Use the AIC and the forward selection method find a final model, outlining each step.

(18)

Table 3. Log Likelihoods for Models fitted to the cement data.

model	logLik
heat $\sim$ aluminate + silicate + ferrite + dicalcium	-26.92
heat $\sim$ aluminate + silicate + ferrite	-26.95
heat $\sim$ aluminate + silicate + dicalcium	-26.93
heat $\sim$ aluminate + ferrite + dicalcium	-27.31
heat $\sim$ silicate + ferrite + dicalcium	-29.73
heat $\sim$ aluminate + silicate	-28.16
heat $\sim$ aluminate + ferrite	-48.00
heat $\sim$ aluminate + dicalcium	-29.82
heat $\sim$ silicate + ferrite	-40.96
heat $\sim$ silicate + dicalcium	-45.76
heat $\sim$ ferrite + dicalcium	-35.37
heat $\sim$ aluminate	-48.21
heat $\sim$ silicate	-46.04
heat $\sim$ ferrite	-50.98
heat $\sim$ dicalcium	-45.87
heat $\sim$ 1	-53.17

c) Give a brief description of the LASSO (least absolute shrinkage and selection operator) method for model fitting in multiple regression models.

(7)

[33]

4. A motor insurance company is conducting an analysis of historic customer retention data. The response variable is a binary indicator of whether that customer renewed their car insurance at the end of their policy period or did not [1=renewed, 0=did not renew]. Potential predictors were: the age group of the customer in years [20–29, 30–39, 40–49, 50–59, over 60]; the type of cover ['comprehensive' or '3rd party fire & theft'] and the engine size of the insured vehicle [coded A, B and C in order of increasing engine size]. These data were analysed in R using the `glm(.)` function to model the probability of renewal. Partial output from R is given below.

```
>summary(fit_cars)

Call:
glm(formula = renewed ~ factor(age) + factor(cover) + factor(engine_size),
    family = binomial(), data = car_insurance)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  1.04557    0.21558   4.850 1.23e-06
factor(age)30-39             0.09835    0.20762   0.474  0.63571
factor(age)40-49             0.08053    0.22047   0.365  0.71491
factor(age)50-60             0.68795    0.24296   2.831  0.00463
factor(age)over 60           0.29764    0.22217   1.340  0.18034
factor(cover)Comprehensive   0.31110    0.13029   2.388  0.01695
factor(engine_size)B        -0.15981    0.15282  -1.046  0.29566
factor(engine_size)C        -0.23794    0.18263  -1.303  0.19262
```

a) Give the general formulation of the likelihood for a logistic regression model fitted to binary data, explaining the terms used and making explicit reference to this example. (12)

b) Find the estimated odds ratio that a customer in the 40-49 age group will renew over a customer in the 20-29 age group - all other variables being equal. (4)

c) Discuss the evidence for the predictor 'cover' being related to the response. (5)

d) Predict the probability that a customer with the following values of the predictors would renew: age= 40-49, cover = comprehensive, engine size= A. (5)

e) State what type of customer the intercept is modelling in this case. Calculate the odds that such a customer will renew and determine a 95% confidence interval for this odds. (7)

[33]