13/5/2019

18.30 - 20.30pm

MATH 9903 Linear & General Regression Models

Basement 1, Kevin Street

S9002/1
MATH9903

TECHNOLOGICAL UNIVERSITY DUBLIN

KEVIN STREET CAMPUS

DT9002 PG Cert Applied Statistics

Year 1

SUMMER EXAMINATIONS 2018/19

MATH 9903 Linear and Generalised Regression Models

Dr J Condon
Dr C Hills
Dr K Hayes

Attempt three questions.
All questions carry equal marks.

Approved calculators may be used
New Cambridge Statistical Tables are provided

1. A pharmaceutical company is conducting research on a newly developed drug compound. They administer different doses (recorded in milligrammes, mg) of the drug to 21 rats and record the concentration of the drug in the bloodstream of each rat after an eight hour period (recorded in micro-grammes per litre of blood, μg/l). These data are read into an R data frame called `drug_data`. The following regression mode is fitted to these data:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

where y_i is the concentration after 8 hours, and x_i is the dose administered.

Partial output from R is shown below.

```
> head(drug_data)
   dose concentration
1   1.0           7.20
2   1.2           7.53
3   1.5           7.70
4   1.8           8.05
5   2.2           7.82
6   2.4           7.55

> summary(fit)...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.8935     0.5875  10.032  8.5e-09
dose           1.9399     0.7036   2.757   0.0130
I(dose^2)     -0.5008     0.1940  -2.581   0.0188

Residual standard error: 0.2096 on 18 degrees of freedom
Multiple R-squared:  0.3455,  Adjusted R-squared:  0.2728
F-statistic: 4.752 on 2 and 18 DF,  p-value: 0.02203
```

a) Give the precise R code for fitting the regression model shown in the output. Give a brief geometric description of this model. (6)

b) Interpret the '`Multiple R-squared`' value shown in the output. (3)

c) Calculate the predicted concentration for a rat administered a dose of 1.6 mg. (4)

question continued overleaf...

d) Discuss the evidence in the output shown for the inclusion of x_i^2 as a predictor in the model. (6)

e) R is used to calculate the following 95% **prediction** interval for concentration for a rat administered a 2mg dose:

```
      fit       lwr      upr
1 7.770207 7.304045 8.23637
```

i) Give the precise R code that could be used to calculate this interval. (5)

ii) Interpret this interval in the context of the data presented. (5)

iii) Discuss the difference between a prediction and confidence interval using these data as an example. (4)

[33]

2. An agricultural experiment was conducted to estimate the crop yield of two different grain (types 'a' and 'b') grown under two irrigation levels ('low' or 'normal'). Each of the treatment combinations, {a, low}, {a, normal}, {b, low} and {b, normal}, were applied to 10 experimental plots each and the crop yield was recorded for each plot at the end of the experiment.

a) A model is fitted to these data with the following output:

```
> fit_grain_1=lm(yield~factor(grain_type) +factor(irrigation)
   ,data=grain_data)
> summary(fit_grain_1)
Call:
lm(formula = yield ~ factor(grain_type) + factor(irrigation),
    data = grain_data)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                108.950      1.389  78.450  < 2e-16
factor(grain_type)b         -1.100      1.604  -0.686  0.49702
factor(irrigation)normal     5.000      1.604   3.118  0.00352

Residual standard error: 5.071 on 37 degrees of freedom
Multiple R-squared:  0.216, Adjusted R-squared:  0.1736
F-statistic: 5.096 on 2 and 37 DF,  p-value: 0.0111
```

i) Calculate the estimated mean crop yields under the four treatments from this output clearly indicating which mean refers to which treatment. (6)

ii) Using the output shown, give your conclusions concerning the need for the two predictors in the model from this analysis. (5)

b) A second analysis is carried on the same data and this analysis gives the new output which is shown overleaf.

i) Discuss the differences between **fit_grain_1** and **fit_grain_2** in terms of the assumed relationship between the predictors under both models. (6)

ii) Re-calculate the estimated mean crop yields under the four treatments from this new output and sketch a suitable plot to illustrate those four means. (8)

iii) Imagine that a pairwise comparison of means is to be conducted from this analysis. Determine how many different pairs will be compared. Calculate the experimentwise error rate for these pairwise comparisons and describe how the Bonferroni adjustment of the resulting p-values would be applied. (8)

question continued overleaf...

```
> fit_grain_2=update(fit_grain_1,.~.  +factor(grain_type):
   factor(irrigation))
> summary(fit_grain_2)

Call:
lm(formula = yield ~ factor(grain_type) + factor(irrigation)
  + factor(grain_type):factor(irrigation), data = grain_data)

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      111.800      1.319  84.742  < 2e-16
factor(grain_type)b               -6.800      1.866  -3.645 0.000838
factor(irrigation)normal          -0.700      1.866  -0.375 0.709728
factor(grain_type)b:
factor(irrigation)normal          11.400      2.639   4.320 0.000117

Residual standard error: 4.172 on 36 degrees of freedom
Multiple R-squared:  0.4837,  Adjusted R-squared:  0.4407
F-statistic: 11.24 on 3 and 36 DF,  p-value: 2.374e-05
```

[33]

3. a) Consider a regression analysis where there are 16 possible predictors for inclusion in the model. Calculate the number of possible regression models that could be fitted in this case. Give a general formula for calculating the same number for the general case with k predictors. (4)

b) A data set comprises fuel consumption (mpg) of 32 cars and 4 aspects of their automobile design as follows:

mpg:	Miles/(US) gallon (the response variable)
cyl:	Number of cylinders
disp:	Engine Displacement (cu.in.)
hp:	Gross horsepower
wt:	Weight (1000 lbs)

Regression models were being considered with mpg as the response variable and the 4 predictors variables of cyl, disp, hp and wt.

The table below shows the log likelihood values for a series of linear regression models fitted to these data. Use the AIC and the forward selection method identify a suggested final model, outlining each step. (15)

model	logLik
mpg ~ cyl+disp+hp+wt	-72.17
mpg ~ cyl+hp+wt	-72.74
mpg ~ cyl+disp+wt	-73.78
mpg ~ disp+hp+wt	-74.32
mpg ~ cyl+disp+hp	-79.01
mpg ~ cyl+wt	-74.01
mpg ~ hp+wt	-74.33
mpg ~ disp+wt	-78.08
mpg ~ cyl+disp	-79.57
mpg ~ disp+hp	-80.31
mpg ~ cyl+hp	-80.78
mpg ~ wt	-80.01
mpg ~ cyl	-81.65
mpg ~ disp	-82.10
mpg ~ hp	-87.62
mpg ~ 1	-102.38

question continued overleaf...

c) The Lasso algorithm fits regression models using the following penalised least squares criterion:

$$Q_{\text{Lasso}} = \sum_{i=1}^{n} \{y_i - \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\}^2 + \lambda \sum_{j=1}^{p-1} |\beta_j|$$

i) Outline how a value of λ could be found using k-fold cross validation. (9)

ii) Briefly describe how the Lasso model may be considered as a model selection method. (5)

[33]

4. A cinema conducts a customer survey to determine what factors might be related to people joining their members club. They survey 378 customers and record the following information:

- Would they strongly consider joining their members club? (yes or no)
- gender,
- age in years,
- day of the week they visited the cinema grouped as Mon-Tue, Wed-Thurs, Weekend (i.e. Fri-Sun).

Logistic regression is used to model the probability that the customer would strongly consider becoming a member. Part of the output from fitting this model using R is shown below.

```
> fit=glm(join~factor(gender)+age+factor(visit),family=binomial
   (),data=cinema)
> summary(fit)

Call:
glm(formula = join ~ factor(gender) + age + factor(visit),
   family = binomial(),
    data = cinema)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6992  -1.0286  -0.8147   1.1731   1.6626

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.02157    0.59540  -1.716 0.086206
factor(gender)Male      -0.20246    0.21529  -0.940 0.347030
age                      0.05356    0.01512   3.543 0.000396
factor(visit)Wed-Thur   -0.94024    0.35675  -2.636 0.008400
factor(visit)Weekend    -1.06170    0.31776  -3.341 0.000834
```

a) Give the general formulation of the logistic regression model for binary data, explaining the terms used and making explicit reference to this particular example. (5)

b) Find the estimated odds ratio that a female would strongly consider becoming a member over a male - all other variables being equal. (5)

question continued overleaf...

c) Discuss the evidence for the predictor 'age' being related to the response. (7)

d) Predict the probability that a customer with the following values of the predictors would strongly consider becoming a member: sex= male, age = 35, visit= Mon-Tue. (8)

e) Describe how a 95% confidence interval for the fitted probability predicted in part (d) could be calculated using R code. (8)

[33]