MATH6157 2016-2017

Applied Statistical Modelling

Assessed Assignment 1

*This assignment should be handed in by **25th April 2017 at 3pm**. It counts for 50% of your final mark for this module. Please hand the assignment to the Student office in the foyer of Building 58 near room 2032, or use the "drop in box"(a cabinet that looks like a locker with a post slot in). Make sure your work is clearly marked with your name, student ID, and the course number.*

Please also submit a copy of your final report using the Turnitin assignment submission tool on the course blackboard. You are reminded of the University's academic integrity policy.

In performing the assignment, you should include

- **appropriate** output from R, or similar statistical analysis tool;
- any graphics or plots that you make reference to.

Marks will be given for

- appropriate visual representations of the data;
- correct use of statistical techniques with appropriate explanations;
- overall quality and clarity of the report.

1. **[10 marks]** The following table shows the weight (in kilograms) at various ages (in weeks) of a baby girl. It is intended that the baby's weight is a response variable, and the age a regressor variable.

Weight	Age	Weight	Age
2.99	1	5.05	12
3.04	2	5.35	13
3.23	3	5.62	14
3.60	5	5.95	16
3.98	6	6.09	17
4.50	9	6.28	20
4.60	10	6.69	22
4.95	11		

   (a) Plot these data in a suitable form and fit a simple linear regression model. **[2 marks]**

   (b) Produce residual plots for the model, and hence or otherwise check the assumptions of the model. **[2 marks]**

   (c) Fit a quadratic model to these data. Carry out the usual model checks. Does this seem a reasonable model? **[2 marks]**

   (d) Add a cubic term as well. Carry out the usual model checks. Do you think this is a reasonable model? **[1 marks]**

   (e) Is it possible to improve your model still further? Justify your answer. **[1 marks]**

   (f) The baby's parents want to predict the weight of the girl at 26 weeks. Predict the weight for each of the three models. Find the 95% prediction interval for the prediction from each model, and comment briefly on how reliable this prediction interval is likely to be. **[2 marks]**

2. **[15 marks]** The file `steam.txt` (on the course blackboard) contains data on the following variables, which were collected over 25 successive months at a chemical manufacturing plant.

Variable	Description of variable
Y	Pounds of steam used monthly
CG	Pounds of Crude Glycerine made
CD	no. of calendar days per month
OD	no. of operating days per month
X	Average atmospheric temperature (degrees F)

Use the variable Y (or transformations of it) as the dependent variable.

(a) Compare the following three methods for deciding on which regressor variables should be included in the model

i. Ridge regression
ii. Lasso regression
iii. Backwards stepwise regression
iv. Forwards stepwise regression

Which method produces the best model? Justify your choice. **[9 marks]**

(b) Can this model be further improved? Justify your answer. **[2 marks]**

(c) Using your preferred model, find a 95% confidence interval for a month with 20 operating days out of 30 calendar days, in which 0.7 pounds of Crude Glycerine are made. **[1 marks]**

(d) Write a short report for the plant manager indicating, in non-technical terms, which factors are important in determining the pounds of steam used monthly. **[3 marks]**

3. **[15 marks]** The Hadley Centre Central England Temperature (HadCET) dataset is the longest instrumental record of temperature in the world. The mean monthly data series began in 1659, and the daily in 1772.

A summary of daily and monthly data collected from this observatory are available in the course blackboard as `cetdl1772on.dat` and `cetml1659on.dat`. (These are a snapshot of the data available originally at `http://www.metoffice.gov.uk/hadobs/hadcet/data/download.html`, and the notes on the formats of the file there should be read.)

(a) Assume that the daily data on maximum temperature from 1901 to 2000 are independently distributed according to a Gumbel distribution. Find a value for a limit which the temperature will only exceed once every ten years, and plot the data in an appropriate form to show this. **[4 marks]**

(b) Using time series modelling, or otherwise, find an appropriate model for the mean Central England Temperature for January 2001 to December 2016. **[8 marks]**

(c) Use the model to find prediction and confidence intervals for the month of July 2017. Use these intervals to write a long range weather forecast of July's temperature as if written for the general public. **[3 marks]**

4. **[10 marks]** 100 students at a university are given two entrance exams, and interest lies in whether either or both of these exams can predict whether a student gets a first class degree.

   The data for the 100 students is available in the file `exams.csv` on the course blackboard.

   The responses (y) were either ″1″ or ″0″ corresponding to the student getting an upper second degree or higher. The results of the two exams are listed as (x_1) and (x_2).

   Fit the following logistic model to this data:

$$\mathrm{E}(\pi^*) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2,$$

   where $\pi = P(Y = 1)$ and $\pi* = \log[\pi/(1-\pi)]$ and the logistic model has the usual assumptions.

   (a) Test the significance of the model overall and of each parameter in the model above. Test all these using $\alpha = 0.01$. **[2 marks]**

   (b) Now fit the following new model to this data:

$$\mathrm{E}(\pi^*) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

   with the same assumptions Using both the original and the new model, predict the probability of a getting an upper second degree (or higher) with marks of 62 in both exams. Comment on the meaning of your answer. **[3 marks]**

   (c) Compare the PRESS statistics for your original and new model. Interpret these statistics. **[1 marks]**

   (d) Use the model deviances to make an appropriate statistical test of whether the original or new model fits significantly better. **[2 marks]**

   (e) For your final model, examine the residual plots to see whether the assumptions of the model were satisfied. Comment on any problems. **[2 marks]**