MA2261 Linear Statistical Models - DLI, Year 2022-2023

Coursework 3

INSTRUCTIONS AND DEADLINE:

Please submit *electronically* one piece of written/typed work per person in a single PDF file by **Friday 26 May 2023 at 4pm UK time/23:00 China time**.

Please use this page as the cover page for your submission. Write below your student ID and sign it.

Student ID:

Signature:

MARKING CRITERIA:

- This problem sheet is worth 100 points. Scores for each main question are indicated at the beginning of each.
- Clearly justify and explain your answers. If you are using the R software for calculations, a printout of your answers without a full explanation of the formulas you are using and your reasoning will not score full marks.
- A true/false question answered without justification will get zero marks.
- Computational mistakes will be penalized more in coursework than in exam marking, since you have plenty of time and tools to check your calculations when doing the coursework.

Question 1 [30 marks]

In a maternity ward, both the weight of newborn babies and the length of the mother's pregnancy are recorded. The data below records information from 20 births, including the baby's weight (y) in kilograms and the length (x) of pregnancy in weeks.

Case Nr.	Weeks x	Weight y	Case Nr.	Weeks x	Weight y
1	40	2.97	11	40	2.94
2	40	3.16	12	38	2.75
3	36	2.63	13	42	3.21
4	37	2.85	14	39	2.82
5	41	3.29	15	40	3.13
6	37	2.63	16	37	2.54
7	38	3.18	17	36	2.41
8	40	3.42	18	38	2.99
9	40	3.32	19	39	2.88
10	36	2.73	20	40	3.23

Table 1: Weight of baby and length of pregnancy

Assume the significance level is $\alpha = 0.05$. Answer the following questions:

a) **[5 marks]** Calculate $\bar{x}$, $\bar{y}$, S_{xx}, S_{yy}, S_{xy}.

b) **[4 marks]** Calculate the point estimates for the model parameters a (the intercept), b (the slope), σ^2 (the error variance) for the simple linear regression model fitted to the data.

c) **[10 marks]** Using the information that the variation between groups of repeated observations is $SSB = 1.153$, calculate the ANOVA table for the data in Table 1.

d) **[6 marks]** Illustrate the meaning of each of the terms you calculated in part c).

e) **[5 marks]** Test the hypothesis that the simple linear regression model is true.

Question 2 [70 marks]

An experiment is conducted to determine the optimal factors for a flux-cored arc welding process on metals. Three factors - current, voltage, and the dimensions of the electrodes - affect the tensile strength of the welded joints. The goal is to identify the settings that maximize resistance to fatigue cycles under a predetermined load. The following table illustrates the settings used in the welding process. The variables are:

1) A = electrical current intensity (Ampere)

2) V = electrical voltage (Volt)

3) D = electrode diameter (mm)

4) R = fatigue resistance of the joint (cycles).

The chosen settings are

Current (Amp)	250	275	300
Voltage (Volt)	25	27.5	30
Electrode diameter (mm)	2.5	5	7.5

The experiment involves carrying out the welding process with all possible combinations of the three variables: A, V, and D. Each of these variables can assume one of three distinct values, resulting in 'combinations with repetition'. Therefore, there will be $3^3 = 27$ distinct cases.

We assign a value for each combination using the following formulas:

$$XA = (A - 275)/25$$

$$XV = (V - 27.5)/2.5$$

$$XD = (D - 5)/2.5$$

The results of the 27 experiments are shown in the following Table 2.

XA	XV	XD	R (cyles)
-1	-1	-1	1348
0	-1	-1	2828
1	-1	-1	7272
-1	0	-1	676
0	0	-1	2044
1	0	-1	3136
-1	1	-1	340
0	1	-1	884
1	1	-1	2280
-1	-1	0	740
0	-1	0	2396
1	-1	0	6368
-1	0	0	532
0	0	0	1240
1	0	0	2140
-1	1	0	236
0	1	0	664
1	1	0	1768
-1	-1	1	584
0	-1	1	1268
1	-1	1	4000
-1	0	1	420
0	0	1	876
1	0	1	1132
-1	1	1	180
0	1	1	440
1	1	1	720

Table 2. Results of the experiment

Recall that a complete second order polynomial in the variables x, y, z is a polynomial that contains all the terms in x, x^2, y, y^2, z, z^2, xy, xz, yz.

Assume the significance level is $\alpha = 0.05$. Answer the following questions:

a) **[10 marks]** Fit a polynomial regression model to the data in Table 1 consisting of a complete second order polynomial model in the three continuous variables XA, XD, XV.

Comment on the results you obtained, in particular about the statistical significance of each term. Can the model be simplified? Justify your answer.

b) **[10 marks]** Using the fitted model from part a), plot the residuals versus the fitted values. Comment on the results you obtained in terms of the validity of the model. Justify your answer.

c) [**10 marks**] Perform a transformation of the response variable R into the natural logarithm $\ln R$, obtaining a new table for the observed $\ln R$, together with the given XA, XD, XV. Fit to these data a complete second order polynomial model in the three continuous variables XA, XD, XV.

Comment on the results you obtained, in particular about the statistical significance of each term. Can the model be simplified?

d) [**10 marks**] Using the fitted model from part c), plot the residuals versus the fitted values. Comment on the results you obtained in terms of the validity of the model, and compare with your conclusions in part b). Justify your answer.

e) [**10 marks**] Based on the analysis in parts a) to d), draw your statistical conclusions regarding the selection of a valid model that is both a good fit and as simple as possible. Justify your answer.

f) [**10 marks**] For the model that you selected in part e), calculate the point estimate and the 95% confidence interval for the *percentage* increase in the mean response R, corresponding to an increase of 0.1 in XA, while XD and XV are kept constant. Justify your answer.

g) [**10 marks**] The aim of the experiment is to identify the combination of current, voltage, and electrode diameter that results in the maximum fatigue strength of the weld. Use the model estimates from the model chosen in part e) to determine this combination.