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.

Solution to Question 1

a) [**5 marks**] $\bar{x} = 38.7, \quad \bar{y} = 2.954, \quad S_{xx} = 60.2, \quad S_{yy} = 1.5353, \quad S_{xy} = 7.834.$

b) [**4 marks**]

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = 0.1301, \qquad \hat{a} = \bar{y} - \hat{b}\bar{x} = -2.082 \ .$$

$$RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 0.5158, \qquad \hat{\sigma}^2 = \frac{RSS}{n-2} = 0.0287, \qquad \hat{\sigma} = 0.1693 \ .$$

c) [**10 marks**] Given $SSB = 1.153$, we have

$$\begin{aligned}
SST &= S_{yy} = 1.5353 \\
SSE &= RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 0.5158 \\
SSM &= SST - SSE = 1.0195 \\
SSW &= SST - SSB = 0.3823 \\
SSL &= SSE - SSW = 0.1335
\end{aligned}$$

The corresponding mean squares are given by

$$\begin{aligned}
MSE &= \frac{SSE}{N-2} = \frac{SSE}{18} = 0.0287 \\
MSM &= SSM \\
MSW &= \frac{SSW}{N-k} = \frac{SSW}{13} = 0.0294 \\
MSB &= \frac{SSB}{k-1} = \frac{SSB}{6} = 0.1922 \\
MSL &= \frac{SSL}{k-2} = \frac{SSL}{5} = 0.0267
\end{aligned}$$

In summary, the ANOVA table is

	Source	df	SS	MS	F
SSM	Regression (Model)	1	1.0195	1.0195	
SSL	Lack of fit	5	0.1335	0.0267	0.9083
SSW	Pure Error (Residual)	13	0.3823	0.0294	
SST	Total	19	1.5353		

d) [**6 marks**] SST is the total sum of squares, representing the total variation of Y about its mean.

$SSE = RSS$ is the error variation (or residual sum of squares) and represents the residual variation in Y after fitting the regression line.

SSM represents the amount of variation in Y explained by x.

SSB is based on the variation between groups of repeated observations.

SSW is the pure error sum of squares and is based on the variation within groups of repeated observations.

SSL is the lack of fit sum of squares and represents the part of the error variation which is due to lack of fit.

e) **[5 marks]**

$$F = \frac{SSL/(k-2)}{SSW/(N-k)} = \frac{SSL/5}{SSW/13} = 0.9083 \sim F_{5,13}$$

The critical region is $(3.025, +\infty)$, hence the model is a good fit.

Solution to Question 2

a) **[10 marks]** The model (R_fit) of a complete second order polynomial model in the three continuous variables XA, XV, XD is

$$\begin{aligned} R = &\beta_0 + \beta_1 XA + \beta_2 XV + \beta_3 XD + \beta_4 XA^2 + \beta_5 XV^2 + \beta_6 XD^2 \\ &+ \beta_7 XA * XV + \beta_8 XA * XD + \beta_9 XV * XD + \varepsilon \end{aligned}$$

We can use R programming to produce the fitted model,

```
> T=read.table("weld.txt")
> XA=T[,1]
> XV=T[,2]
> XD=T[,3]
> R=T[,4]
> XA2=XA^2
> XV2=XV^2
> XD2=XD^2
> R_fit=lm(R ~ XA +XV +XD +XA2 +XV2 +XD2 +XA*XV +XA*XD +XV*XD)
> summary(R_fit)
```

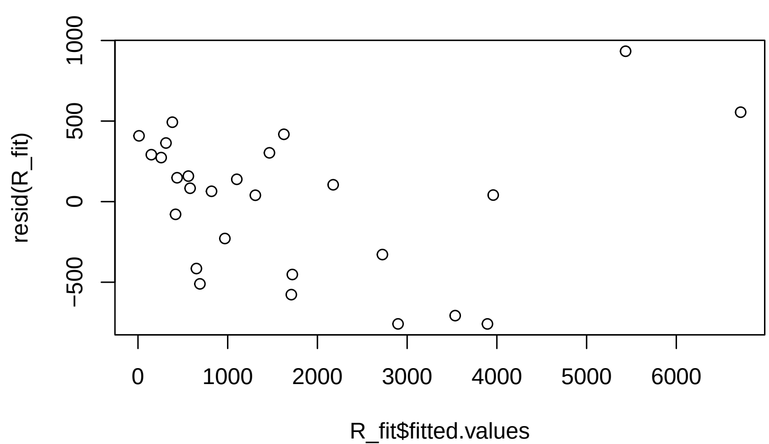
```
Call:
lm(formula = R ~ XA + XV + XD + XA2 + XV2 + XD2 + XA * XV + XA *
    XD + XV * XD)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  1101.33     276.88    3.978 0.000973 ***
XA           1320.00     128.17   10.299 1.00e-08 ***
XV          -1071.78     128.17   -8.362 1.99e-07 ***
XD           -621.56     128.17   -4.850 0.000150 ***
XA2           477.33     221.99    2.150 0.046222 *
XV2           551.33     221.99    2.484 0.023733 *
XD2           -96.67     221.99   -0.435 0.668717
XA:XV        -913.00     156.97   -5.816 2.06e-05 ***
XA:XD        -471.33     156.97   -3.003 0.008010 **
XV:XD         286.00     156.97    1.822 0.086103 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 543.8 on 17 degrees of freedom
Multiple R-squared:  0.9379,Adjusted R-squared:  0.905
F-statistic: 28.52 on 9 and 17 DF,  p-value: 1.562e-08
```

The model can be simplified by deleting the terms of $XD2$ and $XV : XD$. The corresponding p-values of both predictors are 0.668717 and 0.086103 respectively, which are greater than the significance level 0.05. Thus, it implies that the predictors $XD2$ and $XV : XD$ are not statistically significant and therefore can be deleted.

b) **[10 marks]** The plot of residuals versus the fitted values for R_fit is shown below,

```
> plot(R_fit$fitted.values,resid(R_fit))
```



The plot of residuals versus fitted values shows a random pattern, which should imply that the model is valid. However, most of the points are clustered together, casting doubts on the validity of the model (fan-shaped plot).

c) **[10 marks]** The model (lnR_fit) of a complete second order polynomial model after logarithmic transformation of the dependent variable R is

$$\log(R) = \beta_0 + \beta_1 XA + \beta_2 XV + \beta_3 XD + \beta_4 XA^2 + \beta_5 XV^2 + \beta_6 XD^2 \\ + \beta_7 XA * XV + \beta_8 XA * XD + \beta_9 XV * XD + \varepsilon$$

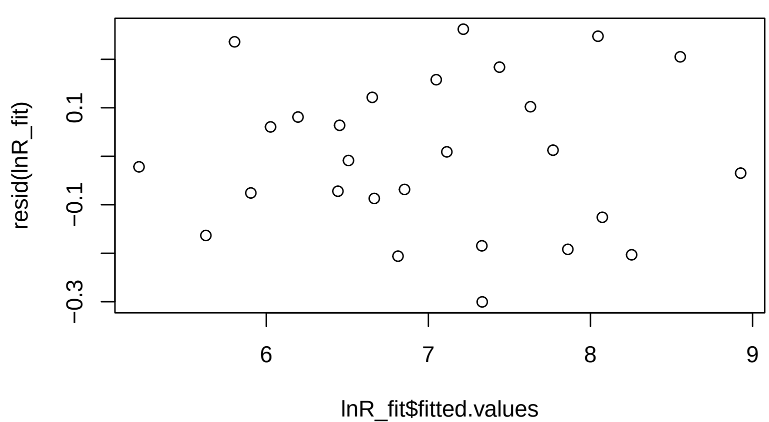
We can use R programming to produce the fitted model,

```
> lnR_fit=lm(log(R) ~ XA +XV +XD +XA2 +XV2 +XD2 +XA*XV +XA*XD +XV*XD)
> summary(lnR_fit)


Call:
lm(formula = log(R) ~ XA + XV + XD + XA2 + XV2 + XD2 + XA * XV +
    XA * XD + XV * XD)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.11377    0.09869  72.078  < 2e-16 ***
XA           0.83238    0.04569  18.219 1.37e-12 ***
XV          -0.63099    0.04569 -13.811 1.14e-10 ***
XD          -0.39249    0.04569  -8.591 1.36e-07 ***
XA2         -0.08570    0.07913  -1.083    0.294
XV2          0.02422    0.07913   0.306    0.763
XD2         -0.06746    0.07913  -0.852    0.406
XA:XV       -0.03824    0.05595  -0.683    0.504
XA:XD       -0.06841    0.05595  -1.223    0.238
XV:XD       -0.02083    0.05595  -0.372    0.714
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1938 on 17 degrees of freedom
Multiple R-squared:  0.9725,Adjusted R-squared:  0.9579
F-statistic: 66.73 on 9 and 17 DF,  p-value: 1.73e-11
```

None of the second order terms and intersection terms is statistically significant as their p-values are all greater than 0.05. Therefore, the model can be simplified by deleting terms of $XA2$, $XV2$, $XD2$, $XA:XV$, $XA:XD$, $XV:XD$.

d) **[10 marks]** The plot of residuals versus the fitted values for lnR_fit is shown below,

```
> plot(lnR_fit$fitted.values,resid(lnR_fit))
```



The plot of residuals versus fitted values shows a random pattern, and points are not clustered together. Hence, it indicates a valid model and the model lnR_fit is a better of fit than the R_fit in part b).

e) **[10 marks]** As is shown above, the model 'lnR_fit' is a better of fit than the model 'R_fit'. Therefore, for the response $\log(R)$, being not relevant the second order and intersection terms, we proceed to calculate the reduced first order linear model (lnR_reduced) with only predictors XA, XV, XD,

$$\log(R) = \beta_0 + \beta_1 XA + \beta_2 XV + \beta_3 XD + \varepsilon$$

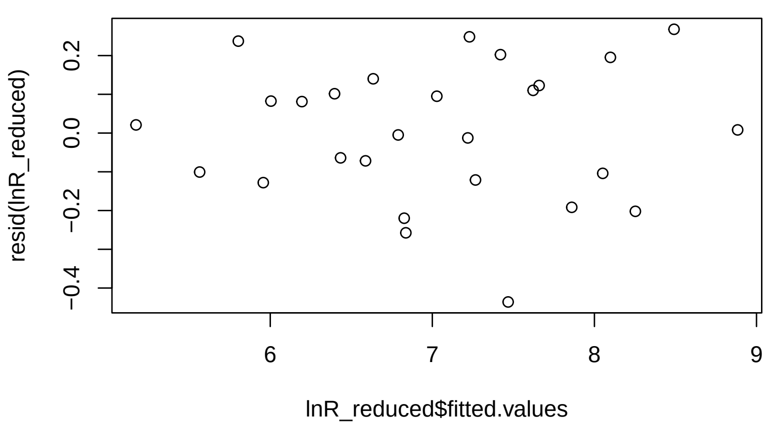
We can use R programming to produce the fitted model,

```
> lnR_reduced=lm(log(R) ~ XA +XV +XD)
> summary(lnR_reduced)
```

```
Call:
lm(formula = log(R) ~ XA + XV + XD)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.02781    0.03572 196.726  < 2e-16 ***
XA           0.83238    0.04375  19.025 1.43e-15 ***
XV          -0.63099    0.04375 -14.422 5.20e-13 ***
XD          -0.39249    0.04375  -8.971 5.69e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1856 on 23 degrees of freedom
Multiple R-squared:  0.9658,Adjusted R-squared:  0.9614
F-statistic: 216.8 on 3 and 23 DF,  p-value: < 2.2e-16
```

The plot of residuals versus the fitted values for lnR_reduced is shown below,

```
> plot(lnR_reduced$fitted.values,resid(lnR_reduced))
```



Since the p-values for all three predictors are less than 0.05, we can conclude that they are statistically significant in affecting the response $\log(R)$. In addition, the pattern of residuals versus fitted values shows still a random behaviour and not clustered together, which indicates the reduced model is valid.

We can then compare the full model 'lnR_fit' with the reduced model 'lnR_reduced'. We want to test the null hypothesis

$$H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$

against

$$H_1 : \text{ at least one of } \beta_4,\ \beta_5,\ \beta_6,\ \beta_7,\ \beta_8,\ \beta_9 \text{ is not } 0$$

```
> anova(lnR_reduced,lnR_fit)


Analysis of Variance Table

Model 1: log(R) ~ XA + XV + XD
Model 2: log(R) ~ XA + XV + XD + XA2 + XV2 + XD2 + XA * XV + XA * XD +
    XV * XD
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     23 0.79252
2     17 0.63871  6   0.15381 0.6823 0.6663
```

From the comparison between the reduced model and the full model, we see that $SS_{extra} = 0.15381$, $F_{6,17} = 0.6823$, p-value$= 0.6663 > 0.05$. Therefore, we accept H_0. The full model does not improve the quality of fit, so we choose the reduced model.

f) **[10 marks]** For the reduced model 'lnR_reduced', we have

$$\begin{aligned}\log\frac{R(XA+0.1,XV,XD)}{R(XA,XV,XD)} &= \log R(XA+0.1,XV,XD) - \log R(XA,XV,XD)\\ &= \beta_0 + \beta_1(XA+0.1) + \beta_2 XV + \beta_3 XD - \beta_0 - \beta_1 XA - \beta_2 XV - \beta_3 XD\\ &= 0.1\beta_1\end{aligned}$$

Thus,

$$\frac{R(XA+0.1,XV,XD)}{R(XA,XV,XD)} = e^{0.1\beta_1}$$

and the percentage increase in mean response R is

$$\frac{R(XA+0.1,XV,XD) - R(XA,XV,XD)}{R(XA,XV,XD)} = e^{0.1\beta_1} - 1$$

Therefore the estimated percentage increase is

$$e^{0.1\hat{\beta}_1} - 1 = e^{0.1\times 0.83238} - 1 = 0.0868 = 8.68\%$$

Since the 95% C.I. for β_1 is

$$\left(\hat{\beta}_1 \pm t_{0.025,23}\, se(\hat{\beta}_1)\right) = (0.83238 \pm 2.069 \times 0.04375)) = (0.7419,\ 0.9229)$$

```
> confint(lnR_reduced, "XA", level = 0.95)
```

```
       2.5 %    97.5 %
XA 0.7418749 0.9228934
```

Hence, the corresponding 95% C.I. for $e^{0.1\beta_1} - 1$ is

$$\left(e^{0.1\times0.7419} - 1,\ e^{0.1\times0.9229} - 1\right) = (0.077, 0.0967)$$

g) [**10 marks**] The following is the last fragment of R code that calculates the fitted values of the response R by the reduced model 'lnR_reduced',

```
> fitted_value=exp(lnR_reduced$fitted.values)
```

```
        1         2         3         4         5         6
1365.0043 3137.8622 7213.2954  726.2696 1669.5434 3837.9345
        7         8         9        10        11        12
 386.4219  888.3039 2042.0267  921.8835 2119.2194 4871.6464
       13        14        15        16        17        18
 490.5010 1127.5603 2592.0275  260.9779  599.9342 1379.1244
       19        20        21        22        23        24
 622.6129 1431.2581 3290.1659  331.2699  761.5209 1750.5787
       25        26        27
 176.2568  405.1778  931.4199
```

The maximum fitted value is 7213.2954, which corresponds to the triplet ($XA = 1$, $XV = -1$, $XD = -1$). Taking into account of the transformation formula of the three variables A, V, D, we can easily calculate the required values:

- Electrical current intensity A = 300 Ampere
- Electrical voltage V = 25 Volt
- Electrode diameter D = 2.5 mm