MAS programmes - Statistical Analysis (Autumn Term)
AUTUMN 2019

Assignment

Deadline: Monday, 20th January, 2020

Total marks: [40]. Marks are shown in boxes [ ]. There are 2 questions in this assignment.

1. Consider a two-stage nested design in which the data are classified according to factor L, which occurs at a levels, and factor M, nested within L, which occurs at b levels. There are r replications taken within each level of M (nested within each level of L). Factor L is considered to be fixed, whereas factor M is considered to be random.

   Consider the model equation

   $$z_{lmn} = \mu + \delta_l + \gamma_{(l)m} + \epsilon_{(lm)n}$$

   where z_{lmn} is the value of the observation for the n-th replication taken within the m-th level of M, nested within the l-th level of L; the $\{\delta_l\}$ are constants such that $\sum_{l=1}^{a} \delta_l = 0$, $\gamma_{(l)m} \sim \text{NID}(0, \sigma_M^2)$, $\epsilon_{(lm)n} \sim \text{NID}(0, \sigma^2)$, with the collections $\{\gamma_{(l)m}\}$ and $\{\epsilon_{(lm)n}\}$ being independent of each other.

   (a) State the conditions that characterize the collection of data within this type of design. What do the terms on the right hand side of the model equation represent? Briefly describe a scenario (not used in lectures, or in this question) in which the use of such an experimental design might be appropriate. [5]

   Define $\overline{z}_{...}$, $\overline{z}_{l..}$, and $\overline{z}_{lm.}$, to be the sample means corresponding to the entire set of observations, the observations occurring under the l-th level of factor L, and the observations occurring under the m-th level of factor M (nested under the l-th level of factor L), respectively, and let $z_{...}$ be the total for the entire set of observations. Further define

   $$F = \sum_{l=1}^{a}\sum_{m=1}^{b}\sum_{n=1}^{r}(z_{lmn} - \overline{z}_{lm.})^2, \qquad G = r\sum_{l=1}^{a}\sum_{m=1}^{b}(\overline{z}_{lm.} - \overline{z}_{l..})^2$$

   $$H = br\sum_{l=1}^{a}(\overline{z}_{l..} - \overline{z}_{...})^2, \qquad K = \sum_{l=1}^{a}\sum_{m=1}^{b}\sum_{n=1}^{r} z_{lmn}^2 - \frac{z_{...}^2}{abr}$$

. . . continued

(b) State an identity that is satisfied by the terms F, G, H, and K. [2]

(c) Give expressions for the expectations of the following expressions

i) $H/(a-1)$
ii) $G/\{a(b-1)\}$
iii) $F/\{ab(r-1)\}$.

[3]

(d) Consider the following pair of hypotheses:

$$H_{0L} : \delta_1 = \delta_2 = \ldots = \delta_a = 0$$

$$H_{0M} : \sigma_M^2 = 0.$$

How do the expectations considered in part (c) change under (i) H_{0L}, (ii) H_{0M} (iii) both H_{0L} and H_{0M}? [3]

(e) A pharmaceutical company has 6 different laboratories with each one producing the same medicine. The company wants to check on the sources of variation for the amount of Tetrasodium Pyrophosphate in the medicine. A random selection of 3 batches (of standard size) of the medicine produced from each laboratory were analyzed in order to ascertain the concentration of the chemical (in appropriate units). The experiment to ascertain the concentration was performed twice on each batch, thus yielding two independent readings. An abbreviated excerpt of the R output for these data is presented below.

```
> study.aov<-aov(conc~Lab/Bat, data=study)
> summary(study.aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
Lab          5  53.83  10.767   5.602 0.00276 **
Lab:Bat     12  57.60   4.800   2.497 0.03870 *
Residuals   18  34.59   1.922
```

With the help of statistical tables, comment on the sources of variation in the concentration. What suggestions would you give to the company to help them minimize the variation in the concentrations? [7]

2. Three varieties of potatoes (`A`,`B`, and `C`) were planted on 3 plots at each of four locations (`1`,`2`,`3`, and `4`) in accordance with the *completely randomized 2-factor factorial design*.

   At the end of a particular season, the yields obtained from each plot were noted and recorded in *bushels*. The data are presented below.

		Location			
		1	2	3	4
variety	A	15 19 12	17 10 13	9 12 6	14 8 11
	B	20 24 18	24 18 22	12 15 10	21 16 14
	C	22 17 14	26 19 21	10 5 8	19 15 12

   (a) Specify the conditions that characterize the collection of data in a 2-factor factorial design. Explain how you would carry out the planting of the potatoes in order to fulfill the requirements of the experimental design. [3]

   (b) Consider the `R` output below which has been used to construct the data frame for this analysis. Crop yield, location of the crop, and variety of potato, have been stored in `yield`, `loc`, and `variety`, respectively.

```
> yield <- c(15, 19, 12, 20, 24, 18, 22, 17, 14, 17, 10, 13, 24, 18, 22,
     +26, 19, 21, 9, 12, 6, 12, 15, 10, 10, 5, 8, 14, 8, 11, 21, 16, 14,
     +19, 15, 12)
> loc <- factor(*)
> variety <- factor(**)
> potato <- data.frame(yield, loc, variety)
```

   Write down viable and consistent code that could be placed in the positions labelled `*` and `**` in the above output. [3]

   (c) Using appropriate notation, describe the form of the linear statistical model, along with the corresponding least squares estimates of the parameters under the *sum-to-zero* constraints and structure of the *analysis of variance table*, that would be most relevant for analyzing the above data, if:

   i) an interaction term is included; [3]

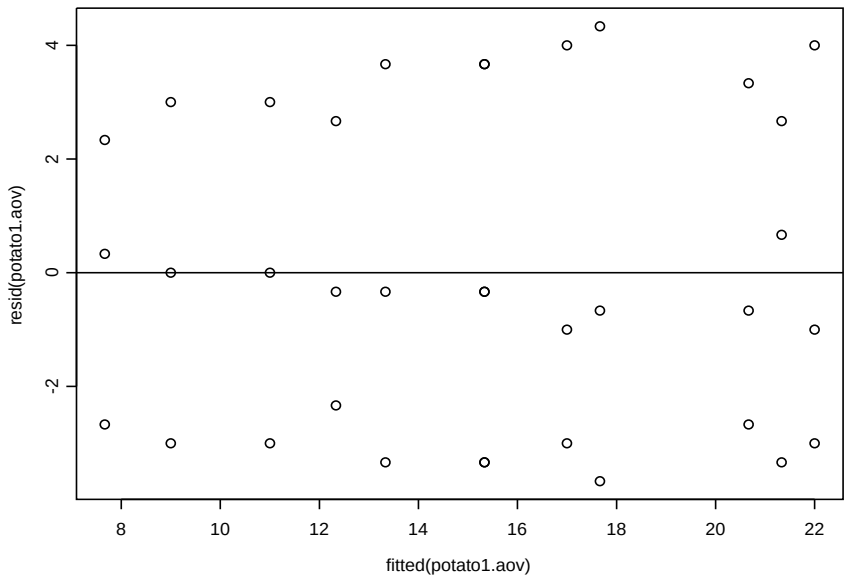
   ii) no interaction term is included; [3]

   [A correct statement of the results based on reasonable intuition will suffice: there is no need, necessarily, to derive the results from first principles].

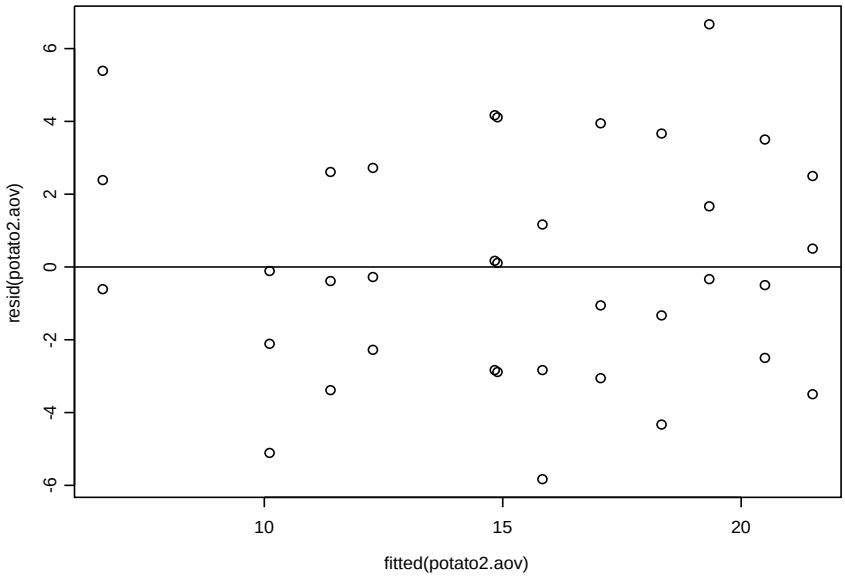
. . . continued

(d) Some analysis in R was carried out and is presented below and overleaf.

```
> potato1.aov <- aov(yield ~ loc + variety + loc:variety, data = potato)
> summary(potato1.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
loc          3  468.2  156.07  14.556 1.31e-05 ***
variety      2  196.2   98.11   9.150  0.00111 **
loc:variety  6   78.4   13.07   1.219  0.33090
Residuals   24  257.3   10.72
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> plot(fitted(potato1.aov), resid(potato1.aov))
> abline(h=0)
>
>
> potato2.aov <- aov(yield ~ loc + variety, data = potato)
> summary(potato2.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
loc          3  468.2  156.07  13.944 7.16e-06 ***
variety      2  196.2   98.11   8.766    0.001 **
Residuals   30  335.8   11.19
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> plot(fitted(potato2.aov), resid(potato2.aov))
> abline(h=0)
```



. . . continued



i) Which linear statistical model should we endorse as providing the most appropriate fit to the data? Justify your answer. [4]

ii) With reference to your answer to part (d)(i), in conjunction with some or all of the R output given below and overleaf, what recommendation can you give to the farmer of the land who wants to maximize crop yield? [4]

```
> model.tables(potato1.aov)
Tables of effects

 loc
loc
     1      2      3      4
 2.667  3.667 -5.556 -0.778

 variety
variety
      A       B       C
-3.0556  2.6111  0.4444

 loc:variety
   variety
loc A       B       C
  1  0.5000  0.1667 -0.6667
  2 -2.5000 -0.1667  2.6667
  3  2.3889  0.0556 -2.4444
  4 -0.3889 -0.0556  0.4444
```

. . . continued

```
> model.tables(potato1.aov, type="means")
Tables of means
Grand mean

15.22222

 loc
loc
     1      2      3      4
17.889 18.889  9.667 14.444

 variety
variety
     A      B      C
12.167 17.833 15.667

 loc:variety
   variety
loc A      B      C
  1 15.333 20.667 17.667
  2 13.333 21.333 22.000
  3  9.000 12.333  7.667
  4 11.000 17.000 15.333
```

Note: `model.tables` **as applied to** `potato1.aov` **provides the basis upon which, with appropriate discernment, the required information can be extracted to address this part of the question, irrespective of the model that was endorsed in part (d)(i).**

Important Note:
• Please read the current version of the *Mathematics & Statistics Coursework Policy*. Copies can be obtained from the departmental website, or failing that, in hardcopy from the programme administrator.