STAT 461 Lab 4 - Estimating ANOVA Model Parameters in R

1. ANOVA Model and Parameter Estimates

Recall from lectures that the one-way ANOVA model with effects coding is written

$$Y_{it} = \mu + \tau_i + \epsilon_{it}, \quad i = 1, 2, \ldots, v \quad t = 1, 2, \ldots, r_i$$

$$\epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2)$$

Any estimable function of the mean parameters can be written as

$$\sum_{i=1}^{v} b_i(\mu + \tau_i)$$

and the least-squares (LS) estimate of such an estimable function is obtained by replacing each model treatment mean $(\mu + \tau_i)$ with it's corresponding treatment sample mean $\bar{Y}_{i\cdot}$.

$$\overline{Y}_{i\cdot} = \frac{1}{r_i} \sum_{t=1}^{r_i} Y_{it\cdot}$$

So the LS estimate of an estimable function $\sum_{i=1}^{v} b_i(\mu + \tau_i)$ is

$$\sum_{i=1}^{v} b_i(\widehat{\mu} + \widehat{\tau_i}) = \sum_{i=1}^{v} b_i \overline{Y}_{i\cdot}$$

2. Greenhouse Example

Consider a greenhouse / fertilizer experiment where the primary goal is to determine which fertilizers help plants to grow the fastest. 24 seeds were placed in 24 pots, and each pot was randomly assigned a treatment such that 6 pots received no fertilizer, 6 pots received fertilizer 1, 6 pots received fertilizer 2, and 6 pots received fertilizer 3. After waiting a month, the plants that had grown from the 24 seeds were measured, and the height of each plant was recorded. This experiment is taken from the STAT 502 online notes https://onlinecourses.science.psu.edu/stat502/node/148. The resulting data are

```
fert=c(rep("Control",6),rep("F1",6),rep("F2",6),rep("F3",6))
height=c(21,19.5,22.5,21.5,20.5,21,
         32,30.5,25,27.5,28,28.6,
         22.5,26,28,27,26.5,25.2,
         28,27.5,31,29.5,30,29.2)
greenhouse=data.frame(fert,height)
greenhouse
```

```
##       fert height
## 1  Control   21.0
## 2  Control   19.5
## 3  Control   22.5
## 4  Control   21.5
```

```
## 5  Control   20.5
## 6  Control   21.0
## 7        F1   32.0
## 8        F1   30.5
## 9        F1   25.0
## 10       F1   27.5
## 11       F1   28.0
## 12       F1   28.6
## 13       F2   22.5
## 14       F2   26.0
## 15       F2   28.0
## 16       F2   27.0
## 17       F2   26.5
## 18       F2   25.2
## 19       F3   28.0
## 20       F3   27.5
## 21       F3   31.0
## 22       F3   29.5
## 23       F3   30.0
## 24       F3   29.2
```

3. Estimation using Summary Statistics

We could write out the ANOVA model as

$$Y_{it} = \mu + \tau_i + \epsilon_{it}, \quad i = c, 1, 2, 3 \quad t = 1, 2, \ldots, 6$$

$$\epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2)$$

where the c subscript will indicate a plant receiving the "control" treatment of no fertilizer.

Consider estimating the effect of the `F1`, fertilizer 1, over that of the control group. To be more specific, and to phrase this in proper statistical terminology, we could say: "We want to estimate the difference in the mean height of a plant receiving `F1` and the mean height of a plant receiving no fertilizer." In terms of our statistical model, we want to estimate

$$\begin{aligned} E\left[Y_{1t} - Y_{ct}\right] &= (\mu + \tau_1) - (\mu + \tau_c) \\ &= \tau_1 - \tau_c \end{aligned}$$

so the LS estimate would be

$$\begin{aligned} \widehat{\tau}_1 - \widehat{\tau}_c &= 1 \cdot \overline{Y}_{1\cdot} - 1 \cdot \overline{Y}_{c\cdot} \\ &= \sum_i b_i \overline{Y}_{i\cdot} \\ &\quad \text{where } b_c = -1,\ b_1 = 1,\ b_2 = 0,\ b_3 = 0 \end{aligned}$$

So to obtain this estimate, we need the sample means from the 6 plant heights receiving each treatment. As we did in Homework #1, we can get these means using subsets.

```
mean.c=mean(height[fert=="Control"])
mean.c
```

```
## [1] 21
```

```
mean.1=mean(height[fert=="F1"])
mean.1
```

```
## [1] 28.6
```

```
mean.2=mean(height[fert=="F2"])
mean.2
```

```
## [1] 25.86667
```

```
mean.3=mean(height[fert=="F3"])
mean.3
```

```
## [1] 29.2
```

```
grand.mean=mean(height)
grand.mean
```

```
## [1] 26.16667
```

From this we see that

- $\bar{Y}_{c\cdot} = (\hat{\mu} + \hat{\tau}_c)$=21
- $\bar{Y}_{1\cdot} = (\hat{\mu} + \hat{\tau}_1)$=28.6
- $\bar{Y}_{2\cdot} = (\hat{\mu} + \hat{\tau}_2)$=25.8666667
- $\bar{Y}_{3\cdot} = (\hat{\mu} + \hat{\tau}_3)$=29.2
- $\bar{Y}_{\cdot\cdot}$ =26.1666667

And so we can obtain an estimate for $\tau_1 - \tau_c$ by:

$$\bar{Y}_{1\cdot} - \bar{Y}_{c\cdot} = 28.6 - 21 = 7.6$$

4. Estimation using aov in R

We can also obtain the LS estimates of estimable functions $\sum_i b_i(\mu + \tau_i)$ using the `aov` command in R. The `aov` procedure will be used for many other analyses in this class, including hypothesis testing and residual analysis to check model assumptions. In this Lab, we will only focus on introducing the syntax of `aov`, and `lsmeans` and using them for estimation.

Before running the code below, we need to install the `lsmeans` package. This needs to be done once for each computer you are using. To install a package in R, use the `install.packages` command:

```
install.packages("lsmeans")
```

```
## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
```

```
##
## The downloaded source packages are in
##  '/tmp/RtmpCA3NPc/downloaded_packages'
```

After a package is installed, you can load it into your R workspace using

```
library(lsmeans)
```

After loading `lsmeans`, you will be able to use the functions in this R package.

4.1 A minimal ANOVA analysis with `aov` and `lsmeans`

The following R code applies `aov` to the greenhouse data:

```
aov.greenhouse=aov(height~fert)
```

This takes as inputs a formula of the form $Y \sim X$, where Y is the name of the response variable and X is the name of the treatment vector. The resulting object `aov.greenhouse` can be used for many analyses. To get treatment means out of an aov object, we use the `lsmeans` command ("lsmeans" stands for "Least-Squares estimate of Means".

```
lsmeans(aov.greenhouse,"fert")
```

```
##  fert      lsmean        SE df lower.CL upper.CL
##  Control 21.00000 0.7131698 20 19.51235 22.48765
##  F1      28.60000 0.7131698 20 27.11235 30.08765
##  F2      25.86667 0.7131698 20 24.37902 27.35431
##  F3      29.20000 0.7131698 20 27.71235 30.68765
##
## Confidence level used: 0.95
```

The arguments to lsmeans are first an `aov` object, and second a list of the names of treatment vectors. So far we only have one treatment vector, but later in the class we will have more. The output from `lsmeans` is a matrix, with the first column showing the treatment names, the second column showing the estimated sample means ($\bar{Y}_{i\cdot}$), and the rest of the columns related to hypothesis tests we'll discuss later.

4.2 Estimating Contrasts Using `lsmeans`

As we discussed in lecture, any estimable function of μ and $\{\tau_1, \tau_2, \ldots, \tau_v\}$ can be written as

$$\sum_{i=1}^{v} b_i(\mu + \tau_i)$$

and the least-squares (LS) estimate of such an estimable function is obtained by replacing each model treatment mean $(\mu + \tau_i)$ with it's corresponding treatment sample mean $\bar{Y}_{i\cdot}$.

Let's consider two estimable contrasts:

1. $\tau_1 - \tau_c = -1(\mu + \tau_c) + 1(\mu + \tau_1) + 0(\mu + \tau_2) + 0(\mu + \tau_3)$
2. $(\tau_1 + \tau_2 + \tau_3)/3 - \tau_c = -1(\mu + \tau_c) + 1/3(\mu + \tau_1) + 1/3(\mu + \tau_2) + 1/3(\mu + \tau_3)$

The first contrast is the difference in mean height between a plant getting F1 and a plant getting no fertilizer. The second contrast is the difference in mean height between a plant getting any fertilizer and a plant getting no fertilizer.

To estimate contrasts in R, we follow three steps:

A. Fit an ANOVA model to the data usin `aov`.

B. Make an lsmeans object using the `lsmeans` command.

C. Use the `contrast` command to estimate the contrasts.

The following code estimates the two contrasts above:

```
aov.greenhouse = aov(height~fert) ## fit ANOVA model
lsm.greenhouse = lsmeans(aov.greenhouse,"fert") ## Make lsmeans object
contrast(lsm.greenhouse,list("1.minus.Control"=c(-1,1,0,0),
                             "Fert.minus.Control"=c(-1,1/3,1/3,1/3)))
```

```
##  contrast           estimate        SE df t.ratio p.value
##  1.minus.Control    7.600000 1.0085744 20   7.535  <.0001
##  Fert.minus.Control 6.888889 0.8234975 20   8.365  <.0001
```

Note the format of the `contrast` command. The first argument is the name of the `lsmeans` object. The second argument is a list of the contrasts you want to estimate, with each one in the form of

"Contrast Name" = c(b_c, b_1, b_2, b_3)

where b_i are the numbers used to write the contrast as

$$\sum_{i=1}^{v} b_i(\mu + \tau_i).$$

Homework Assignment

1. Consider a completely randomized desgin with observations on three treatments coded 1,2,3. For the one-way ANOVA model, determine which of the following are estimable. For those that are estimable, write out the estimable function as $\sum_{i=1}^{3} b_i(\mu + \tau_i)$ and clearly state b_1, b_2, b_3. Finally, for those that are estimable, state the least squares estimator.

   (a) $\tau_1 + \tau_2 - 2\tau_3$

   (b) $\mu + \tau_3$

   (c) $\tau_1 - \tau_2 - \tau_3$

   (d) $\mu + (\tau_1 + \tau_2 + \tau_3)/3$

2. Recall the soap experiment from Homework 1. Look back at Homework 1 for an explanation of the experiment. The data are the weight lost over 24 hours by different types of soap.

Cube	Regular	Deodorant	Moisturizing
1	-0.30	2.63	1.86
2	-0.10	2.61	2.03
3	-0.14	2.41	2.26
4	0.40	3.15	1.82

   (a) Write out the one-way ANOVA model for this experiment.

   (b) By hand or calculator (without using R), obtain the LS estimate for the mean weight lost by a cube of deodorant soap. Show all calculations.

   (c) Consider estimating the difference in weight loss between regular soap and any other type of soap. That is, consider estimating $\tau_{regular} - (\tau_{deodorant} + \tau_{moisturizing})/2$. Show that this is estimable, and find the LS estimate by hand or calculator. Show all calculations.

   (d) Now use R to obtain the LS estimates in parts (b) and (c). Include your R code and the relevant output in your homework.

3. **Pedestrian light experiment** (Larry Lesher, 1985) This experiment questions whether pushing a certain pedestrian light button had an effect on the wait time before the pedestrian light showed "walk." The treatment factor of interest was the number of pushes of the button, and 32 observations were taken with a mix of 0, 1, 2, and 3 pushes of the button. The waiting times for the "walk" sign are shown in the following table, with $r_0 = 7,\ r_1 = r_2 = 10,\ r_3 = 5$ (where the levels of the treatment factor are coded as 0, 1, 2, 3 for simplicity).

0	1	2	3
38.14	38.28	38.17	38.14
38.20	38.17	38.13	38.30
38.31	38.08	38.16	38.21
38.14	38.25	38.30	38.04
38.29	38.18	38.34	38.37
38.17	38.03	38.34	
38.20	37.95	38.17	
	38.26	38.18	
	38.30	38.09	
	38.21	38.06	

   (a) Plot the waiting times against the number of pushes of the button. What does the plot show?

   (b) Write out the one-way ANOVA model for this experiment.

   (c) Use R to estimate the mean waiting time for each number of pushes.

(d) Show that the contrast $\tau_1 - \tau_0$ is estimable, and use R to find it's LS estimate. This contrast compares the effect of no pushes of the button with the effect of pushing the button once.

(e) Show that the contrast $(\tau_1 + \tau_2 + \tau_3)/3 - \tau_0$ is estimable, and use R to find it's LS estimate. This contrast compares the effect of no pushes of the button with the effect of pushing the button at least once.