Moving between the two Linear Parameterisations in R

Dr.Arnab Chakraborty * Mrinmoy Banik †

31th August 2022

Abstract

So far we have seen how the same linear estimation problem of finding the yields of various crop varieties can formulated in 2 different ways.We had also fitted 2 linear models with and without an intercept term. In this section we will see how to switch from the 'mean-yield-type' parameterisations to any of the 'benchmark-yield-type' ones.
Refer to the previous video:45 Keywords: parameterisation, estimation, linear

1 Recap

In the following agri dataset we had three different types of HYV seeds and the corresponding yields in each case.The general linear model is given by:-

$$Y_{ij} = \alpha_i + \epsilon_{ij} \tag{1}$$

But many a times we use a different formulation by introducing a mean yield intercept term μ:-

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{2}$$

Naturally μ then stands for the avg crop yield in any of the varieties with α_i's indicating the extra yield due to that particular crop variety.
But then we run into the problem of *identifiability*,which can be easily handled by imposing an extra condition on α_i's like:

$$\sum_{i=1}^{3} \alpha_i = 0 \qquad or \qquad \alpha_1 = 0$$

TABLE 1: A small dataset of agricultural outputs in 8 separate plots of land.

Agri Dataset:-

SL.no	Variety:-	Yield:-
1	1	210.3
2	2	245
3	2	248.9
4	3	212.3
5	3	230.4
6	2	250.1
7	1	213.5
8	1	212.4

*Department of Applied Statistical Division, Indian Statistical Institute,Kolkata Email:http://www.isical.ac.in/ arnabc/

†This portion of the book is latex-ed and edited by him,Student of B.Stat 3rd yr at ISI,BS2012

2 R Code and Explanation

First we need to set our working directory correctly and import the agri.csv dataset.Then fit a linear model, *fit1* using the lm() function.In the 1st case we use the formula $yield \sim variety$ which lets R compute the intercept term(μ) by default.
But since the design matrix is only of rank 3, R drops the 1st variety column while fitting the model as we can see from it's output below in FIGURE 2 (this is equivalent to setting $\alpha_1 = 0$ in equation (2)).

In fit2 on the other hand,

$$yield \sim variety - 1$$

is fitted forcing R to drop the intercept term.
By looking at the design matrices we can see they are both linearly independent in both cases.

```
#import the dataset
agri=read.csv('Agri.csv',header=TRUE)
#convert variety column into factor
agri$variety=factor(agri$variety)
#print the datset
agri
#show the dimensions
dim(agri)
#print summary
summary(agri)
#plot the box plots
plot(agri)
#fit the 1st model
fit1=lm(yield~variety,agri)
#print the estimated coefficients
fit1
#print the model matrix
model.matrix(fit1)
#fit the 2nd model
fit2=lm(yield~variety-1,agri)
#print he model coefficients
fit2
#print the model matrix
model.matrix(fit2)
```

FIGURE 1: The dataset in csv format is imported and two models *fit1* & *fit2* are fitted with lm() in R.

Now observe that the intercept in fit1 and the mean yield or coefficient of variety1 in fit2 are essentially same.(the small difference is due to numerical approximation.)This is because variety1 is the benchmark crop in fit1 so the intercept μ in this case equal to the mean yield of variety1 and the rest two coefficients are excess yield of varieties 2 & 3 when compared to that of variety1.

Hence the two models, fit1 & fit2 are equivalent or bijective parametrization. So the mean yield of variety2 is 35.993 **units greater than the benchmark while variety3's yield is only** 9.283 **units more making variety2 the highest yielding crop variety followed b variety2 and variety1 is the least productive.**

Lastly we can easy switch from one model to the other by subtracting the mean yield of any of the crop variety we want to set as benchmark from the rest and setting it's own $\alpha_i = 0$.

```
> #fit the 1st model
> fit1=lm(yield~variety,agri)
> #print the estimated coefficients
> fit1

Call:
   lm(formula = yield ~ variety, data = agri)

Coefficients:
   (Intercept)      variety1      variety3
248.00        -35.93        -26.65

> #print the model matrix
> model.matrix(fit1)
   (Intercept) variety1 variety3
1            1        1        0
2            1        0        0
3            1        0        0
4            1        0        1
5            1        0        1
6            1        0        0
7            1        1        0
8            1        1        0
```

(A) variety1 is dropped by R to match the rank of the design matrix

```
> #fit the 2nd model
> fit2=lm(yield~variety-1,agri)
> #print he model coefficients
> fit2

Call:
   lm(formula = yield ~ variety - 1, data =
    agri)

Coefficients:
   variety2  variety1  variety3
248.0     212.1     221.3

> #print the model matrix
> model.matrix(fit2)
   variety2 variety1 variety3
1         0        1        0
2         1        0        0
3         1        0        0
4         0        0        1
5         0        0        1
6         1        0        0
7         0        1        0
8         0        1        0
```

(B) With no intercept term, all three varieties are fit

FIGURE 2: Fitting the two parametrisations with & without the intercept

3 Interpretation

The estimated coefficients of α_i's in fit2 are essentially the mean yield of each variety. While in case of fit1, the $\alpha_1 = 0$ so the intercept term basically signifies the benchmark yield of crop 1 which is same as the estimate of the coefficient of variety1 in fit2.

The other two coefficients of variety2 and variety3 in fit1 is equal to $\alpha_2 - \alpha_1$ & $\alpha_3 - \alpha_1$ respectively from fit2. Similarly if we want R to drop variety 2 instead we can use the *relevel()* command as shown in FIGURE 3.

```
#relevel variety around variety2
agri$variety=relevel(agri$variety,ref=2)
#fit the 3rd model
fit3=lm(yield~variety,agri)
#print the estimated coefficients
fit3
#print the model matrix
model.matrix(fit3)
```

```
> model.matrix(fit3)
   (Intercept) variety1 variety3
1            1        1        0
2            1        0        0
3            1        0        0
4            1        0        1
5            1        0        1
6            1        0        0
7            1        1        0
8            1        1        0
```

FIGURE 3: relevel(agri$variety , ref=2) forces R to centre around variety2, hence now it drops variety2 equivalently setting $\alpha_2 = 0$

4 Conclusion

We see there are several ways to interprete the same (mathematically and statistically) linear model. Which one of them is best depends on the situation, but it doesn't change the final inference like here variety2 was superior in yield followed by variety3 and variety1.

5 Appendix

The complete R code for this section is provided below for reference:-

```
library(faraway)
setwd("C:/My Data/B.Stat/Semester 5/Linear Models/Assignments/Codes")
#################################video 46#####################################
#import the dataset
agri=read.csv('Agri.csv',header=TRUE)
#convert variety column into factor
agri$variety=factor(agri$variety)
#print the datset
agri
#show the dimensions
dim(agri)
#print summary
summary(agri)
#plot the box plots
plot(agri)
#fit the 1st model
fit1=lm(yield~variety,agri)
#print the estimated coefficients
fit1
#print the model matrix
model.matrix(fit1)
#fit the 2nd model
fit2=lm(yield~variety-1,agri)
#print he model coefficients
fit2
#print the model matrix
model.matrix(fit2)
#relevel variety around variety2
agri$variety=relevel(agri$variety,ref=2)
#fit the 3rd model
fit3=lm(yield~variety,agri)
#print the estimated coefficients
fit3
#print the model matrix
model.matrix(fit3)
```

[————————]