BIOE598 Case Study 2: Find minimal set of amino acid required for a bacterium

Reporter: Zong Fan

Email: zongfan2@illinois.edu

1. Objective

- Use 96 experiments to predict the smallest combination of amino acid for bacteria to grow.

- Experiments is split into 2 rounds.

2. Round 1: Method

- Use 32 runs in 2-level Resolution III Fractional Factorial (FF) Design
- Map Amino acid to following symbols:

amino acid	aa01	aa02	aa03	aa04	aa05	aa06	aa07	aa08	aa09	aa10	aa11	aa12	aa13	aa14	aa15	aa16	aa17	aa18	aa19	aa20
symbol	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	U

So the Defineing relation for this design:
I = ABF = ACG = BCH = ADJ = BDK = BCDL = ABCDM = AEN =
BEO = BCEP = ABCEQ = BDER = ABDES = CDET = ACDEU

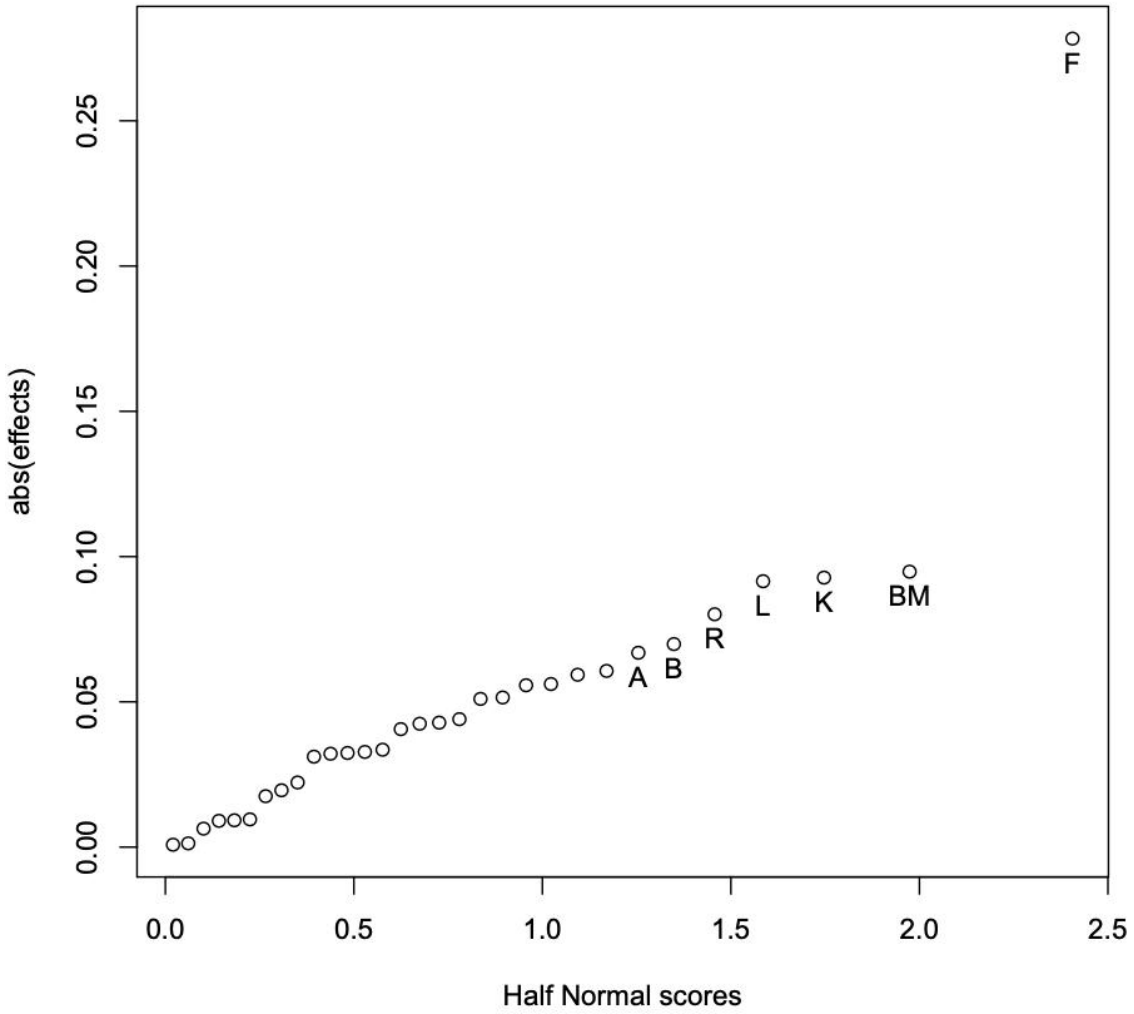
3. Round 1: Results

Set fitness threshold = 0.3, then runs with positive fitness are:

run	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	U	fitness
19	1	-1	-1	-1	1	-1	-1	1	-1	1	-1	-1	1	-1	1	1	1	1	1	1	0.35
25	1	1	1	1	-1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	0.51
26	-1	-1	-1	1	-1	1	1	1	-1	-1	1	-1	1	1	-1	1	1	-1	1	-1	0.60
13	1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	0.61
22	1	1	-1	-1	-1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1	0.66
10	-1	-1	1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	-1	1	0.72
17	-1	-1	-1	-1	-1	1	1	1	1	1	-1	1	1	1	-1	1	-1	1	-1	1	0.74
20	-1	-1	-1	-1	1	1	1	1	1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	0.89
15	1	1	-1	1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	0.96
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.98
1	-1	-1	1	-1	-1	1	-1	-1	1	1	1	-1	1	1	1	-1	-1	1	1	-1	1.00
24	-1	-1	1	-1	1	1	-1	-1	1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1.00
29	-1	-1	-1	1	1	1	1	1	-1	-1	1	-1	-1	-1	1	-1	-1	1	-1	1	1.00

3.1 Significant effects from Half-normal plot

Since we only have 32 runs for 20 main effects, there is no DoF left for confidence intervals. We use Half-normal plot instead.



Only F has relatively large coefficient. For fitness threshold = 0.3, F = 1 in 8/13 runs; if the threshold is set as 0.5, F = 1 in 8/8 runs. It indicates that F may be required for high fitness.

For the rest potential significant factors, A, B, R, L, K are main effects. So we need mirror image design to clear these main effects in the second run.

BM is the TWI term.

Therefore, A, B, F, K, L, M, R are the selected factors to investigate in the second round.

4. Round 2: Method

I. Use first 32 runs as mirror image of the first round with all factor flipped.

II. Use 32 runs in Resolution V half-fractional design for factors A, B, K, L, M, R. Set F=1 and the rest factors as -1

Design II's defining relation is:

I=ABKLMR=BKLMR=AKLMR=ABLMR=ABKMR=ABKLR=ABKLM

5. Round 2: Results

Postive results of mirror design I

Run	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	U	fitness
32	-1	-1	-1	1	1	1	1	1	-1	-1	1	-1	-1	-1	1	-1	-1	1	-1	1	1.00
34	-1	1	1	-1	-1	1	1	-1	-1	1	-1	-1	-1	1	-1	-1	1	1	1	1	0.64
37	1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	1	-1	1	-1	-1	1	1	-1	0.34
38	1	-1	-1	-1	-1	1	1	-1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	1	0.74
39	-1	1	-1	-1	1	1	-1	1	-1	1	1	1	1	-1	-1	-1	-1	-1	1	1	0.98
43	1	-1	1	1	-1	1	-1	1	-1	1	-1	1	1	-1	1	-1	1	-1	-1	1	0.58
45	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	0.60
47	-1	-1	1	-1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	0.55
50	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1	1	-1	1	-1	0.47
51	-1	1	1	1	-1	1	1	-1	1	-1	1	1	-1	1	-1	-1	-1	-1	-1	-1	0.66
53	-1	1	-1	-1	-1	1	-1	1	-1	1	1	1	-1	1	1	1	1	1	-1	-1	0.38
55	1	-1	-1	1	1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	0.66
60	1	1	-1	-1	-1	-1	1	1	1	1	1	-1	1	1	1	-1	1	-1	-1	1	0.39

Postive results of resolution V design II

run	A	B	F	K	L	M	R	fitness
89	1	-1	1	-1	1	1	1	1.00
93	-1	-1	1	1	1	1	1	0.77
95	1	-1	1	1	1	1	-1	0.77
84	1	-1	1	1	-1	1	1	0.77
78	1	1	1	1	-1	-1	1	0.73
71	1	-1	1	1	-1	-1	-1	0.70
88	1	1	1	1	-1	1	-1	0.68
92	-1	1	1	-1	1	-1	-1	0.65
86	-1	-1	1	-1	1	1	-1	0.61
91	1	1	1	-1	1	1	-1	0.60
96	1	-1	1	-1	1	-1	-1	0.60
83	1	-1	1	1	1	-1	1	0.59
76	-1	-1	1	1	1	-1	-1	0.53
73	1	1	1	1	1	1	1	0.52
69	-1	-1	1	-1	1	-1	1	0.51
70	1	1	1	-1	1	-1	1	0.50
80	-1	-1	1	1	-1	1	-1	0.46
75	-1	1	1	1	1	1	-1	0.43
81	-1	1	1	1	1	-1	1	0.41
72	-1	1	1	-1	1	1	1	0.39
90	-1	1	1	1	-1	-1	-1	0.30

5.1 Significant Effects

1. Combine data of round 1 and both round 2 I, II

Significant Factors and effect sizes

factor	estimate
C:F	0.36
F	0.20
A:G	0.12
C:J	0.10
G:H	0.10
C:D	0.10
K	0.09
L	0.07
A:S	0.07
C:L	-0.08
K:L	-0.14
A:H	-0.39

(A,B are near-significant)

2. Use only data of round 2 II

Significant Factors and effect sizes

factor	estimate
L	0.14
K	0.11
A	0.08
B	-0.07
K:L	-0.13

Note: F is set to 1 in this design

So the mirrored design confirms that F, K, L are significant main effects.

L, K are positive while their interaction are negative. Perhaps use either of them is enough for good fitness.

5.2 Minimum set of factors

- Analyze the result of round 2 II design. We find the minimum number of required factors are 3.

run	A	B	F	K	L	M	R	fitness
71	1	-1	1	1	-1	-1	-1	0.70
92	-1	1	1	-1	1	-1	-1	0.65
86	-1	-1	1	-1	1	1	-1	0.61
96	1	-1	1	-1	1	-1	-1	0.60
76	-1	-1	1	1	1	-1	-1	0.53
69	-1	-1	1	-1	1	-1	1	0.51
80	-1	-1	1	1	-1	1	-1	0.46
90	-1	1	1	1	-1	-1	-1	0.30

The combination could be:
AFK, BFL, FLM, AFL, FKL, FLR, FKM, BFK

6. Conclusion

- The significant main effects are mainly F, K, L, A, B.
- Factor F has the largest positive conbribution, while interaction of K and L may have negative impact.
- The minimum set of animo acid to guarantee bacteria fitness could be AFK, BFL, FLM, AFL, FKL, FLR, FKM, BFK.
- We choose the largest 4 (AFK, BFL, FLM, AFL) as the final evaluation run.

7. Appendix: Code

```
library("FrF2")
library("daewr")
library("leaps")

# first round
# 32 run with 3 resoulution
setwd("/Users/zongfan/Downloads")
des1 <- FrF2(nruns=32, nfactors=20, res.min=3)

# load result of first resoulution
data <- read.csv("casestudy2_result_round1.csv",
skip=1)
data <- data[,-1:-1]
# high fitness threshold
thres<-0.5
data_p <- na.omit(data[data$fitness>thres,])
data_n <- na.omit(data[data$fitness<thres,])

# change column names
names(data)[1:20] <- names(des1)
model <- lm( fitness ~ (.)^2, data=data)
cfs <- na.omit(coef(model))[-1:-1]
labels <- names(cfs)
daewr::halfnorm(cfs, labels, alpha=0.25,
refline=FALSE)

# select parameters with exhaustive
modpbr <- regsubsets(fitness~(.)^2, data=data,
method="exhaustive", nvmax=4, nbest=4,
really.big=TRUE)
rs <- summary(modpbr)
plot(c(rep(1:5,each=4)), rs$adjr2)
plot(modpbr, scale="r2")

# second round
# select A, B,M,L,K,R, set F=1, and rest values as -1
# resolution V
des2 <- FrF2(32, 6, res.min=5)
```

Round 1 Design

```
Call:
FrF2(32, 20, res.min = 3)

Experimental design of type  FrF2
32  runs

Factor settings (scale ends):
   A  B  C  D  E  F  G  H  J  K  L  M  N  O  P  Q  R  S  T  U
1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
2  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1

Design generating information:
$legend
 [1] A=A B=B C=C D=D E=E F=F G=G H=H J=J K=K L=L M=M N=N O=O P=P Q=Q R=R S=S T=T
[20] U=U

$generators
 [1] F=AB   G=AC   H=BC   J=AD   K=BD   L=BCD  M=ABCD N=AE   O=BE   P=BCE  Q=ABCE
[12] R=BDE  S=ABDE T=CDE  U=ACDE


Alias structure:
$main
 [1] A=BF=CG=DJ=EN=LM=PQ=RS=TU B=AF=CH=DK=EO              C=AG=BH=KL=OP
 [4] D=AJ=BK=HL=OR             E=AN=BO=HP=KR              F=AB=GH=JK=NO
 [7] G=AC=FH=KM=OQ             H=BC=DL=EP=FG=JM=NQ=RT=SU J=AD=FK=HM=OS
[10] K=BD=CL=ER=FJ=GM=NS=PT=QU L=AM=CK=DH=OT              M=AL=GK=HJ=OU
[13] N=AE=FO=HQ=KS             O=BE=CP=DR=FN=GQ=JS=LT=MU P=AQ=CO=EH=KT
[16] Q=AP=GO=HN=KU             R=AS=DO=EK=HT              S=AR=HU=JO=KN
[19] T=AU=HR=KP=LO             U=AT=HS=KQ=MO

$fi2
 [1] AH=BG=CF=DM=EQ=JL=NP=RU=ST AK=BJ=CM=DF=ES=GL=NR=PU=QT
 [3] AO=BN=CQ=DS=EF=GP=JR=LU=MT BL=CD=ET=FM=GJ=HK=NU=PR=QS
 [5] BM=CJ=DG=EU=FL=NT=PS=QR    BP=CE=DT=FQ=GN=HO=JU=LR=MS
 [7] BQ=CN=DU=EG=FP=JT=LS=MR    BR=CT=DE=FS=GU=JN=KO=LP=MQ
 [9] BS=CU=DN=EJ=FR=GT=LQ=MP    BT=CR=DP=EL=FU=GS=JQ=MN
[11] BU=CS=DQ=EM=FT=GR=JP=LN
```

	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	U
1	-1	-1	1	-1	-1	1	-1	-1	1	1	1	-1	1	1	1	-1	-1	1	1	-1
2	-1	-1	1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	-1	1
3	1	1	1	1	-1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
4	1	1	1	-1	-1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	1	-1	1	-1	-1	-1	1	-1	-1	1	1	1	-1	1	1	1	-1	-1	1	1
7	1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1
8	1	-1	1	1	1	-1	1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	1	1
9	-1	1	1	1	-1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	-1	1
10	1	-1	1	-1	1	-1	1	-1	-1	1	1	1	1	-1	-1	-1	1	1	-1	-1
11	-1	-1	-1	1	1	1	1	1	-1	-1	1	-1	-1	-1	1	-1	-1	1	-1	1
12	1	1	-1	-1	-1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1
13	-1	-1	1	-1	1	1	-1	-1	1	1	1	-1	-1	-1	-1	1	1	-1	-1	1
14	-1	-1	1	1	1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	-1	1	1	-1
15	1	1	-1	1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
16	-1	1	-1	-1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	-1	1
17	-1	-1	-1	-1	-1	1	1	1	1	1	-1	1	1	1	-1	1	-1	1	-1	1
18	-1	1	1	-1	1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1	-1	1
19	-1	-1	-1	-1	1	1	1	1	1	1	-1	1	-1	-1	1	-1	1	-1	1	-1
20	-1	1	1	1	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1	1	-1
21	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	1	-1	-1	1
22	1	-1	-1	1	-1	-1	-1	1	1	-1	1	1	-1	1	-1	-1	1	1	1	1
23	1	1	-1	1	-1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1
24	-1	1	-1	1	-1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	1	-1
25	1	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1
26	-1	-1	-1	1	-1	1	1	1	-1	-1	1	-1	1	1	-1	1	1	-1	1	-1
27	-1	1	-1	-1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	1	-1
28	1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	1	1	1	1	1	-1	-1
29	1	1	1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
30	1	-1	-1	1	1	-1	-1	1	1	-1	1	1	1	-1	1	1	-1	-1	-1	-1
31	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	1	-1
32	1	-1	-1	-1	1	-1	-1	1	-1	1	-1	-1	1	-1	1	1	1	1	1	1

class=design, type= FrF2

Round 2 Design

```
Call:
FrF2(32, 6, res.min = 5)

Experimental design of type  FrF2
32  runs

Factor settings (scale ends):
   A  B  C  D  E  F
1 -1 -1 -1 -1 -1 -1
2  1  1  1  1  1  1

Design generating information:
$legend
[1] A=A B=B C=C D=D E=E F=F

$generators
[1] F=ABCDE


Alias structure:
[[1]]
[1] no aliasing among main effects and 2fis
```

```
The design itself:
    A  B  C  D  E  F
1   1  1 -1 -1 -1 -1
2   1  1 -1 -1  1  1
3  -1 -1 -1 -1  1  1
4  -1  1 -1 -1  1 -1
5  -1 -1 -1  1 -1  1
6   1  1 -1  1 -1  1
7   1 -1  1 -1 -1 -1
8  -1  1 -1  1  1  1
9   1  1  1  1  1  1
10  1  1  1  1 -1 -1
11 -1  1  1  1  1 -1
12 -1 -1  1  1 -1 -1
13 -1 -1  1 -1 -1  1
14  1  1  1 -1 -1  1
15 -1  1 -1 -1 -1  1
16 -1 -1  1 -1  1 -1
17 -1  1  1  1 -1  1
18  1 -1 -1 -1  1 -1
19  1 -1  1  1 -1  1
20  1 -1  1 -1  1  1
21 -1  1  1 -1  1  1
22 -1 -1 -1  1  1 -1
23 -1 -1 -1 -1 -1 -1
24  1  1  1 -1  1 -1
25  1 -1 -1  1  1  1
26 -1  1  1 -1 -1 -1
27  1  1 -1  1  1 -1
28 -1  1 -1  1 -1 -1
29 -1 -1  1  1  1  1
30  1 -1 -1 -1 -1  1
31  1 -1  1  1  1 -1
32  1 -1 -1  1 -1 -1
class=design, type= FrF2
```

Round 2 Model fitting results

```
Call:
lm.default(formula = fitness ~ block + (.)^2, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14792 -0.06621  0.00000  0.06621  0.14792

Coefficients: (137 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4000169  0.0702346   5.695 1.18e-05 ***
block       -0.0693837  0.0444203  -1.562 0.133237
A           -0.0453823  0.0222101  -2.043 0.053770 .
B           -0.0411837  0.0222101  -1.854 0.077798 .
C           -0.0250129  0.0222101  -1.126 0.272790
D           -0.0149339  0.0222101  -0.672 0.508667
E           -0.0197321  0.0222101  -0.888 0.384375
F            0.1980946  0.0222101   8.919 1.38e-08 ***
G            0.0354755  0.0222101   1.597 0.125145
H            0.0262290  0.0222101   1.181 0.250826
J            0.0151907  0.0222101   0.684 0.501482
K            0.0868641  0.0222101   3.911 0.000804 ***
L            0.0708419  0.0222101   3.190 0.004408 **
M            0.0036226  0.0222101   0.163 0.871994
N            0.0254591  0.0222101   1.146 0.264573
O           -0.0008149  0.0222101  -0.037 0.971078
P            0.0197095  0.0222101   0.887 0.384910
Q           -0.0402334  0.0222101  -1.811 0.084393 .
R            0.0135169  0.0222101   0.609 0.549323
S            0.0122544  0.0222101   0.552 0.586946
T           -0.0181738  0.0222101  -0.818 0.422394
U            0.0438486  0.0222101   1.974 0.061650 .

Residual standard error: 0.1777 on 21 degrees of freedom
Multiple R-squared:  0.9363,  Adjusted R-squared:  0.7116
F-statistic: 4.168 on 74 and 21 DF,  p-value: 0.00028
```

Fitting result of main effects combining round1 and round2 data with blocking factor

```
Call:
lm.default(formula = fitness ~ (.)^2, data = data2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.256314 -0.089142  0.007951  0.093334  0.177932

Coefficients: (6 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.40366    0.02714  14.871 3.55e-11 ***
A            0.07886    0.02714   2.905 0.009855 **
B           -0.06819    0.02714  -2.512 0.022378 *
F                NA         NA      NA       NA
K            0.10855    0.02714   3.999 0.000929 ***
L            0.14492    0.02714   5.339 5.43e-05 ***
M            0.03514    0.02714   1.295 0.212732
A:B         -0.01100    0.02714  -0.405 0.690307
A:F              NA         NA      NA       NA
A:K          0.02379    0.02714   0.876 0.393006
A:L         -0.03470    0.02714  -1.278 0.218283
A:M          0.03069    0.02714   1.131 0.273891
B:F              NA         NA      NA       NA
B:K         -0.01436    0.02714  -0.529 0.603581
B:L         -0.05594    0.02714  -2.061 0.054941 .
B:M         -0.04805    0.02714  -1.770 0.094602 .
F:K              NA         NA      NA       NA
F:L              NA         NA      NA       NA
F:M              NA         NA      NA       NA
K:L         -0.13460    0.02714  -4.959 0.000119 ***
K:M          0.02863    0.02714   1.055 0.306290
L:M          0.02103    0.02714   0.775 0.449076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1552 on 17 degrees of freedom
Multiple R-squared:  0.8505,  Adjusted R-squared:  0.7185
F-statistic: 6.445 on 15 and 17 DF,  p-value: 0.0002239
```

Fitting result of main effects with only round2 II design data