Analysis of drug combination experiment on Lymphoma treatment

Mira
University of Georgia
Athens, GA 30602

October 9, 2020

1 Introduction

For this experiment we are taking a look at 7 FDA approved chemo-therapeutics. These drugs are include: Drug A(aclitaxel), Drug B (doxorubicin), Drug C(mitoxantrone), Drug D(cisplatin), Drug E(etoposide), Drug F (fluorouracil), and Drug G(ethotrexate). The goal of this experiment is to learn which combinations of these drugs will do best in the treatment of Lymphoma cancer. Lymphoma is a cancer mutating and over producing lymphocytes(cells that fight infections). To perform this experiment in the lab, Raji cells were taken from the human Lymphoma cell line in-vitro. Efficacy of treatment is measured by the rate of inhibition of cancer cells. Due to the expensive nature of this experiment, only 32 batched were produced. Statistical analysis could not be down without more trials because all of the combinations were not able to be tested.

2 Methods

In order to perform analysis on which combination of drugs would be most beneficial in fighting Lymphoma, we need to replicate the experiment many more times. 128 times to be exact. This is because we have two possible levels for each drug. So, 2^7 is 128 total trials that are needed. In order to complete this analysis we will need to do a 1/4 fraction factorial $2^7 - 2$ design. Using a fractional design instead of a full factorial is cheaper and will give the same effect as a full factorial design

2.1 Resolution and Defining Relation

The resolution of this design is 4, since the shortest defining relation has a length of 4:

$$2_{IV}^{7-2} = I = ABCF \tag{1}$$

2.2 Alias Structure

The alias structures created by the defining relation :

I=ABCF=ABDEG=CDEFG

A=BCF=BDEG=ACDEFG B=ACF=ADEG=BCDEFG
C=ABF=DEFG=ABBCDEF D=ABEG=CEFG=ABDF
E=ABDG=CDFG=ABCEF F=CDEG=ABDEFG=ABC
G=ABCFG=CDEF=ABDE AB=DEG=ABCDEFG=CF
AC=BCDEG=BF=ADEFG AD=BEG=BCDF=ACEFG
AE=BDG=BCEF=ACDFG AF=ACDEG=BDFG=BC
AG=BDE=BCFG=ACDEF BD=ACDF=BCEFG=AEG
BE=ADG=BCDEFG=ACEF BG=ACEG=BCDEF=ADE
CD=ABCEG=ABDF=EFG CE=DFG=ABEF=ABCDG
CG=DEF=ABEG=ABCDE DE=ABG=ABCDEF=CFG

DF=CEG=ABCD=ABEFG DG=ABC=ABCDFG=CEF
EF=CDG=ABCE=ABDFG EG=CDF=ABCEFG=ABD
FG=CDE=ABDEF=ABCG

3 Analysis

3.1 Factor Screening

Through running my R code to determine the alias structure, I was able to pull out all the important factors that may be included in the model. The next step is to fit a linear regression with all the important factors and plot them into a half-normal plot.

Regression Model:

Y∼A+B+C+D+E+F+G+A:B+A:C+A:D+A:E+A:F+A:G+B:D+B:E+B:G+C:D+C:E+C:G+D:E+D:F+D:G+E:F+E:G+F:G+A:C:D+A:C:E+A:C:G+A:D:F+A:E:F+A:F:G

Half Normal Plot with 10 significant factors:

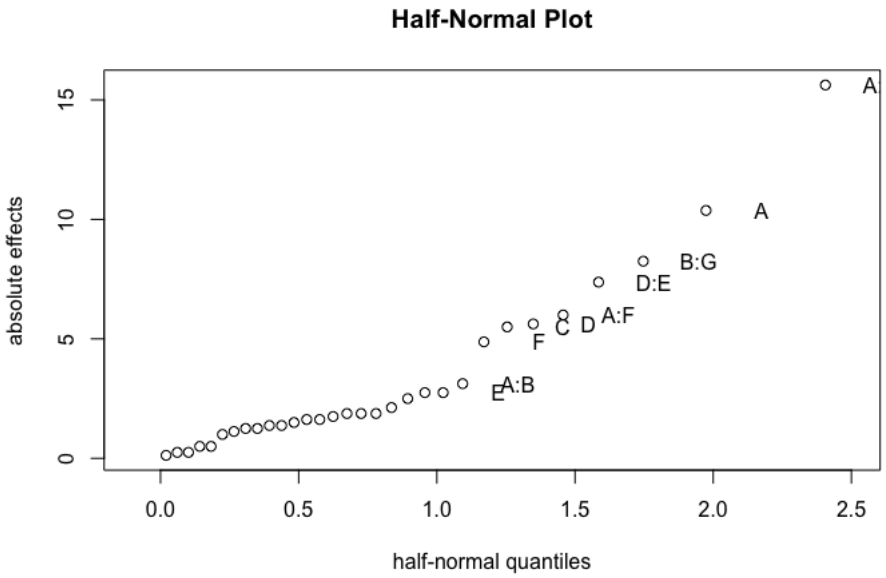

Figure 1: Half Normal Plot

From the half normal plot, we gathered that the significant factors from out linear model above are:

A, C, D, E, F, A:B, A:C, A:F, B:G, D:E

The main effect factors B and G do not show as significant according to the half normal plot. A:B is OKAY to include in our model since the Hereditary Principle states that at

least one parent factor should be included in the model. Based on this, the interaction term B:G must be removed from the model or at least on main effect factor must be included in the model.

After completing summary statistics on multiple model cases, the model with B:G, B and G will be the final model chosen as it has the lowest p-value and highest R^2 as well as R^2 adjusted. Note that B and G are not completely significant but are very close. I will keep this in mind, but proceed with caution. E does not show to be significant, but if we take it out the interaction term A:B becomes insignificant, and the R^2 value goes down, so I will keep it in model.

4 Results

The Final Model:

$$Y\sim \text{A+B+C+D+E+F+G+A:B+B:G+A:C+A:F+D:E}$$

4.1 Residual Analysis on Model:

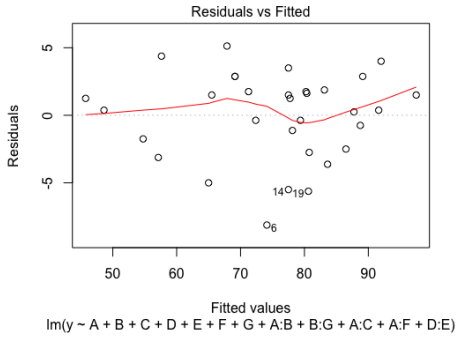

Figure 2: Residual vs Fitted Values

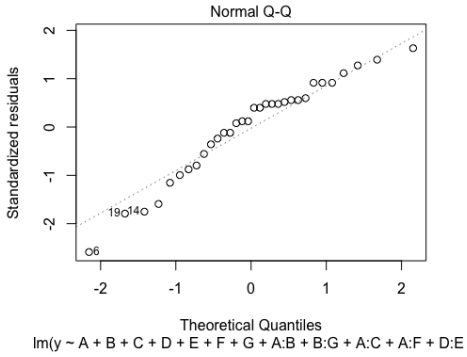

Figure 3: Normal Q-Q Plot.

Disregarding a few outliers such as 6,14, and 19, the two plots look good. Residual vs fitted shows random scatter around 0 with no distinct pattern. Also, the Normal q-q plot looks as a straight line given the exception of the three outliers.

4.2 Summary of Model:

```
Call:
lm(formula = y ~ A + B + C + D + E + F + G + A:B + B:G + A:C +
    A:F + D:E, data = data)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.1250 -1.9375  0.8125  1.7813  5.1250

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.2500     0.7206 104.421  < 2e-16 ***
A             5.1875     0.7206   7.198 7.75e-07 ***
B             1.3750     0.7206   1.908 0.071617 .
C             2.7500     0.7206   3.816 0.001166 **
D             2.8125     0.7206   3.903 0.000957 ***
E            -1.3750     0.7206  -1.908 0.071617 .
F             2.4375     0.7206   3.382 0.003126 **
G             0.7500     0.7206   1.041 0.311060
A:B           1.5625     0.7206   2.168 0.043047 *
B:G           4.1250     0.7206   5.724 1.62e-05 ***
A:C          -7.8125     0.7206 -10.841 1.41e-09 ***
A:F          -3.0000     0.7206  -4.163 0.000528 ***
D:E          -3.6875     0.7206  -5.117 6.13e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.077 on 19 degrees of freedom
Multiple R-squared:  0.9404,Adjusted R-squared:  0.9028
F-statistic: 24.99 on 12 and 19 DF,  p-value: 4.308e-09
```

An R^2 value of .94 is very high and our p-value is also very low at 4.308e-09. We can see that G is not significant in our model, however I am choosing to keep it in because it gives the highest adjusted R^2.

4.3 Predictions for all possible treatments using fitted model

1	2	3	4	5	6	7	8	9	10	11	12
69.625	65.750	60.500	76.750	65.875	58.750	81.375	66.250	76.750	63.750	79.500	85.000
13	14	15	16	17	18	19	20	21	22	23	24
68.375	91.625	81.250	81.500	79.625	70.750	83.250	66.625	72.000	86.875	87.750	81.000
25	26	27	28	29	30	31	32	33	34	35	36
67.250	74.000	82.625	88.000	68.875	73.750	53.250	79.750	86.875	71.375	88.125	89.750
37	38	39	40	41	42	43	44	45	46	47	48
68.500	77.125	77.375	60.125	47.875	70.375	90.000	57.625	63.750	87.000	68.625	66.625
49	50	51	52	53	54	55	56	57	58	59	60
44.625	57.875	89.625	73.875	91.500	79.875	77.000	74.875	90.875	75.625	70.875	68.375
61	62	63	64	65	66	67	68	69	70	71	72
55.750	92.750	49.625	101.000	86.750	76.875	56.125	92.625	90.875	85.750	55.875	73.125

73	74	75	76	77	78	79	80	81	82	83	84
84.875	49.250	74.250	82.625	55.500	68.500	79.000	80.750	65.500	52.875	74.375	78.625
85	86	87	88	89	90	91	92	93	94	95	96
79.125	58.000	94.000	64.125	77.250	82.125	99.875	85.000	89.500	97.875	82.500	75.500
97	98	99	100	101	102	103	104	105	106	107	108
75.625	62.625	76.625	67.000	83.875	57.500	81.500	66.125	78.500	69.000	95.125	76.625
109	110	111	112	113	114	115	116	117	118	119	120
87.375	84.375	83.875	76.750	58.375	89.750	89.875	85.250	81.625	76.750	72.750	78.625
121	122	123	124	125	126	127	128				
98.250	77.875	85.625	68.750	78.750	47.500	45.000	72.500				

Using table above of the 128 predicted values from each combination level, we found the maximum predicted response and determined the ideal levels for each if the 7 main effect factors:

Table 1: ideal factor settings for maximized Y: coded values

A	B	C	D	E	F	G
1	1	-1	1	-1	-1	1

Table 2: ideal factor settings for maximized Y: actual values

A	B	C	D	E	F	G
$3\mu M$	$.05\mu M$	$1\mu M$	$100\mu M$	$100\mu M$	$0.1\mu M$	$5\mu M$

5 Conclusion

The experiment we performed included 7 factors with 2 levels each and we decided to use a 1/4 fractional factorial design. The model and estimates we obtained are shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	75.2500	0.7206	104.42	0.0000
A	5.1875	0.7206	7.20	0.0000
B	1.3750	0.7206	1.91	0.0716
C	2.7500	0.7206	3.82	0.0012
D	2.8125	0.7206	3.90	0.0010
E	-1.3750	0.7206	-1.91	0.0716
F	2.4375	0.7206	3.38	0.0031
G	0.7500	0.7206	1.04	0.3111
A:B	1.5625	0.7206	2.17	0.0430
B:G	4.1250	0.7206	5.72	0.0000
A:C	-7.8125	0.7206	-10.84	0.0000
A:F	-3.0000	0.7206	-4.16	0.0005
D:E	-3.6875	0.7206	-5.12	0.0001

From this we can write out our model in terms of actual values for the factors.

Final Model with actual values for Factors:

Y=20.933+16.298A-396.875B+40.833C+0.163D+0.07E+337.5F-2.719G+78.125AB-17.361AC -120AF+103.125BG0.001821DE

Using this model and Table 2, we are concluded that the combination of drugs in Table 2 will yield the highest rate of inhibition of cancer cells and therefore be the most effective treatment for fighting Lymphoma cancer. All 7 drugs: paclitaxel at $3\mu M$, doxorubin at $.05\mu M$, mitoxantrone at $1\mu M$, cisplatin at $100\mu M$, etoposide at $100\mu M$, fluorouracil at $.1\mu M$, and methotrexate at $5\mu M$ are to be used in combination for optimal treatment.

6 Appendix

R code:

```
data
attach(data)
g0 = lm(y~.^7,data)
alias(g0)
options(max.print=1000000)


g=lm(y~A+B+C+D+E+F+G+A:B+A:C+A:D+A:E+A:F+A:G+B:D+B:E+B:G+C:D+C:E+C:G+
D:E+D:F+D:G+E:F+E:G+F:G+A:C:D+A:C:E+A:C:G+ A:D:F+A:E:F+A:F:G, data=data)
summary(g)
2*coef(g)[-1]  # estimated factorial effects

source("http://www.stat.ucla.edu/~hqxu/stat201A/R/halfnormal.R")
halfnormalplot(2*coef(g)[-1], label=T, n=8)
halfnormalplot(2*coef(g)[-1], label=T, n=10)

#checking multiple models
#Hereditery Principle, atleast one parent factor should be included in the model
g2=lm(y~A+C+D+E+F+A:B+A:C+A:F+D:E,data=data) #NO B:G, B, G, or B+G
g3=lm(y~A+B+C+D+E+F+A:B+B:G+A:C+A:F+D:E,data=data) #B:G and B
g4=lm(y~A+B+C+D+E+F+G+A:B+B:G+A:C+A:F+D:E,data=data) #B:G and B and G
g5=lm(y~A+C+D+E+F+A:B+A:C+A:F+B:G+D:E,data=data) #Only B:G
#I'm not checking of B:G and G only because B
factor takes higher precedent. So if I'm going to have
#one or the other, I will choose higher level (B)
summary(g2)
summary(g3)
summary(g4)
summary(g5)
#Based on my F stat p-value and R^2, g4 does the best when B:G and both parent factors
#BG and G are included in the model. However, the P-value of the two estimates do
```

```
#not show significant.I will keep this in mind, but proceed with caution. E does not show
#to be significant, but if we take it out the interaction term A:B becomes insignificant,
#and the R^2 value goes down, so I will keep it in model.

#g4 is final model!!
plot(g4,1:2)
#disregarding a few outliers such as 6,14, and 19, the two plots look good. Residual vs
#fitted shows random scatter around 0 with no distinct pattern. Also, the Normal q-q plot
#looks as a straight line given the exception of the three outliers.

#final model to make predication for all treatment
library(FrF2)
design = FrF2(nruns = 128, nfactors = 7)
design = data.matrix(design)
design = 2*(design-1.5)
design
yall = predict(g4, newdata = data.frame(design))
yall
design[which(yall == max(yall)),c(1,2,3,4,5,6,7)]
library(xtable)
xtable(summary(g4))
```

6.1 Initial tables and data

Table 1: Ranges of doses for drugs

Drug A	Drug B	Drug C	Drug D	Drug E	Drug F	Drug G
1-3 μM	0.01-0.05 μM	0.1-1 μM	10-100 μM	10-100 μM	0.05-0.1 μM	1-5 μM

Data taken through in-vitro experiment (32 batches)

"A" "B" "C" "D" "E" "F" "G" "y"
"1" 1 1 1 -1 1 1 -1 72
"2" -1 -1 1 -1 -1 1 1 79
"3" 1 -1 1 -1 -1 -1 -1 73
"4" 1 1 1 1 -1 1 -1 78
"5" -1 1 1 -1 1 -1 1 79
"6" 1 -1 1 1 1 -1 -1 66
"7" 1 1 1 -1 -1 1 1 79
"8" -1 1 -1 1 1 1 -1 62
"9" -1 -1 -1 -1 -1 -1 1 47
"10" -1 1 -1 -1 1 1 1 72
"11" 1 1 -1 1 1 -1 1 92
"12" -1 -1 -1 -1 1 -1 -1 54
"13" 1 -1 -1 -1 1 1 1 77
"14" -1 1 -1 1 -1 1 1 72
"15" 1 1 -1 -1 1 -1 -1 80
"16" -1 -1 1 -1 1 1 -1 92
"17" 1 1 -1 1 -1 -1 -1 96

"18" 1 -1 -1 1 -1 1 1 84
"19" -1 -1 1 1 1 1 1 75
"20" 1 -1 -1 -1 -1 1 -1 82
"21" 1 1 -1 -1 -1 -1 1 88
"22" -1 1 1 -1 -1 -1 -1 60
"23" 1 -1 -1 1 1 1 -1 85
"24" 1 -1 1 -1 1 -1 1 72
"25" -1 1 1 1 -1 -1 1 88
"26" -1 -1 -1 1 -1 -1 -1 67
"27" -1 1 -1 -1 -1 1 -1 53
"28" 1 1 1 1 1 1 1 82
"29" -1 -1 -1 1 1 -1 1 49
"30" -1 1 1 1 1 -1 -1 73
"31" -1 -1 1 1 -1 1 -1 99
"32" 1 -1 1 1 -1 -1 1 81