Christina Hartnett

AMS 582

Fractional Design Experiment

Introduction

The data provided was synthetically generated using ten independent variables, A through J. The corresponding values to each variable must be between -1 and 1. The dependent variable, Y, is defined by $Y = f(a,b,\ldots,j) + \varepsilon$. The error term, ε, follows a Normal(0 , σ^2) distribution. The primary purpose of this report is to create a design experiment to analyze the given data and find $f(a, b, \cdots, k)$. The data set consisted of a $2^{10-2} = 2^8$ design which has 256 runs. This design has Resolution VI. Using R studio, this report will find the model using linear regression, analytical strategies and multiple comparisons.

Methods

The data request sent into the TA was 256 runs that consisted only of -1's and 1's. I felt that this request was the most cost-effective method and was my best chance to save costs later. This csv file was created using FrF2 in RStudio, setting "Randomize" to False. The data returned now also contained a dependent variable, Y. The task is to now find the model that was used to generate these Y values. A 2^{10-2} fractional design was picked because it has a resolution VI design. This 1/4 fraction design will have the defining relationship: I = ABCDEFI = ABCGHJ. Because our design is 2^{10-2} we know that 2 of our variables are generated by the other variables. For instance, I is equal to ABCDEF and J is equal to ABCGH. The resolution VI design has three different abilities; it can estimate main effects that are unconfounded by four or less factor interactions, it can estimate two-factor interaction effects that are unconfounded by three or less factor interactions and it can estimate three-factor interaction interactions. The one issue with estimating three-factor interaction effects is that they could possibly be confounded by other three-factor interactions. Below are the three-factor aliases (determined using our defining relationship) that we need to watch out for. If we had one of these interactions in the design, more data would need to be requested from the TA to verify which should actually be included.

```
A:B:C = G:H:J
A:B:G = C:H:J
A:B:H = C:G:J
A:B:J = C:G:H
A:C:G = B:H:J
A:C:H = B:G:J
A:C:J = B:G:H
A:G:H = B:C:J
A:G:J = B:C:H
A:H:J = B:C:G
```

Methods used to determine the model included the linear regression and stepwise regression. I preformed these methods using “lm”, “stepAIC” and “aov” in R. StepAIC uses the Akaike Information Criteria to decide which model is best. When picking between two models we will choose the one with the lower AIC value. Models with unnecessary variables are penalized in this method. The stepAIC can be performed using an input direction; backward, forward or both. All of these methods returned the same variables with the same interactions.

```
> summary(stepAIC(fit23))
Start:  AIC=4467.36
y ~ C:J + A:E:F

        Df  Sum of Sq         RSS    AIC
<none>                 9479173828 4467.4
- C:J    1  856756541 10335930368 4487.5
- A:E:F  1 4152483698 13631657525 4558.4

Call:
lm.default(formula = y ~ C:J + A:E:F, data = design.num)

Residuals:
     Min       1Q   Median       3Q      Max
-17263.1  -4153.1   -123.8   4046.9  21157.7

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  89950.3      382.6 235.124 < 0.0000000000000002 ***
C:J           1829.4      382.6   4.782           0.00000295 ***
A:E:F         4027.5      382.6  10.528 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6121 on 253 degrees of freedom
Multiple R-squared:  0.3457,    Adjusted R-squared:  0.3406
F-statistic: 66.85 on 2 and 253 DF,  p-value: < 0.00000000000000022
```

To further help determine my model, I looked at the ANOVA table of all three factor and below interactions using the aov function in R. If the P-values in the ANOVA table were significant, less than .0001, then they were included in the model (C:J and A:E:F). I decided to only include very significant variables in my final model and eliminate significant p-values that were < .01 but > .001 (H). According to Occam's Razor Principle we should chose a simpler model to a complex one when they produce a very similar result. We want to select an explanation with the smallest number of assumptions.

Results

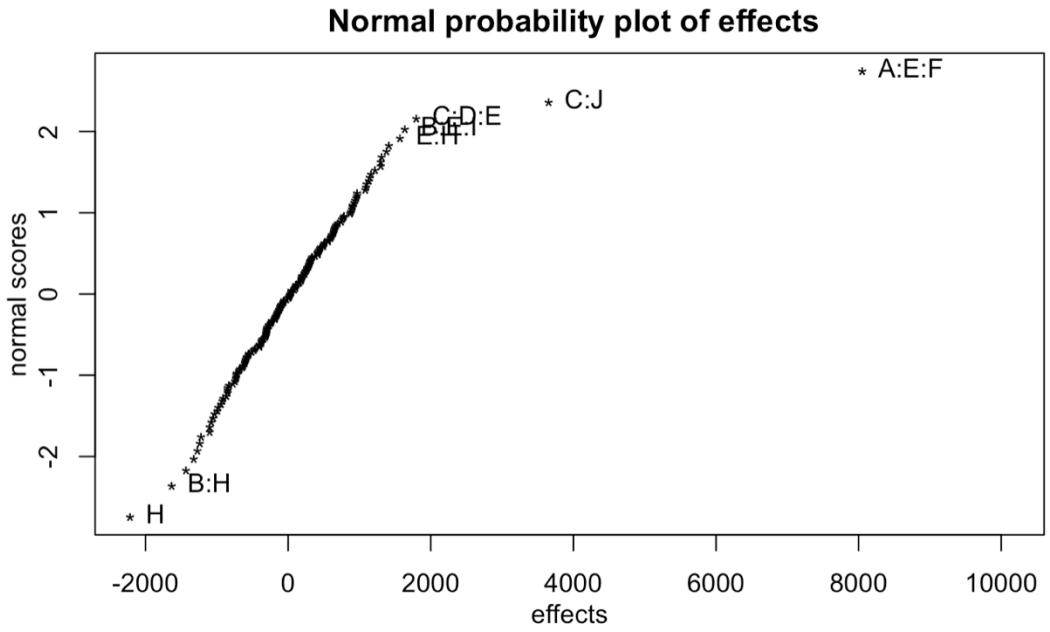
The estimated final model is **$y = (C:J) + (A:E:F) + \epsilon$.** Below describes which variables are in or out of the model.

Variables	Model
A	In
B	Out
C	In
D	Out
E	In
F	In
G	Out
H	Out
I	Out
J	In

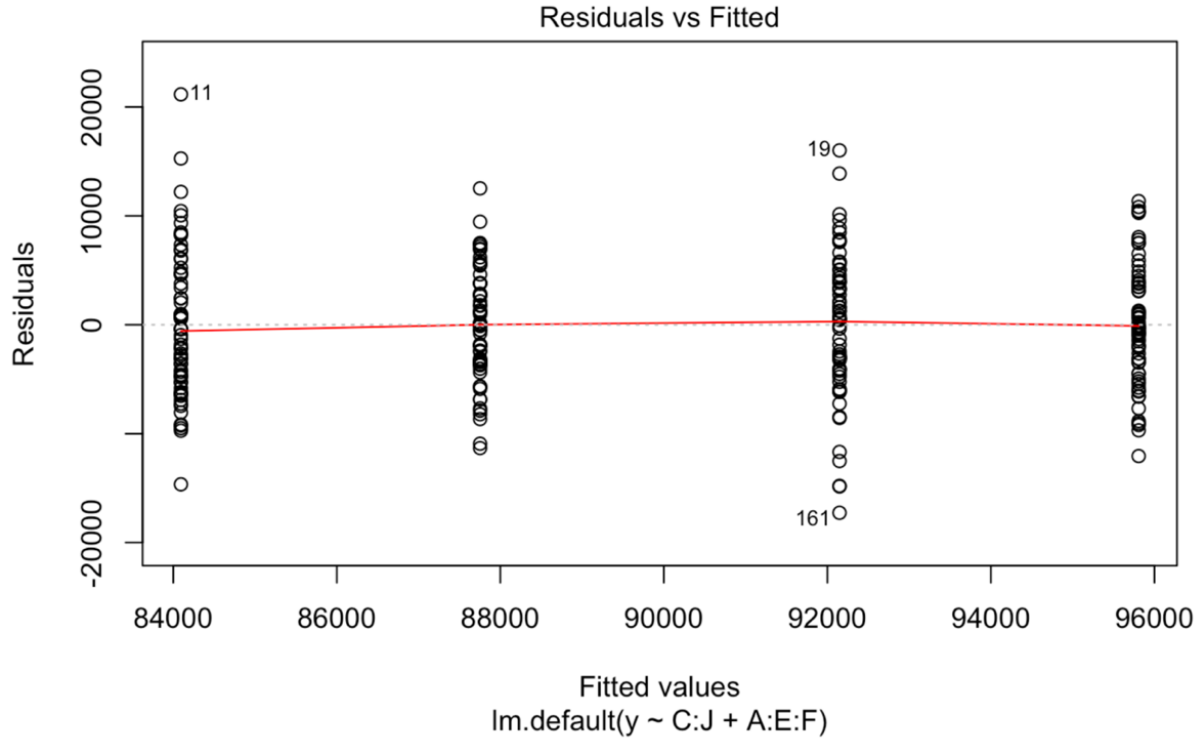
Below is the ANOVA table of the model. It shows the interactions have P-values that are highly significant. We can see that C:J and A:E:F are highly significant for my model.

```
Call:
lm.default(formula = y ~ C:J + A:E:F, data = design.num)

Residuals:
     Min       1Q   Median       3Q      Max
-17263.1  -4153.1   -123.8   4046.9  21157.7

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  89950.3      382.6 235.124 < 0.0000000000000002 ***
C:J           1829.4      382.6   4.782           0.00000295 ***
A:E:F         4027.5      382.6  10.528 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6121 on 253 degrees of freedom
Multiple R-squared:  0.3457,    Adjusted R-squared:  0.3406
F-statistic: 66.85 on 2 and 253 DF,  p-value: < 0.00000000000000022


Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq F value                Pr(>F)
C:J         1  856756541  856756541  22.867           0.00000295 ***
A:E:F       1 4152483698 4152483698 110.830 < 0.00000000000000022 ***
Residuals 253 9479173828   37467090
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

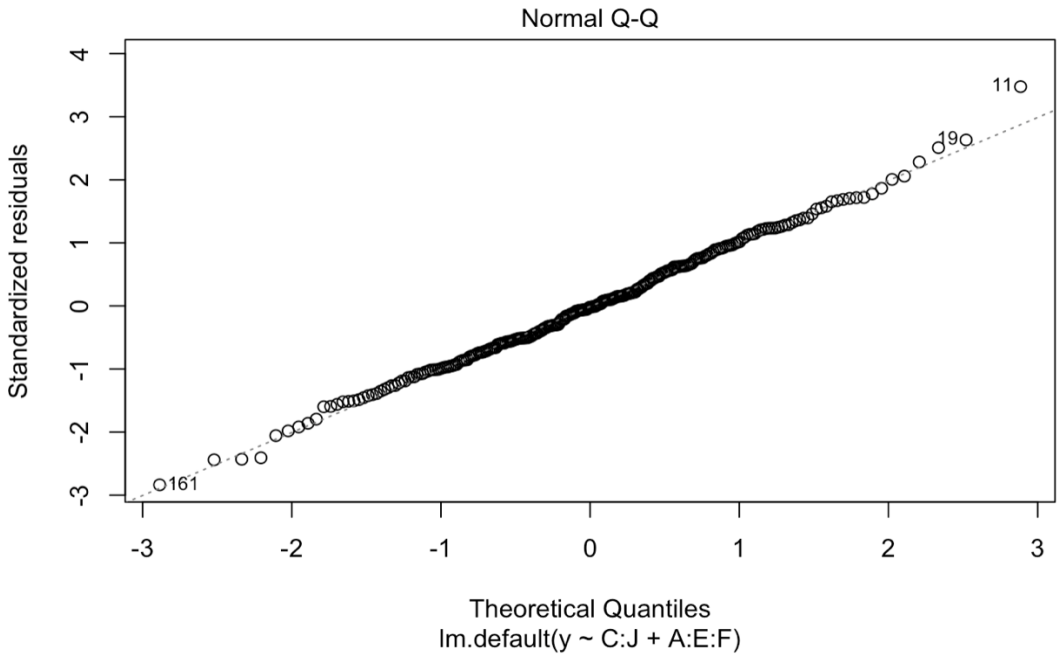
We can also look at the Daniel plot to assess the significance of variables. Below we can see that C:J and A:E:F are still our most significant in this plot. The Daniel plot is also referred to as a half normal plot. For the best results using a Daniel plot we should include as many effects in the model as possible. The Daniel plot also works best for larger designs.



Another method to further test the model is to check the Residuals vs. Fitted plot to ensure that there is a straight line close to 0. It shows that the variances of the error terms are equal.



The normal Q-Q plot was used to check for normality of the residuals. It is a fairly straight line following the normal distribution we assume the data is normal.



Conclusions:

The estimated model is y = 89950.3 + 1829.4 * (C:J) + 4027.5 * (A:E:F) + ϵ. The variables in the final model are A, C, E, F, J. These were decided based on their significance and the significance of their interaction. This estimated model returns an adjusted R-Squared of .3406. This adjusted R-squared is large enough to ensure that it is a good fit for the given data. There were some limitations which could have affected the accuracy of this model. The first limitation is the cost of the experiments. I decided to choose a larger number of runs which has a high cost. This method I chose increased the risk of having to request more data at more cost. In the end I did not have to request a second run, so I saved cost that way. I was able to test all third order interactions with the amount of data requested so I felt confident in the amount of data I requested. I felt that the p-values of the interactions I choose were significant enough to include in the final model. In the end, despite the limitations, I feel confident that the amount of data requested was useful in determining the model.

References:

"Fractional Factorial Design." *Wikipedia*, Wikimedia Foundation, 6 Nov. 2020, en.wikipedia.org/wiki/Fractional_factorial_design.

Gromping, Ulrike. "R Package FrF2 for Creating and Analyzing Fractional Factorial 2-Level Designs." Journal of Statistical Software, 2014.