Experimental Designs

# Factorial Design

Final Project

Author: Bita Nezamdoust

Instructor: Dr. Qin Xi

Fall 2018

Georgia State University

# Introduction

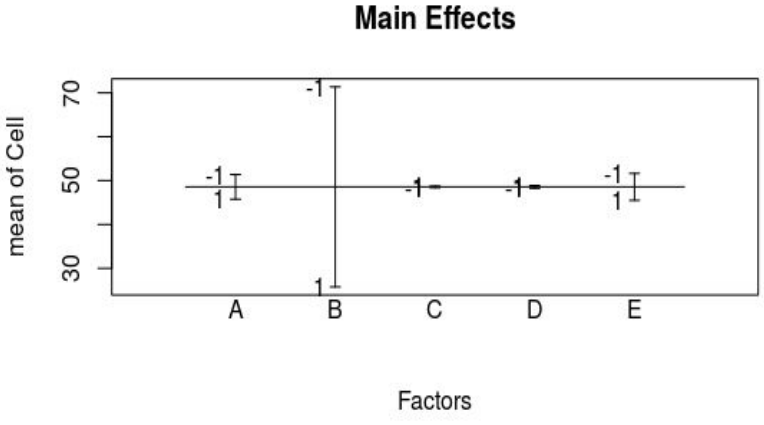
A special type of cell is incubated in a biology lab. The amount of cells are measured after a specific time period. Given the same initial amount of cells, the final amount of cells may vary due to factors such as temperatures, diluted concentration, and so on. The experimenters want to choose a combination of the levels of these factors such that they can obtain the maximum final amount of cells. They choose five factors which may affect the amount of cells. The factors are denoted by A, B, C, D and E. Two levels are selected for each factor and are denoted by "+" and "-".

In this study, I will analyze the data to look for the most important main effects or interaction effects that are influencing the final amount of cells and I will attempt to produce models to be used to explain and also predict the amount of cells on the basis of the levels of the desired factors. The purpose is to find the best combination of these factors which can produce the maximum amount of cells. I will analyze four different datasets obtained from different factorial design experiments for this study and appropriate designs are made which will be described in the following.

# Data Analysis

## Design 1

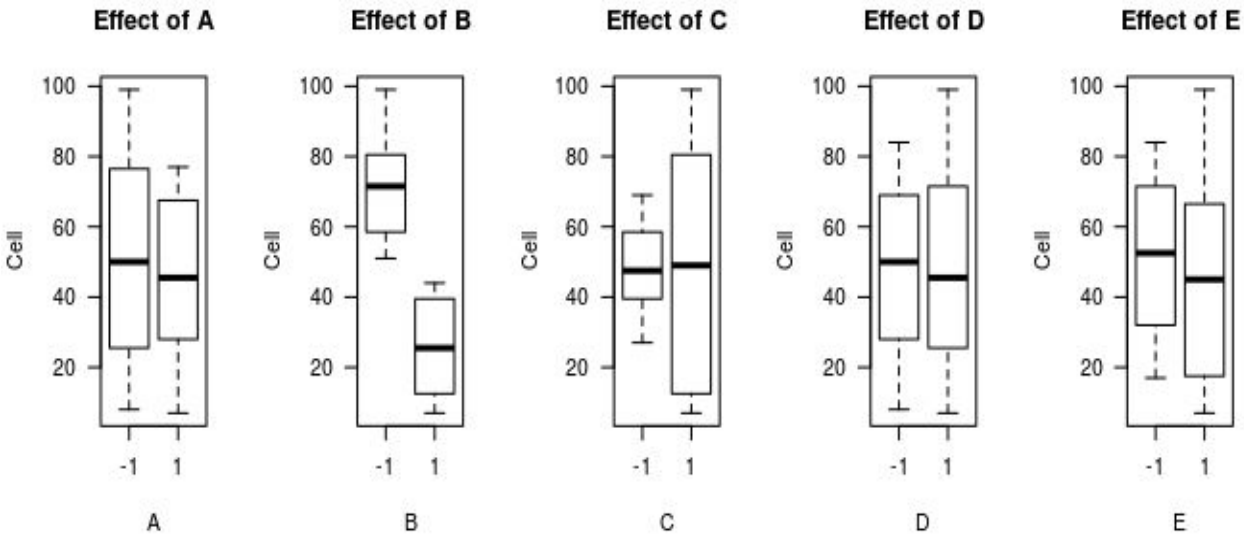
A complete replicate of a $2^5$ experimental design requires 32 runs in one replicate. However, due to low budget, in design 1, the data is collected on 16 observations of six variables, one response variable a.k.a Cells and five factors called A, B, C, D, and E. That means, the data that we have in hand is a fraction of the complete factorial experiment on which we can make a *fractional factorial design*. More specifically, a *one-half fractional design* of $2^5$ experiments a.k.a a $2^{5-1}$ design is performed.




Initially, a boxplot of the several two-level factors against the response is drawn. According to the plot, the most variation in the response seems to be caused by factor B, while there is little

variation caused by the other four factors. I will investigate this in fuller length in my analysis.

Main effects can also be shown in a main effect plot. Consistent with the box plot, effect B seems to be most significant.

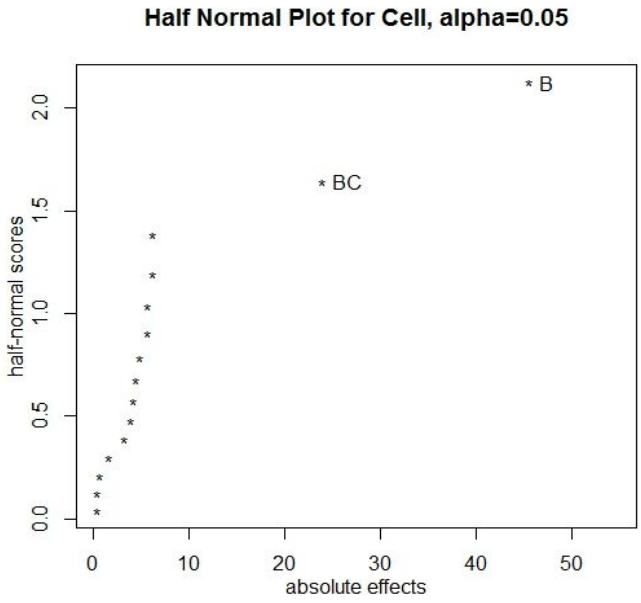



To begin the analysis, I investigate the alias structure for the design. The aliases are determined by using a *defining relation.* Multiplying any effect by the defining relation yields the aliases for that effect. In this design, the defining relation is $I = ABCDE$ which is obtained by multiplying $A$ by the first term found below in the aliases. Using this defining relation, this is a $2^{5-1}$ design with a resolution V, or $2_V^{5-1}$ . Higher choice of resolution is always more desirable that is why we chose this defining relation. While can be done by hand, R can also be used to produce the resulting aliases. The results are shown below:

```
A = B:C:D:E, B = A:C:D:E, C = A:B:D:E, D = A:B:C:E, E = A:B:C:D
A:B = C:D:E, A:C = B:D:E, B:C = A:D:E, A:D = B:C:E, B:D = A:C:E
C:D = A:B:E, A:E = B:C:D, B:E = A:C:D, C:E = A:B:D, D:E = A:B:C
```

After diagnosing the aliases, to do the analysis, I project the fractional design to a full $2^4$ factorial design with four factors A,B,C,D, where the factor E is aliased with the effect ABCD as can be seen above. I now use *half-normal probability plots* to determine which main or interaction effects are the most significant.

The plot shows that main effect B (aliased with ACDE) and BC (aliased with ADE) are the most important effects in the design by Lenth's method.

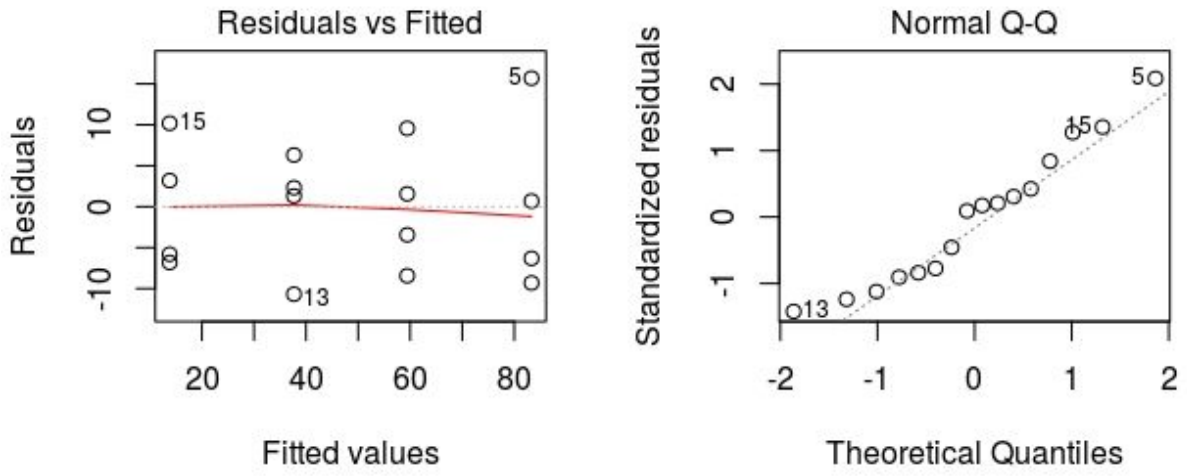



The main effects are aliased with the four-factor effects, and the two-factor effects are aliased with the three-factor effects. So to estimate B and BC, we have to assume that all the three-factor effects and above are negligible. Having this assumption, we then can claim that between effects B and ACDE, and also BC and ADE, only B and BC are significant.

Having neglected the three-factor effects, I test three permutations of full models consisting of three out of the five initial effects. I use ANOVA test to investigate the factor effects in each model. According to the results, the model can be made using either one effect B or two effects B and B:C as most significant predictors. I use ANOVA test of comparison to compare such two models, one model with two effects and one reduced model with one effect. Since the p-value is too large (p-value = 0.5137), I fail to reject the null hypothesis representing the reduced model and accept the reduced model as the best fit: Cell~B+BC .

The model is significant with F-statistic = 46.9 and p-value= 6.674e-07. The multiple R-squared = 0.92 and adjusted R-squared = 0.90.

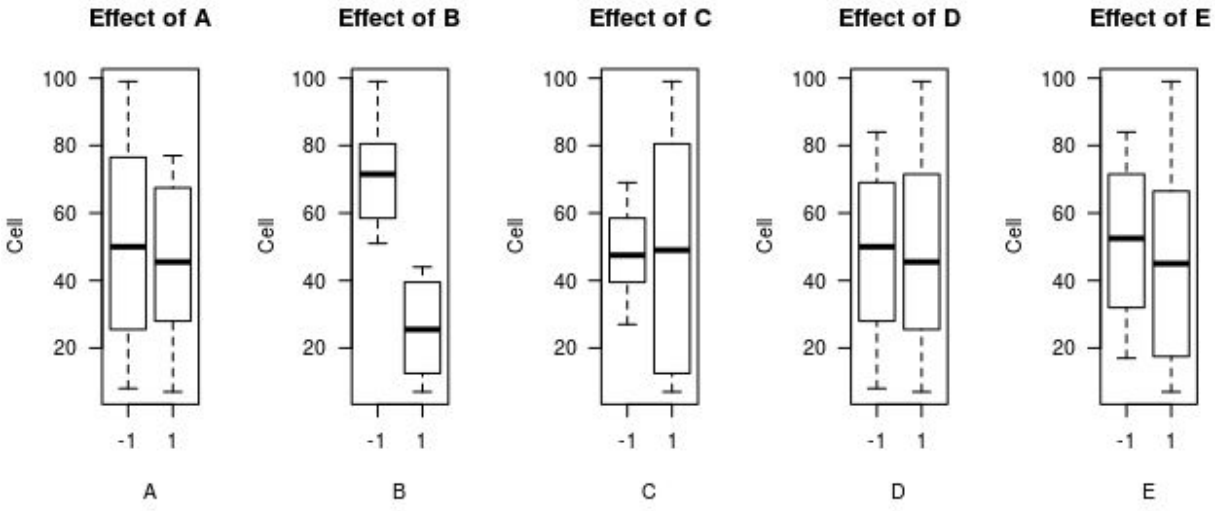
Finally, the model adequacy was checked using residual plots. As can be seen below, the residuals are normally distributed and the error standard deviation is constant as desired, with a few potential outliers.



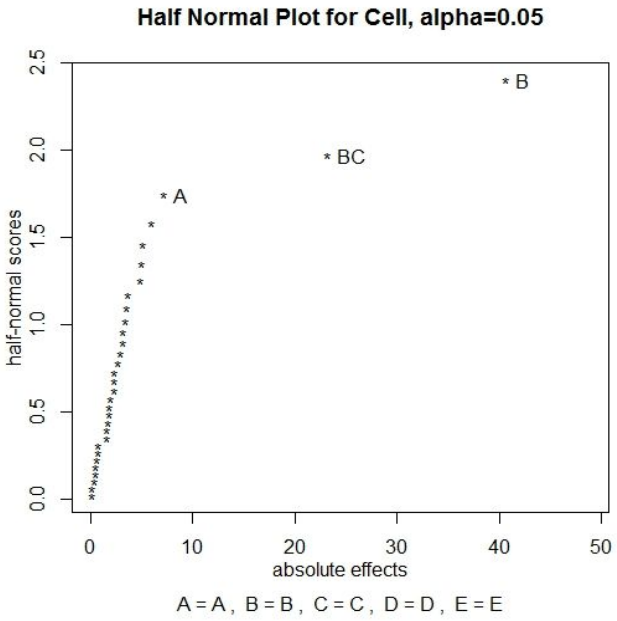

## Design 2

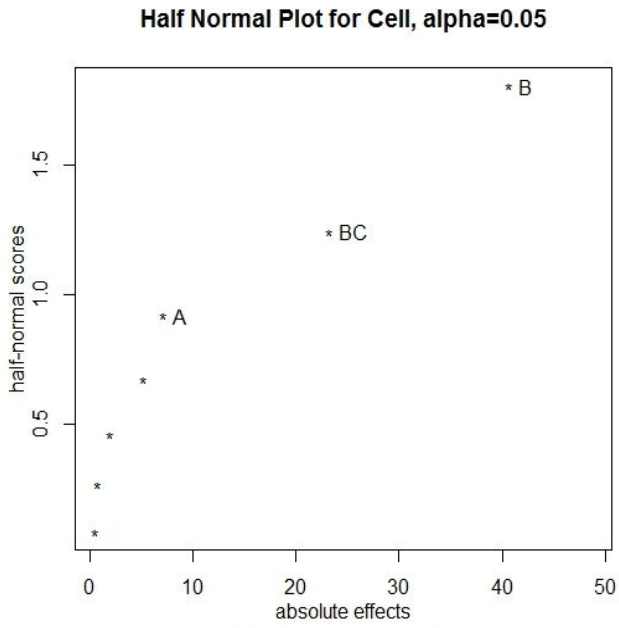
For the second design, the follow-up experiments with 16 runs is performed. The full fold-over technique is supposed to be used in the design of the follow-up experiments. I have combined the data from the follow up experiment with the data from the first design and the resulting data consists of 32 observation of the six variables. With 32 observations, we have a full $2^5$ model. This is verified by checking the aliases and not detecting any alias effects.

The boxplot of the data does not indicate much change from design 1 and effect B still seems significant.



After checking the alias effects, I then use half-normal probability plot to find important effects. The plot for a full five-factor model, as seen to the right, shows that main effects A, B and BC are the most important effects in the design by Lenth's method. The results are consistent when I draw the plot for the full three-factor model as shown below. Using the three-factor model, I will project the factorial design to a full $2^3$ factorial design.. I try three different full model fits with three-factor combinations of A,B,C,D,E and use ANOVA output to identify significant factors. I then use ANOVA comparison test to test the significance of keeping A and BC in the model. With significant p-values of 6.209e-08 and 0.01865 I verify that BC and A, respectively, should stay in the model.



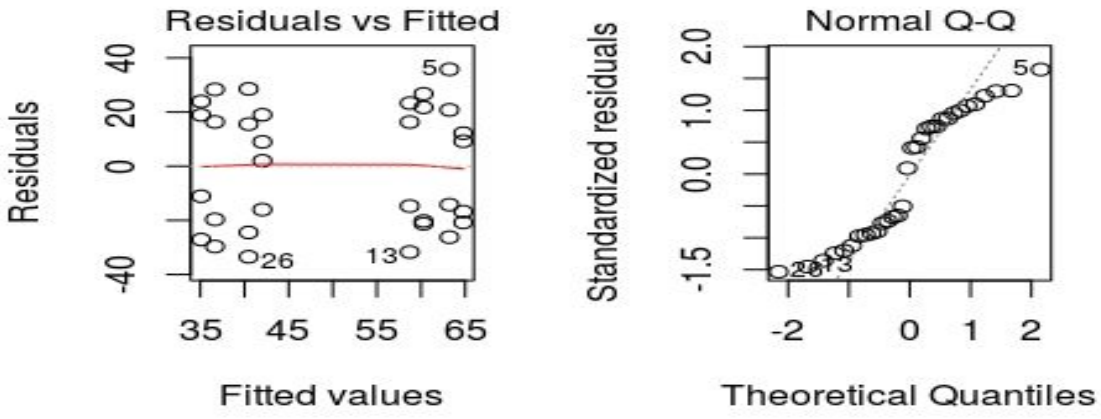




In addition, I have made a block column to account for the facts that the two experiments are not performed is one day. So two block have been assigned for the two days and using ANOVA their effect was tested. As a result, with a p-valued of 0.275, the model with block effect was deemed insignificant as compared to th model without the block effect, so the block effect was removed from the model.

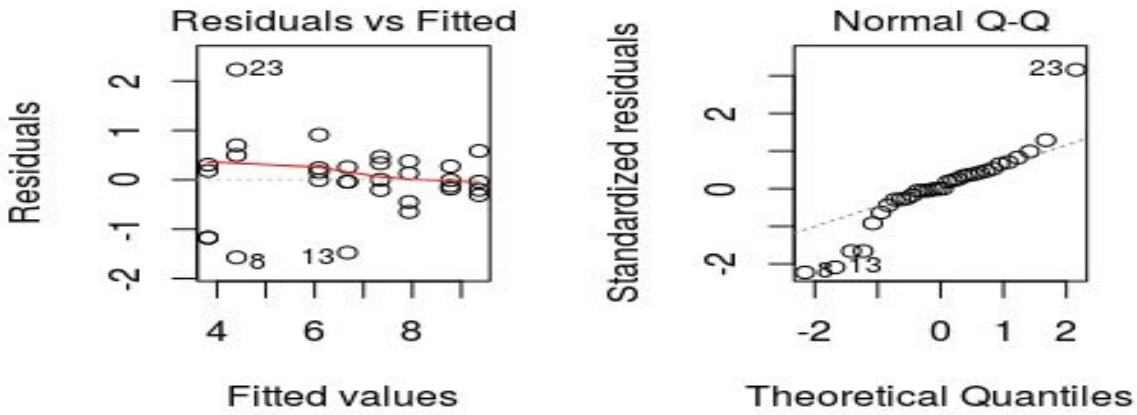
Finally, I use the three final terms to fit the model and check the model adequacy. However, sing Cell~A+B+AB model, the residual plot as shown below does not seem to meet the assumptions of normality and constant variance of errors. Therefore, to improve the model, I transform the response by taking the square root of the

variable Cell. The residual plots as can be seen are more reliable in meeting the assumptions with only a few potential outliers. So the final model is $\sqrt{Cell}$~A+B+BC.

"Residual plots of Cell~A+B+BC"




"Residual plots of $\sqrt{Cell}$~A+B+BC"

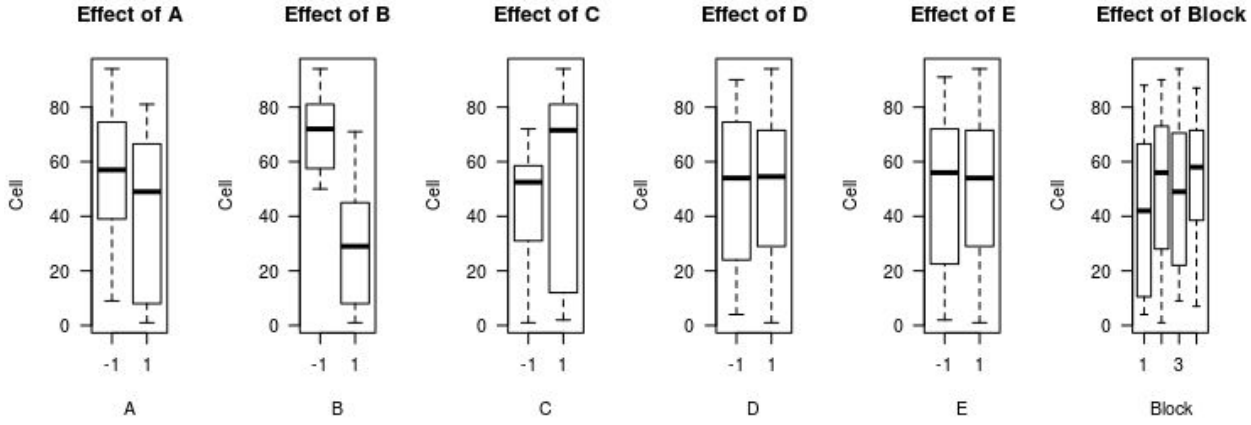



In comparison to design 1, the model made in design 2 has in fact deteriorated in terms of R-squared which have decreased from 0.92 to 0.69 and residual error which have increased from 8.68 to 14.49. Also, it is noteworthy that the interaction effects cannot be measured as properly in design 2 because we don't have any replicates and the sum of residual errors is large. While it is included in the model, it should be dealt with caution.

## Design 3

For the third design, as the team obtains more financial support to do the experiments, a full factorial experiment with two replicates is performed. However, due to the limitations of incubators, only 16 experiments can be run on the same day. Hence, the 64 runs were assigned to four days which will make up four blocks to be considered in the model. The assignment of the first replicate was based on the effect ABCDE which was selected to be confounded with block effects. The assignment of the second replicate was based on effect ABCD.

The blocks are made accordingly and the boxplot of all main effects and the block effect is drawn.




There is variability in the two levels of effect B and partly C, and the variability can also be seen among the four levels of the blocks. Further analysis is needed to investigate their significance.

To make the model, a full model of all effects and the block effect is fit. Using the ANOVA test, significant effects can be identified and the model is re-fit using the significant effects. As for the block effect, it can be seen in the output that the mean squares for the block is equal to 1116.2 in the model, while the mean squares of error is 2607.6 which is larger. That means the variation between the blocks is smaller than the variation within the blocks, which means the blocking is not very significant and can be removed from the model.

The model is refit and the ANOVA test shows that all the effects are significant. Alternatively, I have tried to examine the effects using half-normal probability plots and fit a shorter model with significant

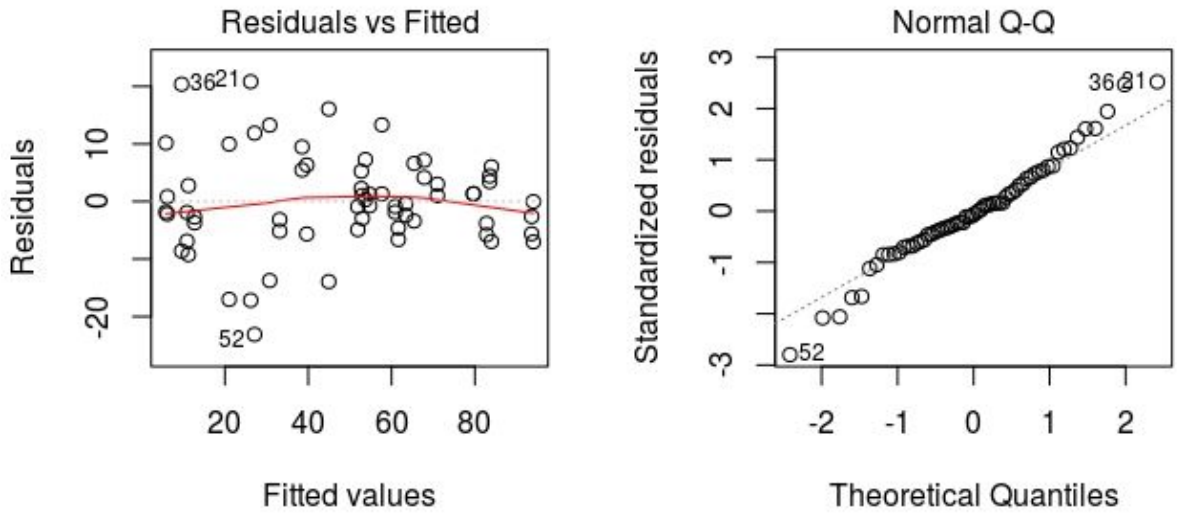
terms. I then used ANOVA to compare the two models and with a small p-value, the former model was accepted.

Therefore, the final model using the ANOVA method is as follows:

Cell = 48.4 - 7.1A - 21.7B + 3.6C - 0.6D - 0.1E - 3.3AB - 3.0AC - 8.4BC - 2.6AD + 4.3CD + 2.8CE + 0.01BD - 2.0AE + 1.8BE - 2.9ACD -4.2BCD - 2.7ABE

With effect B being the strongest, having the largest coefficient.

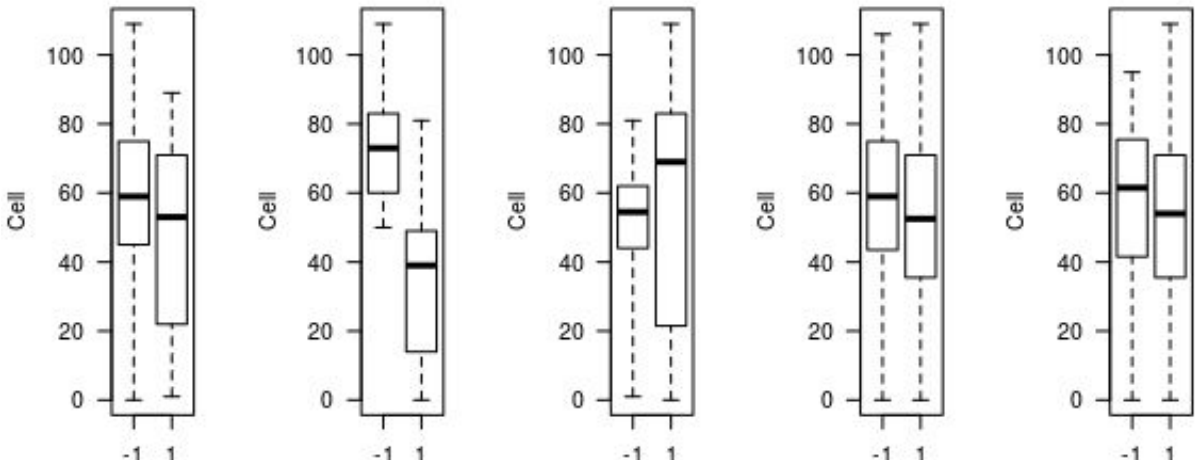
Finally, the model adequacy was checked using residual plots. As can be seen below, the residuals are normally distributed and the error standard deviation is quite constant with a few potential outliers.



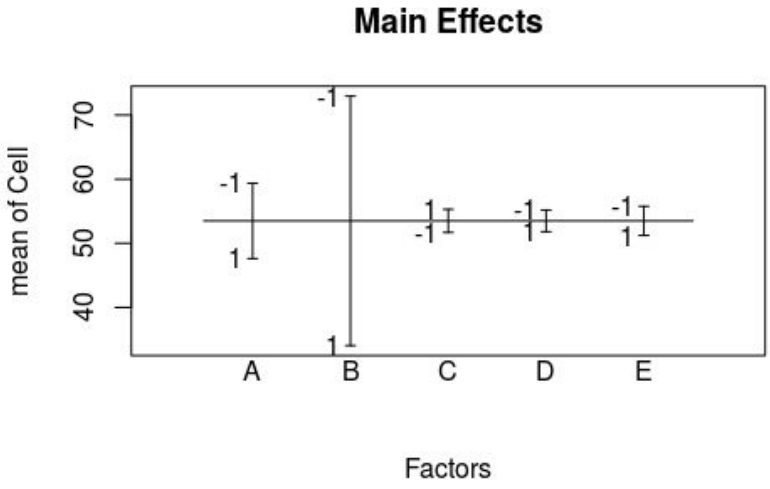

The model have improved in comparison to the model in design 2 both in r-squared (0.91) and residual errors (9.75), however, it is not as good a fit as design 1.

## Design 4

In the fourth design, with new budget, a new incubator was purchased which allows 320 experiments simultaneously. Hence, a full factorial experiment with 10 replicates is designed.




There is variability in the two levels of effect B and partly C. The plot of main effects is also made which is consistent with the box plots.



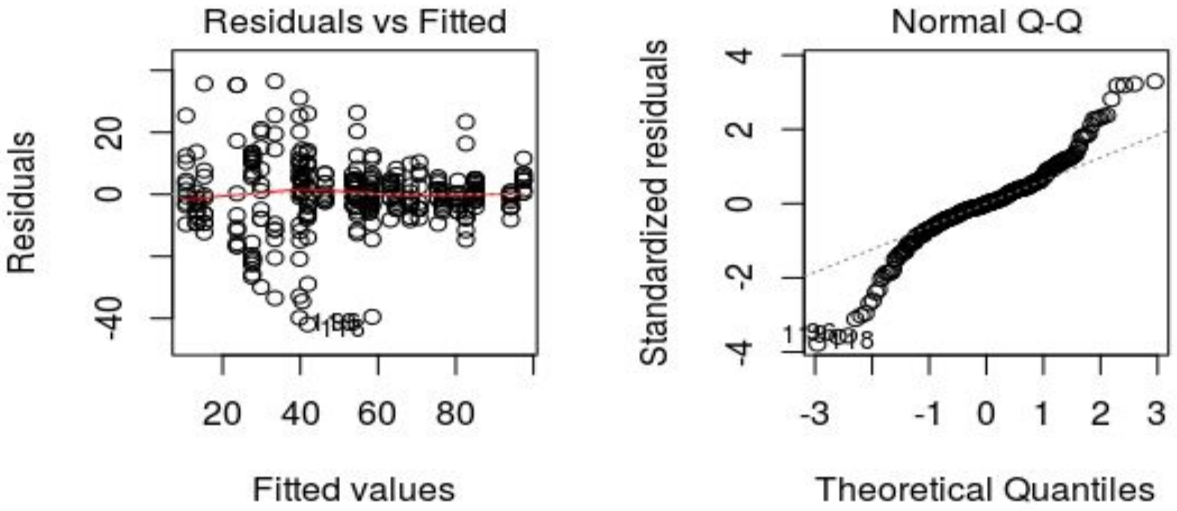

To make the model, a full model of all effects is fit. Using the ANOVA test, significant effects can be identified and the model is re-fit using the significant effects. The model is refit and the anova test shows that all the effects are significant.

The final model using the ANOVA method is as follows:

Cell = 53.5 - 5.9A - 19.4B + 1.8C - 1.7D - 2.3E - 2.5AB - 1.8AC - 9.9BC - 1.9BD + 1.3CD - 2.1AE + 2.6CE + 1.4DE - 0.8AD + 0.6BE + 1.5ACD - 2.4ACD - 1.6ABE

With effect B being the strongest having the largest coefficient.

Finally, the model adequacy was checked using residual plots. As can be seen below, the residuals are normally distributed while the error standard deviation is quite, but not completely constant.

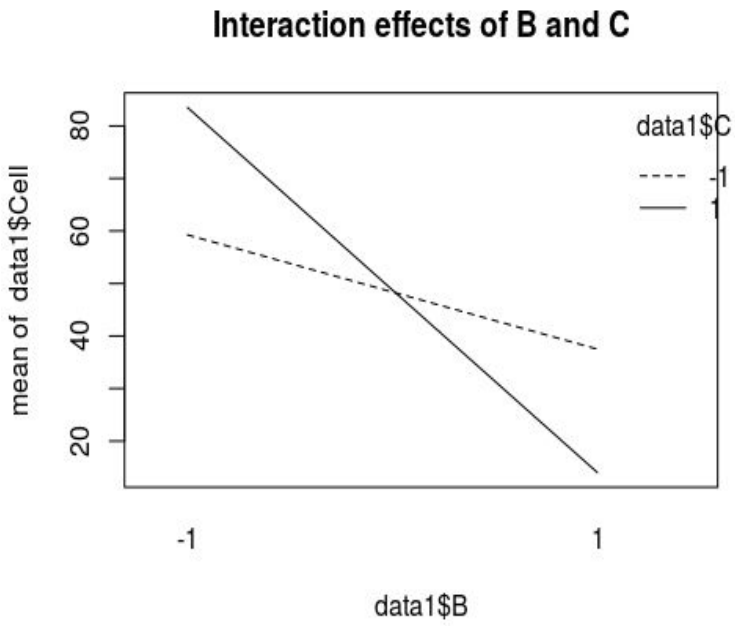



Final comparison of the four models will be given in the conclusion section.

## Conclusion

Comparing the four obtained model as follows, we can conclude that the first model, Cell~B+BC, with the lowest residual error of 8.683 and highest multiple R-squared of 0.92 is the best model. The result is also consistent with another criteria I used called PRESS residuals and model 1 has the smallest press score equal to 1608.4 which makes it the most desirable model.

Furthermore, the interaction plot of B and C using design 1 indicates that in order to maximize the amount of cells, as desired, both levels of B and C should be associated with the sign "-". It may also be noted that approaching the "+" level of both variables causes for the strength of the effects of the two treatments two reverse, which is worthwhile to take into consideration.




Model 1:

Residual standard error: 8.683 on 12 degrees of freedom

Multiple R-squared: 0.9214, Adjusted R-squared: 0.9018

F-statistic: 46.9 on 3 and 12 DF, p-value: 6.674e-07

Model 2:

Residual standard error: 14.49 on 29 degrees of freedom

Multiple R-squared: 0.6917, Adjusted R-squared: 0.6704

F-statistic: 32.53 on 2 and 29 DF, p-value: 3.892e-08

Model 3:

Residual standard error: 9.75 on 46 degrees of freedom

Multiple R-squared: 0.9107, Adjusted R-squared: 0.8778

F-statistic: 27.61 on 17 and 46 DF, p-value: < 2.2e-16

Model 4:

Residual standard error: 11.44 on 301 degrees of freedom

Multiple R-squared: 0.8199, Adjusted R-squared: 0.8091

F-statistic: 76.13 on 18 and 301 DF, p-value: < 2.2e-16