Design and Analysis of Experiments

Course notes for STAT 568

Adam B Kashlak
Mathematical & Statistical Sciences
University of Alberta
Edmonton, Canada, T6G 2G1

March 27, 2019

Contents

Preface **1**

1 One-Way ANOVA **2**
1.0.1 Terminology . . . 2
1.1 Analysis of Variance . . . 3
1.1.1 Sample size computation . . . 6
1.1.2 Contrasts . . . 7
1.2 Multiple Comparisons . . . 8
1.3 Random Effects . . . 9
1.3.1 Derivation of the F statistic . . . 11
1.4 Cochran's Theorem . . . 12

2 Multiple Factors **15**
2.1 Randomized Block Design . . . 15
2.1.1 Paired vs Unpaired Data . . . 17
2.1.2 Tukey's One DoF Test . . . 18
2.2 Two-Way Layout . . . 20
2.2.1 Fixed Effects . . . 20
2.3 Latin Squares . . . 21
2.3.1 Graeco-Latin Squares . . . 23
2.4 Balanced Incomplete Block Designs . . . 24
2.5 Split-Plot Designs . . . 26
2.6 Analysis of Covariance . . . 30

3 Multiple Testing **33**
3.1 Family-wise Error Rate . . . 34
3.1.1 Bonferroni's Method . . . 34
3.1.2 Sidak's Method . . . 34
3.1.3 Holms' Method . . . 35
3.1.4 Stepwise Methods . . . 35
3.2 False Discovery Rate . . . 36
3.2.1 Benjamini-Hochberg Method . . . 37

4 Factorial Design **39**
4.1 Full Factorial Design . . . 40
4.1.1 Estimating effects with regression . . . 41
4.1.2 Lenth's Method . . . 43
4.1.3 Key Concepts . . . 44
4.1.4 Dispersion and Variance Homogeneity . . . 45
4.1.5 Blocking with Factorial Design . . . 46
4.2 Fractional Factorial Design . . . 48
4.2.1 How to choose a design . . . 49
4.3 3^k Factorial Designs . . . 52
4.3.1 Linear and Quadratic Contrasts . . . 54
4.3.2 3^{k-q} Fractional Designs . . . 55
4.3.3 Agricultural Example . . . 58

5 Response Surface Methodology **63**
5.1 First and Second Order . . . 64
5.2 Some Response Surface Designs . . . 65
5.2.1 Central Composite Design . . . 65
5.2.2 Box-Behnken Design . . . 67
5.2.3 Uniform Shell Design . . . 68
5.3 Search and Optimization . . . 69
5.3.1 Ascent via First Order Designs . . . 69
5.4 Chemical Reaction Data Example . . . 70

6 Nonregular, Nonnormal, and other Designs **74**
6.1 Prime Level Factorial Designs . . . 74
6.1.1 5 level designs . . . 75
6.1.2 7 level designs . . . 78
6.1.3 Example of a 25-run design . . . 78
6.2 Mixed Level Designs . . . 79
6.2.1 2^n4^m Designs . . . 81
6.2.2 36-Run Designs . . . 84
6.3 Nonregular Designs . . . 86
6.3.1 Plackett-Burman Designs . . . 86
6.3.2 Aliasing and Correlation . . . 89
6.3.3 Simulation Example . . . 91
6.A Paley's Contruction of H_N . . . 94

Preface

> It's the random factors. You can't be sure, ever. All of time and space for them to happen in. How can we know where to look or turn? Do we have to fight our way through every possible happening to get the thing we want?
>
> ---
>
> *Time and Again*
> Clifford D Simak (1951)

The following are lecture notes originally produced for a graduate course on experimental design at the University of Alberta in the winter of 2018. The goal of these notes is to cover the classical theory of design as born from some of the founding fathers of statistics. The proper approach is to begin with an hypothesis to test, design an experiment to test that hypothesis, collected data as needed by the design, and run the test. These days, data is collected en masse and often subjected to many tests in an exploratory search. Though, understanding how to design experiments is still critical for being able to determine which factors effect the observations.

These notes were produced by consolidating two sources. One is the text of Wu and Hamada, *Experiments: Planning, Analysis, and Optimization*. The second is lecture notes and lecture slides from Dr. Doug Wiens and Dr. Linglong Kong, respectively.

Adam B Kashlak
Edmonton, Canada
January 2018

Additional notes on multiple testing were included based on the text *Large-Scale Inference* by Bradley Efron, which is quite relevant to the many hypothesis tests considered in factorial designs.

ABK, Jan 2019

Chapter 1

One-Way ANOVA

Introduction

We begin by considering an experiment in which k groups are compared. The primary goal is to determine whether or not there is a significant difference among all of the groups. The secondary goal is then to determine which specific pairs of groups differ the most.

One example would be sampling n residents from each of the $k = 10$ Canadian provinces and comparing their heights perhaps to test whether or not stature is effected by province. Here, the province is the single *factor* in the experiment with 10 different *factor levels*.

Another example would be contrasting the heights of $k = 3$ groups of flowers where group one is given just water, group two is given water and nutrients, and group three is given water and vinegar. In this case, the *factor* is the liquid given to the flowers. It is also often referred to as a *treatment*. When more than one factor is considered, the treatment refers to specific levels of all factors.

1.0.1 Terminology

In the design of experiments literature, there is much terminology to consider. The following is a list of some of the common terms:

- Size or level of a test: the probability of a false positive. That is, the probability of falsely rejecting the null hypothesis.
- Power of a test: the probability of a true positive. That is, the probability of correctly rejecting the null hypothesis.
- Response: the dependent variable or output of the model. It is what we are interested in modelling.

- Factor: an explanatory variable or an input into the model. Often controlled by the experimenter.
- Factor Level: the different values that a factor can take. Often this is categorical.
- Treatment: the overall combination of many factors and levels.
- Blocking: grouping subjects by type in order to understand the variation between the blocks versus the variation within the blocks.
- Fixed effects: When a factor is chosen by the experimenter, it is considered fixed.
- Random effects: When a factor is not controlled by the experimenter, it is considered random.

One example comes from the `Rabbit` dataset from the `MASS` library in `R`. Here, five rabbits are given drugs (MDL) and placebos in different dosages and the effect of their blood pressure is recorded. The blood pressure would be the *response*. The *factors* are drug, dosage, and rabbit where the *factor levels* for drug are {MDL, placebo}, the levels for dosage are {6.25, 12.5, 25, 50, 100, 200}, and the levels for rabbit are {R1,...,R5}. A specific *treatment* could be rabbit R3 with a dosage of 25 of placebo.

1.1 Analysis of Variance[1]

In statistics in general, analysis of variance or ANOVA is concerned with decomposing the total variation of the data by the factors. That is, it determines how much of the variation can be explained by each factor and how much is left to random noise.

We begin with the setting of one-way fixed effects. Consider a sample of size $N = nk$ and k different treatments. Thus, we have k different groups of size n. Each group is given a different treatment, and measurements y_{ij} for $i = 1, \ldots, k$ and $j = 1, \ldots, n$ are collected. The one-way ANOVA is concerned with comparing the between group variation to the within group variation, which is the variation explained by the treatments vs the unexplained variation.

Remark 1.1.1 (Randomization)**.** *In the fixed effects setting in practise, the N subjects are randomly assigned to one of the k treatment groups. However, this is not always possible for a given experiment.*

[1] See Wu & Hamada Section 2.1

The model for the observations y_{ij} is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where μ is the global mean and τ_i is the effect of the ith category or treatment. The ε_{ij} are random noise variables generally assumed to be iid $\mathcal{N}\left(0, \sigma^2\right)$ with σ^2 unknown. Based on this model, we can rewrite the observation y_{ij} as

$$y_{ij} = \hat{\mu} + \hat{\tau}_i + r_{ij} \tag{1.1.1}$$

where

$$\hat{\mu} = \bar{y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}, \quad \hat{\tau}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} - \frac{1}{N}\sum_{l=1}^{k}\sum_{j=1}^{n} y_{lj}, \quad r_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

Equation 1.1.1 can be rearranged into

$$y_{ij} - \bar{y}_{\cdot\cdot} = (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot}).$$

This, in turn, can be squared and summed to get

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2,$$

which is just the *total sum of squares*, SS_{tot}, decomposed into the sum of the *treatment sum of squares*, SS_{tr}, and the *error sum of squares*, SS_{err}.

Under the assumption that the errors ε_{ij} are normally distributed, the usual F statistic can be derived to test the hypothesis that

$$H_0 : \tau_1 = \ldots = \tau_k \quad \text{vs} \quad H_1 : \exists i_1, i_2 \text{ s.t. } \tau_{i_1} \neq \tau_{i_2}.$$

Indeed, under this model, it can be shown that $SS_{\text{err}} \sim \chi^2\left(N - k\right)$ and that under the null hypothesis $SS_{\text{tr}} \sim \chi^2\left(k - 1\right)$. Hence, the test statistic is

$$F = \frac{SS_{\text{tr}}/(k-1)}{SS_{\text{res}}/(N-k)} \sim F\left(k - 1, N - k\right).$$

Often, for example in `R`, all of these terms from the one-way ANOVA experiment are represented in a table as follows:

	DoF	Sum Squares	Mean Squares	F value	p-value
Treatment	$k-1$	SS_{tr}	$SS_{\text{tr}}/(k-1)$	F	$\text{P}\left(> F\right)$
Residuals	$N-k$	SS_{err}	$SS_{\text{err}}/(N-k)$		

Remark 1.1.2 (Degrees of Freedom)**.** *In an intuitive sense, the degrees of freedom (DoF) can be thought of as the difference in the sample size and the number of estimated parameters. The DoF for the total sum of squares is just $N-1$ where N is the total sample size and -1 is for estimating $\bar{y}_{\cdot\cdot}$. Similarly, for SS_{tr}, the DoF is $k-1$ corresponding to the k group means minus the one global mean. The DoF for the remainder is $N-k$, which is the sample size minus the number of group means.*

This model falls under the general linear model setting

$$Y = X\beta + \varepsilon$$

where $\beta^{\mathrm{T}} = (\mu, \tau_1, \ldots, \tau_k)$ is the vector of parameters, Y and ε are the N-long vectors of y_{ij} and ε_{ij}, respectively, and X is the design matrix. In the context of the above model, the $N \times (k+1)$ matrix X is of the form

$$X = \begin{pmatrix} 1 & 1 & 0 & \cdots & \\ 1 & 0 & 1 & 0 & \cdots \\ \vdots & & & \ddots & \\ 1 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & \\ \vdots & & & & \\ \vdots & & & & \\ 1 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Using the usual least squares estimator from linear regression, we could try to estimate the parameter vector β as

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y.$$

The problem is that $X^{\mathrm{T}}X$ is not an invertible matrix. Hence, we have to add a constraint to the parameters to make this viable.

Remark 1.1.3 (Identifiability)**.** *The easiest way to see the problem with the model as written is that you could have a global mean $\mu = 0$ and have group means τ_i or you could have a global mean $\mu = 1$ and have group means $\tau_i - 1$, which would give identical models. That is, the parameters are not identifiable.*

The common constraint to apply is to require that $\sum_{i=1}^{k} \tau_i = 0$ as this is satisfied by the estimators $\hat{\tau}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$. Here, the interpretation of the parameters is as before where μ is the global mean and τ_i is the offset of the ith group. However, now the parameter $\tau_k = -\sum_{i=1}^{k-1} \tau_i$. As a result, the new design matrix is now of dimension

$N \times k$ and is of the form

$$X = \begin{pmatrix} 1 & 1 & 0 & \cdots & \\ 1 & 0 & 1 & 0 & \cdots \\ \vdots & & & \ddots & \\ 1 & 0 & \cdots & 0 & 1 \\ 1 & -1 & \cdots & \cdots & -1 \\ 1 & 1 & 0 & \cdots & \\ \vdots & & & & \\ \vdots & & & & \\ 1 & 0 & \cdots & 0 & 1 \\ 1 & -1 & \cdots & \cdots & -1 \end{pmatrix}.$$

This new X allows for $X^{\mathrm{T}}X$ to be invertible. Furthermore, the hypothesis test is now stated slightly differently as

$$H_0 : \tau_1 = \ldots = \tau_{k-1} = 0 \quad \text{vs} \quad H_1 : \exists i \text{ s.t. } \tau_i \neq 0.$$

Note that the above equations all hold even if the k groups for unequal sample sizes. In that case, $N = \sum_{i=1}^{k} n_i$.

Example

Imagine we have $k = 4$ categories with $n = 10$ samples each. The categories will be labelled A, B, C, and D. The category means are, respectively, -1, -0.1, 0.1, and 1, and the added noise is $\varepsilon \sim \mathcal{N}(0, 1)$. A model can be fit to the data via the `aov()` function in `R`. The result is a table such as

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
label	3	19.98	6.662	8.026	0.000319	***
Residuals	36	29.88	0.830			

The significant p-value for the F test indicates that we can (correctly!) reject the null hypothesis that the category means are all equal.

1.1.1 Sample size computation

In practise, researchers are often concerned with determining the minimal sample size to achieve a certain reasonable amount of statistical power. To compute the sample size exactly is usually impossible as it is based on unknowns. However, if guesses are available from similar past or pilot studies, then a sample size estimate can be computed.

Such a computation can be performed by the `R` function `power.anova.test()`. As an example, if the number of groups is $k = 6$, the between group variation is

2, the within group variation is 4, the size of the test is $\alpha = 0.05$, and the desired power is 0.9, then the sample size for each of the 6 groups is $n = 7.57$, which will round up to 8.

```
power.anova.test(groups = 6, between.var = 2,
                                          within.var = 4, power=0.9)
```

1.1.2 Contrasts[2]

When an ANOVA is run in `R`, it defaults to taking the first alphabetical factor level to use as the *intercept* or reference term. Instead, we can use the `contrast()` command to tell `R` to apply the above sum-to-zero constraint instead. Many more complicated contrasts can be used for a variety of testing purposes.

Continuing from the above example with $n = 10$ and $k = 4$, we can use the `summary.lm()` function to look at the parameter estimates for the four categories.

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
(Intercept)	-0.8678	0.2881	-3.012	0.00472	**
labB	1.0166	0.4074	2.495	0.01731	*
labC	0.8341	0.4074	2.047	0.04799	*
labD	1.9885	0.4074	4.881	2.16e-05	***

In this table, the Intercept corresponds to the mean of category A. Meanwhile, the labB, labC, and labD estimates correspond to the difference between the mean of category A and the means of categories B, C, and D, respectively.

We can use `contr.sum` to construct a sum-to-zero contrast. The result of refitting the model is

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
(Intercept)	0.09203	0.14405	0.639	0.52697	
lab1	-0.95982	0.24950	-3.847	0.00047	***
lab2	0.05682	0.24950	0.228	0.82115	
lab3	-0.12570	0.24950	-0.504	0.61747	

Now, the Intercept estimate corresponds to the global mean $\bar{y}_{\cdot\cdot}$, which is generally preferable. Furthermore, the estimates for lab1,2,3 correspond to the difference between the category means for A,B,C and the global mean–e.g. check that $0.092 - 0.959 = -0.867$ for category A. The t-tests in the table are now testing for whether or not the category mean is equal to the global mean.

[2] See Wu & Hamada Section 2.3

1.2 Multiple Comparisons[3]

Imagine that we have run the above hypothesis test and have rejected the null hypothesis. A natural follow up question is *which specific τ_i is non-zero?* This is equivalent to asking which pairs of τ_i, τ_j have significant differences.

To compare the means of two different groups of size n_i and n_j, we can use the t-statistic

$$t_{ij} = \frac{\bar{y}_{i\cdot} - \bar{y}_{j\cdot}}{\sqrt{(n_i^{-1} + n_j^{-1})SS_{\text{err}}/(N-k)}} \sim t\,(N-K)$$

and the usual two sample t-test. Thus, we can reject the hypothesis that the group means are equal at the α level if $|t_{ij}| > t_{N-k,\alpha/2}$.

However, if we were to run such a test for each of the $\tilde{k} = k(k-1)/2$ pairings, then the probability of a false positive would no longer be α but much larger. One approach to correcting this problem is the *Bonferroni* method.

The Bonferroni method simply states that one should run all of the tests at a new level $\alpha' = \alpha/\tilde{k}$. As a result, the probability of at least one false positive is

$$\begin{aligned}
\mathrm{P}\left(\exists i,j \ s.t.\ |t_{ij}| > t_{N-k,\alpha'} \,\middle|\, H_0\right) &= \mathrm{P}\left(\bigcup_{i,j}\{|t_{ij}| > t_{N-k,\alpha'}\} \,\middle|\, H_0\right) \\
&\leq \sum_{i,j} \mathrm{P}\left(|t_{ij}| > t_{N-k,\alpha'} \,\middle|\, H_0\right) \\
&= \tilde{k}\frac{\alpha}{\tilde{k}} = \alpha.
\end{aligned}$$

This method is quite versatile. However, it is also quite conservative in practise. Often, it is recommended to instead use the *Tukey* method.

In Tukey's method, the same test statistic, t_{ij}, is used. However, instead of comparing it to a t distribution, it is instead compared to a distribution known as the Studentized Range distribution:[4] reject the hypothesis that groups i and j do not differ if $|t_{ij}| \geq \frac{1}{\sqrt{2}} q_{k,N-k,\alpha}$.

The value $q_{k,N-k,\alpha}$ comes from first assuming that $n_1 = \ldots = n_k = n$, and noting that, for some constant c,

$$\mathrm{P}\left(\exists i,j \ s.t.\ |t_{ij}| > c \,|\, H_0\right) = \mathrm{P}\left(\max_{i,j}\{|t_{ij}|\} > c \,\middle|\, H_0\right).$$

The distribution of the $\max_{i,j}\{|t_{ij}|\}$ can be shown to be related to the Studentized Range distribution and c is set to be $q_{k,N-k,\alpha}/\sqrt{2}$ where

$$\mathrm{P}\left(\frac{\sqrt{n}(\max_{i=1,\ldots,k}\bar{y}_{i\cdot} - \min_{i=1,\ldots,k}\bar{y}_{i\cdot})}{\sqrt{SS_{\text{err}}/(N-k)}} > q_{k,N-k,\alpha} \,\middle|\, H_0\right) = \alpha.$$

[3] See Wu & Hamada Section 2.2

[4] https://en.wikipedia.org/wiki/Studentized_range

When the categories are balanced–i.e. when $n_1 = \ldots = n_k = n$–then the error rate for Tukey's method is exactly α. However, it can still be applied when the categories are not balanced. In general, Tukey's method will result in tighter confidence intervals than Bonferroni's method. However, it may not be applicable in the more complicated settings to come. Tukey confidence intervals can be computed in `R` via the function `TukeyHSD()`.

Example

Continuing again with the above example, if `TukeyHSD` is used to construct simultaneous confidence intervals for the differences of the means of the four categories, the result is:

	diff	lwr	upr	p adj
B-A	1.0166422	-0.08068497	2.1139694	0.0777351
C-A	0.8341240	-0.26320318	1.9314511	0.1900675
D-A	1.9885274	0.89120026	3.0858546	0.0001230
C-B	-0.1825182	-1.27984537	0.9148089	0.9695974
D-B	0.9718852	-0.12544193	2.0692124	0.0981594
D-C	1.1544034	0.05707629	2.2517306	0.0360519

Note that the most significant p-value comes from the difference between category A and D, which is reasonable as those groups have means of -1 and 1, respectively.

1.3 Random Effects[5]

The difference between fixed and random effects can be confusing. The models are the same, but the interpretation is different. For example, returning to the `Rabbit` data. The dosage of the blood pressure drug is a fixed effect; it is chosen by the experimenter, and we are interested in the difference in effect between two different dosages. Contrasting, the rabbit itself is a random effect; it is selected at random from the population, and we do not care about the difference between two specific animals, but instead about the overall variation in the population.

For a random effects model, we begin as before with

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

with μ fixed and ε_{ij} iid $\mathcal{N}\left(0, \sigma^2\right)$ random variables. However, now the τ_i are treated as iid random variables with distribution $\mathcal{N}\left(0, \nu^2\right)$. The τ_i and ε_{ij} are assumed to be independent of each other.

In this case, we are not interested in estimating τ_i, but instead with calculating ν^2, which is the between-group variance. Hence, we want to test the hypothesis

$$H_0 : \nu^2 = 0 \quad \text{vs} \quad H_1 : \nu^2 > 0,$$

[5] See Wu & Hamada Section 2.5

which, in words, is whether or not there is any variation among the categories. The test statistic is identical to the fixed effects setting:

$$F = \frac{SS_{\text{tr}}/(k-1)}{SS_{\text{res}}/(N-k)} \sim F\left(k-1, N-k\right).$$

As we do not care about the individual category means in this setting, we also do not care about comparison tests like Tukey's method. Instead, we are interested in estimating the value of ν^2. As in the fixed effects setting,

$$SS_{\text{err}}/\sigma^2 \sim \chi^2\left(N-k\right).$$

Hence, $SS_{\text{err}}/(N-k)$ is an unbiased estimator of σ^2. Next, we will compute $\mathrm{E}\left(SS_{\text{tr}}/(k-1)\right)$ assuming $n_1 = \ldots = n_k = n$ for simplicity.[6]

$$\begin{aligned}
\mathrm{E}\left(SS_{\text{tr}}\right) &= \mathrm{E}\left(\sum_{i=1}^{k} n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2\right) \\
&= n\sum_{i=1}^{k} \mathrm{E}\left[(\tau_i - \bar{\tau})^2 + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2 + 2(\tau_i - \bar{\tau})(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})\right] \\
&= nk\left[(1-k^{-1})\nu^2 + n^{-1}(1-k^{-1})\sigma^2 + 0\right] \\
&= (k-1)n\nu^2 + (k-1)\sigma^2.
\end{aligned}$$

Hence, $\mathrm{E}SS_{\text{tr}}/(k-1) = n\nu^2 + \sigma^2$. Furthermore, we have the unbiased estimator

$$\hat{\nu}^2 = n^{-1}\left[\frac{SS_{\text{tr}}}{k-1} - \frac{SS_{\text{err}}}{N-k}\right].$$

Lastly, we can use this estimator to construct a confidence interval for the global mean μ about the estimator $\hat{\mu} = \bar{y}_{\cdot\cdot\cdot}$. That is,

$$\begin{aligned}
\mathrm{Var}\left(\hat{\mu}\right) = \mathrm{Var}\left(\mu + \bar{\tau} + \bar{\varepsilon}\right) &= 0 + k^{-1}\nu^2 + (nk)^{-1}\sigma^2 \\
&= (nk)^{-1}(n\nu^2 + \sigma^2) = (nk)^{-1}\mathrm{E}SS_{\text{tr}}/(k-1) \approx \frac{SS_{\text{tr}}}{nk(k-1)}.
\end{aligned}$$

Hence, we get the following $1-\alpha$ confidence interval:

$$|\mu - \hat{\mu}| \leq t_{k-1,\alpha/2}\sqrt{SS_{\text{tr}}/(nk(k-1))}.$$

[6] See Equation 2.45 in Wu & Hamada for the necessary correction when the n_i do not all coincide.

Example

Once more continuing the with example, we could choose to consider the labels A,B,C,D as random effects instead of fixed effects. Using the `lmer()` function from the `lme4` results in

Random effects:

Groups	Name	Variance	Std.Dev.
label	(Intercept)	0.5832	0.7637
Residual		0.8300	0.9111

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	0.09203	0.40810	0.226

One can check by hand that the variance of the label factor is $\hat{\nu}^2$ and that the standard error for the intercept term is $SS_{\text{tr}}/(nk(k-1))$.

1.3.1 Derivation of the F statistic

In both the fixed and random effects settings, we arrive at the same F statistic to test the hypothesis. Namely,

$$F = \frac{SS_{\text{tr}}/(k-1)}{SS_{\text{res}}/(N-k)} \sim F\left(k-1, N-k\right).$$

Why is this necessarily the case?

Fixed Effects

In the fixed effects setting, we are interested in testing the null hypothesis that the τ_i are all equal vs the alternative that at least one pair of τ_i and τ_j are not equal. We have, in general, that $SS_{\text{err}}/\sigma^2 \sim \chi^2\left(N-k\right)$. Under the null hypothesis, we have that $SS_{\text{tr}}/\sigma^2 \sim \chi^2\left(k-1\right)$ and that these two random variables are independent. This can be solved for directly or one can use Cochran's theorem. Hence, the above F does indeed follow an F distribution under the null hypothesis.

Under the alternative hypothesis, we can show that $SS_{\text{tr}}/\sigma^2 \sim \chi^2\left(k-1,\theta\right)$ where $\chi^2\left(k-1,\theta\right)$ is a *non-central chi squared distribution*[7] with non-centrality parameter $\theta > 0$. Whereas the mean of a $\chi^2\left(k-1\right)$ random variable is $k-1$, the mean of a $\chi^2\left(k-1,\theta\right)$ random variable is $k-1+\theta$. Hence, it is shifted by θ.

Consequently, under the alternative hypothesis, the statistic F has a *non-central F distribution*[8] with the same non-centrality parameter θ. While, the standard F distribution with DoFs (d_1, d_2) has mean $d_2/(d_2-2)$, which is approximately 1

[7] https://en.wikipedia.org/wiki/Noncentral_chi-squared_distribution
[8] https://en.wikipedia.org/wiki/Noncentral_F-distribution

for large sample sizes, the mean of a non-central F distribution with parameters (d_1, d_2, θ) is

$$\left(\frac{d_2}{d_2 - 2}\right)\left(\frac{d_1 + \theta}{d_1}\right).$$

Hence, as the the non-centrality parameter θ grows–i.e. as we move farther from the null hypothesis setting–the mean of the non-central F increases providing statistical power to reject the null.

Remark 1.3.1. *It seems reasonable that one could try to use the Neyman-Pearson lemma to show that this is a most powerful test. However, I have not tried that at this time.*

Random Effects

The random effects case follows the fixed effects case almost identically. The main difference is in the beginning. We desire to test test the null hypothesis that $\nu^2 = 0$ against that alternative that it is greater than zero where ν^2 is the variance of the τ_i.

Under the null hypothesis, the fact that $\nu^2 = 0$ implies that all of the τ_i are equal to zero as they have the degenerate distribution $\mathcal{N}(0, 0)$. Thus, we again arrive at an F distribution. Under the alternative hypothesis, $\nu^2 > 0$, which allows for the τ_i to differ, we can follow the same argument as above to arrive at a non-central F distribution. Thus, we can use the same F statistic in both the fixed and random effects settings.

1.4 Cochran's Theorem[9]

Cochran's theorem allows us to decompose sums of squared normal random variables, as often occurs in statistics, to get independent chi-squared random variables. This ultimately allows for the ubiquitously used F-tests. The theorem is as follows.

Theorem 1.4.1 (Cochran's Theorem, 1934[10])**.** *Let $Z_1, \ldots, Z_m$ be independent and identically distributed $\mathcal{N}(0, \sigma^2)$ random variables and $Z \in \mathbb{R}^m$ be the vector of Z_i. For $k = 1, \ldots, s$, let A_k be an $m \times m$ symmetric matrix with ijth entry a_{ij}^k. Let Q_k be the following quadratic form*

$$Q_k = Z^{\mathrm{T}} A_k Z = \sum_{i,j=1}^{m} a_{ij}^k Z_i Z_j.$$

Lastly, let

$$\sum_{i=1}^{m} Z_i^2 = \sum_{k=1}^{s} Q_k.$$

[9] See Wu & Hamada Section 2.4

[10] https://doi.org/10.1017/S0305004100016595

Denoting the rank of A_k by r_k, Consequently, $\sum_{k=1}^{s} r_k = m$ if and only if the Q_k are independent random variables with $Q_k/\sigma^2 \sim \chi^2(r_k)$.

To prove this theorem, we begin with some lemmas. The first lemma concerns the distribution of a single quadratic form.

Lemma 1.4.2. *Let $Z_1, \ldots, Z_m$ be m iid $\mathcal{N}(0,1)$ random variables. Let $A \in \mathbb{R}^{m\times m}$ be a symmetric matrix–i.e. $A = A^{\mathrm{T}}$. Define $Q = Z^{\mathrm{T}}AZ = \sum_{i,j=1}^{m} a_{i,j} Z_i Z_j$. Then,*

$$Q = \sum_{i=1}^{m} \lambda_i W_i^2$$

where the λ_i are the eigenvalues of A and the W_i are iid $\mathcal{N}(0,1)$ random variables.

Proof. By the spectral theorem for symmetric matrices, we can write $A = U^{\mathrm{T}}DU$ where D is the diagonal matrix of real eigenvalues $\lambda_1, \ldots, \lambda_m$ and U is the orthonormal matrix of eigenvectors–i.e. $UU^{\mathrm{T}} = U^{\mathrm{T}}U = 1$. As the columns of U form an orthonormal basis for $\mathbb{R}^m$, we have that $W = UZ \sim \mathcal{N}(0, I_m)$. Hence,

$$Q = Z^{\mathrm{T}}AZ = Z^{\mathrm{T}}U^{\mathrm{T}}DUZ = (UZ)^{\mathrm{T}}D(UZ) = W^{\mathrm{T}}DW = \sum_{i=1}^{m} \lambda_i W_i^2.$$

□

Remark 1.4.3. *Note that if $W_i \sim \mathcal{N}(0,1)$, then $W_i^2 \sim \chi^2(1)$. Hence, Q from Lemma 1.4.2 is a weighted sum of chi-squared random variables.*

Lemma 1.4.4. *Given the set up of Lemma 1.4.2, if $\lambda_1 = \ldots = \lambda_r = 1$ and the $\lambda_{r+1} = \lambda_m = 0$ from some $0 < r < m$, then*

$$Q \sim \chi^2(r) \quad \text{and} \quad Q' = Z^{\mathrm{T}}(I - A)Z \sim \chi^2(m - r),$$

and Q and Q' are independent.

Proof. First, if $\lambda_1 = \ldots = \lambda_r = 1$ with the other eigenvalues being zero, then by Lemma 1.4.2

$$Q = \sum_{i=1}^{m} \lambda_i W_i^2 = \sum_{i=1}^{r} W_i^2 \sim \chi^2(r)$$

as it is the sum of r independent $\chi^2(1)$ random variables.

Secondly, $I - A$ is also diagonalized by U as

$$I - A = U^{\mathrm{T}}U - U^{\mathrm{T}}DU = U^{\mathrm{T}}(I - D)U.$$

The diagonal entries of $I - D$ are r zeros and $m - r$ ones. Hence, as before,

$$Q' = \sum_{i=r+1}^{m} W_i^2 \sim \chi^2(m - r).$$

Lastly, Q is a function of $W_1, \ldots, W_r$ and Q' is a function of $W_{r+1}, \ldots, W_m$. We can write $Q = f(W_1, \ldots, W_r)$ and $Q' = g(W_{r+1}, \ldots, W_m)$ for some functions $f : \mathbb{R}^r \to \mathbb{R}$ and $g : \mathbb{R}^{m-r} \to \mathbb{R}$. As the W_i are independent random variables, Q and Q' are functions on independent random variables and, hence, also independent. □

Proof (Cochran's Theorem). Without loss of generality, we set $\sigma^2 = 1$. Given the set up of Theorem 1.4.1, assume first that the Q_k are independent random variables with $Q_k \sim \chi^2(r_k)$. Then, we have that

$$\sum_{i=1}^{m} Z_i^2 \sim \chi^2(m) \quad \text{and} \quad \sum_{k=1}^{s} Q_k \sim \chi^2\left(\textstyle\sum_{k=1}^{s} r_k\right)$$

as the degrees of freedom add with independent chi-squared random variables are summed. Hence, $\sum_{k=1}^{s} r_k = m$.

Next, assume that $\sum_{k=1}^{s} r_k = m$. We begin with considering $Q_1 = Z^{\mathrm{T}} A_1 Z$. Denote $A_{-1} = \sum_{k=2}^{s} A_k$. From Lemma 1.4.4, we have that if A_1 is an $m \times m$ symmetric matrix then A_1 and $A_{-1} = I_m - A_1$ are simultaneously diagonalizable. That is,

$$I_m = A + A_{-1} = U^{\mathrm{T}} D_1 U + U^{\mathrm{T}} D_{-1} U = U^{\mathrm{T}}(D_1 + D_{-1})U.$$

Furthermore, $D_1 + D_{-1} = I_m$ where $\operatorname{rank}(D_1) = r_1$ and $\operatorname{rank}(D_{-1}) = m - r_1$. This implies that only r_1 and $m - r_1$ of the diagonal entries of D_1 and D_{-1} are non-zero, respectively. Hence, without loss of generality, we can reorder the basis vectors in U such that

$$D_1 = \begin{pmatrix} I_{r_1} & 0 \\ 0 & 0_{m-r_1} \end{pmatrix} \quad \text{and} \quad D_{-1} = \begin{pmatrix} 0_{r_1} & 0 \\ 0 & I_{m-r_1} \end{pmatrix}.$$

From Lemma 1.4.4, we have that Q_1 and $Q_{-1} = \sum_{k=2}^{s} Q_k$ are independent random variables with distributions $\chi^2(r_1)$ and $\chi^2(m - r_1)$, respectively.

Now, looking at $Q_2, \ldots, Q_s$, we have that

$$U(A_2 + \ldots + A_s)U^{\mathrm{T}} = \begin{pmatrix} 0_{r_1} & 0 \\ 0 & I_{m-r_1} \end{pmatrix}.$$

Thus, we can simultaneously diagonalize A_2 and $A_{-2} = \sum_{k=3}^{s} A_k$ and proceed as above to get that

$$D_2 = \begin{pmatrix} 0_{r_1} & 0 & 0 \\ 0 & I_{r_2} & 0 \\ 0 & 0 & 0_{m-r_1-r_2} \end{pmatrix} \quad \text{and} \quad D_{-2} = \begin{pmatrix} 0_{r_1} & 0 & 0 \\ 0 & 0_{r_2} & 0 \\ 0 & 0 & I_{m-r_1-r_2} \end{pmatrix}$$

and that $Q_2 \sim \chi^2(r_2)$ is independent of $Q_{-2} \sim \chi^2(m - r_1 - r_2)$. This process can be iterated to get the desired result. □

Chapter 2

Multiple Factors

Introduction

This chapter continues from the previous but whereas before we only considered a single factor, now we will consider multiple factors. Having more than one factor leads to many new issues to consider, which mainly revolve around proper experimental design to tease out the effects of each factor as well as potential cross effects.

2.1 Randomized Block Design[1]

As mentioned in the terminology section of Chapter 1, *Blocking* is a key concept in this course. The goal is to group similar subjects into blocks with the idea that the within block variance should be much smaller than the between block variance. The purpose of such blocking is to remove or control unwanted variability to allow for better estimation and testing of the actual factor under consideration. Lastly, a blocking is called *complete* if every treatment is considered in every block.

Some examples of blocking are as follows. Blocking by gender in a drug trial could be used to remove the variation between the difference ways the drug effects men vs women. In an agricultural experiment, one could block by fields to take into account soil differences, amount of sunlight, and other factors that we are not explicitly considering.

First, we consider a model with two factors: a treatment factor and a blocking factor. Let b be the number of blocks and k be the number of treatments. Then, we write

$$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}$$

where $y_{ij} \in \mathbb{R}$ is the measurement for block i and treatment j, μ is the global mean, β_i is the effect of block i, τ_j is the effect of treatment j, and ε_{ij} is the random noise, which is again assumed to be iid $\mathcal{N}\left(0, \sigma^2\right)$. Note that the sample size here is

[1] See Wu & Hamada Section 3.2

$N = bk$, which is the complete randomized block design. The experiment could be replicated n times to achieve a new sample of size $N = nbk$ if desired.

Remark 2.1.1 (Constraints / Contrasts)**.** *As before with the one-way ANOVA, we will require some constraint on the β_i and the τ_j to allow for the model to be well-posed. Unless otherwise stated, we will use the sum-to-zero constraint for both terms. That is,*

$$\sum_{i=1}^{b} \beta_i = 0 = \sum_{j=1}^{k} \tau_j.$$

Decomposing each y_{ij} into a sum of block, treatment, and residual effects gives us

$$y_{ij} = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}),$$

which, as before, can be squared and summed to get

$$\begin{aligned} SS_{\text{tot}} &= \sum_{i=1}^{b}\sum_{j=1}(y_{ij} - \bar{y}_{\cdot\cdot}) = \\ &= \sum_{i=1}^{b} k(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \sum_{j=1}^{k} b(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^{b}\sum_{j=1}^{k}(\bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2 = \\ &= SS_{\text{bl}} + SS_{\text{tr}} + SS_{\text{err}}. \end{aligned}$$

As we are interested in determining whether or not the treatment has had any effect on the observed y_{ij}, the hypotheses to test are identical to those from the previous chapter. Namely,

$$H_0 : \tau_1 = \ldots = \tau_k \quad \text{vs} \quad H_1 : \exists j_1, j_2 \text{ s.t. } \tau_{j_1} \neq \tau_{j_2}.$$

By Cochran's Theorem, under the null hypothesis, the terms SS_{bl}, SS_{tr}, and SS_{err} are all independent chi squared random variables with degrees of freedom $b-1$, $k-1$, and $(b-1)(k-1)$, respectively. The test statistic for the above hypothesis test is another F statistic of the form

$$F = \frac{SS_{\text{tr}}/(k-1)}{SS_{\text{err}}/[(b-1)(k-1)]} \sim F\left(k-1, (b-1)(k-1)\right) \quad \text{under } H_0.$$

This can all be summarized as before in a table.

	DoF	Sum Squares	Mean Squares	F value
Block	$b-1$	SS_{bl}	$SS_{\text{bl}}/(b-1)$	F_{bl}
Treatment	$k-1$	SS_{tr}	$SS_{\text{tr}}/(k-1)$	F_{tr}
Residuals	$(b-1)(k-1)$	SS_{err}	$SS_{\text{err}}/[(b-1)(k-1)]$	

Remark 2.1.2 (Multiple Comparisons)**.** *Similarly to the case of one-way ANOVA, a post-hoc Tukey test can be used to construct simultaneous confidence intervals for every paired difference* $\tau_{j_1} - \tau_{j_2}$ *for* $1 \le j_1 < j_2 \le k$*. In Chapter 1, we divided by* $\sqrt{n_i^{-1} + n_j^{-1}}$*. Now, as there are* b *observations for each treatment, we replace that with* $\sqrt{b^{-1} + b^{-1}}$ *and get a t-statistic of the form*

$$t_{j_1 j_2} = \frac{\bar{y}_{\cdot j_1} - \bar{y}_{\cdot j_2}}{\sqrt{SS_{err}/[(b-1)(k-1)]}} \frac{1}{\sqrt{2b^{-1}}}.$$

Example

Consider the setting where $b = 6$ and $k = 4$ with

$$\beta = (-3, -2, -1, 1, 2, 3)/3, \text{ and } \tau = (0, 0, 1, 2)$$

with $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.5625)$. Such data was generated in `R` and run through the `aov()` function to get the following table:

	Df	Sum Sq	Mean Sq	F value	Pr($>F$)	
block	5	14.46	2.893	4.577	0.00981	**
treat	3	15.80	5.266	8.332	0.00169	**
Residuals	15	9.48	0.632			

Two box plots of the data partitioned by blocking factor and by treatment factor is displayed in Figure 2.1.

If we were to ignore the blocking factor, then the treatment will yield a less significant test statistic.

	Df	Sum Sq	Mean Sq	F value	Pr($>F$)	
treat	3	15.80	5.266	4.398	0.0157	*
Residuals	20	23.94	1.197			

2.1.1 Paired vs Unpaired Data[2]

An example of a common randomized block design is a known as a paired comparison design. This occurs when the blocks are of size $k = 2$. This implies that each block contains two subjects, which can be randomly assigned to a treatment or control group.

For example, if you wanted to test for the effect difference between a drug and a placebo, you could consider twins where one sibling is randomly given the drug and the other is given the placebo. Due to a similar genetic makeup of each pair of twins,

[2] See Wu & Hamada Section 3.1

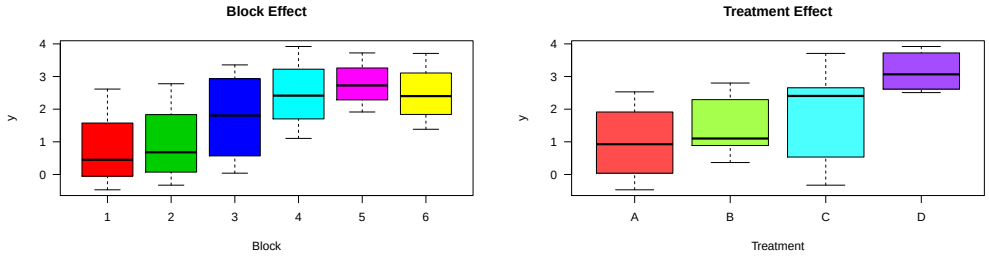

Figure 2.1: Box plots of blocking effects and treatment effects on the observed data.

such a study will have more power in detecting the effect of the drug as opposed to a similar study where the entire set of subjects is partitioned into two groups, control and test, at random. Another example could be human eyes. Instead of randomly assigning b subjects into either a control or a test group, each of the subjects could have a single eye randomly assigned to one group and the other eye assigned to the other group.

From the randomized block design notation, a paired t-test would have the statistics

$$t_{\text{paired}} = \frac{\bar{y}_{\cdot 1} - \bar{y}_{\cdot 2}}{\sqrt{s^2/b}}$$

where $s^2 = (b-1)^{-1} \sum_{i=1}^{b} ((y_{i1} - y_{i2}) - (\bar{y}_{1\cdot} - \bar{y}_{2\cdot}))^2$ is the sample standard deviation of the differences, which has a $t\,(b-1)$ distribution under the null hypothesis.

In comparison, an unpaired test would take on the form of one-way ANOVA from Chapter 1. The test statistic would be

$$t_{\text{unpaired}} = \frac{\bar{y}_{\cdot 1} - \bar{y}_{\cdot 2}}{\sqrt{(s_1^2 + s_2^2)/b}}$$

where s_1^2 and s_2^2 are the sample variance of the y_{i1} and the y_{i2}, respectively. This statistic will have a $t\,(2b-2)$ distribution under the null hypothesis.

If the pairing is effective, then we should have $s^2 < s_1^2 + s_2^2$. This would make the test statistic larger and, hence, more significant. In the unpaired case, the DoFs increases, so the decrease in the variance estimate must be large enough to counteract the increase in the DoFs.

2.1.2 Tukey's One DoF Test[3]

In Tukey's article, he proposes to consider "when the analysis has been conducted in terms where the effects of rows and columns are not additive" where *row* and *column* would refer to the two factors in the model.

[3] *One Degree of Freedom for Non-Additivity*, JW Tukey, Biometrics, Vol. 5, No. 3 (Sep., 1949) or https://en.wikipedia.org/wiki/Tukey%27s_test_of_additivity

Consider the two factor model

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where we only have $N = k_A k_B$ observations. Without replicating this experiment, we do not have enough observations to estimate the interaction effects $(\alpha\beta)_{ij}$. Often, researchers will claim that such terms are negligible due to "expertise" in the field of study. However, Tukey has provided us with a way to test for *non-additivity* interactions between factors A and B.

As noted, we cannot test a general interaction term in a model such as

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$$

without replicating the experiment. However, we can consider a model of the form

$$y_{ij} = \mu + \alpha_i + \beta_j + \lambda\alpha_i\beta_j + \varepsilon_{ij},$$

and running an hypothesis test $H_0 : \lambda = 0$ and $H_1 : \lambda \neq 0$. As we have already estimated the α_i and β_j, we only require a single DoF to estimate and test λ. Doing the usual sum of squares decomposition results in

Term	Formula	DoF
SS_A	$k_B \sum_{i=1}^{k_A} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$k_A - 1$
SS_B	$k_A \sum_{j=1}^{k_B} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$	$k_B - 1$
SS_λ	$\left[\sum_{i=1}^{k_A}\sum_{j=1}^{k_B} y_{ij}\hat{\alpha}_i\hat{\beta}_j\right]^2 \left[\sum_i \hat{\alpha}_i^2 \sum_j \hat{\beta}_j^2\right]^{-1}$	1
SS_{err}	$SS_{\text{tot}} - SS_A - SS_B - SS_\lambda$	$(k_A - 1)(k_B - 1) - 1$

Hence, the F-statistic[4] is

$$\frac{SS_\lambda}{SS_{\text{err}}/[(k_A - 1)(k_B - 1) - 1]} \sim F\left(1, (k_A - 1)(k_B - 1) - 1\right).$$

This test is implemented in the `dae` library with the `tukey.1df()` function.

Tukey's recommendation is

> "The occurrence of a large non-additivity mean square should lead to consideration of a transformation followed by a new analysis of the transformed variable. This consideration should include two steps: (a) inquiry whether the non-additivity was due to analysis in the wrong form or to one or more unusually discrepant values; (b) in case no unusually discrepant values are found or indicated, inquiry into how much of a transformation is needed to restore additivity."

This procedure can be run on the `Rabbit` data with the commands

[4] Note that I have not verified that the SS_λ formula is correct.

(1) `m = aov( BPchange ~ Error(Animal) + Dose:Treatment, data=Rabbit )`

(2) `tukey.1df(md,data=Rabbit)`

which gives an F statistic of 2.39 and a p-value of 0.129. Hence, we do not reject the $\lambda = 0$ assumption.

2.2 Two-Way Layout

In the first chapter, we were concerned with one-way layouts of the form

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}.$$

Now, we will generalize this to two-way layout–with higher order layouts also possible–as

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijl}.$$

Here, we have two factors, say A and B, with k_A and k_B factor levels, respectively. Then, μ is still the global mean, α_i for $i = 1, \ldots, k_A$ is the effect of the ith level of factor A, similarly β_j for $j = 1, \ldots, k_B$ is the effect of the jth level of factor B, and lastly, γ_{ij} is the interaction effect between the ith level of A and the jth level of B. Meanwhile, ε_{ijl} is once again the iid $\mathcal{N}\left(0, \sigma^2\right)$ random noise for $l = 1, \ldots, n$. Once again for simplicity, we will assume all of the treatments–combinations of factor levels–are observed the same number of times, n.

Remark 2.2.1 (Blocks vs Experimental Factors)**.** *Note that two-way layout is different than randomized block design mainly as a blocking factor is not treated as an experiment factor. A block is a way of splitting your data into more homogeneous groups in order to (hopefully) reduce the variance and achieve a more powerful test. An additional experimental factor results in $k_A \times k_B$ different treatments to consider and apply randomly to n subjects each. Furthermore, we are generally concerned with interaction effects between experimental factors, but not concerned with interaction effects concerning a blocking factor.*

2.2.1 Fixed Effects[5]

Besides the added complexity of an additional factor and the interaction terms, we can proceed as in the one-way setting to decompose

$$\sum_{i=1}^{k_A}\sum_{j=1}^{k_B}\sum_{l=1}^{n}(y_{ijl} - \bar{y}_{\cdots})^2 = SS_A + SS_B + SS_{A\times B} + SS_{\text{err}}$$

[5] See Wu & Hamada Section 3.3

where

$$SS_A = nk_B \sum_{i=1}^{k_A} (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2 \qquad SS_B = nk_A \sum_{j=1}^{k_B} (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$$

$$SS_{A\times B} = n \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2 \qquad SS_{\text{err}} = \sum_{i,j,l} (\bar{y}_{ijl} - \bar{y}_{ij\cdot})^2.$$

Under the null hypothesis that the $\alpha_i, \beta_j, \gamma_{ij}$ are all zero, we have that these sums of squares are all chi squared distributed by Cochran's theorem. The degrees of freedom are

$$SS_A \sim \chi^2 (k_A - 1), \qquad SS_B \sim \chi^2 (k_B - 1),$$
$$SS_{A\times B} \sim \chi^2 ((k_A - 1)(k_B - 1)), \qquad SS_{\text{err}} \sim \chi^2 (nk_Ak_b - k_Ak_B)$$

Thus, if we want to test for the the significance of factor A, factor B, or the interaction effects, we can construct an F test based on taking the corresponding sum of squares and dividing by the error sum of squares after normalizing by the DoFs, of course.

Note that if three F tests are run to test for the effects of A, B, and A×B, then we should adjust for multiple testing using Bonferroni's method. That is, declare a test to be significant at the α level if the p-value is smaller than $\alpha/3$.

2.3 Latin Squares[6]

A k dimensional Latin Square is a $k \times k$ array of symbols $a_1, \ldots, a_k$ such that each row and column contains exactly one instance of each of the a_i. An example would be a Sudoku puzzle where $k = 9$.[7] Such a combinatorial object can be used in an experiment with precisely 2 blocking factors and 1 experimental factor when all three of these factors take on precisely k levels. Consequently, there are $N = k^2$ total observations in such an design. The term *Latin* is used as the experimental factor levels are often denoted by the letters A, B, C, etc resulting in a table that may look like

		Block 2			
		1	2	3	4
	1	A	B	C	D
Block	2	B	A	D	C
1	3	C	D	A	B
	4	D	C	B	A

[6] See Wu & Hamada Section 3.6

[7] https://en.wikipedia.org/wiki/Latin_square

Note that the resulting matrix does not necessarily have to be symmetric.

In this setting, we have a model of the form

$$y_{ijl} = \mu + \alpha_i + \beta_j + \tau_l + \varepsilon_{ijl}$$

where α_i is the effect of the ith row, β_j is the effect of the jth column, τ_l is the effect of the lth letter, and ε_{ijl} is random $\mathcal{N}\left(0, \sigma^2\right)$ noise. Note that we do not have enough data in this setting to consider interaction terms. If this experiment were replicated n times, then we could consider such terms.

As with the previous models, we get a sum of squares decomposition. Note that the sum over i, j, l only contains k^2 terms as given a row i and a column j, the choice for l has already been made.

$$\begin{aligned} SS_{\text{tot}} = \sum_{(i,j,l)} (\bar{y}_{ijl} - \bar{y}_{\cdots})^2 = SS_{\text{row}} + SS_{\text{col}} + SS_{\text{tr}} + SS_{\text{err}} = \\ = \sum_{i=1}^{k} k(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2 + \sum_{j=1}^{k} k(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2 + \sum_{l=1}^{k} k(\bar{y}_{\cdot\cdot l} - \bar{y}_{\cdots})^2 + \\ + \sum_{(i,j,k)} (y_{ijl} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot l} + 2\bar{y}_{\cdots})^2 \end{aligned}$$

The degrees of freedom for the four terms are $k-1$ for SS_{row}, SS_{col}, and SS_{tr} and is $(k^2-1) - 3(k-1) = (k-2)(k-1)$ for SS_{err}.

Using Cochran's theorem, we can perform an F test for the hypotheses

$$H_0 : \tau_1 = \ldots = \tau_k \qquad H_1 : \exists l_1 \neq l_2 \ s.t. \ \tau_{l_1} \neq \tau_{l_2}.$$

Namely, under the null hypothesis, we have that

$$F = \frac{SS_{\text{tr}}/(k-1)}{SS_{\text{err}}/(k-1)(k-2)} \sim F\left(k-1, (k-1)(k-2)\right).$$

Given that we reject H_0, we can follow this test with a post-hoc Tukey test to determine which differences $\tau_i - \tau_j$ are significant.

Remark 2.3.1 (Why would we want to do this?)**.** *In the below data example, there is little difference between blocking and not blocking. The key point to remember is that a "good" blocking will reduce variation and result in a more significant test statistic. Thus, it can sometimes be the difference between rejecting and not rejecting H_0.*

Example: Fungicide

From the text of C. H. Goulden, *Methods of Statistical Analysis*, and also from the `R` package `agridat`, we consider dusting wheat crops with sulphur for the purposes

of controlling the growth of a certain fungus. In this experiment, $k = 5$ and the blocking factors are literally the rows and columns of a plot of land partitioned into 25 subplots. The responses y_{ijl} are the yields of the crop in bushels per acre. Each treatment–A,B,C,D,E–corresponds to a different dusting method of the sulphur including treatment E which is no dusting. The design and the data are, respectively,

$$\begin{pmatrix} B & D & E & A & C \\ C & A & B & E & D \\ D & C & A & B & E \\ E & B & C & D & A \\ A & E & D & C & B \end{pmatrix} \qquad \begin{pmatrix} 4.9 & 6.4 & 3.3 & 9.5 & 11.8 \\ 9.3 & 4.0 & 6.2 & 5.1 & 5.4 \\ 7.6 & 15.4 & 6.5 & 6.0 & 4.6 \\ 5.3 & 7.6 & 13.2 & 8.6 & 4.9 \\ 9.3 & 6.3 & 11.8 & 15.9 & 7.6 \end{pmatrix}.$$

Running an ANOVA in `R` gives the following output:

	Df	Sum Sq	Mean Sq	F value	Pr($>F$)	
row	4	46.67	11.67	4.992	0.0133	*
col	4	14.02	3.50	1.500	0.2634	
trt	4	196.61	49.15	21.032	2.37e-05	***
Residuals	12	28.04	2.34			

A post-hoc Tukey test indicates significant differences between (A, C), (B, C), (D, C), (E, C). Hence, treatment C, which was "dust weekly" led to increases in yield over all other cases. If the same experiment is run without considering the blocking factors, the same conclusions are reached, but the test statistics are less significant.

Note that the p-values for the blocking factors are generally not of interest. However, The mean squares does quantify the variation explained by that blocking factor, which can be used to determine whether or not the blocking was successful.

From Goulden (1957),

> "The column mean square is not significant. This is probably due to the shape of the plots. They were long and narrow; hence the columns are narrow strips running the length of the rectangular area. Under these conditions the Latin square may have little advantage on the average over a randomized block plan."

2.3.1 Graeco-Latin Squares[8]

The Graeco-Latin Square is an extension of the Latin square where two blocking factors and two experimental factors with k level each are considered. The sample size is again k^2. The idea is to superimpose two orthogonal Latin squares such that every entry is unique. Generally, the second square uses Greek letters to differentiate

[8] See Wu & Hamada Section 3.7

it from the first.

$$\begin{pmatrix} A & B & C & D \\ B & A & D & C \\ C & D & A & B \\ D & C & B & A \end{pmatrix}, \quad \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \gamma & \delta & \alpha & \beta \\ \delta & \gamma & \beta & \alpha \\ \beta & \alpha & \delta & \gamma \end{pmatrix} \Rightarrow \begin{pmatrix} A\alpha & B\beta & C\gamma & D\delta \\ B\gamma & A\delta & D\alpha & C\beta \\ C\delta & D\gamma & A\beta & B\alpha \\ D\beta & C\alpha & B\delta & A\gamma \end{pmatrix}.$$

The analysis follows exactly as above given the model with one additional term corresponding to the Greek letters:

$$y_{ijlm} = \mu + \alpha_i + \beta_j + \tau_l + \kappa_m + \varepsilon_{ijlm}$$

The degrees of freedom for all of the factors is still $k-1$ and the degrees of freedom for the residuals is now $(k^2-1) - 4(k-1) = (k-3)(k-1)$.

Remark 2.3.2 (Fun Fact)**.** *There are no Graeco-Latin squares of order 6.*

2.4 Balanced Incomplete Block Designs[9]

Returning to the concept of Randomized Block Design from Section 2.1, imagine that we have a single experimental factor and a single blocking factor. Originally, we had b blocks, k treatments, and a total of bk observations. The block size–i.e. the number of observations per block–was precisely k. In the Balanced Incomplete Block Design (BIBD), we consider the case where the block size k is less than t, the total number of treatments. Furthermore, we require the notation r being the number of blocks for each treatment and λ being the number of blocks for each pairs of treatments. For such a design to be *balanced*, every pair of treatments must be considered in the same number of blocks.

As an example, consider

	Treatment			
block	A	B	C	D
1	$y_{1,1}$	$y_{1,2}$	.	.
2	.	$y_{2,2}$	$y_{2,3}$	.
3	.	.	$y_{3,3}$	$y_{3,4}$
4	$y_{4,1}$	.	$y_{4,3}$	.
5	.	$y_{5,2}$	.	$y_{5,4}$
6	$y_{6,1}$	.	.	$y_{6,4}$

Here, we have $t = 4$ treatments, $b = 6$ blocks, $k = 2$ treatments per block, $r = 3$ blocks per treatment, and $\lambda = 1$ block for each pair of treatments.

This type of design occurs when only so many treatments can be applied to a single block before that block is saturated. For example, if one were to apply

[9] See Wu & Hamada Section 3.8

rust protective coatings to steel girders, then there is only so much surface area per girder to apply differing treatments. In taste testing scenarios, only so many samples can be tasted by a judge before they lose the ability to be discerning. For testing antibiotics, only so many different drugs can be applied to a bacteria culture.

There are some relationships that are required for such a blocking to be valid. They are

$bk = rt$	Both are total sample size
$r(k-1) = \lambda(t-1)$	More equivalent counting
$t > k$	More total treatments than those in a given block
$r > \lambda$	single occurrences greater than paired occurrences
$b > r$	More total blocks than those given a certain treatment
$rk > \lambda t$	follows from above

The model in this setting is identical to that of Section 2.1, but with missing data points.

$$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}.$$

The sum of squares decomposition results in

Term	DoF	Equation
SS_{bl}	$b-1$	$k\sum_{i=1}^{b}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$
SS_{tr}	$t-1$	$k\sum_{j=1}^{t} Q_j^2/\lambda t$
SS_{err}	$bk-b-t+1$	$SS_{\text{tot}} - SS_{\text{tr}} - SS_{\text{bl}}$
SS_{tot}	$bk-1$	$\sum_{i,j}(y_{ij} - \bar{y}_{\cdot\cdot})^2$

where, for the treatment sum of squares,

$$Q_j = r\left(\bar{y}_{\cdot j} - \frac{1}{r}\sum_{i=1}^{b}\bar{y}_{i\cdot}\mathbf{1}[y_{i,j} \text{ exists}]\right).$$

This formula arises from the usual $\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$, but replacing the global average with the average of the block averages for only the r blocks where treatment j occurred.

Once again, we can divide the treatment sum of squares by the error sum of squares to get an F statistic via Cochran's theorem. This can be followed up with a Tukey test for individual differences $\tau_i - \tau_j$.

Example

Continuing with the example from Section 2.1, we remove 12 of the entries from the data table to match the pattern from the example design at the beginning of this section. In this case, `R` produces the following table from `aov()`.

	Df	Sum Sq	Mean Sq	F value	Pr($>F$)	
block	5	19.122	3.824	39.059	0.00623	**
treat	3	2.233	0.744	7.602	0.06489	.
Residuals	3	0.294	0.098			

The treatment effect is very weak with the reduced sample, but perhaps is strong enough to warrant a follow up study.

On the other hand, if we were to ignore the blocking factor. Then the results are not significant at all.

	Df	Sum Sq	Mean Sq	F value	Pr($>F$)
treat	3	8.002	2.667	1.564	0.272
Residuals	8	13.647	1.706		

2.5 Split-Plot Designs[10]

The concept of a split plot design–as well as the name *split plot*–comes directly from agricultural research. Given two experimental factors with k_1 and k_2 levels that we wish to apply to a field of some crops, the ideal was to proceed is to cut the field, or *plot* of land, into k_1k_2 subplots and apply one of the k_1k_2 treatments to each subplot at random. However, such an approach is often impractical if one of the treatments is only able to be applied to large areas of land. As a result, we could instead *split* the *plot* of land into k_1 regions, apply different levels of the first experimental factor to each, then split each of those regions into k_2 subregions and apply different levels of the second experimental factor. In this case, the first factor is sometimes referred to as the *whole plot* factor and the second as the *subplot* factor.

To begin, say we have experimental factors A and B with k_A and k_B levels, respectively. Furthermore, assume the experiment is replicated n times. Thus, the total sample size is nk_Ak_B. If we were able to fully randomize the k_Ak_B treatments, then we would have a two-way fixed effects model:

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijl}.$$

However, as we have instead randomized factor A and then, given a level for A, randomized factor B, the model has to change. Consequently, the specific replication can have an effect on the observation y_{ijl}, which leads to a three-way *mixed* effects model:

$$y_{ijl} = \mu + \alpha_i + \beta_j + \tau_l + (\alpha\beta)_{ij} + (\alpha\tau)_{il} + (\beta\tau)_{jl} + (\alpha\beta\tau)_{ijl}$$

where α_i is the effect of the ith level of factor A, β_j is the effect of the jth level of factor B, τ_l is the effect of the lth replicate, the three paired terms, $(\alpha\beta)_{ij}, (\alpha\tau)_{il}, (\beta\tau)_{jl}$,

[10] See Wu & Hamada Section 3.9

quantify the three pairwise interaction effects, and lastly, $(\alpha\beta\tau)_{ijl}$ quantifies the three-way interaction effect.

Usually, the two experimental factors A and B are treated as fixed effects and the replication factor as a random effect resulting in a so-called *mixed effects model*. We can categorize the eight terms as

	Fixed	Random
Whole Plot	α_i	τ_l, $(\alpha\tau)_{il}$
Sub Plot	β_j, $(\alpha\beta)_{ij}$	$(\beta\tau)_{jl}$, $(\alpha\beta\tau)_{ijl}$

Note that there is no ε_{ijl} (even though there is one in Wu & Hamada, Equation 3.59) as if we consider the degrees of freedom, there are none left to estimate an ε_{ijl} term:

Term	Total	μ	α	β	τ	$(\alpha\beta)$
Dofs	nk_Ak_B	1	k_A-1	k_B-1	$n-1$	$(k_A-1)(k_B-1)$
Term	$(\alpha\tau)$			$(\beta\tau)$		$(\alpha\beta\tau)$
Dofs	$(k_A-1)(n-1)$			$(n-1)(k_B-1)$		$(k_A-1)(k_B-1)(n-1)$

Hence, if we were to include such a term, it would not be identifiable. Instead we consider two error terms based on the random effects:

$$\varepsilon_{il}^{\text{whole}} = (\alpha\tau)_{il} \quad \text{and} \quad \varepsilon_{ijl}^{\text{sub}} = (\beta\tau)_{jl} + (\alpha\beta\tau)_{ijl}.$$

For split-plot designs, as with the previous, we are interested in testing for whether or not there are differences in the effects of the factor levels. Applying the sum-to-zero constraints,

$$\sum_{i=1}^{k_A} \alpha_i = 0, \quad \sum_{j=1}^{k_B} \beta_j = 0, \quad \text{and } \sum_{i,j} (\alpha\beta)_{ij} = 0,$$

results in three null hypotheses of interest,

$$H_{01} : \alpha_i = 0 \ \forall i, \quad H_{02} : \beta_j = 0 \ \forall j, \quad \text{and } H_{03} : (\alpha\beta)_{ij} = 0 \ \forall i, j.$$

To test these, we decompose the total sum of squares as usual:

$$SS_{\text{tot}} = SS_A + SS_B + SS_R + SS_{A\times B} + SS_{A\times R} + SS_{B\times R} + SS_{A\times B\times R},$$

which can be rewritten as

$$SS_{\text{tot}} = SS_A + SS_B + SS_R + SS_{A\times B} + SS_{\text{whole}} + SS_{\text{sub}}.$$

Finally, we can write out the F statistics in the following table:

Sum Sq	DoF	F stat
SS_A	$k_A - 1$	$\frac{SS_A/df_A}{SS_{\text{whole}}/df_{\text{whole}}}$
SS_R	$n - 1$	
SS_{whole}	$(k_A - 1)(n - 1)$	
SS_B	$k_B - 1$	$\frac{SS_B/df_B}{SS_{\text{sub}}/df_{\text{sub}}}$
$SS_{A\times B}$	$(k_A - 1)(k_B - 1)$	$\frac{SS_{A\times B}/df_{A\times B}}{SS_{\text{sub}}/df_{\text{sub}}}$
SS_{sub}	$k_A(k_B - 1)(n - 1)$	

Remark 2.5.1 (Example of two-way vs three-way)**.** *The two-way model would be used if, for example, we had $k_A k_B$ human subjects who were each randomly assigned a specific level factors of A and B, and then this experiment was replicated with n groups of subjects.*

In the three-way model, we are assuming that each of the n replicates corresponds to a single subject, like a plot of land, hence, there could be replicate effects that unlike in the previous setting are not removed by randomization.

Example

As an example, we can look at the `stroup.splitplot` dataset from the `R` package `agridat`. This is simulated data in the form of a split plot design.

The dataset has $N = 24$ observations, y_{ijl}, with whole-plot factor A with 3 levels, sub-plot factor B with 2 levels, and $n = 4$ replicates. The design looks like

A1			
r1	r2	r3	r4
B1,B2	B1,B2	B1,B2	B1,B2

for treatment A1 with two other identical blocks for A2 and A3.

First, we will analyse the data as a two-way fixed effects model,

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijl},$$

using the `R` command `aov( y~a*b, data=stroup.splitplot )`. This yields three F statistics none of which are significant.

	Df	Sum Sq	Mean Sq	F value	Pr($> F$)
a	2	326.6	163.29	1.874	0.182
b	1	181.5	181.50	2.083	0.166
a:b	2	75.3	37.63	0.432	0.656
Residuals	18	1568.5	87.14		

Next, we can include the replications as a random effect,

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \tau_l + \varepsilon_{ijl},$$

using the command `aov( y∼a*b + Error(rep), data=stroup.splitplot )`. The result is now a significant result for factors A and B, but not for the interaction term.

Error: rep						
	Df	Sum Sq	Mean Sq	F value	Pr($>F$)	
Residuals	3	1244	414.5			
Error: Within						
	Df	Sum Sq	Mean Sq	F value	Pr($>F$)	
a	2	326.6	163.29	7.537	0.00542	**
b	1	181.5	181.50	8.377	0.01112	*
a:b	2	75.3	37.63	1.737	0.20972	
Residuals	15	325.0	21.67			

Notice that the degrees of freedom for the *within* residuals has reduced by 3 and a large amount of the error sum of squares is now contained in the `rep` factor.

Lastly, we can fit the split plot design,

$$y_{ijl} = \mu + \alpha_i + \tau_l + \varepsilon_{\text{whole}} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{\text{sub}},$$

using the command `aov(y∼a*b + Error(rep/a),data=stroup.splitplot)`. As a result, the interaction term a:b looks more significant whereas factor A looks less significant.

Replication Effect						
Error: rep						
	Df	Sum Sq	Mean Sq	F value	Pr($>F$)	
Residuals	3	1244	414.5			
Whole Plot Terms						
Error: rep:a						
	Df	Sum Sq	Mean Sq	F value	Pr($>F$)	
a	2	326.6	163.29	4.07	0.0764	.
Residuals	6	240.8	40.13			
Sub Plot Terms						
Error: Within						
	Df	Sum Sq	Mean Sq	F value	Pr($>F$)	
b	1	181.50	181.50	19.389	0.00171	**
a:b	2	75.25	37.63	4.019	0.05658	.
Residuals	9	84.25	9.36			

Often, the significance of the whole plot factor A will be overstated if a split plot design is not analysed as such.

2.6 Analysis of Covariance[11]

Sometimes when make observations y_{ij}, we are faced with a variable x_{ij} that cannot be controlled and varies with every observation. We can include $x_{i,j}$, referred to as a *covariate*, in the model in order to account for it. One common type of covariate is size such as the body weight of a test subject or the population of a city. In the below example, we consider the number of plants in a plot as a covariate of the yield of the plot.

The model we consider combines simple linear regression with the one-way ANOVA:

$$\text{Model 1:} \qquad y_{ij} = \mu + \tau_i + \beta x_{ij} + \varepsilon_{ij}$$

where μ is the global mean, τ_i is the ith treatment effect, $x_{ij} \in \mathbb{R}$ is the covariate, β is the unknown regression coefficient, and, as usual, ε_{ij} is the iid normal noise.

To test whether or not the treatment terms, τ_i's, have any detectable effect on the y_{ij}, we compare the above model to the simpler model

$$\text{Model 0:} \qquad y_{ij} = \mu + \beta x_{ij} + \varepsilon_{ij}.$$

Assume, as before, we have k treatment levels, n observations of each treatment, and a total sample of size N. Following the usual procedure from linear regression, we can compute the residual sum of squares for each model to get RSS_0 and RSS_1 with degrees of freedom $N-2$ and $(N-2)-(k-1) = N-k-1$, respectively.[12] The resulting F statistic is

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(k-1)}{\text{RSS}_1/(N-k-1)},$$

which under the null hypothesis–i.e. that all of the τ_i coincide–has an F distribution with degrees of freedom $k-1$ and $N-k-1$, respectively.

The above Model 1 assumes that the slope β is independent of treatment category i. We could allow for variable slopes as well with

$$\text{Model 2:} \qquad y_{ij} = \mu + \tau_i + (\beta + \alpha_i)x_{ij} + \varepsilon_{ij}$$

where the α_i term also has $k-1$ degrees of freedom. Note that as μ and τ_i are the global and category mean effects, we have β and α_i as the global and category slope effects. Testing would proceed as before taking into account the new RSS_2 with degrees of freedom $N-2-2(k-1) = N-2k$.

Remark 2.6.1 (**Warning!**). *All of the previous models when put into R's* `aov()` *function are independent of the variable order. This is because when the factors*

[11] See Wu & Hamada Section 3.10

[12] Note that the covariate x_{ij} is considered as a continuous variable and has only 1 degree of freedom.

are translated into a linear model, the columns of the design matrix are orthogonal. However, when dealing with a continuous regressor x_{ij}, the results of `summary.aov()` *are dependent on the variables' order. Hence, we compare nested models as above to test for the significance of the experimental factor.*

Example

As an example, we consider the `cochran.beets` dataset from the `agridat` library. This dataset considers sugarbeet yield based on $k = 7$ different fertilizer treatments. It also includes a blocking factor with $b = 6$ levels and a single continuous covariate, number of plants per plot.

Ignoring the blocking factor and covariate and using a one-way ANOVA results in a very significant p-value for the fertilizer factor.

	Df	Sum Sq	Mean Sq	F value	Pr($> F$)	
fert	6	112.86	18.809	22.28	1.3e-10	***
Residuals	35	29.55	0.844			

Next we take the ANCOVA approach by comparing the simple linear regression,

$$(\text{yield}) = \mu + \beta(\text{plants}) + \varepsilon$$

to the model with variable means based on the fertilizer and then to the model with variable means and variable slopes. This results in the following table from `R`.

Model 1: yield ~ plants
Model 2: yield ~ fert + plants
Model 3: yield ~ fert * plants

	Res.Df	RSS	Df	Sum of Sq	F	Pr($> F$)	
1	40	28.466					
2	34	20.692	6	7.7734	2.1822	0.07495	.
3	28	16.623	6	4.0694	1.1424	0.36431	

The p-value for the fertilizer effect is now no longer significant at the 0.05 level.

We can also compare the same three models as previously done by also incorporating the blocking factor. Now, we get

Model 1: yield ~ block + plants
Model 2: yield ~ fert + block + plants
Model 3: yield ~ fert * plants + block

	Res.Df	RSS	Df	Sum of Sq	F	Pr($> F$)
1	35	9.4655				
2	29	6.9971	6	2.4684	1.891	0.1256
3	23	5.0038	6	1.9933	1.527	0.2138

Here, the blocking resulted in a reduction in the sum of squares for comparing model 2 to model 1. The conclusion is that the fertilizer effect disappears once the covariate is taken into account.

Chapter 3

Multiple Testing

Introduction

Multiple testing is a ubiquitous problem throughout statistics. Consider flipping 10 coins; if we observe 10 heads, then we may conclude that the coins are not fairly weighted as the probability of getting all 10 heads assuming fair coins is 1/1024. However, if we were to repeat this experiment with 100 sets of ten coins, then the probability that at least one set yields 10 heads is

$$1 - \left(\frac{1023}{1024}\right)^{100} \approx 0.093.$$

Hence, there is a 9.3% chance that we may incorrectly conclude that a set of coins is not fairly weighted.

Similarly, running a single hypothesis test of size α means that the probability that we falsely reject the null hypothesis is α. If m independent tests are conducted with size α, then the probability of at least one false positive is

$$1 - (1 - \alpha)^m \gg \alpha.$$

In the following chapter, we consider factorial designs where $m = 2^k$ independent hypothesis tests can be conducted. This leads to the potential for many false positives unless we correct our methodology. This chapter contains some methods for correcting for multiple testing.

For the remainder of this chapter, consider that we have conducted m independent hypothesis tests that have yielded a set of p-values $p_1, \ldots, p_m \in [0, 1]$. The ith null hypothesis is denoted H_{0i} while the ith alternative hypothesis is denoted H_{1i}. We assume that under the null hypothesis that $p_i \sim \text{Uniform}\,[0, 1]$. Given that we want a false positive rate of α, without correcting, we would reject the ith null hypothesis if $p_i < \alpha$.

3.1 Family-wise Error Rate

There are many proposed methods to control the *Family-wise Error Rate* (FWER), which is the probability of at least one false positive:

$$FWER = \mathrm{P}\left(\text{reject any } H_{0i} \,|\, H_{0i}\right).$$

Thus, instead of having a test size α such that $\mathrm{P}\left(\text{reject } H_{0i} \,|\, H_{0i}\right) \leq \alpha$, we want to devise a methodology such that $FWER < \alpha$. What follows are some methods for doing just that.

3.1.1 Bonferroni's Method

As already mentioned in Chapter 1, Bonferroni's method is an effective but extremely conservative method for multiple testing correction. Its strength is that it requires few assumptions; its weakness is that it is too aggressive often resulting in too many false negatives.

Bonferroni's method simply says that if we want the FWER to be no greater than α, then we should reject H_{0i} if $p_i \leq \alpha/m$. Indeed, let $I_0 \subset \{1, \ldots, m\}$ be the set of indices such that H_{0i} is true. Then, applying a union bound or Boole's inequality, we have

$$FWER = \mathrm{P}\left(\bigcup_{i \in I_0} \{p_i \leq \alpha/m\}\right) \leq \sum_{i \in I_0} \mathrm{P}\left(p_i \leq \alpha/m\right) = |I_0| \frac{\alpha}{m} \leq \alpha.$$

While we will mostly assume our hypothesis tests to be independent–i.e. $p_i \overset{\text{iid}}{\sim} \text{Uniform}\,[0,1]$ under H_{0i}–this derivation works regardless of any dependency or lack thereof among the hypothesis tests.

3.1.2 Sidak's Method

Sidak's method is similar to Bonferroni. In this method, we reject H_{0i} if $p_i \leq 1-(1-\alpha)^{1/m}$. Unlike Bonferroni's method, Sidak's requires the hypothesis tests to be independent. To see this, note that the above rejection region can be equivalently writen as Don't reject H_{0i} if $(1-\alpha)^{1/m} \leq 1 - p_i$. Therefore,

$$\begin{aligned}
1 - (FWER) &= \mathrm{P}\left(\bigcap_{i \in I_0} \{p_i \geq (1-\alpha)^{1/m}\}\right) \\
&= \mathrm{P}\left(p_i \geq (1-\alpha)^{1/m}\right)^{|I_0|} \\
&= (1-\alpha)^{|I_0|/m} \\
FWER &= 1 - (1-\alpha)^{|I_0|/m} \leq 1 - (1-\alpha) = \alpha.
\end{aligned}$$

3.1.3 Holms' Method

Holms' Method is the first stepwise method we will consider. It specifically implements Bonferroni's method in a stepwise fashion. In this approach, we first order the p-values to get

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

We then reject hypothesis H_{0i} if, for all $j = 1, \ldots, i$,

$$p_{(j)} \leq \alpha/(m - j + 1).$$

Hence, we can only reject H_{0i} if we have already rejected all H_{0j} for $j < i$. Furthermore, when considering H_{01}, we check whether or not $p_{(1)} \leq \alpha/m$, which is just Bonferroni's method. Then, the threshold for rejection is relaxed as i increases. Thus, this approach is necessarily more powerful than regular Bonferroni as the standard Bonferroni only rejects hypotheses when $p_{(i)} \leq \alpha/m$ regardless of order. It can also be shown that this method still achieves a $FWER \leq \alpha$. Indeed, denote $\hat{i}$ to be the index such that H_{0i} is rejected for all $i \leq \hat{i}$, and let i_0 be the smallest index such that H_{0i_0} should not be rejected. Then, (step 1) applying Bonferroni in reverse, (step 2) noting that $|I_0| < m - i_0 + 1$, and (step 3) noting that the event H_{0i_0} not being rejected implies that $\hat{i} < i_0$ gives that

$$\begin{aligned} 1 - \alpha \leq \mathrm{P}\left(p_{(i)} > \frac{\alpha}{|I_0|} \text{ for all } i \in I_0\right) = \mathrm{P}\left(p_{(i_0)} > \frac{\alpha}{|I_0|}\right) \leq \\ \leq \mathrm{P}\left(p_{(i_0)} > \frac{\alpha}{m - i_0 + 1}\right) \leq \mathrm{P}\left(\hat{i} < i_0\right) = \mathrm{P}\left(\text{No null is rejected}\right). \end{aligned}$$

Thus, this proceedure makes sense.

3.1.4 Stepwise Methods

Step-down Methods

Holms' method is a specific type of step-down method using Bonferroni's method iteratively. Such approaches to multiple testing iteratively consider the p-values from smallest to largest and only reject hypothesis H_{0i} if all hypotheses H_{0j} for $j < i$ have already been rejected.

In general, let $I \subset \{1, \ldots, m\}$. And let H_{0I} be the joint null that all of H_{0i} are true for $i \in I$. That is, the intersection $\bigcap_{i \in I} H_{0i}$, and thus rejection of H_{0I} implies that at least one H_{0i} is rejected for $i \in I$. If $J \supset I$, then not rejecting H_{0J} implies not rejecting H_{0I}. Assume for every subset I, we have a level-α nonrandomized test function—i.e. some $\phi(x; I) = 0, 1$ such that $\mathrm{P}\left(\phi(x; I) = 1 \,|\, H_{0I}\right) \leq \alpha$ where 1 indicates rejection of H_{0I}. The simultaneous test function is

$$\Phi(x; I) = \min_{J \supseteq I} \{\phi(x; J)\},$$

which implies that we reject H_{0I} at level α only when all H_{0J} are rejected at level α. Hence, if I_0 contains the indices of all true null hypotheses, then for any $I \subseteq I_0$,

$$\text{if } \mathrm{P}\left(\phi(x; I_0) = 1\right) \leq \alpha, \text{ then } \mathrm{P}\left(\Phi(x; I) = 1\right) \leq \alpha.$$

Thus, the test Φ simultaneously controls the probability of falsely rejecting any H_{0I} comprised of true nulls.

To see this idea in practice, we return to Holms' method. Here, Bonferroni's method is ϕ and Holms' method is Φ. Indeed, assuming that the p-values are ordered $p_1 \leq p_2 \leq \ldots \leq p_m$, consider the indices $I_i = \{i, i+1, \ldots, m\}$. The cardinality of this set is $|I_i| = m - i + 1$, so applying Bonferroni to this set says reject any H_{0k} for $k \in I_i$ if $p_k < \alpha/(m - i + 1)$. Since the p-values are ordered, this means that we reject H_{0I_i} if $p_i < \alpha/(m - i + 1)$. As $I_j \supseteq I_i$ for any $j \leq i$, we have the simultaneous rule which says

$$\text{reject } H_{0I_i} \text{ only if we reject all } H_{0I_j}, \forall j \leq i,$$

which equivalently is

$$\text{reject } H_{0I_i} \text{ only if } p_j < \alpha/(m - j + 1), \forall j \leq i,$$

and this is Holms' method.

Step-up Methods

Similar to step-down procedures, there are step-up procedures where we begin with $p_{(n)}$ and accept null hypothesis H_{0i} only if we have first accepted all H_{0j} for $j > i$.

3.2 False Discovery Rate

As an alternative to FWER, we have the False Discovery Rate (FDR). To understand these ideas, consider that we have run m hypothesis tests where m_0 are from the null and m_1 are from the alternative. Furthermore, assume that we rejected r null hypotheses. This results in the following table:

	Decided Null	Decided Alternative	
True Null	$m_0 - a$	a	m_0
True Alternative	$m_1 - b$	b	m_1
	$m - r$	r	m

Here, we have r rejected null hypotheses with a of them false rejections and b of them true rejections. Using this notation, we have

$$\begin{aligned} FWER &= \mathrm{E}\left(a/m_0\right) = \mathrm{P}\left(a > 0\right) \\ FDR &= \mathrm{E}\left(a/r\right) \end{aligned}$$

3.2.1 Benjamini-Hochberg Method

The Benjamini-Hochberg Method, published in 1995 in JRSSB with currently more than 50,000 citations on Google Scholar, is one of the most important advancements beyond standard FWER control, which was the standard approach for most of the 20th century.

The method is in its essence quite simple. Given ordered p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$, we choose a $q \in (0,1)$ which will be the desired FDR. Then, we find the maximal index i such that

$$p_{(i)} \leq \frac{i}{m} q$$

and then we reject all H_{0j} for $j \leq i$. This proceedure is validated by the following theorem.

Theorem 3.2.1. *Given m ordered p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$, if the p-values corresponding to null hypotheses are independent then*

$$FDR = \mathrm{E}\,(a/r) = \pi_0 q \leq q \text{ where } \pi_0 = |I_0|/m.$$

Here, π_0 is the proportion of true null hypotheses to all tests.

Proof. For $t \in (0,1)$, let $r(t)$ be the total number of p-values less than t and let $a(t)$ be the number of p-values from null hypotheses less than t. Hence, the false discovery proportion at t is

$$FDP(t) = a(t)/\max\{r(t), 1\}.$$

Also, let $Q(t) = mt/\max\{r(t),1\}$ and $t_q = \sup\{t : Q(t) \leq q\}$. Considering the ordered p-values, $r(p_{(i)}) = i$ and $Q(p_{(i)}) = mp_{(i)}/i$. Therefore, the Benjamini-Hochberg method can be rewritten as

$$\text{Reject } H_{0(i)} \text{ for } p_{(i)} \leq t_q.$$

Under the null hypothesis, p-values are uniformly distributed on $[0,1]$. Therefore $\mathrm{E}a(t) = t|I_0|$, which is $t|I_0|$ p-values on average will be in the interval $[0,t]$. Now, defining $A(t) = a(t)/t$, we have that for $s < t$

$$\begin{aligned}
&\mathrm{E}\,(a(s) \,|\, a(t)) && = (s/t)a(t) && \text{and}\\
&\mathrm{E}\,(A(s) \,|\, A(t)) && = A(t) && \text{and}\\
&\mathrm{E}\left(A(s) \,|\, A(t')\,\forall t' \geq t\right) && = A(t)
\end{aligned}$$

This makes $A(t)$ a martingale as t goes from 1 to 0. Considering t_q as a stopping time—that is, as t decreases from 1 to 0, then t_q is the first time that $Q(t)$ drops below q—we can apply the *Optional Stopping Theorem*[1] to get that

$$\mathrm{E}A(t_q) = \mathrm{E}A(1) = \mathrm{E}a(1) = |I_0|.$$

[1] A powerful result from the theory of martingales: https://en.wikipedia.org/wiki/Optional_stopping_theorem

Therefore,

$$FDP(t_q) = \frac{a(t_q)}{\max\{r(t_q), 1\}} = a(t_q)\frac{Q(t_q)}{mt_q} = \frac{a(t_q)}{t_q}\frac{q}{m},$$

and finally

$$\mathrm{E}[FDP(t_q)] = q\frac{|I_0|}{m} \leq q.$$

□

Remark 3.2.2. *Note that Benjamini-Hochberg requires the p-values to be independent. A more conservative approach that works under arbitrary dependence is to find the maximal index i such that*

$$p_{(i)} \leq \frac{1}{c(m)}\frac{i}{m}q$$

where $c(m) = \sum_{i=1}^{m} i^{-1}$ and then we reject all H_{0j} for $j \leq i$.

Chapter 4

Factorial Design

Introduction

In this chapter, we will consider designs where we wish to consider k factors all with 2 different levels. Hence, there are 2^k total treatments to consider, which is *full factorial design*. As one might expect, this can quickly become impractical as k grows. Hence, we will often consider *fractional factorial design* where some subset of treatments is "optimally" selected.

Data from a factorial design can be displayed as

Factor			
A	B	C	Data
-	-	-	y_1
-	-	+	y_2
-	+	-	y_3
-	+	+	y_4
+	-	-	y_5
+	-	+	y_6
+	+	-	y_7
+	+	+	y_8

where - and + refer to the two levels of each of the $k = 3$ factors. For each of the $2^3 = 8$ treatments, we have a some observed data. If each factor level occurs in the same number of treatments, then the design is *balanced*. If two factors have all of their level combinations occurring an equal number of times, then they are said to be *orthogonal*. If all factor pairs are orthogonal, then the design is said to be *orthogonal*.

When running such an experiment, the ideal situation is to randomize the order of the rows above and thus test the treatments in a random order. Often some factors are harder to change than others, hence we can also consider the case of *restricted randomization* where, for example, factor A may be set to - and the levels

of B and C are randomized. Then A is set to + and B and C are randomized again. An experiment can also be *replicated*, say n times, to collect multiple y_{ij} for $j = 1, \ldots, n$. Each time an experiment is replicated, it should be randomized again.

4.1 Full Factorial Design[1]

A first consideration is to just treat such a design as a multi-way fixed effects model. The problem is that without replication, we have too few observations to take an ANOVA approach. And even with replication, we are faced with $2^k - 1$ hypothesis tests, which require a fix for multiple testing. The Bonferroni method is too inefficient in this setting, so other techniques are developed. First, however, we need to define the *main* and *interaction* effects.

Consider the experiment with k factors, $A_1, \ldots, A_k$, each with 2 levels, $\{-, +\}$, such that every combination of levels–i.e. treatments–is tested once. The data size is then $N = 2^k$. We estimate the *main effect* of factor A_i by

$$\text{ME}(i) = \bar{y}(i+) - \bar{y}(i-)$$

where $\bar{y}(i+)$ is the average of all of the 2^{k-1} observations such that factor A_i is at level + and similarly for $\bar{y}(i+)$. As we are averaging over all other factor levels, a significant main effect should be *reproducible* in the sense that it has a direct effect on the response regardless of other factor levels.

Next, we can consider the *conditional main effect* of one factor given another. That is, for example, we could compute the main effect of A_i given A_j is at level $-$.

$$\text{ME}(i|j+) = \bar{y}(i + |j+) - \bar{y}(i - |j+)$$

where $\bar{y}(i - |j+)$ is the average of all of the 2^{k-2} observations with A_i at level i and A_j at level +.

We are also interested in the interaction effects between two or more factors, which we can write as $\text{INT}(i, j, \ldots)$ with as many arguments as desired. In the case of two-way interactions, we have that

$$\text{INT}(i, j) = \frac{1}{2} \left[\text{ME}(i|j+) - \text{ME}(i|j-) \right].$$

Note that this function is symmetric in the arguments–i.e. $\text{INT}(i, j) = \text{INT}(j, i)$.

Continuing down the rabbit hole, we can define conditional interaction effects for A_i and A_j conditionally on A_l taking a certain level:

$$\text{INT}(i, j|l+) = \frac{1}{2} \left[\text{ME}(i|j+, l+) - \text{ME}(i|j-, l+) \right].$$

[1] See Wu & Hamada Sections 4.3.1 & 4.3.2

This allows us to write down three-way interaction terms as

$$\text{INT}(i,j,l) = \frac{1}{2}\left[\text{INT}(i,j|l+) - \text{INT}(i,j|l-)\right].$$

This function is similarly invariant under permutations of the arguments. This conditioning process can be iterated to consider m-way interactions with

$$\text{INT}(1,2,\ldots,m) = \frac{1}{2}\left[\text{INT}(1,2,\ldots,m-1|m+) - \text{INT}(1,2,\ldots,m-1|m-)\right].$$

4.1.1 Estimating effects with regression[2]

We can write such a factorial design in the framework of linear regression. To do this, we require the xor operator $\oplus$ such that

a	b	$a \oplus b$
0	0	1
1	0	0
0	1	0
1	1	1

In this case, we have k regressors $x_1, \ldots, x_k \in \{0,1\}$. The model is

$$y = \beta_0 + \sum_{i=1}^{2^k} \beta_i \left[\bigoplus_{j \in \mathcal{J}_i} x_j\right]$$

where $\mathcal{J}_i = \{j = 1, \ldots, k \mid \lfloor i2^{-j} \rfloor \mod 2 = 1\}$

The estimators $\hat{\beta}_i$ can be computed as usual.[3] If we wished to take the standard sum of squares approach, assume the experiment has been replicated n times. Each main effect and each interaction effect would have a single degree of freedom totalling $2^k - 1$ in all. Hence the DoFs for error sum of squares would be $(n2^k - 1) - (2^k - 1) = 2^k(n-1)$. Hence, if we do not replicate the experiment, then we cannot estimate the variance via the SS_{err}. Furthermore, if we do replicate the experiment n times, we are faced with a multiple testing problem as a result of the $2^k - 1$ hypothesis tests performed. We could apply Bonferroni's method in this case. However, that approach is often inefficient–i.e. too conservative–when a large number of tests is considered.

[2] See Wu & Hamada Sections 4.4 & 4.5

[3] There is also a method due to Frank Yates that computes the least squares estimates for the 2^k parameters. https://en.wikipedia.org/wiki/Yates_analysis

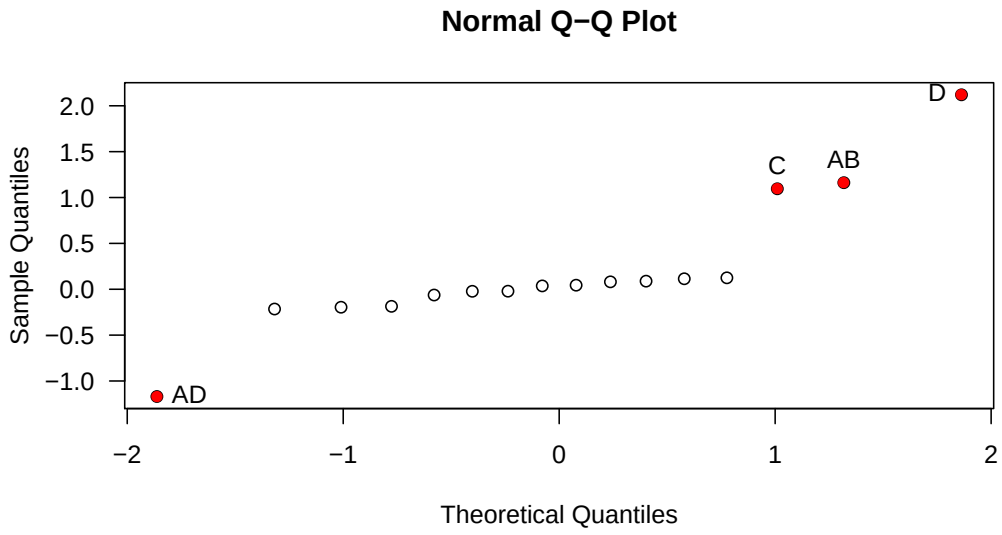

Figure 4.1: Applying `qqnorm()` to the coefficient estimates from the example.

Example

As an example, consider the 2^4 unreplicated experiment with factors A, B, C, and D. Data was randomly generated with the C effect of 1, the D effect of 2, the AB effect of 1, and the AD effect of -1. The data has normal errors with $\sigma = 0.1$. This results in the following least squares estimates.

Effect	A	B	C	D	AB	AC	BC	AD
Truth	·	·	1	2	1	·	·	-1
Estimate	0.0810	-0.0209	1.10	2.12	1.16	-0.196	0.0367	-1.17
Effect	BD	CD	ABC	ABD	ACD	BCD	ABCD	μ
Truth	·	·	·	·	·	·	·	·
Estimate	0.115	-0.215	-0.022	-0.186	0.125	0.0443	0.0884	-0.0626

These coefficient estimates can be plugged into R's `qqnorm` function to produce a normal qq-plot for the 16 values as displayed in Figure 4.1. In this example, we can see the four significant effects standing out from the rest. However, in practice, the significant effects may not be as obvious.

4.1.2 Lenth's Method[4]

As we cannot rely on sums of squares and ANOVA for testing for the significance of a main or interaction effect, other approaches are considered. Wu & Hamada Section 4.8, as well as other sources, consider looking at plots to determine the effect qualitatively. But as this is a bit ad-hoc compared to proper testing, we move onward and consider Lenth's method.

Consider the unreplicated full factorial experiment–i.e. $n = 1$. If we wish to test all $2^k - 1$ effects, then there are no degrees of freedom remaining to estimate the variance σ^2. Consequently, Lenth (1989) proposed the *psuedo standard error*, which is a robust estimator in terms of the median. In this case, robust means that if a few of the effects, θ_i, are non-zero then they will not skew the estimator of the standard error. Let $\hat{\theta}_i$ be the estimated effect of the ith treatment for $i = 1, \ldots, 2^k - 1$, then

$$\text{PSE} = 1.5\text{median}\left\{|\hat{\theta}_i| \ : \ |\theta|_i < 2.5\text{median}\{|\hat{\theta}_i|\}\right\},$$

which says, compute the median of $|\hat{\theta}_i|$, then consider on the $|\hat{\theta}_i|$ that are less than 2.5 times the median, and take their median and multiply by 1.5. The claim is that this is a consistent estimator under H_0 given the usual normality assumption.

Claim 4.1.1. *Under 2^k factorial design with iid $\mathcal{N}\left(0, \sigma^2\right)$ errors, if $\theta_1 = \ldots = \theta_{2^k-1} = 0$, then $PSE \xrightarrow{P} \sigma$ as $n \to \infty$.*

For testing for the significance of a given treatment, it is recommended to construct a "t-like" statistic

$$t_{\text{PSE},i} = \hat{\theta}_i / PSE,$$

which can be compared to tabulated critical values.

Note that we still have to correct for multiple testing in this case. This leads to two type of thresholds for determining whether or not $t_{\text{PSE},i}$ is significant. The uncorrected threshold–or *individual error rate* (IER) from Wu & Hamada or *marginal error rate* (ME) from the `R` package `BsMD`–is a real number such that

$$\text{P}\left(|t_{\text{PSE},i}| > \text{IER}_\alpha \,|\, H_0\right) = \alpha \quad \forall i = 1, \ldots, 2^k - 1.$$

The corrected threshold–or *experimentwise error rate* (EER) from Wu & Hamada or *simultaneous margin of error* (SME) from `R`–is a real number such that

$$\text{P}\left(\max_{i=1,\ldots,2^k-1} |t_{\text{PSE},i}| > \text{IER}_\alpha \,\middle|\, H_0\right) = \alpha.$$

Remark 4.1.2 (Advice from Wu & Hamada)**.** *In their text, Wu & Hamada claim that it is preferable to use the uncorrected thresholds as (1) the corrected threshold is too conservative and (2) it is better to include an uninteresting effect (false positive) than miss a significant one (false negative).*

[4] See Wu & Hamada Sections 4.9

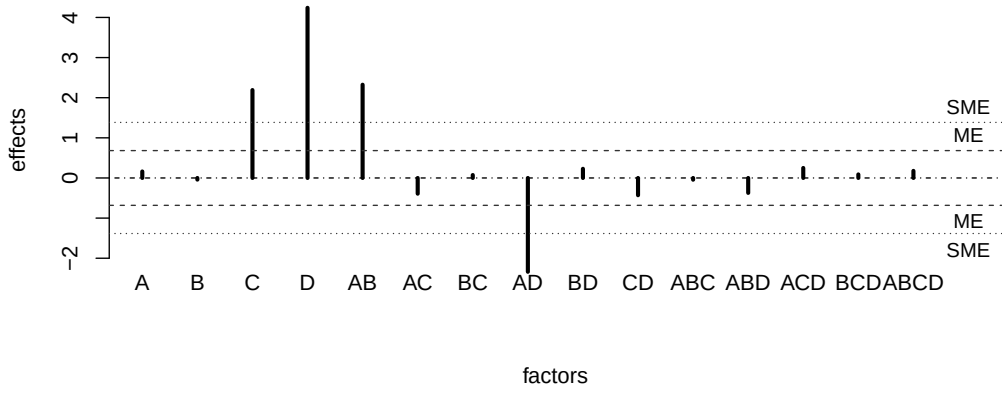

Figure 4.2: Results from applying Lenth's method with $\alpha = 0.05$ to our 2^4 factorial design. Here, ME is the uncorrected threshold and SME is the threshold corrected for multiple testing. The significant effects are C, D, AB, and AC.

Remark 4.1.3 (Advice from Lenth)**.** *According to RV Lenth, "Lenths method should not be used in replicated experiments, where the standard error can be estimated more efficiently by traditional analysis-of-variance methods."*

Example

Applying Lenth's method via the `BsMD` library in `R` with the function `LenthPlot()` to the data from the previous example, we have the plot in Figure 4.2. Here, $\alpha = 0.05$, the psuedo standard error is 0.265, the ME or IER is 0.682, and the SME or EER is 1.38. The result is that we correctly detect the four significant effects.

4.1.3 Key Concepts[5]

There are a few principles of Factorial Design that are worth mentioning:

1. *Hierarchy*: Lower order effects are often more important than higher order effects.

2. *Sparsity*: The number of significant terms is generally small.

3. *Heredity*: Interaction terms are often only significant if one of there interacting factors is solely significant.

[5] See Wu & Hamada Sections 4.6

The first point says that we should probably focus on estimating the lower order effects before investing in data collection for the higher order effects. Though, every problem has its own peculiarities. The second point is common assumption in many statistical modelling settings. The third point gives intuition about which factors to test.

4.1.4 Dispersion and Variance Homogeneity[6]

Thus far, we have been concerned with mean effects between different treatments. We can also consider the variances if the experiment has been replicated. Indeed, for the 2^k treatments, if the experiment is replicated $n > 1$ times, then we can compute sample variances for each treatment

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2.$$

Furthermore, under the normality assumption, we have that

$$\frac{(n-1)s_i^2}{\sigma^2} \sim \chi^2\,(n-1)\,,$$

and thus $X_i^2 = s_i^2/\sigma^2$ is a random variable with mean 1 and variance $2/(n-1)$.

We then consider $\log X^2$. The mean of this random variable lies in the interval $[-(n-1)^{-1}, 0]$. The upper bound follows from Jensen's Inequality,

$$\mathrm{E}\log X^2 \leq \log \mathrm{E}X^2 = 0$$

and the concavity of the logarithm. The lower bound comes from Taylor's Theorem,

$$\mathrm{E}\log X^2 \geq \mathrm{E}\left((X^2-1) - \frac{1}{2}(X^2-1)^2\right) = \mathrm{E}(X^2-1) - \frac{1}{2}\mathrm{Var}\left(X^2\right) = -\frac{1}{n-1}.$$

Hence, the claim in Wu & Hamada that "$\mathrm{E}\log X^2$ is approximately zero" is asymptotically true even if it is necessarily negative.

The variance of $\log X^2$ can be approximated via the delta method–which is basically just Taylor's theorem again. That is, for a continuously differentiable function $h : \mathbb{R} \to \mathbb{R}$,

$$\mathrm{Var}\left(h(X^2)\right) \approx \left[h'(\mathrm{E}X^2)\right]^2 \mathrm{Var}\left(X^2\right).$$

Hence, for the log transform, $\mathrm{Var}\left(\log X^2\right) \approx 2/(n-1)$. This is often referred to as a *variance stabilizing transform*.

Therefore, as $\log s_i^2 = \log X^2 + \log \sigma^2$, we have that the mean $\mathrm{E}\log s_i^2 \approx \log \sigma^2$ and $\mathrm{Var}\left(\log s_i^2\right) \approx 2/(n-1)$ for n sufficiently large.[7] The critical property of $\log s_i^2$

[6] See Wu & Hamada Sections 4.11 & 4.13

[7] Wu & Hamada say $n \geq 10$.

is that the variance does not depend on the value of σ^2. Hence, we can use this to test for variance homogeneity.

The desired hypothesis test is as follows. Consider m categories with $n_1, \ldots, n_m$ observations each, which are not necessarily the same. Let σ_i^2 be the unknown variance of the ith category. Then we wish to test

$$H_0 : \sigma_1^2 = \ldots = \sigma_m^2, \qquad H_1 : \exists i, j \text{ s.t. } \sigma_i^2 \neq \sigma_j^2.$$

Sample variances can be computed as $s_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$. From here, many tests are possible.

Wu & Hamada make the unsubstantiated claim that $\log s_i^2$ has an approximate normal distribution with mean $\log \sigma_i^2$ and variance $2(n_i - 1)^{-1}$. Then, they construct a chi squared test based on this assumption.[8] Let $z_i = \log s_i^2$ and $\bar{z}$ be the mean of the z_i weighted by the observations per category. That is, $\bar{z} = \frac{1}{N-m} \sum_{i=1}^{m} (n_i - 1) z_i$. Then,

$$\sum_{i=1}^{m} \left(\frac{n_i - 1}{2} \right) (z_i - \bar{z})^2 \xrightarrow{\text{d}} \chi^2 (m - 1).$$

Hence, we can compare the test statistic to a threshold based on the chi squared distribution.

Bartlett's test for homogeneity of variance is also based around $\log s_i^2$. However, it is derived from the likelihood ratio test.[9] Let $N = \sum_{i=1}^{m} n_i$ be the total sample size. The test statistic for Bartlett's method is

$$B = \frac{(N - m) \log \bar{s}^2 - \sum_{i=1}^{m} (n_i - 1) \log s_i^2}{1 + \frac{1}{3(m-1)} \left[\sum_{i=1}^{m} (n_i - 1)^{-1} - (N - m)^{-1} \right]}.$$

The statistic $B \xrightarrow{\text{d}} \chi^2 (m - 1)$, which is a consequence of Wilks' Theorem that says that the log likelihood ratio converges in distribution to a chi squared distribution. This test can be performed in `R` by the function `bartlett.test()`. Note that as Bartlett's test is based on the likelihood ratio, it is dependent on the assumption of normality. If the data is non-normal, then other methods should be considered.

A family of tests for variance homogeneity which contains many standard approaches can be found in *On the Admissibility and Consistency of Tests for Homogeneity of Variances*, Cohen & Marden (1989).

4.1.5 Blocking with Factorial Design

When blocking in a full factorial design without replication, we are forced to sacrifice the ability to estimate one or more of the interaction effects. That is, if we have

[8] The approximate normality of $\log s_i^2$ is demonstrated by Bartlett & Kendall (1946), *The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation*, where they demonstrate that the higher order cumulants of $\log s_i^2$ decay quickly to zero as $n \to \infty$. Note that the normal distribution is the only distribution with zero cumulants of all orders greater than two.

[9] See Snedecor & Cochran, *Statistical Methods*.

a 2^k factorial design and a desired blocking that splits the data into two blocks of equal size 2^{k-1}, Then, one of the interaction terms, say 1268, will always take on a value of + in one block and - in the others. Meanwhile, all other effects will be balanced. This is referred to as *confounding*.

As an example, consider the 2^3 design,

	Treatments			Interactions				
	1	2	3	12	13	23	123	Block
1	.	.	.	+	+	+	.	
2	.	.	+	+	.	.	+	
3	.	+	.	.	+	.	+	
4	.	+	+	.	.	+	.	
5	+	.	.	.	.	+	+	
6	+	.	+	.	+	.	.	
7	+	+	.	+	.	.	.	
8	+	+	+	+	+	+	+	

Any blocking scheme filled in on the right that is *balanced*, so that each on the levels of 1,2,3 are considered equal number of times, has to coincide with one of the interaction terms.

For a single blocking factor, the usual approach is to confound it with the highest order interaction term as, given the hierarchy assumption, it is usually one of the the least important effects to consider.

More complications arise when you block more than once. Two separate blocks that split the data in half can be combined into one block that splits the data into four pieces. Then we can consider three different separations: $B_1 = -/+$, $B_2 = -/+$, and $B_1B_2 = -/+$. In this case, we actually have a mathematical group structure. Specifically, $(\{0,1\}^k, \oplus)$ is a finite albelian group and we can choose a few elements (blocks) which will generates a subgroup. For example, if block one confounds 123–we write $B_1 = 123$–and if $B_2 = 12$, then $B_1B_2 = (12)(123) = 3$.[10] This implies that we also confound the main effect of factor 3, which is generally not desirable.

To determine whether one blocking scheme is better than another with respect to confounded interactions, we say that one blocking has less *aberration* than another if the smallest order in which the number of confounded effects differs is smaller in the first blocking scheme than the second. That is, for a blocking $\mathbf{B^1} = (B_1^1, \ldots, B_m^1)$ and $\mathbf{B^2} = (B_1^2, \ldots, B_l^2)$, we can define the function $\psi : \mathbb{B} \times \{1, \ldots, k\} \to \mathbb{Z}$, where $\mathcal{B}$ is the space of blocking schemes, by

$$\psi(\mathbf{B^1}, i) = \text{The number of confounded factors of order } i \text{ by block } \mathbf{B^1}.$$

[10]Here, we are effectively xoring the columns 12 and 123 of the above table to get column 3. This group is actually known as the Klein-4 group https://en.wikipedia.org/wiki/Klein_four-group. In general, these groups will all be Boolean groups as they are products of copies of the cycle group $\mathbb{Z}/2\mathbb{Z}$.

Then, find the smallest i such that $\psi(\mathbf{B^1}, i) \neq \psi(\mathbf{B^2}, i)$. If $\psi(\mathbf{B^1}, i) < \psi(\mathbf{B^2}, i)$, then we say that $\mathbf{B^1}$ has less aberration than $\mathbf{B^2}$. If $\mathbf{B^1}$ has no more aberration than any other blocking scheme, then it is said to be a *minimal aberration blocking scheme.*

4.2 Fractional Factorial Design

Often, it is not practical to consider a full 2^k factorial experiment as k becomes larger as the number of observations required, even before replication, grows exponentially. Hence, we need to consider the consequences of testing 2^k treatments with only 2^{k-p} observations.

As an example, consider the 2^3 design from the beginning of this chapter, but with only 2^{3-1} treatments considered:

Main			Interaction				
A	B	C	AB	AC	BC	ABC	Data
							~~y_1~~
							~~y_2~~
-	+	-	-	+	-	+	y_3
-	+	+	-	-	+	-	y_4
+	-	-	-	-	+	+	y_5
+	-	+	-	+	-	-	y_6
							~~y_7~~
							~~y_8~~

Notice that the interaction AB is only considered at the - level and thus cannot be estimated. Notice further that the main effects of A and B are such that

$$ME(A) = \bar{y}(A+) - \bar{y}(A-) = \bar{y}(B-) - \bar{y}(B+) = -ME(B).$$

Hence, we cannot separate the estimation of the A and B main effects. This is known as *aliasing.* Similarly, C and ABC are aliased, and AC and BC are aliased. Hence, we only have three independent terms to estimate, which is reasonable as the total sum of squares will have $2^2 - 1 = 3$ degrees of freedom.

In general, we can define an aliasing relation by writing $\mathbf{I} = AB$, which is when column A and column B are xored together, we get the identity element. This is sometimes referred to as the *defining relation.* In this case, we can only estimate the effect of A if the effect of B is negligible and vice versa. This is an extreme case, however. Often, if the defining relation is $\mathbf{I} = ABCDE$, then each main effect is aliased with a four-way interaction effect. If all four-way interactions are negligible, as they often our assuming the hierarchy principle, then we can still estimate the main effects without interference.

Generally, prior knowledge is necessary to guess which effects will be likely to be important and which are negligible and can be aliased. There are some properties to consider when choosing a design as will be discussed in the following subsection.

Example

Continuing with the same example from before, we assume instead of conducting a 2^4 full factorial experiment, that we conduct a 2^{4-1} fractional factorial experiment with the relation $\mathbf{I} = ABCD$. Hence, each of the two factor effects is aliased with another two factor effect. Running the eight data points through `R` gives

Effect	A/BCD	B/ACD	C/ABD	D/ABC	AB/CD	AC/BD	BC/AD
Truth	·	·	1	2	1	·	-1
Estimate	-0.465	0.716	1.50	1.49	0.971	-0.05	-1.11
1000 reps	-0.500	0.499	1.498	1.5	0.995	-0.001	-0.993

The third row gives the results of replicating this experiment 1000 times and averaging the results. Certainly, there are issues with the aliasing.

4.2.1 How to choose a design

First, we will require some terminology. A main or two-factor effect is referred to as `clear` if none of its aliased effects are main or two-factor effects. It is referred to as *strongly clear* if none of its aliased effects are one, two, or three way effects. If we have a 2^{k-q} design, then we will have q defining relations that will generate a group. This is called the *defining contrast subgroup*, which consists of 2^q elements being the $2^q - 1$ strings of factors and identity element.

Example 4.2.1. *Considering the* 2^{5-2} *design, if we have the relations* $\mathbf{I} = 123$ *and* $\mathbf{I} = 345$*, then we would also have*

$$\mathbf{I} = (123)(345) = 1245.$$

Thus, we have the group with four elements $\{\mathbf{I}, 123, 345, 1245\}$.

Given the defining contrast subgroup, the *resolution* is the smallest length of any of the elements. In the above example, the resolution would be 3. One criterion for design selection is to choose the design 2^{k-q} design with the largest resolution. This is known as the *maximal resolution* criterion. This is based around the idea of hierarchy, which is that the lower order effects are probably more significant, so we do not want to lose the ability to estimate those effects. A design of resolution r is such that effects involving i factors cannot be aliased with any effect with less than $R - i$ factors. For example, resolution 4 implies that the main and two-way effects are not aliased, which is generally desirable.

Similar to blocking in the full factorial setting, we can also consider aberration between two 2^{k-q} designs and the minimal aberration design. The development here is completely analogous to that for blocking.

Follow up testing and de-aliasing

From the above example, if we have a 2^{4-1} design with $\mathbf{I} = ABCD$, then we cannot differentiate between the significance of two aliased terms. If the significance is between C and ABD, then we often yield to the hierarchy principle and choose C and the significant effect. But how do we decide between AB and CD?

In general, if we have a 2^{k-q} design, then the *fold-over* technique requires us to run a second but slightly different 2^{k-q} design in order to *augment* the first. In the above example, this would merely be considering the other 2^{4-1} design with $\mathbf{I} = -ABCD$, which would be the full factorial design when combined with the original data.

In a more interesting setting, consider the 2^{5-2} design and assume the defining relations

$$\mathbf{I} = 123 = 145 = 2345.$$

In this case, 1 is aliased with 23 and with 45 while 2,3,4,5 are each aliased with a two factor interaction. The corresponding table is

	1/23/45	2/13/345	3/12/245	4/15/235	5/14/234	run
1	.	.	+	.	+	.
2	.	.	+	+	.	.
3	.	+	.	.	+	.
4	.	+	.	+	.	.
5	+	.	.	.	.	.
6	+	.	.	+	+	.
7	+	+	+	.	.	.
8	+	+	+	+	+	.

If we wished to de-alias 1 with 23 and 45, then we could rerun the same design but flip the sign of all of the entries in column 1. Namely, we change the relation $1 = 23 = 45$ into $-1 = 23 = 45$. The result is

	-1	2/-13/345	3/-12/245	4/-15/235	5/-14/234	run
1	+	.	+	.	+	+
2	+	.	+	+	.	+
3	+	+	.	.	+	+
4	+	+	.	+	.	+
5	.	.	.	.	.	+
6	.	.	.	+	+	+
7	.	+	+	.	.	+
8	.	+	+	+	+	+

Combining these two tables into one 2^{5-1} design results in de-aliasing factor 1 from the two factor interactions. Similarly, we also de-alias factors 2,3,4,5 from the two

factor interactions. As this negation removes the relations $\mathbf{I} = 123$ and $\mathbf{I} = 145$, we are left with only $\mathbf{I} = 2345$. Thus, main effects 2,3,4,5 are only aliased with three-way terms and 1 is only aliased with the five way term 12345.

Note that as we have performed a second 2^{5-2} design, we could add an additional factor in the place of "run" in the tables to test a 2^{6-2} design. We could alternatively use this term as a blocking factor to compare the first and second runs.

Measures of Optimality

Perhaps we do not wish to test a second set of 2^{k-q} treatments either because it is too costly and impractical or because we only have a few aliased terms that require de-aliasing. We could consider adding on a few more rows to our design. Note that the aliases derived from the defining relations, $\mathbf{I} = 123 = 145 = 2345$, are

$$\begin{array}{lll} 1 = 23 = 45 = 12345, & 2 = 13 = 1245 = 345, & 3 = 12 = 1345 = 245 \\ 4 = 1234 = 15 = 235, & 5 = 1235 = 14 = 234, & 24 = 134 = 125 = 35, \\ 25 = 135 = 124 = 34, & I = 123 = 145 = 2345 & \end{array}$$

Considering the 2^{5-2} design from before, we can consider adding two more treatments such as

	1	2	3	4	5	24	25	run
1	.	.	+	.	+	+	.	.
2	.	.	+	+	.	.	+	.
3	.	+	.	.	+	.	+	.
4	.	+	.	+	.	+	.	.
5	+	.	.	.	.	+	+	.
6	+	.	.	+	+	.	.	.
7	+	+	+	.	.	.	.	.
8	+	+	+	+	+	+	+	.
9	.	.	.	.	.	+	+	+
10	.	+	.	.	+	.	+	+

As each of the factors 1,...,5 can be considered at the levels $\{-, +\}$, we have 32 possible treatments minus the 8 already considered for each. This results in $\binom{24}{2} = 276$ possibilities. Note that the *run* column allows us to block the first set of treatments and the second for comparison if desired.

The obvious question is, which treatments are optimal to include? A collection of different methods is detailed on Wikipedia.[11] In Wu & Hamada, they consider *D-optimality* which seeks to maximize the determinant of $X_d^{\mathrm{T}} X_d$ where X_d is the $(2^{k-q} + t) \times k$ matrix with t being the additional number of treatments considered. The two treatments presented in the above table were selected by this criteria. Note that the choice of treatments is not necessarily unique.

[11] https://en.wikipedia.org/wiki/Optimal_design

A list of possible optimality criteria is presented in the following table.

Name	Formula	Intuition
D-optimality	$\max\{X_d^{\mathrm{T}}X_d\}$	Minimize volume of confidence ellipsoid
A-optimality	$\min\{\mathrm{trace}([X_d^{\mathrm{T}}X_d]^{-1})\}$	Minimize $\mathrm{Var}\,(\beta) = \sigma^2(X_d^{\mathrm{T}}X_d)^{-1}$
T-optimality	$\max\{\mathrm{trace}(X_d^{\mathrm{T}}X_d)\}$	Similar to A-optimality
G-optimality	$\min\max_i[X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}]_{i,i}$	Minimize maximal variance of predictions

Instead of optimizing globally, we may also be interested in de-aliasing two specific effects from one another. To achieve this, consider writing the design matrix X_d in block form as

$$X_d = (X_1 \; X_2)$$

where X_2 corresponds to the $(2^{k-q}+t)\times 2$ matrix with two columns corresponding to two effects to be de-aliased. Consequently,

$$X_d^{\mathrm{T}}X_d = \begin{pmatrix} X_1^{\mathrm{T}}X_1 & X_1^{\mathrm{T}}X_2 \\ X_2^{\mathrm{T}}X_1 & X_2^{\mathrm{T}}X_2 \end{pmatrix}$$

and using the formula for inverting a block matrix[12] gives

$$(X_d^{\mathrm{T}}X_d)^{-1} = \begin{pmatrix} \star & \star \\ \star & (X_2^{\mathrm{T}}X_2 - X_2^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}X_2)^{-1} \end{pmatrix}$$

Hence, we can optimize over two specific effects, or alternatively a general subset of more than two effects, by choosing the new treatments to optimize

$$\max\{X_2^{\mathrm{T}}X_2 - X_2^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}X_2\}.$$

This is referred to as D_s-optimality. It is similar to how D-optimality looks for the design that optimizes

$$\max\{X_d^{\mathrm{T}}X_d\} = \min\{(X_d^{\mathrm{T}}X_d)^{-1}\}.$$

4.3 3^k Factorial Designs[13]

In some cases, we may be interested in testing three instead of two factor levels. Hence, in this section, we will replace $\{-,+\}$ with $\{0,1,2\}$ to denote the levels that each of the k factors can take. In some sense, very little has changed from the previous 2^k setting. The main difference it added complexity with respect to testing effects and aliasing when considering fractional designs.

[12] https://en.wikipedia.org/wiki/Block_matrix#Block_matrix_inversion
[13] See Wu & Hamada Section 6.3

When we considered a 2^2 design with factors A and B, we were able to estimate the main effects of A and B as well as the interaction term $A \times B$. In that setting, all of these terms have one degree of freedom.

The 3^2 analogue is slightly different. We can still consider the two main and one interaction effect. However, now the main effects have 2 degrees of freedom while the interaction effect has $4 = (3-1)^2$. The main effect of factor A considers the differences among the three levels of $A = 0, 1, 2$. Writing the interaction term as $A \times B$ can be a bit misleading as it is actually comparing the responses at

$$A + B = 0, 1, 2 \mod 3$$

and

$$A + 2B = 0, 1, 2 \mod 3,$$

which are displayed in the following tables.

A+B		A		
		0	1	2
	0	0	1	2
B	1	1	2	0
	2	2	0	1

A+2B		A		
		0	1	2
	0	0	1	2
B	1	2	0	1
	2	1	2	0

The above two tables are 3×3 Latin squares. Furthermore, replacing $(0, 1, 2)$ on the left with (α, β, γ), we can see that they are orthogonal Latin squares which can be combined into

$$\begin{pmatrix} \alpha 0 & \beta 1 & \gamma 2 \\ \beta 2 & \gamma 0 & \alpha 1 \\ \gamma 1 & \alpha 2 & \beta 0 \end{pmatrix}$$

Thus, the interaction term can be thought of as a Graeco-Latin square testing two treatments referring to the above two equations.

As a consequence of this, the sum of squares for $A \times B$ can be further decomposed into sums of squares for two terms denoted AB and AB^2 corresponding, respectively, to $A+B$ and $A+2B$ mod 3. Recalling the Latin squares design, each of these effects has $3 - 1 = 2$ degrees of freedom.

These ideas can be extended in the 3^3 design to the three way interaction term $A \times B \times C$. This term will have 8 degrees of freedom and can be decomposed into four effects with 2 DoFs.

$$\begin{aligned} ABC: &\quad A + B + C = 0, 1, 2 \mod 3 \\ ABC^2: &\quad A + B + 2C = 0, 1, 2 \mod 3 \\ AB^2C: &\quad A + 2B + C = 0, 1, 2 \mod 3 \\ AB^2C^2: &\quad A + 2B + 2C = 0, 1, 2 \mod 3 \end{aligned}$$

Note that we set the coefficient of the first factor equal to 1. Otherwise, we will have repeats as, for example,

$$2A + B + 2C = 2(A + 2B + C) = 0, 1, 2 \mod 3$$

showing that A^2BC^2 is the same effect as AB^2C.

This method of decomposing the interaction terms is referred to as the *orthogonal component system*. This is because all of the terms in the decomposition are orthogonal, which is always a desirable property. In a 3^k design, we have $3^k - 1$ degrees of freedom and each of these terms requires 2. Hence, there are $(3^k - 1)/2$ effects to estimate.

Remark 4.3.1. *Even though we can estimate all of these effects, they may quite hard to interpret in practise. If an F-test or Lenth's method implies significance for AB^2C^2, what does that mean?*

4.3.1 Linear and Quadratic Contrasts[14]

One benefit to the 3^k design over the 2^k design is the ability to estimate quadratic effects. That is, when we only have two levels $\{-, +\}$ to consider, the best we can do to claim that one results in larger responses than the other. However, in the 3^k setting, we can look for increases and decreases and even use such models to find an optimal value for a given factor with respect to minimizing or maximizing the response.

The main effect of factor A in a 3^k design has two degrees of freedom and compares the difference among the responses for $A = 0, 1,$ and 2. However, we can decompose this main effect into a linear and quadratic contrast with

$$A_l = (-1, 0, 1)/\sqrt{2}, \text{ and } A_q = (1, -2, 1)/\sqrt{6}.$$

Let $\bar{y}_{A0}, \bar{y}_{A1}, \bar{y}_{A2}$ denote the average–i.e. averaged over all other factor levels– response for $A = 0, 1, 2$, respectively. Then, the linear contrast considers $\bar{y}_{A2} - \bar{y}_{A0}$. If this value is significantly removed from zero, then there is evidence of an increase or decrease between the two extreme values of $A = 0$ and $A = 2$. This is, in fact, just the combination of two linear contrasts

$$\bar{y}_{A2} - \bar{y}_{A0} = (\bar{y}_{A2} - \bar{y}_{A1}) - (\bar{y}_{A1} - \bar{y}_{A0}).$$

The quadratic contrast is orthogonal to the linear contrast. It tests for the case where $\bar{y}_{A1}$ is either significantly larger or smaller than $\bar{y}_{A0} + \bar{y}_{A2}$. Each of these contrasts has 1 degree of freedom.

[14] See Wu & Hamada Section 6.6

Now when considering an interaction term like $A \times B$, we can forego testing AB and AB^2 and instead test the four terms $(AB)_{ll}, (AB)_{lq}, (AB)_{ql}, (AB)_{qq}$ each with one degree of freedom. These contrasts can be written as

$$
\begin{aligned}
(AB)_{ll}: &\quad \frac{1}{2}[(y_{22} - y_{20}) - (y_{02} - y_{00})] = \frac{1}{2}[(y_{22} - y_{02}) - (y_{20} - y_{00})] \\
(AB)_{lq}: &\quad \frac{1}{2\sqrt{3}}[(y_{22} - 2y_{12} + y_{02}) - (y_{20} - 2y_{10} + y_{00})] \\
(AB)_{ql}: &\quad \frac{1}{2\sqrt{3}}[(y_{22} - 2y_{21} + y_{20}) - (y_{02} - 2y_{01} + y_{00})] \\
(AB)_{qq}: &\quad \frac{1}{6}[(y_{22} - 2y_{21} + y_{20}) - 2(y_{12} - 2y_{11} + y_{10}) + (y_{02} - 2y_{01} + y_{00})]
\end{aligned}
$$

Compare the above formulae to the following table to understand what they are testing.

y_{02}	y_{12}	y_{22}
y_{01}	y_{11}	y_{21}
y_{00}	y_{10}	y_{20}

The linear-linear contrast is looking for a difference in the linear contrasts of A conditional on $B = 0$ or 2. This is equivalent to looking for a difference in the linear contrasts of B conditional on $A = 0$ or 2. A significant value could indicate, for example, that the response is increasing in A when $B = 0$ but decreasing in A when $B = 2$.

The linear-quadratic contrasts are similar in interpretation. $(AB)_{ql}$ tests for differences between the quadratic contrast of A when $B = 0$ and when $B = 2$. Similarly, $(AB)_{lq}$ tests the same thing with the roles of A and B reversed. The quadratic-quadratic term is a bit harder to interpret. It is effectively looking for quadratic changes in the quadratic contrasts of one factor conditional on another.

4.3.2 3^{k-q} Fractional Designs[15]

Even more so than in the 2^k factorial design, the number of required observations for a 3^k design grows quite rapidly. As a result, 3^{k-q} fractional factorial designs are often preferred when $k \geq 4$.

Similarly to the 2^k setting, we need to define a defining relation such as, for a 3^4 design,

$$AB^2C^2D = \mathbf{I}.$$

However, our modular arithmetic is now performed in mod 3 instead of mod 2. Hence, the above relationship is equivalent to

$$A + 2B + 2C + D = 0 \mod 3.$$

[15] See Wu & Hamada Section 6.4

This equation can be used to derive all of the alias relations, but doing so is quite tedious. For example, Adding A to both sides of the above equation gives

$$\begin{aligned} A &= 2A + 2B + 2C + D \mod 3 \\ &= 2(A + B + C + 2D) \mod 3 \end{aligned}$$

and adding $2A$ gives

$$2A = 0 + 2B + 2C + d \mod 3.$$

Hence, we have $A = ABCD^2 = B^2C^2D$. This process can be continued for the other factors in order to construct the 13 aliasing relations.

In general, a single defining relation with take a 3^k to a 3^{k-1} design. The number of orthogonal terms is reduced from

$$\frac{3^k - 1}{2} \Rightarrow \frac{(3^k - 1)/2 - 1}{3} = \frac{3^{k-1} - 1}{2}.$$

Hence, we lose the defining the relation, and the remaining effects are aliased into groups of three.

If we have two defining relations like

$$\mathbf{I} = AB^2C^2D = ABCD,$$

then we will also have an additional two relations. This can be seen by sequentially adding the defining relations. Including $\mathbf{I} = AB^2C^2D$ implies that all remaining terms are aliased in groups of 3. Hence, $ABCD$ is aliased with

$$(ABCD)(AB^2C^2D) = A^2B^3C^3D^2 = AD$$

and

$$(A^2B^2C^2D^2)(AB^2C^2D) = A^3B^4C^4D^3 = BC.$$

Hence, including a second defining relation $\mathbf{I} = ABCD$ immediately includes AD and BC as well. Thus, our defining contrast subgroup is

$$\mathbf{I} = AD = BC = ABCD = AB^2C^2D.$$

We can count effects to make sure all are accounted. In a 3^4 design, we have $(3^4 - 1)/2 = 40$ effects. In a 3^{4-1} design, we have $(40 - 1)/3 = 13$ aliased groups with 3 effects each resulting in $13 \times 3 = 39$ plus 1 for the defining relation gives 40. In a 3^{4-2} design, we have $(13 - 1)/3 = 4$ aliased groups with 9 effects each resulting in $4 \times 9 = 36$ plus the 4 defining relations gives 40 again.

Partial Aliasing

The above aliasing relations apply to the orthogonal components system, but what is the effect of such aliasing on the linear and quadratic contrasts?

If a design has resolution 5, then all of the main and two-way effects are not aliased with any other main or two-way effects. Thus, their linear and quadratic contrasts are also free of aliasing with each other. Note, however, that they can still be aliased with higher order terms.

For more complex aliasing relationships, consider the 3^{3-1} design with defining relation $\mathbf{I} = ABC$. Then we have the following four aliased groups,

$$\begin{array}{cccccc} 1 & A & = & BC & = & AB^2C^2 \\ 2 & B & = & AC & = & AB^2C \\ 3 & C & = & AB & = & ABC^2 \\ 4 & AB^2 & = & AC^2 & = & BC^2 \end{array}$$

and a design matrix that has 9 rows such that $A + B + C = 0 \mod 3$

	A	B	C	AB^2
1	0	0	0	0
2	0	1	2	2
3	0	2	1	1
4	1	2	0	2
5	1	0	2	1
6	1	1	1	0
7	2	1	0	1
8	2	0	1	2
9	2	2	2	0

Each of these four terms comes with 2 degrees of freedom.

We can furthermore decompose the above into a design matrix of linear and quadratic contrasts by mapping $(0, 1, 2) \to (-1, 0, 1)$ in the linear case and mapping $(0, 1, 2) \to (1, -2, 1)$ in the quadratic case. The interaction term columns are filled in with the product mod 3 of the corresponding main effects columns.

	A_l	A_q	B_l	B_q	C_l	C_q	$(AB)_{ll}$	$(AB)_{lq}$	$(AB)_{ql}$	$(AB)_{qq}$
1	-1	1	-1	1	-1	1	1	-1	-1	1
2	-1	1	0	-2	1	1	0	2	0	-2
3	-1	1	1	1	0	-2	-1	-1	1	1
4	0	-2	1	1	-1	1	0	0	-2	-2
5	0	-2	-1	1	1	1	0	0	2	-2
6	0	-2	0	-2	0	-2	0	0	0	1
7	1	1	0	-2	-1	1	0	-2	0	-2
8	1	1	-1	1	0	-2	-1	1	-1	1
9	1	1	1	1	1	1	1	1	1	1

However, as we only have 2 degrees of freedom from AB^2, we must select two of the four to estimate in practice. Note that if you add (mod 3) the columns of, for example, A_l and A_q that you will get the above column for A.

We know from before that C is aliased with AB but what about the relationships between C_l, C_q, and $(AB)_{ll}$,$(AB)_{lq}$, $(AB)_{ql}$,$(AB)_{qq}$? We can compute the correlation between each pair of vectors[16] to get

	C_l	C_q	$(AB)_{ll}$	$(AB)_{lq}$	$(AB)_{ql}$	$(AB)_{qq}$
C_l	1	.	.	$\sqrt{1/2}$	$\sqrt{1/2}$	.
C_q	.	1	$\sqrt{1/2}$	.	.	$-\sqrt{3/5}$
$(AB)_{ll}$	.	$\sqrt{1/2}$	1	.	.	.
$(AB)_{lq}$	$\sqrt{1/2}$	.	.	1	.	.
$(AB)_{ql}$	$\sqrt{1/2}$	.	.	.	1	.
$(AB)_{qq}$	.	$-\sqrt{3/5}$	.	.	.	1

Hence, none of the columns coincide implying that we can estimate the terms simultaneously. However, the columns are not orthogonal meaning that the estimation is not as efficient as if they were orthogonal. This is the partial aliasing situation. In Wu & Hamada, they recommend using a stepwise regression approach on the entire set of 6 main linear and quadratic effects and 18 two-way interactions.

4.3.3 Agricultural Example

In the `agridat` library, there is the dataset `chinloy.fractionalfactorial`, which considers a 3^{5-1} fractional factorial design with an additional blocking variable. The five factors considered are denoted N, P, K, B, and F.[17]

We first run the commands

```
dat <- chinloy.fractionalfactorial
dat <- transform(dat,N=factor(N), P=factor(P),
   K=factor(K), B=factor(B), F=factor(F))
```

to tell R that the levels of N,P,K,B,F are to be treated as factors.

If we were to ignore the aliasing and just include as many terms as possible with

```
md1 = aov( yield~(N+P+K+B+F)^5,data=dat )
```

then we would get sums of squares for the 5 main effects, the 10 two-way effects, and only 6 of the three-way effects. However, an additional peculiarity is that $K \times B$, $K \times F$, and $B \times F$ all have only two degrees of freedom instead of four. Also,

[16] Correlation of x and y is $\frac{\langle x,y \rangle}{\|x\|\|y\|}$

[17] The original paper associated with this dataset can be found at https://doi.org/10.1017/S0021859600044567

the corresponding three-way terms with N attached only have four instead of eight degrees of freedom.

The defining relation for the aliasing in this experiment is $\mathbf{I} = PK^2B^2F$ or, in modular arithmetic form,

$$P + 2K + 2B + F = 0 \mod 3.$$

This results in the following aliased relations for the main effects,

N	$=$	NPK^2B^2F	$=$	NP^2KBF^2	P	$=$	$PKBF^2$	$=$	KBF^2
K	$=$	PB^2F	$=$	PKB^2F	B	$=$	PK^2F	$=$	PK^2BF
F	$=$	$PK^2B^2F^2$	$=$	PK^2B^2					

For the second order effects, N does not occur in the defining relation and thus all Nx terms are aliased with 4th and 5th order interactions. For the terms involving P, we have

PK	$=$	PBF^2	$=$	KB^2F	PK^2	$=$	$PKBF^2$	$=$	BF^2
PB	$=$	PKF^2	$=$	KB^2F^2	PB^2	$=$	PB^2KF^2	$=$	KF^2
PF	$=$	$PKFB$	$=$	KB	PF^2	$=$	PKB	$=$	KBF

We can see that BF^2, KB, and KF^2 are each aliased with a two-way term involving P. Hence, the general interactions $B \times F$, $K \times B$, and $K \times F$ each lose 2 degrees of freedom in the ANOVA table.

In the original study, the authors were only concerned with main and two-way interaction effects and chose to save the remaining degrees of freedom to estimate the residual sum of squares. Ignoring the blocking factor for the moment, we can test for significant effects by

```
md2 = aov( yield~(N+P+K+B+F)^2,data=dat )
```

This results in N, P, B, F, and $P \times F$ all significant at the 0.05 level–without any multiple testing correction, that is.

If we set the contrasts to `contr.poly`, then we can look at the linear and quadratic effects. This can be achieved in `R` by adding the argument

```
split=list(P=list(L=1,Q=2),F=list(L=1,Q=2))
```

to the `summary()` command. For the $P \times F$ term, we get from the ANOVA table that

		DoF	Sum Sq	Mean Sq	F value	Pr(>F)	
P:F		4	6.109	1.527	2.674	0.047424	*
	L.L	1	5.816	5.816	10.182	0.002938	**
	Q.L	1	0.197	0.197	0.344	0.560909	
	L.Q	1	0.000	0.000	0.000	0.994956	
	Q.Q	1	0.096	0.096	0.169	0.683578	

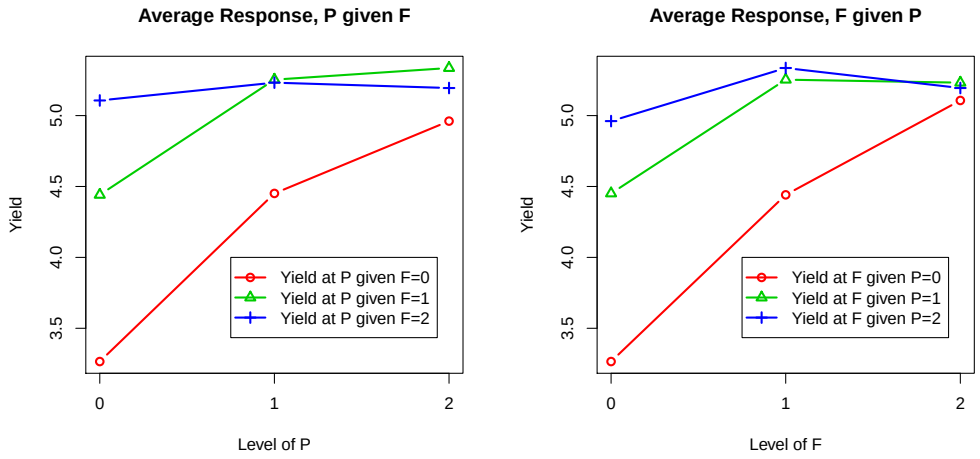

Figure 4.3: Plots of the interaction between factors P and F. In the left plot, we see changes in P conditional on F. In the right plot, we see the reverse conditioning.

This indicates that the significance of the interaction between P and F is contained in the linear-linear contrast. We can look directly at the 3×3 matrix of average responses across different levels of P and F to see the interaction behaviour.

P\F	0	1	2
0	3.27	4.44	5.11
1	4.45	5.25	5.23
2	4.96	5.36	5.19

The rows and columns of this table are plotted in Figure 4.3.

Considering all of the main and two-way effects, we can use the `model.matrix()` command to extract the design matrix. We can further use the `cor()` command to compute the correlations between the terms. In Figure 4.4, we can see that some of the polynomial contrasts have non-zero correlation. Running a backwards variable selection via the `step()` function reduces the model containing all two-way interactions to just

```
yield ~ N + P + K + B + F + N:P + P:F
```

One can check that the polynomial contrasts in this reduced model are all orthogonal. Note that this does not necessarily have to happen.

This dataset also contains a blocking factor which has 9 levels. If we add it into the model and try to model all two-way interactions with

```
md3 = aov( yield~block+(N+P+K+B+F)^2,data=dat ),
```

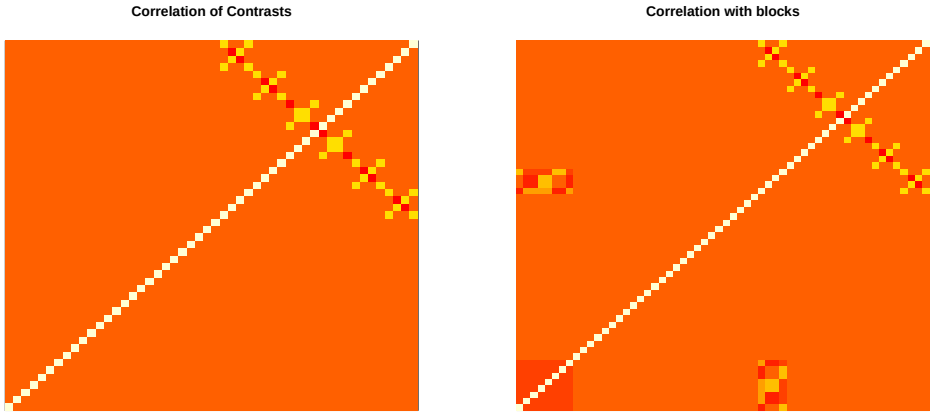

Figure 4.4: Correlations of the polynomial contrasts (Left) and correlations with blocks (Right). Most are zero. However, some of the two-way contrasts have non-zero correlation.

then we see that the residual degrees of freedom has reduced from 36 to 30 and that the degrees of freedom for $P \times K$ has dropped from 4 to 2, which accounts for the 8 DoFs required for the blocking factor. Including the three-way interactions in our model allows us to use `R` see that we lose 2 DoFs from each of $P \times K$, $N \times P \times B$, $N \times B \times F$, $N \times P \times F$ by running the command

```
summary(aov( yield~Error(block)+(N+P+K+B+F)^5,data=dat )).
```

Two confounded factors by the blocking relations as detailed in Chinloy et al (1953) are PK and NPB^2. From here, we can construct the remaining confounded factors using. Combining these two factors gives two more

$$\begin{aligned} (PK)(NPB^2) &= NP^2KB^2 \\ (P^2K^2)(NPB^2) &= NK^2B^2. \end{aligned}$$

Finally, we have to consider the two other terms aliased with the above four by the relation $\mathbf{I} = PK^2B^2F$.

PK	$=$	PBF^2	$=$	KB^2F
NPB^2	$=$	$NP^2K^2B^2F$	$=$	NKB^2F^2
NP^2KB^2	$=$	NB^2F	$=$	$NPK^2B^2F^2$
NK^2B^2	$=$	$NPKBF$	$=$	NPF

Thus, the two DoFs that we see being taken away come from the interactions PK, NPB^2, NB^2F, and NPF. In the correlation plot of Figure 4.4, we have correlation between the $P \times K$ polynomial terms and the block factor.

If we similarly apply backwards variable selection to the model with the blocking variable included, then the result is that more interaction terms remain. However, the polynomial contrasts are still orthogonal to one another.

```
yield ~ block + N + P + K + B + F +
                          N:P + N:K + N:B + N:F + P:B + P:F + B:F.
```

Problem of reordering your terms

As a result of the partial aliasing, we miss significant results if we were to fit the effects in a different order. To see this, running

```
md2 = aov( yield~(N+P+K+B+F)^2,data=dat )
```

as above yields an ANOVA table with a slightly significant `P:F` term with 4 degrees of freedom and, subsequently, a $(PF)_{ll}$ term that is very significant.

If instead, we ran

```
md2 = aov( yield~(N+B+P+K+F)^2,data=dat ),
```

then the `P:F` term will only have 2 degrees of freedom and a p-value of 0.147. The polynomial contrasts will no longer be significant.

		DoF	Sum Sq	Mean Sq	F value	Pr(>F)	
P:F		2	2.308	1.154	2.020	0.147	
	L.L	1	2.207	2.207	3.864	0.057	.
	Q.L	1	0.101	0.101	0.176	0.677	
	L.Q	1					
	Q.Q	1					

We can, however, solve this problem using the `R` function `step()`, which will result in the same submodel as before retaining the term $P : F$ but not the term $B : K$. The original 1953 paper on this dataset came before the invention of AIC. In that case, they directly identify that the mean sum of squares for the aliased $PF = KB$ term is larger than the rest and then look at each interaction more closely to determine which is significant.

Chapter 5

Response Surface Methodology

Introduction

In this chapter, we consider the regression setting where given k input variables, we are interested in modelling the response variable y as

$$y = f(x_1, \ldots, x_k) + \varepsilon$$

where f is some function to be specified and ε is iid mean zero variance σ^2 noise. Here, the variables x_i will be treated as *coded* factors. If the factor takes on three levels, then it is generally mapped to $\{-1, 0, 1\}$. If it takes on five levels, it is often mapped to $\{-\alpha, -1, 0, 1, \alpha\}$ for some $\alpha > 1$. For example, if an experiment were taking place in an oven, the temperature factor of $\{325, 350, 375\} \to \{-1, 0, 1\}$. If we wanted five levels, we could choose $\alpha = 1.6$ and consider temperatures $\{310, 325, 350, 375, 390\} \to \{-1.6, -1, 0, 1, 1.6\}$.

The goal of response surface methods is to model the function $f(\cdot)$ as a polynomial surface. While higher orders are possible, we will consider only first and second order. These tools can be used to choose treatments sequentially with the goal of optimizing with respect to the response. In general, an experimenter can use a first order model to get a local linear approximation of $f(\cdot)$ which provides an estimate of the slope of the surface. Second order models can be used to test the local curvature of the surface. Before proceeding with such an optimization, it is often wise to run a factorial or other design for the purposes of variable selection as optimization for large values of k can be quite costly and time consuming.

Response surface methods are often used to improve some process in, say, and engineering context. As an example, we could try to optimize cake baking with respect to the variables temperature, baking time, and sugar content where the response y could be a score given by a panel of judges. As obtaining a single y is time consuming, we would want to search the parameter space as efficiently as possible.

Note that this is only a brief overview of the subject of response surface methods. Entire textbooks are dedicated to the subject.

5.1 First and Second Order[1]

A first order model for $f(\cdot)$ is of the form

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon,$$

which is effectively a linear regression but where the uncoded values of x_i were chosen carefully with respect to some experimental design. This could be, for example, a 2^{k-q} fractional factorial design, which, by considering different combinations of ± 1 allows for estimation of the β_i.

The second order model is of the form

$$\begin{aligned} y &= \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \varepsilon \\ &= \beta_0 + x^{\mathrm{T}} b + x^{\mathrm{T}} B x \end{aligned}$$

where $b = (\beta_1, \ldots, \beta_k)$ and B is the $k \times k$ matrix with ith diagonal entry β_{ii} and ijth off-diagonal entry $\frac{1}{2}\beta_{ij}$ with $\beta_{ij} = \beta_{ji}$ for $i > j$. In order to estimate all of the parameters in the second order model, more treatment combinations need to be considered. These will be discussed in the following section. For now, assume a design exists that allows us to estimate the β_i and the β_{ij}.

Given such a second order model fitted locally to our experimental data, we can solve for the critical or stationary point x_s where the derivative is equal to zero. Namely,

$$0 = b + 2Bx_s, \text{ or } x_s = -B^{-1}b/2$$

As B is a real symmetric matrix, the spectral theorem allows us to write B as $U^{\mathrm{T}}DU$ where U is the orthonormal matrix of eigenvectors and D is the diagonal matrix of eigenvalues $\lambda_1 \geq \ldots \geq \lambda_k$.

To understand the behaviour of the fitted value $\hat{y}$ about the critical point x_s, we can rewrite

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + x^{\mathrm{T}} b + x^{\mathrm{T}} B x \\ &= \hat{\beta}_0 + (x_s + z)^{\mathrm{T}} b + (x_s + z)^{\mathrm{T}} B (x_s + z) \\ &= \hat{\beta}_0 + {x_s}^{\mathrm{T}} b + {x_s}^{\mathrm{T}} B x_s + z^{\mathrm{T}} b + 2 z^{\mathrm{T}} B x_s + z^{\mathrm{T}} B z \\ &= \hat{y}_s + z^{\mathrm{T}} B z + 0 \end{aligned}$$

[1] See Wu & Hamada Section 10.4

as $z^{\mathrm{T}}b + 2z^{\mathrm{T}}Bx_s = z^{\mathrm{T}}(b + 2Bx_s) = 0$. Furthermore,

$$\begin{aligned}\hat{y} &= \hat{y}_s + z^{\mathrm{T}}Bz \\ &= \hat{y}_s + z^{\mathrm{T}}U^{\mathrm{T}}DUz \\ &= \hat{y}_s + v^{\mathrm{T}}Dv \\ &= \hat{y}_s + \sum_{i=1}^{k} \lambda_i v_i^2.\end{aligned}$$

Hence, the behaviour of $\hat{y}$ about the critical point is determined by the eigenvalues of the matrix B.

If all of the eigenvalues are positive, then the surface is locally elliptic and convex implying that $\hat{y}_s$ is a local minima. If all of the eigenvalues are negative, we conversely have that $\hat{y}_s$ is a local maxima. If the signs of the eigenvalues are mixed, then we have a saddle point and should continue searching for the desired optima. If one or more of the eigenvalues is zero–or very close to zero when compared to the other eigenvalues–then there will be a linear subspace of critical values spanned by the corresponding eigenvectors. For example, if all of the eigenvalues are positive except for λ_k, then we have a minimum for any input value of v_k.

5.2 Some Response Surface Designs

If we are interested in only the first order properties of the response surface–i.e. the slope–then we can use factorial, fractional factorial, or similar designs. In order to test for second order proprieties–i.e. curvature–then we will require new factor level combinations chosen carefully to allow us to estimate the β_{ij} terms.

Note that a second order model has $1 + k + \binom{k}{2} + k = (k+1)(k+2)/2$ parameters to estimate. Hence, we will need at least that many observations to estimate the parameters. However, it is often inefficient to use a model with more than the minimal required parameters. Choice in design is often made to minimize the number of required observations especially when these designs are used sequentially.

5.2.1 Central Composite Design[2]

A central composite design consists of three types of points: cube points from a fractional factorial design; central points that lie in the centre of the cube; axial points where all but one factor is at the zero level.

A theorem due to Hartley (1959) is presented in Wu & Hamada as Theorem 10.1, which is

[2] See Wu & Hamada Section 10.7

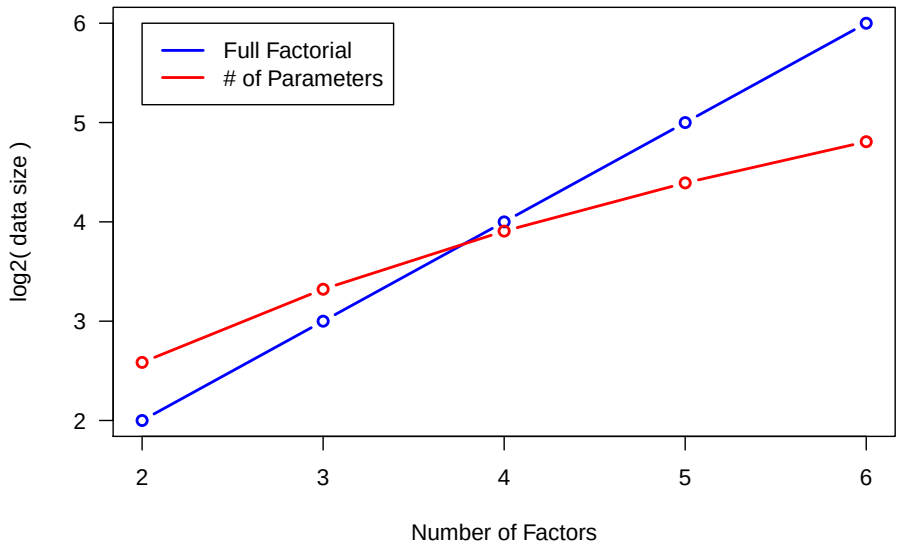

Figure 5.1: A comparison of the number of parameters in a second order model with the data size of a 2^k factorial design on the $\log_2$ scale.

Theorem 5.2.1. *For a central composite design based on a 2^{k-q} fractional factorial design such that no defining relations contain a main factor, all of the β_i and β_{ii} are estimable as well as one β_{ij} for each aliased group of factors.*

The theorem implies that if we want to estimate all of the β_{ij}, then we cannot have any two factor terms aliased with another two factor term. Hence, we must avoid defining relations of order 4. Oddly enough, defining relations of order 3 are allowed even though the alias a main effect with a two-way effect. The reason is that the axial points allow us to estimate the main effects and, hence, de-alias them from the two factor effects. A design with resolution 3 but with no defining relations of length 4 is referred to as *resolution III**.

Cube Points

The cube points are those from a fractional factorial design, which is those whose factors take on the values of ± 1 and hence lie on the corners of an hypercube. As mentioned above in the theorem, ideally we would choose a 2^{k-q} design of resolution III*. Other designs beyond fractional factorial can be considered such as the *Plackett-Burman designs*, which we will discuss in a later chapter.

Central Points

The central points are those such that all factors are set to level zero and hence, the points lie in the centre of the design. Without central points, we merely have a *composite design*. The central points allow for estimation of the second order parameters. Depending on the choice of axial points, multiple replications of the central points may be required. Wu & Hamada recommend anywhere from 1 to 5 replications of this point.

Axial Points

The axial points lie on the main axes of the design. That is, all but a single factor are set to zero and the single factor is tested at the levels $\pm\alpha$ where α is chosen in $[1, \sqrt{k}]$. The addition of axial points adds $2k$ treatments into the design.

If the value of α is chosen to be 1, then these points lie on the faces of the hypercube from the factorial design. A benefit to this is that only 3 factor levels are required–i.e. $\{-1, 0, 1\}$–instead of 5. As the shape of the design in now a cube, such a choice is also useful when optimizing over a cuboidal region, which is often the case.

If the value of α is chose to be $\sqrt{k}$, then all of these and the cube points lie the same distance from the origin, which is a distance of $\sqrt{k}$. As a result, this is sometimes referred to as spheroidal design. Values of α closer to $\sqrt{k}$ allow for more varied observations, which can help with estimation efficiency. However, when $\alpha = \sqrt{k}$, the variance estimate of $\hat{y}$ is infinite without the inclusion of central points to stabilize the calculation.

One useful design criterion is to choose a design that is *rotateable*. This occurs when the variance of the predicted value $\hat{y}$ can be written as a function of $\|x\|_2$. That is, the variance of the prediction is only dependent on how far away the inputs are from the centre of the design. Allegedly, a 2^{k-q} design with resolution 5 and $\alpha = 2^{(k-q)/4}$ is rotateable. Note that while rotateability is a nice property, it is dependent on how the factors are coded and hence is not invariant to such choices.

5.2.2 Box-Behnken Design[3]

In 1960, Box and Behnken proposed a method of combining factorial designs of Section 4.1 with balanced incomplete block designs of Section 2.4. The idea is that if each block tests k treatments, then replace each block with a 2^k factorial design where the untested factors are set to zero.

As an example, we could begin with a BIBD such as

[3] See Wu & Hamada Section 10.7

	Factor			
block	A	B	C	D
1	$y_{1,1}$	$y_{1,2}$	.	.
2	.	$y_{2,2}$	$y_{2,3}$	.
3	.	.	$y_{3,3}$	$y_{3,4}$
4	$y_{4,1}$	.	$y_{4,3}$	.
5	.	$y_{5,2}$	.	$y_{5,4}$
6	$y_{6,1}$	.	.	$y_{6,4}$

However, we would replace each block with a 2^2 factorial design. For example, the first row would become

	Factor			
block	A	B	C	D
1	1	1	0	0
	1	-1	0	0
	-1	1	0	0
	-1	-1	0	0

The total number of treatments to test would be $\#\{\text{blocks}\} \times 2^2$ or 24 in this example.

Geometrically, we are testing points on the faces of the hypercube. However, all of the points are equal in Euclidean norm and hence lie on the same sphere. This is geometrically nice as our points line on the edges of the same cube and the surface of the same sphere. Also, only three factor levels are required for each factor. One additional requirement, however, is that we will need to replicate the central point multiple times in order to compute the variance.

One more property of note is that such a design is said to be *orthogonally blocked*. This is the case where the block effects and parameter effects are orthogonal and hence do not interfere with one another.

The total number of treatments to test becomes too large as the number of factors increases. Hence, other methods are more efficient. A table of the sizes can be found in the Chapter 10 Appendix in Wu & Hamada and also on the Wikipedia page.[4]

5.2.3 Uniform Shell Design[5]

Uniform Shell Designs are due to Doehlert (1970) in a paper by the same name. The idea is that "designs are generated which have an equally spaced distribution of points lying on concentric spherical shells" and that more points can be added to fill in the entire sphere if desired. For k factors, this design considers $k^2 + k$ points

[4] https://en.wikipedia.org/wiki/Box%E2%80%93Behnken_design
[5] See Wu & Hamada Section 10.7

uniformly spaced over the surface of a sphere and a single point at the center of the sphere. The total, $k^2 + k + 1$, is roughly twice that of the minimal requirement of $(k^2 + 3k + 2)/2$ for large values of k. Hence, this method is generally only used for small values of k.

As these designs can be rotated, they are often rotated to minimize the number of factor levels required to consider. While the Box-Behnken and central composite designs with $\alpha = \sqrt{k}$ are also spherical in nature, the important feature of the uniform shell design is the fact that the points are uniformly spread across the sphere's surface. In general, spherical designs can be repeated for spheres of different radii in order to better estimate the change in the response as the inputs move away from the central point.

Remark 5.2.2. *Fun Fact: For $k = 3$, the uniform shell design coincides with the Box-Behnken design.*

5.3 Search and Optimization[6]

Repeated experiments such as those discussed above can be sequentially repeated at different factor levels in order to search the input space for an optimal–minimal or maximal–response.

5.3.1 Ascent via First Order Designs

To find the direction of greatest change in the response surface, we only require estimating the first order components of the model, so a fractional factorial design will suffice at this stage. However, in order to test for overall significance of the local curvature, we will add some central points to the design.

Specifically, consider a 2^{k-q} design with n_c central points included. Recall that the second order model is

$$\begin{aligned} y &= \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \varepsilon \\ &= \beta_0 + x^{\mathrm{T}} b + x^{\mathrm{T}} B x \end{aligned}$$

When testing a central point, we have $y = \beta_0 + \varepsilon$. Hence, the average of the n_c central points is an unbiased estimator of β_0. That is, $\mathrm{E}\bar{y}_c = \beta_0$.

In contrast, consider the 2^{k-q} points from the factorial design. The average of

[6] See Wu & Hamada Section 10.3

those points, denoted $\bar{y}_f$, is

$$\begin{aligned}\bar{y}_f &= 2^{q-k} \sum_{\text{treatments}} \left[\beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \varepsilon \right] \\ &= \beta_0 + 0 + 0 + \sum_{i=1}^{k} \beta_{ii} + \bar{\varepsilon}\end{aligned}$$

as all of the $x_i^2 = 1$ and the other terms cancel with each other. Hence, $\mathrm{E}\bar{y}_f = \beta_0 + \sum_{i=1}^{k} \beta_{ii}$.

The above derivations imply that the difference of the two sample means, $\bar{y}_f - \bar{y}_c$, can be used to test for the overall local curvature of the response surface. That is, if this difference is significantly different from zero, then we have evidence to believe that there are second order effects present. To test this, we can construct a t-statistic,

$$\frac{|\bar{y}_f - \bar{y}_c|}{s\sqrt{2^{q-k} + n_c^{-1}}},$$

where s is the sample standard deviation computed from the n_c central points as this is the only replicated treatment in the design. Hence, these points allow us to construct an unbiased estimate of the variance σ^2. Under the null hypothesis that $\sum_{i=1}^{k} \beta_{ii} = 0$, this test statistic has a $t\,(n_c - 1)$ distribution.

5.4 Chemical Reaction Data Example

In this section, we consider the `ChemReact` dataset the `R` package `rsm`. This dataset contains two factors, time and temperature, and the response is the yield of the chemical reaction. The first block consists of a 2^4 design with $n_c = 3$ center points. The second block contains an additional 3 center points as well as 4 axial points.

First, we have to code the factor levels by running the command

```
CR <- coded.data (ChemReact, x1~(Time - 85)/5, x2~(Temp - 175)/5)
```

As a result, $\alpha \approx \sqrt{2}$ making this a spherical design.

Considering the first block, we can perform a curvature check to test the null hypothesis that $H_0 : \beta_{11} + \beta_{22} = 0$. The t-statistic is

$$\frac{|\bar{y}_f - \bar{y}_c|}{s\sqrt{1/3 + 1/4}} \approx 13.78.$$

Comparing this to the $t\,(2)$, results in a p-value of 0.005 indicating that there is evidence of curvature in this section of the response surface.

A first order model can be fit to the 2^4 factorial design with the command

```
rsm( Yield~FO(x1,x2), data=CR, subset=1:4 )
```

In this case, we get the model

$$\hat{y} = 81.875 + 0.875x_1 + 0.625x_2.$$

We can also include the center points in the first order model,

```
rsm( Yield~FO(x1,x2), data=CR, subset=1:7 )
```

Then, the coefficient estimates for x_1 and x_2 do not change. However, the intercept term is now $\hat{\beta}_0 = 82.8$. Furthermore, the standard error, degrees of freedom, and thus the p-values for the t-tests change. In both cases, we get a vector for the direction of steepest ascent, $(0.814, 0.581)$.

We can further try to estimate the pure quadratic terms–i.e. the coefficients for x_1^2 and x_2^2–with only the factorial and central points.

```
rsm( Yield~FO(x1,x2)+PQ(x1,x2), data=CR, subset=1:7 )
```

However, they will be aliased, and similar to the above t-test, we can only estimate the sum $\beta_{11} + \beta_{22}$ until the second set of data is collected.

Considering the entire dataset, we can fit a full second order model with

```
rsm( Yield~SO(x1,x2), data=CR ).
```

In this case, we get a stationary point at $(0.37, 0.33)$, which corresponds to a time and temperature of $(86.9, 176.7)$. The eigenvalue estimates are -0.92 and -1.31 indicating we have a local maximum.

It is worth noting, however, that performing an F-test on whether or not the quadratic terms improve the fit of the model yields a non-significant p-value. This may indicate that we should continue searching the parameter space. A plot of the response surface and design points is displayed in Figure 5.2.

Design Matrices for Chemical Data

The design matrix can be extracted via the `model.matrix()` function. For this dataset, we get

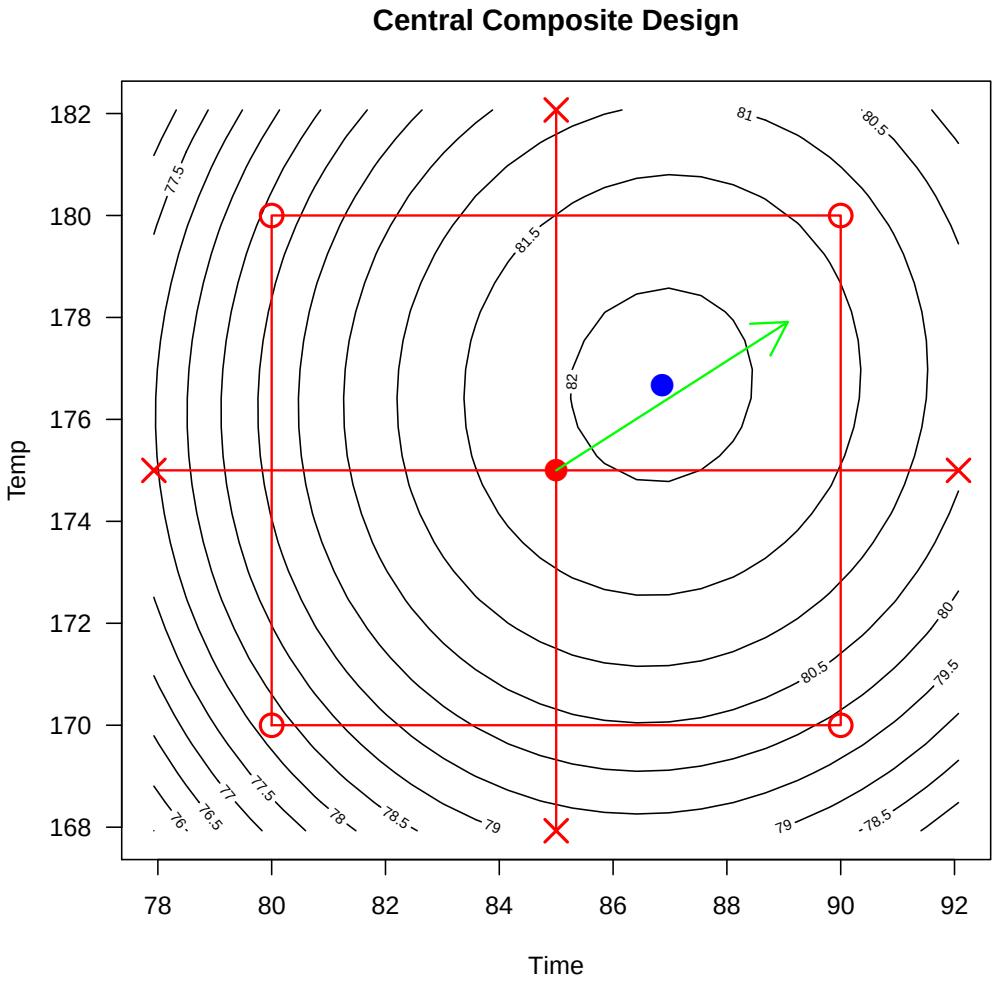

Figure 5.2: A plot of the fitted second order response surface to the chemical reaction dataset. The solid red point is the centre point of the design. The four circle points are from the 2^k design. The x points are the axial points. The solid blue point is the critical point, and the green arrow is direction of greatest increase based on the first order model.

	β_0	β_1	β_2	β_{12}	β_{11}	β_{22}
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	.	.	.	.	.
6	1	.	.	.	.	.
7	1	.	.	.	.	.
8	1	.	.	.	.	.
9	1	.	.	.	.	.
10	1	.	.	.	.	.
11	1	$\sqrt{2}$	.	.	2	.
12	1	$-\sqrt{2}$	.	.	2	.
13	1	.	$\sqrt{2}$	.	.	2
14	1	.	$-\sqrt{2}$	.	.	2

Denoting this matrix as X, we can see the effects of including or excluding the center points from the model. For comparison, the two matrices $X^{\mathrm{T}}X$ for the above data with and without the center points are, respectively,

$$\begin{pmatrix} 14 & . & . & . & 8 & 8 \\ . & 8 & . & . & . & . \\ . & . & 8 & . & . & . \\ . & . & . & 4 & . & . \\ 8 & . & . & . & 12 & 4 \\ 8 & . & . & . & 4 & 12 \end{pmatrix} \text{ and } \begin{pmatrix} 8 & . & . & . & 8 & 8 \\ . & 8 & . & . & . & . \\ . & . & 8 & . & . & . \\ . & . & . & 4 & . & . \\ 8 & . & . & . & 12 & 4 \\ 8 & . & . & . & 4 & 12 \end{pmatrix}.$$

The second matrix here is not invertible. The eigenvalues for the above matrices are, respectively,

$$(26.35, 8, 8, 8, 4, 3.64) \text{ and } (24, 8, 8, 8, 4, 0).$$

Hence, recalling that the variance of the estimator $\hat{\beta}$ is $\mathrm{Var}\left(\hat{\beta}\right) = \sigma^2(X^{\mathrm{T}}X)^{-1}$, we have that the variance is infinite when there are not center points in the model. This is specifically because the designer chose $\alpha = \sqrt{2}$. If a smaller value closer to 1 was chosen, then the variance would not be infinite. However, the addition of center points does stabilize the computation regardless of choice of α.

Chapter 6

Nonregular, Nonnormal, and other Designs

Introduction

In this chapter, we will consider a collection of other designs beyond the scope of what was discussed previously. These include factorial designs with more than 2 or 3 levels, mixed level factorial designs, and nonregular designs such as the Plackett-Burman designs.

Often we will not have precisely 2^k or 3^k observations to work with. If we are given a maximum of say, 20 or 25 measurements, then how best can we design an experiment to test for the significance of all of the factors to be considered?

6.1 Prime Level Factorial Designs[1]

In Chapter 4, we discussed factorial designs at 2 and 3 levels. These same concepts can be extended to factorial designs at r levels for any prime number r. The most common designs beyond 2 and 3 levels are 5 and 7 levels, which will be discussed in subsections below.

Remark 6.1.1. *Note that the requirement that r is prime comes from the fact that a cyclic group of prime order is a finite field, because of the existence of a multiplicative inverse for each element. For example, for $r = 5$, we have that*

$$1 \times 1 = 2 \times 3 = 4 \times 4 = 1 \mod 5.$$

Whereas if $r = 4$, the element 2 does not have such an inverse, since

$$2 \times 1 = 2 \mod 4, \quad 2 \times 2 = 0 \mod 4, \quad 2 \times 3 = 2 \mod 4.$$

[1] See Wu & Hamada Section 7.8

If r is a power of a prime like 4, 8, or 9, then one can use Galois theory to construct a design.

In general, to construct an r^{k-q} fractional factorial design, we begin with $k-q$ orthogonal columns of length r^{k-q} with entries $0, 1, \ldots, r-1$. These columns correspond to the main effects to be tested. Denoting them by $x_1, \ldots, x_{k-q}$, we can construct the interaction effects by summing linear combinations of these columns modulo r as follows:

$$\sum_{i=1}^{k-q} c_i x_i \mod r$$

for $c_i = 0, 1, \ldots, r-1$. Counting the total number of unique columns, we have r^{k-q} choices for the c_i; we subtract 1 for the case when $c_i = 0$ for all i; Then, we impose the restriction that the first non-zero $c_i = 1$ for uniqueness of the factors. As a result, we have a total of

$$\frac{r^{k-q}-1}{r-1}$$

factors to test. Recall that this coincides with the 2 and 3 level settings where we had $2^{k-q}-1$ and $(3^{k-q}-1)/2$ effects, respectively, to test.

As an example, consider the 3^{3-1} design with factors A, B, C in the following table:

	A	B	AB	C= AB^2
1	0	0	0	0
2	0	1	1	2
3	0	2	2	1
4	1	0	1	1
5	1	1	2	0
6	1	2	0	2
7	2	0	2	2
8	2	1	0	1
9	2	2	1	0

6.1.1 5 level designs

In this section, we will consider the special case of the 25-run design, which is a $5^{k-(k-2)}$ fractional factorial design based on 25 observations and k factors taking on 5 levels each. In such a design, we will be able to test $(25-1)/(5-1) = 6$ effects.

Similar to the general case presented above, we begin with two orthogonal columns of length 25 and taking values 0,1,2,3,4. Then, the four interaction columns can be computed via

$$x_1 + cx_2 \mod 5$$

for $c = 1, 2, 3, 4$ giving terms AB, AB^2, AB^3, AB^4. The main effects A and B have 5-1=4 degrees of freedom. The $A \times B$ interaction has the remaining 16 of which 4 are given to each of the sub-interaction terms.

We could treat this as a 5^2 full factorial design. However, if there are more 5-level factors to include in the model, we can add defining relations such as

$$C = AB, \quad D = AB^2, \quad E = AB^3 \quad F = AB^4$$

to treat this as $5^{k-(k-2)}$ for values of $k = 3, 4, 5, 6$. For example, if $k = 3$, we would have three factors A, B, C and the defining relation $I = ABC^4$. Then, for example, the term AB^4 would be aliased with

$$\begin{aligned} AB^4 &= A^2B^5C^4 = AC^2 \\ (AB^4)^2 &= A^3B^9C^4 = A^3B^4C^4 = AB^3C^3 \\ (AB^4)^3 &= A^4B^{13}C^4 = A^4B^3C^4 = AB^2C \\ (AB^4)^4 &= A^5B^{17}C^4 = B^2C^4 = BC^2 \end{aligned}$$

Thus, the aliased group is

$$\{AB^4, AC^2, BC^2, AB^3C^3, AB^2C\}.$$

If we consider the entire 25×6 table–displayed in Table 6.1–the first two factors A and B can be treated as the rows and columns, respectively, for a Latin square design. We can thus include factors C, D, E, and/or F to construct a Latin, Graeco-Latin, or Hyper-Graeco-Latin square design. As there are 4 orthogonal 5×5 Latin square designs, each of the $5^{k-(k-2)}$ designs can be treated in this way for $k = 3, 4, 5, 6$.

Remark 6.1.2. *While a general formula for the number of mutually orthogonal $n \times n$ Latin squares does not exist. Denoting this number as $a(n)$, it is known*[2] *that*

$$a(r^k) = r^k - 1$$

for all primes r and integers $k > 0$. Hence, we can consider any $r^{k-(k-2)}$ factorial design as a hyper-Graeco-Latin square design.

When analyzing such a design, the main effects have 4 degrees of freedom. Hence, if they are ordinal variables, then we can extend the polynomial contrasts considered in the 3-level factorial setting to linear, quadratic, cubic, and quartic interactions. The contrast vectors are

Linear:	$A_1 = (-2, -1, 0, 1, 2)/\sqrt{10}$
Quadratic:	$A_2 = (2, -1, -2, -1, 2)/\sqrt{14}$
Cubic:	$A_3 = (-1, 2, 0, -2, 1)/\sqrt{10}$
Quartic:	$A_4 = (1, -4, 6, -4, 1)/\sqrt{70}$

[2] See https://oeis.org/A001438

	A	B	$C = AB$	$D = AB^2$	$E = AB^3$	$F = AB^4$
1	0	0	0	0	0	0
2	0	1	1	2	3	4
3	0	2	2	4	1	3
4	0	3	3	1	4	2
5	0	4	4	3	2	1
6	1	0	1	1	1	1
7	1	1	2	3	4	0
8	1	2	3	0	2	4
9	1	3	4	2	0	3
10	1	4	0	4	3	2
11	2	0	2	2	2	2
12	2	1	3	4	0	1
13	2	2	4	1	3	0
14	2	3	0	3	1	4
15	2	4	1	0	4	3
16	3	0	3	3	3	3
17	3	1	4	0	1	2
18	3	2	0	2	4	1
19	3	3	1	4	2	0
20	3	4	2	1	0	4
21	4	0	4	4	4	4
22	4	1	0	1	2	3
23	4	2	1	3	0	2
24	4	3	2	0	3	1
25	4	4	3	2	1	0

Table 6.1: A $5^{k-(k-2)}$ factorial design.

The interaction term $A \times B$ can be similarly broken up into 16 contrasts $(AB)_{i,j}$ for $i, j = 1, 2, 3, 4$.

6.1.2 7 level designs

The 7-level design of most interest is the 49-run design similar to the 25-run design considered in the previous subsection. That is, because $7^3 = 343$ is a large number of observations required, and most likely, the desired hypotheses can be tested with a smaller design.

Mathematically, this design is nearly identical to the previous one except now each factor can take on 7 levels making the arithmetic performed modulo 7. We can similarly consider all $7^{k-(k-2)}$ designs are hyper-Graeco-Latin squares. The maximal number of factors to test will be $(49-1)/(7-1) = 8$. Each main effect has 6 degrees of freedom. Hence, we could feasibly consider polynomial contrasts from linear up to 6th degree. Generally, these higher order polynomials are not wise to consider as (1) they may lead to overfitting and (2) are often difficult to interpret.

Remark 6.1.3. *Note that*

$$\frac{r^2 - 1}{r - 1} = r + 1.$$

for integers $r \geq 2$.

6.1.3 Example of a 25-run design

Data was randomly generated based on a 25-run design with $k = 4$ factors to test. The "yield" was generated as a normal random variate with variance 1 and mean

$$\mu(C, D) = D/2 + \phi(C)$$

where $\phi : \{0, 1, 2, 3, 4\} \to \{-1, 1, 0, -1, 1\}$ and $C, D = 0, 1, 2, 3, 4$. That is, the yield is linear in D and cubic in C.

Running a standard ANOVA model only considering main effects gives

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
A	4	0.560	0.140	0.082	0.9856	
B	4	1.561	0.390	0.229	0.9143	
C	4	23.930	5.982	3.515	0.0614	.
D	4	16.504	4.126	2.425	0.1333	
Residuals	8	13.614	1.702			

Here, we see some marginal significance in C and none in D. However, expanding with respect to polynomial contrasts gives

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
A	4	0.560	0.140	0.082	0.9856	
B	4	1.561	0.390	0.229	0.9143	
C	4	23.930	5.982	3.515	0.0614	.
C: P1	1	5.868	5.868	3.448	0.1004	
C: P2	1	0.314	0.314	0.184	0.6790	
C: P3	1	16.564	16.564	9.733	0.0142	*
C: P4	1	1.184	1.184	0.696	0.4285	
D	4	16.504	4.126	2.425	0.1333	
D: L	1	13.125	13.125	7.713	0.0240	*
D: HOT	3	3.379	1.126	0.662	0.5984	
Residuals	8	13.614	1.702			

If we also tried to include $A \times B$ interactions in our model, we would have no extra degrees of freedom to compute the residuals. However, the mean sum of squares for $A \times B$ will still be large. We could quickly apply the hierarchy principal to debunk this as a false positive. The source of this significance comes from the aliasing relations as

$$C = AB, \text{ and } D = AB^2.$$

The terms AB^3 and AB^4 are still orthogonal and thus $A \times B$ still has 8 degrees of freedom remaining.

6.2 Mixed Level Designs

In this section, we will consider mixed 2 & 4 and 2 & 3 level designs. That is, some factors in these models will have two levels and others will have 3 or 4 levels. This will, of course, allow for more flexible modelling and testing scenarios.

Before discussing such designs, we require the notion of a *symmetric orthogonal array*,[3] which is one with the same levels in each columns, and the more general *mixed level orthogonal array*.

Definition 6.2.1 (Symmetric Orthogonal Array)**.** *Given s symbols denoted $1, 2, \ldots, s$, and an integer $t > 0$ referred to as the strength of the table, an $N \times m$ orthogonal array of strength t, denoted*

$$OA(N, s^m, t)$$

is an $N \times m$ table with entries $1, \ldots, s$ such that any set of t columns will every combination of t symbols appearing an equal number of times.

We have already seen many examples of orthogonal arrays, which include the 2^{k-q} and 3^{k-q} designs. For example, a 2^{4-1} design would be an $OA(8, 2^7, t)$. In the

[3] https://en.wikipedia.org/wiki/Orthogonal_array

following table, we have a 2^{4-1} design with defining relation $I = ABCD$ giving us a resolution of 4

Remark 6.2.2 (Resolution and Strength)**.** *In Wu & Hamada, they claim that the strength of an orthogonal array is the resolution minus 1. However, the three columns A, B, and AB will only ever take on 4 of the 8 possible binary combinations. Instead, the strength of the array restricted to the main effects only should be the resolution-1.*

	A	B	C	D	AB	AC	BC
1	.	.	.	.	.	.	.
2	.	.	+	+	.	+	+
3	.	+	.	+	+	.	+
4	.	+	+	.	+	+	.
5	+	.	.	+	+	+	.
6	+	.	+	.	+	.	+
7	+	+	.	.	.	+	+
8	+	+	+	+	.	.	.

Here, for example, the first three columns will contain all 8 unique binary sequences while the last three columns will contain 4 unique sequences repeated twice. Similarly, for 3^{k-q} designs, we have an orthogonal array $OA(3^{k-q}, 3^{3^{k-q}-1}, t)$. The strength of this array is also 1 minus the resolution of the design.

As we will be concerned with mixed-level designs in this section, we need a more general definition of orthogonal arrays.

Definition 6.2.3 (Mixed Level Orthogonal Array)**.** *Given $s_1, \ldots, s_\gamma$ symbols denoted $1, 2, \ldots, s_i$, $m = \sum_{i=1}^{\gamma} m_i$, and an integer $t > 0$ referred to as the strength of the table, an $N \times m$ orthogonal array of strength t, denoted*

$$OA(N, s_1^{m_1} \ldots s_\gamma^{m_\gamma}, t)$$

is an $N \times m$ table with m_i columns with entries $1, \ldots, s_i$ such that any set of t columns will have every combination of t symbols appearing an equal number of times.

A simple example of a mixed level orthogonal array is the $OA(8, 2^4 4^1, 2)$ below. It has $m = 4 + 1$ columns and strength 2 resulting from the inclusion of the A column.

	A	1	2	3	123
1	0	.	.	.	.
2	1	.	.	+	+
3	2	.	+	.	+
4	3	.	+	+	.
5	3	+	.	.	+
6	2	+	.	+	.
7	1	+	+	.	.
8	0	+	+	+	+

6.2.1 2^n4^m Designs[4]

In Wu & Hamada, they detail a "Method of Replacement" for constructing a 2^n4^m design from a 2^N design. To do this, we begin with a symmetric orthogonal design $OA(N, 2^m, t)$ with $t \geq 2$, which is any 2^{k-q} fractional factorial design with resolution 3 or greater.

Consequently, any two columns in $OA(N, 2^m, t)$ will contain all 4 possible combinations of $\{\cdot, +\}$ as otherwise the two columns would be aliased. Denoting these two columns as α and β, there is a third column in the array constructed by the product of columns α and β, which will be denoted as $\alpha\beta$. Considering only these three columns, we have four possible triples, which can be mapped to the levels $0, 1, 2, 3$. Hence, we replace these three 2-level columns with a single 4-level column.

α	β	$\alpha\beta$		A
.	.	.	$\rightarrow$	0
.	+	+	$\rightarrow$	1
+	.	+	$\rightarrow$	2
+	+	.	$\rightarrow$	3

A similar example occurs in the previous two example arrays where the columns AB, AC, and BC where combined into a single 4-level factor and A,B,C,D were relabeled as 1,2,3,123.

This process can be iterated n times to transform an $OA(N, 2^m)$ into an $OA(N, 2^{m-3n}4^n)$. If k is even, then

$$2^k - 1 = 2^{2l} - 1 = 4^l - 1,$$

which is divisible by 3 as $4^l = 1 \mod 3$ for any integer $l \geq 0$. It has been proven that it is possible to, in fact, decompose a 2^k design with k even into $(2^k - 1)/3$ mutually exclusive sets of three columns that can in turn be replaced with a 4-level columns. Hence, any design of the form $OA(2^k, 2^{m-3n}4^n)$ exists for k even, $m = 2^k - 1$, and $n = 0, 1, \ldots, (2^k - 1)/3$.

[4] See Wu & Hamada Sections 7.2, 7.3, & 7.4

For k odd, we have that

$$2^k - 2 = 2^{2l+1} - 2 = 2(4^l - 1).$$

Hence, in this case $2^k - 2$ is divisible by 3. However, it has been proven[5] that actually only $2^k - 5$ mutually exclusive sets of three elements exist when k is odd. Therefore, when k is odd, designs of the form $OA(2^k, 2^{m-3n}4^n)$ exists for $m = 2^k - 1$, and $n = 0, 1, \ldots, (2^k - 5)/3$. This implies that for 16 and 64 run designs, we can accommodate 5 and 21 four-level effects, respectively.

Construction of an $OA(N, 4^5)$ design

Beginning with a full 2^4 factorial design, denoted $OA(2^4, 2^{2^4-1}, 4)$, we can replace each mutually exclusive set of three columns by a single 4-level column as follows.

	1	2	12	3	4	34	13	24	1234	23	124	134	123	14	234
1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	.	.	.	.	+	+	.	+	+	.	+	+	.	+	+
3	.	.	.	+	.	+	+	.	+	+	.	+	+	.	+
4	.	.	.	+	+	.	+	+	.	+	+	.	+	+	.
5	.	+	+	.	.	.	.	+	+	+	+	.	+	.	+
6	.	+	+	.	+	+	.	.	.	+	.	+	+	+	.
7	.	+	+	+	.	+	+	+	.	.	+	+	.	.	.
8	.	+	+	+	+	.	+	.	+	.	.	.	.	+	+
9	+	.	+	.	.	.	+	.	+	.	+	+	+	+	.
10	+	.	+	.	+	+	+	+	.	.	.	.	+	.	+
11	+	.	+	+	.	+	.	.	.	+	+	.	.	+	+
12	+	.	+	+	+	.	.	+	+	+	.	+	.	.	.
13	+	+	.	.	.	.	+	+	.	+	.	+	.	+	+
14	+	+	.	.	+	+	+	.	+	+	+	.	.	.	.
15	+	+	.	+	.	+	.	+	+	.	.	.	+	+	.
16	+	+	.	+	+	.	.	.	.	.	+	+	+	.	+

[5] Wu, C. F. J. *Construction of 2m4 n designs via a grouping scheme.* The Annals of Statistics (1989): 1880-1885.

	(1,2,12)	(3,4,34)	(13,24,1234)	(23,124,134)	(123,14,234)
1	0	0	0	0	0
2	0	1	1	1	1
3	0	2	2	2	2
4	0	3	3	3	3
5	1	0	1	3	2
6	1	1	0	2	3
7	1	2	3	1	0
8	1	3	2	0	1
9	2	0	2	1	3
10	2	1	3	0	2
11	2	2	0	3	1
12	2	3	1	2	0
13	3	0	3	2	1
14	3	1	2	3	0
15	3	2	1	0	3
16	3	3	0	1	2

Counting degrees of freedom, we started with $15 = 2^4 - 1$. We now have $5 \times (4-1) = 15$, which coincides. This is an $OA(2^4, 4^5, 2)$. Furthermore, as 4 is a prime power–i.e. $4 = 2^2$–there are 3 mutually orthogonal 4×4 Latin squares. Hence, in the above design, the first two columns can be considered as the row and column of an hyper-Graeco-Latin square with 3 experimental factors to test.

Choosing and Analysing a Design

Briefly, the concept of aberration can be extended to this setting for choosing which sets for three 2-level columns to collapse into a 4-level column. It is often preferred to have defining relations that include the 4-level factors as they have more degrees of freedom to spare than the 2-level factors. Hence, significance may still be detectable.

For analysis, polynomial contrasts for the 4-level factors can be considered to capture linear, quadratic, and cubic effects. It is recommended though to consider an alternative system of contrasts related to the polynomials:

$$A_1 = (-1, -1, 1, 1)$$
$$A_2 = (-1, 1, 1, -1)$$
$$A_3 = (-1, 1, -1, 1)$$

The reason for this system is that is coincides with the construction of the 4-level factor. Recalling the table from above,

α	β	$\alpha\beta$		A
.	.	.	$\rightarrow$	0
.	+	+	$\rightarrow$	1
+	.	+	$\rightarrow$	2
+	+	.	$\rightarrow$	3

we see that contrasts A_1, A_2, and A_3 correspond, respectively, to the factorial effects α, $\alpha\beta$, and β.

For two 4-level factors, A and B, we cannot decompose them as we did for the prime level factors as the integers mod 4 is not a finite field. To illustrate the problem, for 3-level factors we have

A	0	0	0	1	1	1	2	2	2
B	0	1	2	0	1	2	0	1	2
AB	0	1	2	1	2	0	2	0	1
A^2B^2	0	2	1	2	1	0	1	0	2

The last two rows are identical up to swapping 1 and 2. Thus, the interactions AB and A^2B^2 are equivalent. However, if each of these has 4 levels, then

A	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3
B	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
AB	0	1	2	3	1	2	3	0	2	3	0	1	3	0	1	2
A^2B^2	0	2	0	2	2	0	2	0	0	2	0	2	2	0	2	0

Hence, the term A^2B^2 only takes on 2 levels of a possible 4. Additionally, the term AB^2 will only take on even values when A is even and odd values when A is odd, so every possible combination will not occur.

6.2.2 36-Run Designs[6]

We can construct a 36-run design by combining a 2^{3-1} and a 3^{3-1} design. In the first case, we impose the defining relation $I = ABC$. In the second case, it is $I = DEF^2$. These two designs can be considered as symmetric orthogonal arrays $OA(4, 2^3, 2)$ and $OA(9, 3^4, 2)$.

[6] See Wu & Hamada Section 7.7

	$OA(4, 2^3, 2)$		
	A	B	C=AB
1	.	.	.
2	.	+	+
3	+	.	+
4	+	+	.

	$OA(9, 3^4, 2)$			
	D	E	$F = DE$	DE^2
1	0	0	0	0
2	0	1	1	2
3	0	2	2	1
4	1	0	1	1
5	1	1	2	0
6	1	2	0	2
7	2	0	2	2
8	2	1	0	1
9	2	2	1	0

We can combine these two orthogonal arrays into an $OA(36, 2^3 3^3, 2)$ by applying a "tensor"-like operation. The first set of four rows would be

	A	B	C	D	E	F
1	.	.	.	0	0	0
2	.	+	+	0	0	0
3	+	.	+	0	0	0
4	+	+	.	0	0	0

This can then be repeated eight more times for the other factor levels of D, E, and F.

The defining contrast subgroup of this $2^{3-1}3^{3-1}$ design can be constructed by multiplying the above defining words together to get

$$I = ABC = DEF^2 = ABCDEF^2 = ABCD^2E^2F.$$

This design has 35 degrees of freedom to work with. The 2-level main effects have 1 DoF and the 3-level main effects have 2 DoFs totalling 9. There is one 3×3 interaction, DE^2, which is not aliased with any main effects and has 2 DoFs. There are nine 2×3 interaction factors, which come from the defining words of length 6. This can be seen by choosing one of $\{A, B, C\}$ and pairing it with one of $\{D, E, F\}$. Note that, for example, AD and AD^2 are equivalent effects. Furthermore, each of these has $(3-1) \times (2-1) = 2$ DoFs. In total, thus far, we have $9 + 2 + 2 \times 9 = 29$. The remaining 4 degrees of freedom come from the interactions of $\{A, B, C\}$ with DE^2 each having 2 DoFs. These final three interaction terms can be ignored and their 6 total degrees of freedom reserved for the residual sum of squares.

Linear and quadratic contrasts can be considered for the 3-level factors. Furthermore, interactions between 2 and 3 level factors can be interpreted as, for example with AD, a difference in the linear or quadratic behaviour of D conditional on whether $A = -$ or $A = +$.

Remark 6.2.4 (Other 36-Run Designs)**.** *Similarly to the previous section, we can also construct 36-Run designs by combining one design of* $\{2^2, 2^{3-1}\}$ *with one design of* $\{3^2, 3^{3-1}, 3^{4-2}\}$*.*

6.3 Nonregular Designs

Thus far, every design considered can be classified as *regular* meaning that all of the factorial effects are either orthogonal or completely aliased, which is a correlation of either 0 or ± 1. A *nonregular* design allows for non-zero non-one correlations between factorial effects. This is similar to how polynomial contrasts were shown in the 3^{k-q} designs to have partial aliasing. However, now we are concerned with the actual factorial effects and not correlations between specific contrasts.

The reason for considering such designs is one of economy and efficiency. If we, for example, were interested in a 2^{k-q} factorial design to test 6 main effects and 15 two-way interactions, we would require a 2^{6-1} design with 32 observations whereas we will only be using 21 of the 31 degrees of freedom for parameter estimation. Instead, we could consider a Plackett-Burman design on 24-runs. A reduction in 8 observations can save a lot of resources especially if this design were to be used iteratively in a response surface search procedure.

6.3.1 Plackett-Burman Designs

For designs involving only 2-level factors, we have considered cases only where the number of runs is a power of 2. The design matrix for these designs can be thought of as an *Hadamard matrix*.

Definition 6.3.1 (Hadamard matrix)**.** *An* $n \times n$ *Hadamard matrix, denoted* H_n*, is an orthogonal matrix with entries* ± 1*.*

The simplest example is

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

From this matrix, we can construct any H_n for $n = 2^k$ by successive application of the tensor or Kronecker product for matrices. For an Hadamard matrix of size n, we can construct one of size $2n$ by

$$H_{2n} = H_2 \otimes H_n = \begin{pmatrix} H_n & H_n \\ H_n & -H_n \end{pmatrix}.$$

Using this formula to construct an H_{2^k} matrix gives one where the first column is all ones. We can remove this column–it corresponds to the intercept term in our model–and consider the $2^k \times (2^k - 1)$ matrix as our design matrix. For example,

when $k = 3$, we have

$$H_8 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

We can remove the first column and consider the 2^3 design

	A	B	AB	C	AC	BC	ABC
1	1	1	1	1	1	1	1
2	-1	1	-1	1	-1	1	-1
3	1	-1	-1	1	1	-1	-1
4	-1	-1	1	1	-1	-1	1
5	1	1	1	-1	-1	-1	-1
6	-1	1	-1	-1	1	-1	1
7	1	-1	-1	-1	-1	1	1
8	-1	-1	1	-1	1	1	-1

From here, we can construct fractional factorial designs on 8-runs by including aliasing relations such as $D = ABC$. If we have reason to believe that all of the interactions between the factor are negligible, we could feasibly pack 7 factors into this design.

In the case where the number of runs is $N = 2^k$, the Plackett-Burman design coincides with the set of $2^{(k+q)-q}$ fractional factorial designs for $q = 0, 1, \ldots, 2^k - k - 1$. These correspond to orthogonal arrays $OA(2^k, 2^m, t)$ where m is the number of factors and t is the strength of the design which is the resolution - 1.

If we were interested in 2-level designs with N not a power of 2, we need to work harder to construct the matrix H_N. First of all, N cannot be odd as every column must have an equal number of 1's and -1's. Secondly, N must be divisible by 4. Otherwise, we have, without loss of generality, a first column of all 1's and a second columns of $N/2$ 1's followed by $N/2$ -1's. Any subsequent column must be orthogonal to both of these. Orthogonality with respect to column 1 requires an equal amount of 1's and -1's. Denoting the ith column by X_i, we have

$$0 = X_1 \cdot X_3 = \sum_{i=1}^{N} (X_3)_i.$$

Orthogonality with respect to column 2 requires the sum of the first $N/2$ entries to

equal the sum of the second $N/2$ entries as

$$0 = X_2 \cdot X_3 = \sum_{i=1}^{N/2} (X_3)_i - \sum_{i=N/2+1}^{N} (X_3)_i.$$

Hence, these two conditions can only hold simultaneously if $\sum_{i=1}^{N/2}(X_3)_i = 0$, which is impossible if $N/2$ is odd–i.e. if N is not divisible by 4. This leads to the Hadamard conjecture.

Conjecture 6.3.2. *An $N \times N$ Hadamard matrix exists for for all integers N divisible by 4.*

Thus far, such a matrix has been found for all $N < 668$. The $N = 668$ case is still open at the time of writing these notes. Luckily, for our purposes, we are only interested in smaller Hadamard matrices–namely, those of orders $N = 12, 20, 24, 28, 36, 44$. These are the multiples of 4 that are not powers of 2.

Constructing an Hadamard matrix with N not a power of 2 requires much more sophistication than in the power of 2 case. Some matrices result from the *Paley Construction*, which relies on quadratic residues for finite fields. This will be mentioned in the chapter appendix for completeness, but is, in general, beyond the scope of this course. The end result of the construction is a generating row of length $N-1$. This can then be cyclically shifted to construct an $(N-1)\times(N-1)$ matrix. In turn, this matrix becomes H_N by appending an $(N-1)$-long row of -1's to the bottom and then a column of N 1's on the left.

For example, the generating row for $N = 12$ is

$$(1, 1, -1, 1, 1, 1, -1, -1, -1, 1, -1).$$

This results in

$$H_{12} = \begin{pmatrix}
1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 \\
1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 \\
1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\
1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \\
1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 \\
1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 \\
1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\
1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \\
1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1
\end{pmatrix}$$

The first column can be removed to get an $OA(12, 2^{11}, 2)$. It can be checked, in `R` for example, that $H_{12}H_{12}^{\mathrm{T}} = 12I_{12}$. Note that we can use this to immediately construct

a design with $N = 24, 48$ by using the fact that

$$H_{24} = H_{12} \otimes H_2, \text{ and}$$
$$H_{48} = H_{24} \otimes H_2$$

Remark 6.3.3. *Note that such designs can be generated by the* `pb()` *function in the* `R` *library* `FrF2`*. As an example,*

```
pb(12,randomize=FALSE)
```

which produces the same design as H_{12} above except that the function shifts to the right whereas we shifted to the left.

6.3.2 Aliasing and Correlation

In general, these designs are used only in the case where the interaction effects are negligible. This is because of the complex aliasing that occurs in such a design. We will not discuss this in general, but instead look at an example.

Consider the 12-run design presented above. We can fit 11 factors into this design assuming no interaction effects. If we consider the interactions between A, B, and C, we have the following correlation structure.

	A	B	C	D	E	F	G	H	I	J	K
AB	0	0	-1/3	-1/3	-1/3	1/3	-1/3	-1/3	1/3	1/3	-1/3
AC	0	-1/3	0	1/3	-1/3	-1/3	1/3	-1/3	1/3	-1/3	-1/3
BC	-1/3	0	0	-1/3	-1/3	-1/3	1/3	-1/3	-1/3	1/3	1/3
ABC	0	0	0	$-1/\sqrt{8}$	$1/\sqrt{8}$	$-1/\sqrt{8}$	$1/\sqrt{8}$	$1/\sqrt{8}$	$1/\sqrt{8}$	$1/\sqrt{8}$	$-1/\sqrt{8}$

The correlation matrix for all main effects and two-factor interactions is displayed in Figure 6.1. In that graphic, the red squares correspond to a correlation of $-1/3$ and the blue squares to $+1/3$.

Plackett-Burman designs are peculiar in the sense that while, for example, effects A, B, and AB are orthogonal, main effect A is partially aliased with every two-way interaction that does not include the factor A. Hence, A is partially correlated with $\binom{11-1}{2} = 45$ of the 55 two-way interactions. This is why the main use of these designs is for *screening* many factors to decide which few are important. From a Minitab Blog,

> "Plackett-Burman designs exist so that you can quickly evaluate lots of factors to see which ones are important. Plackett-Burman designs are often called screening designs because they help you screen out unimportant factors."

In general, when fitting a linear model but leaving out non-negligible terms, bias is introduced into the parameter estimates. This can occur in regression when

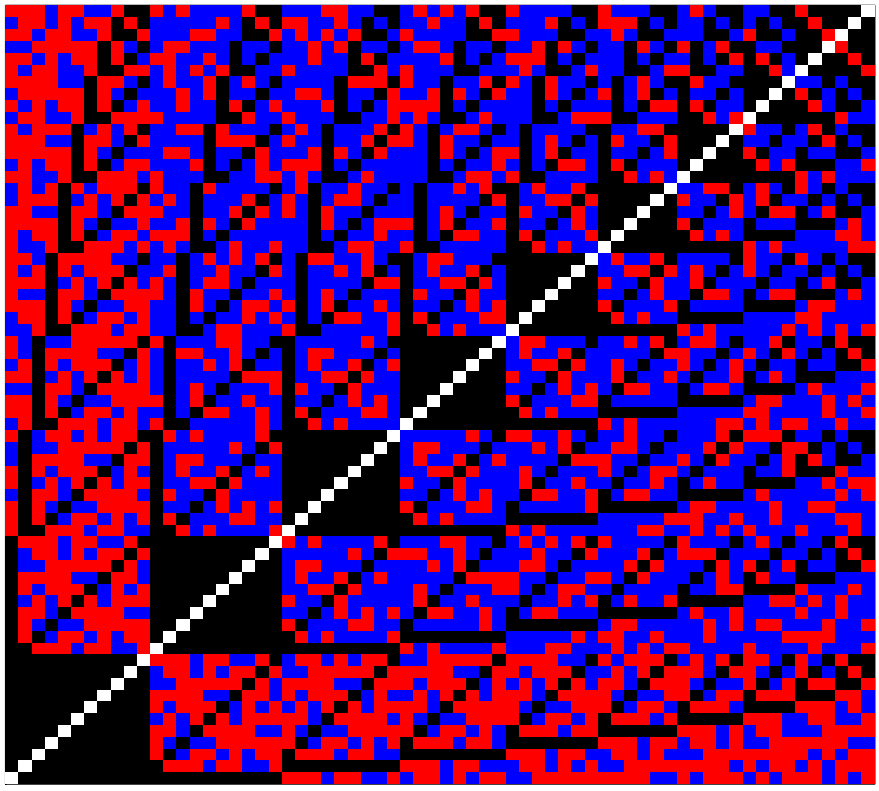

Figure 6.1: Correlation matrix for the 12-Run Plackett-Burman design with main effects and two-way interactions.

using a variable selection technique or similarly in one of these models with partial correlations if the partially correlated terms are ignored. Consider a linear model

$$Y = X\beta_1 + Z\beta_2 + \varepsilon$$

where X is the design matrix containing the effects we are interested in–e.g. the main effects–and Z is the design matrix containing effects that are to be ignored–e.g. interaction effects.

Solving for the least squares estimator for the reduced model

$$Y = X\beta_1 + \varepsilon$$

yields the usual $\hat{\beta}_1 = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$. However, when taking the expected value of $\hat{\beta}_1$, we no longer have an unbiased estimator. Indeed, we have

$$\begin{aligned} \mathrm{E}\hat{\beta}_1 &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{E}Y \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\left[X\beta_1 + Z\beta_2\right] \\ &= \beta_1 + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z\beta_2. \end{aligned}$$

Resulting from this derivation, if the parameters $\beta_2 = 0$ or if the X and Z terms are orthogonal–i.e. $X^{\mathrm{T}}Z = 0$–then our estimate of β_1 is unbiased. Otherwise, we have a biased estimator. The matrix $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z$ is often referred to as the alias matrix.

6.3.3 Simulation Example

Consider a 12-run Plackett-Burman design applied to 6 factors, {A,B,C,D,E,F}, taking values $\{0, 1\}$ where the true model is

$$\text{yield} = (D) - 1.5(F) + 1.25(F \oplus E) + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, 0.25)$.

A first step maybe to just consider the main effects as they are all orthogonal to each other.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
A	1	0.806	0.806	1.612	0.2601	
B	1	0.150	0.150	0.301	0.6069	
C	1	0.198	0.198	0.396	0.5570	
D	1	3.029	3.029	6.057	0.0571	.
E	1	0.212	0.212	0.423	0.5439	
F	1	5.378	5.378	10.754	0.0220	*
Residuals	5	2.501	0.500			

This leads us to assume that one or both of D and F are significant main effects. However, a strong two-way interaction effect could be causing this result due to aliasing.

Applying the hereditary principal, we can assume that any two-way effects of interest involve either D or F. Hence, we can fit models of the form

$$\texttt{output} \sim \texttt{x + D + F + x:D + x:F}$$

for `x` being A, B, C, and E. From here, we see only two significant interaction terms being `B:D` and `E:F`. Thus, we can fit the model

$$\texttt{output} \sim \texttt{B + D + E + F + B:D + E:F}$$

which gives

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
E	1	0.212	0.212	1.367	0.29502	
B	1	0.150	0.150	0.971	0.36965	
D	1	3.029	3.029	19.554	0.00688	**
F	1	5.378	5.378	34.715	0.00200	**
B:D	1	1.575	1.575	10.164	0.02432	*
E:F	1	1.156	1.156	7.458	0.04123	*
Residuals	5	0.775	0.155			

This table is considering type 1 anova. Instead, using type 2 anova, we see the three significant terms appear.

	Sum Sq	Df	F value	Pr(>F)	
E	0.0002	1	0.0011	0.974604	
B	0.0291	1	0.1880	0.682631	
D	4.6727	1	30.1604	0.002733	**
F	2.9780	1	19.2217	0.007126	**
B:D	0.4530	1	2.9242	0.147951	
E:F	1.1555	1	7.4583	0.041231	*
Residuals	0.7746	5			

As neither B nor B:D are significant, we can try refitting the model without those terms giving

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
E	1	0.212	0.212	1.18	0.313412	
D	1	3.029	3.029	16.87	0.004529	**
F	1	5.378	5.378	29.95	0.000933	***
E:F	1	2.398	2.398	13.36	0.008125	**
Residuals	7	1.257	0.180			

When the interaction term is strongly significant

If we were to instead consider a model like

$$\text{yield} = (D) - 1.5(F) + 5(F \oplus E) + \varepsilon,$$

then the strong coefficient for the E:F term makes it hard to find significance in the main effects:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	1	9.38	9.382	1.254	0.314
B	1	6.52	6.518	0.871	0.394
C	1	6.81	6.811	0.910	0.384
D	1	0.18	0.180	0.024	0.883
E	1	0.21	0.212	0.028	0.873
F	1	5.38	5.378	0.719	0.435
Residuals	5	37.41	7.482		

One way to proceed would be to use stepwise regression between the main effects and the two-way effects model using the following commands.

```
mdFirst = aov(yield~.,data=dat);
mdSecond = aov(yield~(.)^2,data=dat);
out=step(mdFirst,scope=list(lower=mdFirst,upper=mdSecond),direction='both');
```

One problem with this is that `step()` will continuing adding variables until to the model is saturated. Backing up from the saturated model, we have the new model

```
output ~ A + B + C + D + E + F + E:F + B:C + A:D .
```

Considering the anova type 2 or 3 table, we find that neither A or A:D is significant. Removing them from this model gives

```
output ~ B + C + D + E + F + E:F + B:C .
```

Now that A and A:D are gone, we see from another anova type 2 or 3 table that B, C, and B:C are no longer significant. Removing them results in the desired model

```
output ~ D + E + F + E:F
```

with the anova 2 or 3 table

	Sum Sq	Df	F value	Pr(>F)	
D	4.654	1	25.9190	0.0014135	**
E	0.212	1	1.1796	0.3134116	
F	5.378	1	29.9555	0.0009327	***
E:F	58.866	1	327.8620	3.875e-07	***
Residuals	1.257	7			

Alternatively, one could try every possible interaction term on its own, which are models of the form

```
output ~ A+B+C+D+E+F + x:y .
```

Then corresponding F-statistics for the models all have DoFs (7,4). The values are

A:B	1.15,	A:C	0.49,	A:D	0.98,	A:E	0.82,	A:F	0.53,
B:C	0.43,	B:D	1.32,	B:E	0.44,	B:F	0.45,	C:D	0.46,
C:E	0.47,	C:F	0.47,	D:E	0.50,	D:F	0.77,	E:F	33.77,

which quickly identifies E:F are a significant term.

6.A Paley's Contruction of H_N

The above construction of the Hadamard matrices H_N only works for N a power of 2. A more powerful construction due to Raymond Paley allows for the construction of H_N for any $N = r^m + 1$ for r a prime number given that $r^m + 1$ is divisible by 4. This construction combined with the Kronecker product allows for the construction of H_N for all $N < 100$ except for $N = 92$. More details on this construction can be found in *The Theory of Error-Correcting Codes* by FJ MacWilliams and NJA Sloane, Chapter 2 Section 3.

First, we require the concept of a *quadratic residue*. For a prime r, the elements of the finite field $k^2 \mod r$ for $k = 1, 2, \ldots, r-1$ are the are the quadratic residues mod r. That is, these are the square numbers in that field. Second, we require the definition of the *Legendre symbol*. For a prime r, the Legendre symbol is a function $\chi(\cdot)$ defined as

$$\chi(k) = \begin{cases} 0 & \text{for } k \text{ divisible by } r \\ 1 & \text{for } k \text{ a quadratic residue mod } r \\ -1 & \text{otherwise} \end{cases}$$

This allows us to construct the *Jacobsthal matrix* Q, which is a skew symmetric $r \times r$ matrix with ijth entry equal to $\chi(i-j)$. Then, letting $\mathbf{1}_r$ be the $r \times r$ matrix of all 1's, we claim that

$$QQ^{\mathrm{T}} + \mathbf{1}_r = rI_r.$$

Indeed, the diagonal entries of QQ^{T} are just $\sum_{i=0}^{r-1} \chi(i)^2 = r - 1$. Furthermore, the off-diagonal ijth entry is

$$\sum_{k=0}^{r-1} \chi(k-i)\chi(k-j).$$

It can be shown that this sum is equal to -1 for all $i \neq j$.[7]

[7] See Theorem 6, MacWilliams & Sloane.

As a result, for $N = r + 1$ and 1_r an r-long column vector of 1's, we can write

$$H_N = \begin{pmatrix} 1 & 1_r^{\mathrm{T}} \\ 1_r & Q - I_r \end{pmatrix}.$$

Then,

$$H_N H_N^{\mathrm{T}} = \begin{pmatrix} r+1 & 1_r^{\mathrm{T}} + 1_r^{\mathrm{T}}(Q^{\mathrm{T}} - I_r) \\ 1_r + (Q - I_r)1_r & \mathbf{1}_r + (Q - I_r)(Q^{\mathrm{T}} - I_r) \end{pmatrix}$$

The off-diagonal entries are zero as it can be shown that there are, in each row and column of Q, precisely $(r-1)/2$ entries of $+1$ and precisely $(r-1)/2$ entries of -1. Using the above claim, the bottom right entries becomes $(r+1)I_r$. Hence, $H_N H_N^{\mathrm{T}} = N I_N$ making it an Hadamard matrix.