Analyzing 2^{K-p} Fractional Factorial Experiments

- We have seen that 2^{K-p} fractional factorial designs are a clever alternative to full 2^K designs for purposes of factor screening.
  - They still explore K factors, but in just a *fraction* of the conditions required by a full 2^K design.
  - This is made possible by *aliasing* and reliance on the *principle of effect sparsity*.
  - However, this aliasing causes *confounding* which can complicate conclusions.
  - We try to mitigate the negative side-effects of confounding by choosing designs with *maximum resolution* and *minimum aberration*.

- It turns out that the analysis of a 2^{K-p} fractional factorial design is not very different from the analysis of a full 2^K factorial design.
  - We visually summarize effects of interest via main and interaction effect plots
  - Regression models are used to test hypotheses of the form

    $$H_0 : \beta = 0$$

    to determine whether a given effect is significantly different from zero

HOWEVER!

- Now we have to deal with confounding. Recall: two effects that are **confounded** cannot be separately estimated.
  - Just 2^{K-p} effects (and hence β's) can be estimated
  - Each of these β's jointly quantifies 2^p different effects
  - It is therefore important to know the complete aliasing structure of the design so as to be fully aware of *which* effects are confounded

- Accounting for this confounding is particularly important when interpreting effect estimates and evaluating their significance.
  - **The $\mathbf{2^{5-2}_{III}}$ Example:** Suppose we find that the main effect of factor A is significant. What can we conclude?

- The uncertainty surrounding this interpretation motivates why we avoid confounding effects that are likely to be significant with other ones that are also likely to be significant.

The Chehalem Example

- Here we consider an example from Montgomery (2019) in which a 2^{8-4} fractional factorial experiment was used in the production of wine to study the influence of a variety of factors on a particular vintage of Pinot Noir.

- In this experiment $K = 8$ factors were investigated each at two levels (the factors and their levels are shown in the table below) which, if a full factorial experiment was used, would have required 256 conditions.

Factor	Low (−)	High (+)
Pinot Noir clone (A)	Pommard	Wadenswil
Oak type (B)	Allier	Troncias
Age of barrel (C)	Old	New
Yeast/skin contact (D)	Champagne	Montrachet
Stems (E)	None	All
Barrel toast (F)	Light	Medium
Whole cluster (G)	None	10%
Fermemtation temperature (H)	Low (75°F max)	High (92°F max)

- To keep the experiment as small as possible a 2^{8-4}_{IV} fractional factorial experiment was performed that required only 16 conditions.

- The response variable in this case is the rating of the wine as determined by 5 raters.

- Thus, 16 different wines were produced (based on the 16 unique combinations of these factors' levels) and $n = 5$ raters tasted and rated each of them (low scores are good, large scores are bad). The design matrix and a summary of the response is provided in the table below.

Condition	A	B	C	D	E	F	G	H	Average Rating
1	−1	−1	−1	−1	−1	−1	−1	−1	9.6
2	+1	−1	−1	−1	−1	+1	+1	+1	10.8
3	−1	+1	−1	−1	+1	−1	+1	+1	12.6
4	+1	+1	−1	−1	+1	+1	−1	−1	9.2
5	−1	−1	+1	−1	+1	+1	+1	−1	9.0
6	+1	−1	+1	−1	+1	−1	−1	+1	15.0
7	−1	+1	+1	−1	−1	+1	−1	+1	5.0
8	+1	+1	+1	−1	−1	−1	+1	−1	15.2
9	−1	−1	−1	+1	+1	+1	−1	+1	2.2
10	+1	−1	−1	+1	+1	−1	+1	−1	7.0
11	−1	+1	−1	+1	−1	+1	+1	−1	8.8
12	+1	+1	−1	+1	−1	−1	−1	+1	2.8
13	−1	−1	+1	+1	−1	−1	+1	+1	4.6
14	+1	−1	+1	+1	−1	+1	−1	−1	2.4
15	−1	+1	+1	+1	+1	−1	−1	−1	9.2
16	+1	+1	+1	+1	+1	+1	+1	+1	12.6

- Because the response variable in this setting is continuous, we use linear regression to analyze the data from this experiment.

- Because only $2^4 = 16$ conditions were used, we can only fit a model with 16 regression coefficients. In the context of a full 2^4 factorial experiment, this would be the model with 4 main effects, 6 two-factor interactions, 4 three-factor interactions and 1 four-factor interaction:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.5000     0.2658  31.985  < 2e-16 ***
A             0.8750     0.2658   3.293 0.001619 **
B             0.9250     0.2658   3.481 0.000906 ***
C             0.6250     0.2658   2.352 0.021772 *
D            -2.3000     0.2658  -8.655 2.27e-12 ***
A:B          -0.3500     0.2658  -1.317 0.192532
A:C           1.3000     0.2658   4.892 7.07e-06 ***
B:C           0.4500     0.2658   1.693 0.095261 .
A:D          -0.8750     0.2658  -3.293 0.001619 **
B:D           1.2250     0.2658   4.610 1.98e-05 ***
C:D           0.3750     0.2658   1.411 0.163063
A:B:C         1.5750     0.2658   5.927 1.35e-07 ***
A:B:D        -0.3000     0.2658  -1.129 0.263168
A:C:D        -1.0000     0.2658  -3.763 0.000367 ***
B:C:D         1.1000     0.2658   4.139 0.000104 ***
A:B:C:D       0.4750     0.2658   1.787 0.078613 .
---
```

E = BCD

F = ACD

G = ABC

H = ABD

I = ABCDEFGH

AE = ABCD

AF = CD

AG = BC

AH = BD

- But this output does not involve the factors E, F, G or H – it only directly references factors A, B, C and D.

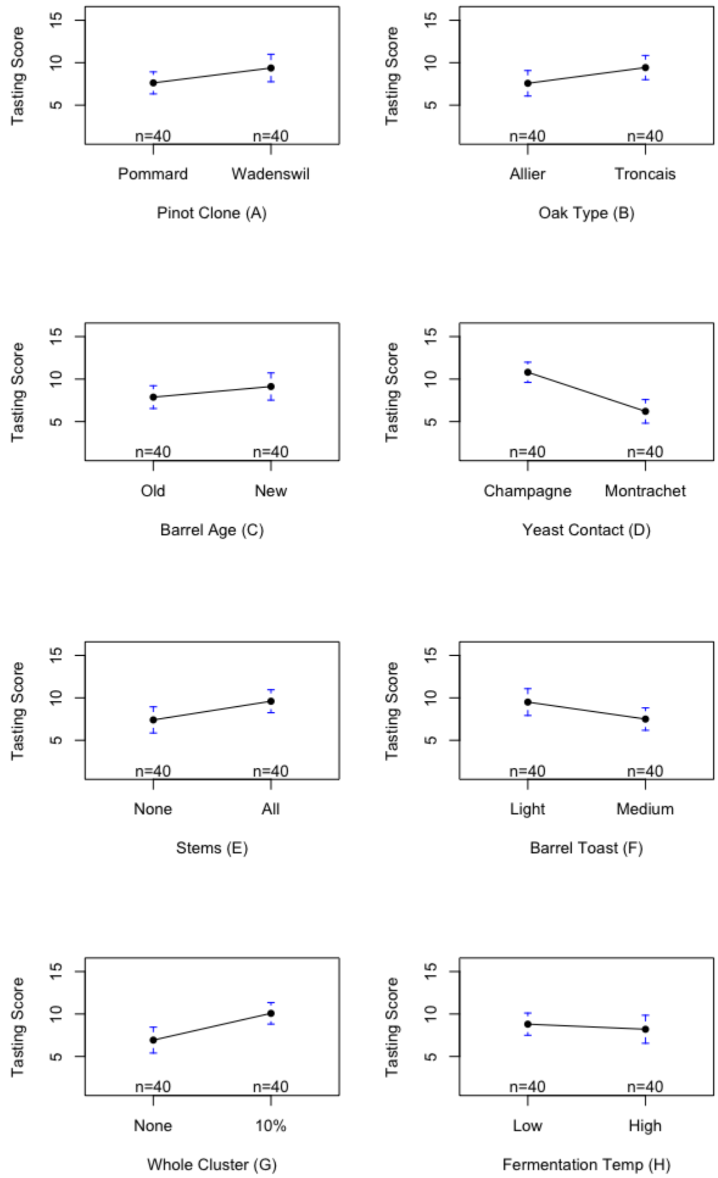
- This is because of confounding.
  - The BCD interaction estimate also corresponds to the main effect of E
  - The ACD interaction estimate also corresponds to the main effect of F
  - The ABC interaction estimate also corresponds to the main effect of G
  - The ABD interaction estimate also corresponds to the main effect of H

- While we cannot technically separate these effects, we assume that the three-factor interactions are neglibible, and hence any significant effect observed is due to the aliased main effect.

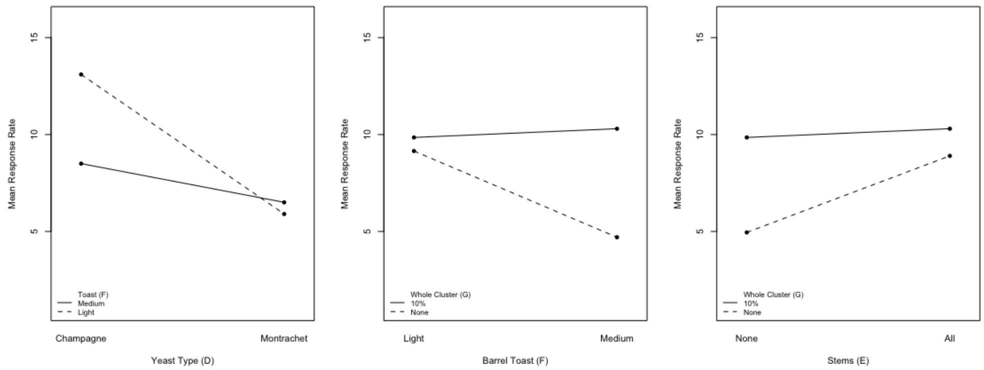
- The same model summary from above is shown below, but this time with factors E, F, G and H referenced instead of the three-factor interactions.

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.5000     0.2658  31.985  < 2e-16 ***
A             0.8750     0.2658   3.293 0.001619 **
B             0.9250     0.2658   3.481 0.000906 ***
C             0.6250     0.2658   2.352 0.021772 *
D            -2.3000     0.2658  -8.655 2.27e-12 ***
E             1.1000     0.2658   4.139 0.000104 ***
F            -1.0000     0.2658  -3.763 0.000367 ***
G             1.5750     0.2658   5.927 1.35e-07 ***
H            -0.3000     0.2658  -1.129 0.263168
A:B          -0.3500     0.2658  -1.317 0.192532
A:C           1.3000     0.2658   4.892 7.07e-06 ***
A:D          -0.8750     0.2658  -3.293 0.001619 **
A:E           0.4750     0.2658   1.787 0.078613 .
A:F           0.3750     0.2658   1.411 0.163063
A:G           0.4500     0.2658   1.693 0.095261 .
A:H           1.2250     0.2658   4.610 1.98e-05 ***
---
```

- The figures below depict main effect plots for all eight factors.



- The figures below depict the interaction effect plots for the three signifcant interactions.



Introduction to Response Surface Methodology

- Effective experimentation is sequential
  - Information gained in one experiment can help to inform future experiments
  - This is the philosophy of **response surface methodology**

- We have seen that the primary purpose of screening experiments is to identify which among a large number of factors are the ones that significantly influence the response variable

- Now we discuss how screening experiments may be followed-up by further experiments whose primary purpose is response optimization
  - We use the **method of steepest ascent/descent** and **response surface designs** to locate optimal settings of the factors that were identified as significant in the screening phase.

Overview of Response Optimization

Coded Factors

- Here we consider $K' \leq K$ design factors which are a subset of the K factors investigated during the screening phase.

- The set of possible values these factors can take on is referred to as the **region of operability**
  - It is this region that we explore and in which we run our experiments to determine the *optimal* operating condition

- Although this region specifies acceptable factor values in their natural units (such as dollars, minutes, percent, etc.), we typically work on a transformed scale.

- Just like in the regression models used in the experiments, we represent each factor by a coded variable x that takes on the values -1 and $+1$ when the factor is at its *low* and *high* levels
  - When the factor is categorical this coding is arbitrary
  - When the factor is numeric the coding arises through the following transformation

$$x = \frac{U - (U_H + U_L)/2}{(U_H - U_L)/2}$$

  - This equation may also be inverted allowing for conversion from the coded units back to the natural units as follows:

$$U = x \times \frac{(U_H - U_L)}{2} + \frac{(U_H + U_L)}{2}$$

- Adopting this notation, the objective of response optimization may be stated as determining the value of $\mathbf{x} = (x_1, x_2, \ldots, x_{K'})^T$ (and hence $\mathbf{U} = (U_1, U_2, \ldots, U_{K'})^T$) at which we expect the response to be optimized.

The Models

- The goal of response optimization may be achieved via **response surface experimentation** where one seeks to characterize the relationship between the expected response E[Y] and the K' design factors

- In the case of a continuous response, we may write this relationship generally as

$$\mathrm{E}[Y] = f(x_1, x_2, \ldots, x_{K'})$$

  and in the case of a binary response

$$\log\left(\frac{\mathrm{E}[Y]}{1-\mathrm{E}[Y]}\right) = f(x_1, x_2, \ldots, x_{K'}).$$

- In both cases, the function $f(x_1, x_2, \ldots, x_{K'})$ respresents the *true* but *unknown* **response surface**

- Because $f(\cdot)$ is unknown, we must fit models that approximate this surface. As usual, we use linear and logistic regression.

- Although many different models may be used to approximate the response surface we exploit Taylor's Theorem and use *low-order polynomials*:
  - First Order model:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{K'} x_{K'}$$

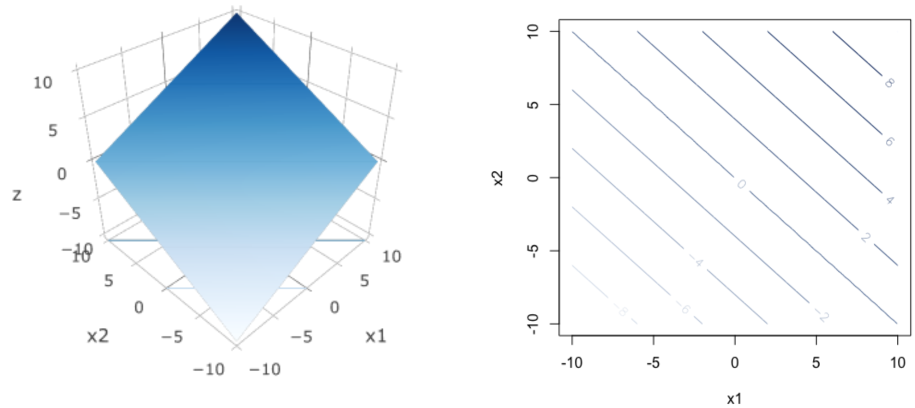
  - First Order + Interaction model:

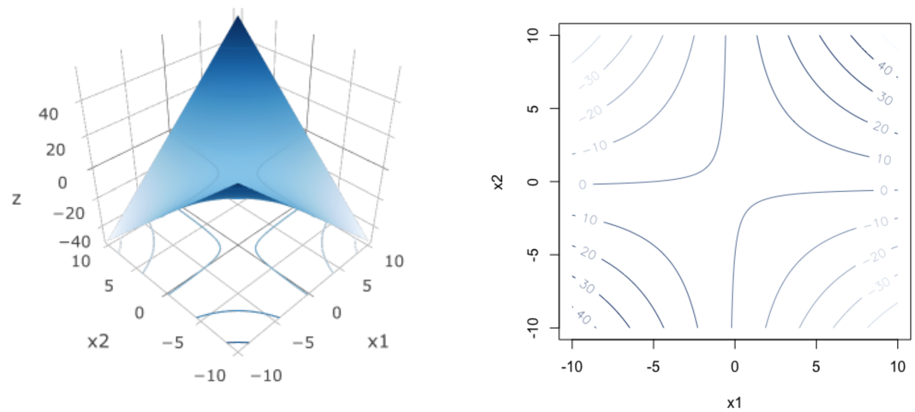
$$\eta = \beta_0 + \sum_{j=1}^{K'} \beta_j x_j + \sum_{j<l} \beta_{jl} x_j x_l$$

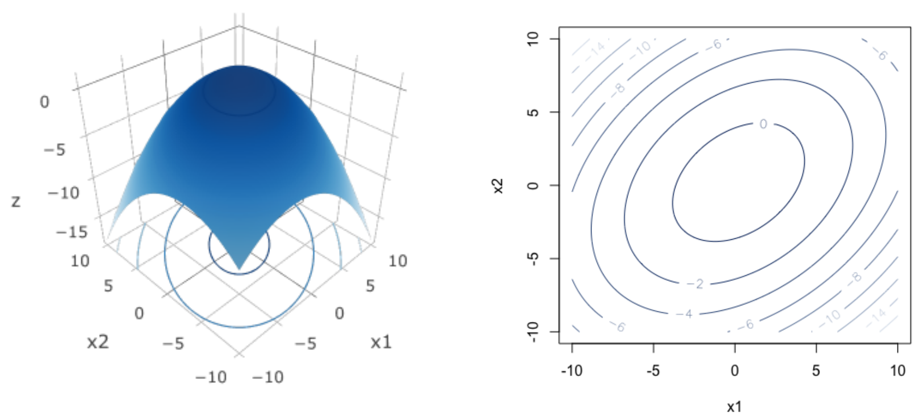
  - Second Order model:

$$\eta = \beta_0 + \sum_{j=1}^{K'} \beta_j x_j + \sum_{j<l} \beta_{jl} x_j x_l + \sum_{j=1}^{K'} \beta_{jj} x_j^2$$

- Examples of such response surfaces (for $K' = 2$) are visualized below:







- We must acknowledge that the approximation of $f(x_1, x_2, \ldots, x_{K'})$ by η (regardless of whether η is first or second order) is likely to be poor when considered across the entire x-space.
  - However, in the small localized region of an experiment, such low-order polynomials should well-approximate $f(\cdot)$.

- Which model is appropriate is dictated by the goal of the experiment.
  - In the context of factor screening we saw that first-order and first-order-plus-interaction models suited our needs
  - But in order to identify maxima/minima we require the second-order model as it is capable of modeling concavity/convexity

Finding the Optimum

- Supposing that sufficient data is collected and the second order model may be fitted, we obtain the estimated response surface

$$\widehat{\eta} = \widehat{\beta}_0 + \sum_{j=1}^{K'} \widehat{\beta}_j x_j + \sum_{j<l} \widehat{\beta}_{jl} x_j x_l + \sum_{j=1}^{K'} \widehat{\beta}_{jj} x_j^2$$

- This expression may be re-written in vector-matrix notation as

$$\widehat{\eta} = \widehat{\beta}_0 + \mathbf{x}^T\mathbf{b} + \mathbf{x}^T\mathbf{B}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{K'} \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_{K'} \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \widehat{\beta}_{11} & \frac{1}{2}\widehat{\beta}_{12} & \cdots & \frac{1}{2}\widehat{\beta}_{1K'} \\ \frac{1}{2}\widehat{\beta}_{12} & \widehat{\beta}_{22} & \cdots & \frac{1}{2}\widehat{\beta}_{2K'} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\widehat{\beta}_{1K'} & \frac{1}{2}\widehat{\beta}_{2K'} & \cdots & \widehat{\beta}_{K'K'} \end{bmatrix}$$

- In order to find the value of $\mathbf{x} = (x_1, x_2, \ldots, x_{K'})^T$ that maximizes/minimizes the expected response, we must find the **stationary point** of the estimated response surface.

- The stationary point is

$$\mathbf{x}_s = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b}$$

which is found by solving

$$\frac{d\widehat{\eta}}{d\mathbf{x}} = \mathbf{b} + 2\mathbf{B}\mathbf{x} = \mathbf{0}$$

$\hat{\beta}_0 + x_s^T b + x_s^T B x_s$

$= \hat{\beta}_0 + (-\frac{1}{2} b^T B^{-1} b) + \frac{1}{4} b^T B^{-1} B B^{-1} b = \hat{\beta}_0 - \frac{1}{4} b^T B^{-1} b = \hat{\beta}_0 + \frac{1}{2} x_s^T b$

- The optimal expected response is
$$\widehat{\eta}_s = \widehat{\beta}_0 + \frac{1}{2}\mathbf{x}_s^T\mathbf{b}$$
in the case of linear regression and
$$\frac{e^{\widehat{\eta}_s}}{1+e^{\widehat{\eta}_s}} = \frac{e^{\widehat{\beta}_0+\frac{1}{2}\mathbf{x}_s^T\mathbf{b}}}{1+e^{\widehat{\beta}_0+\frac{1}{2}\mathbf{x}_s^T\mathbf{b}}}$$
in the case of logistic regression

- For practical implementation of this solution, the stationary point $\mathbf{x}_s$ must be translated into optimal operating conditions in natural units $\mathbf{U}_s$ using the following conversion formula:
$$U = x \times \frac{(U_H - U_L)}{2} + \frac{(U_H + U_L)}{2}$$

BUT!

- For us to be confident that $\mathbf{x}_s$ indeed optimizes $f(\cdot)$, we must be confident that η and, in particular, that $\widehat{\eta}$ adequately represents $f(\cdot)$
  - Since we only expect the second-order approximation to be adequate in a small localized region, it is important that this small localized region contains the true optimum
  - It is quite unlikely that the values of $x_1, x_2, \ldots, x_{K'}$ considered in the screening phase are close to the optimum
  - This is why we needed the method of steepest ascent/descent
    - This intermediate phase of experimentation helped us determine roughly where the region of the optimum lies

Optional Exercises:

- Calculations: 11, 12, 14, 15
- R Analysis: 12(b)-(d), 21(b)-(c), 26(b)
- Communication: 1(g)