Advanced Engineering Statistics
- Section 5 -

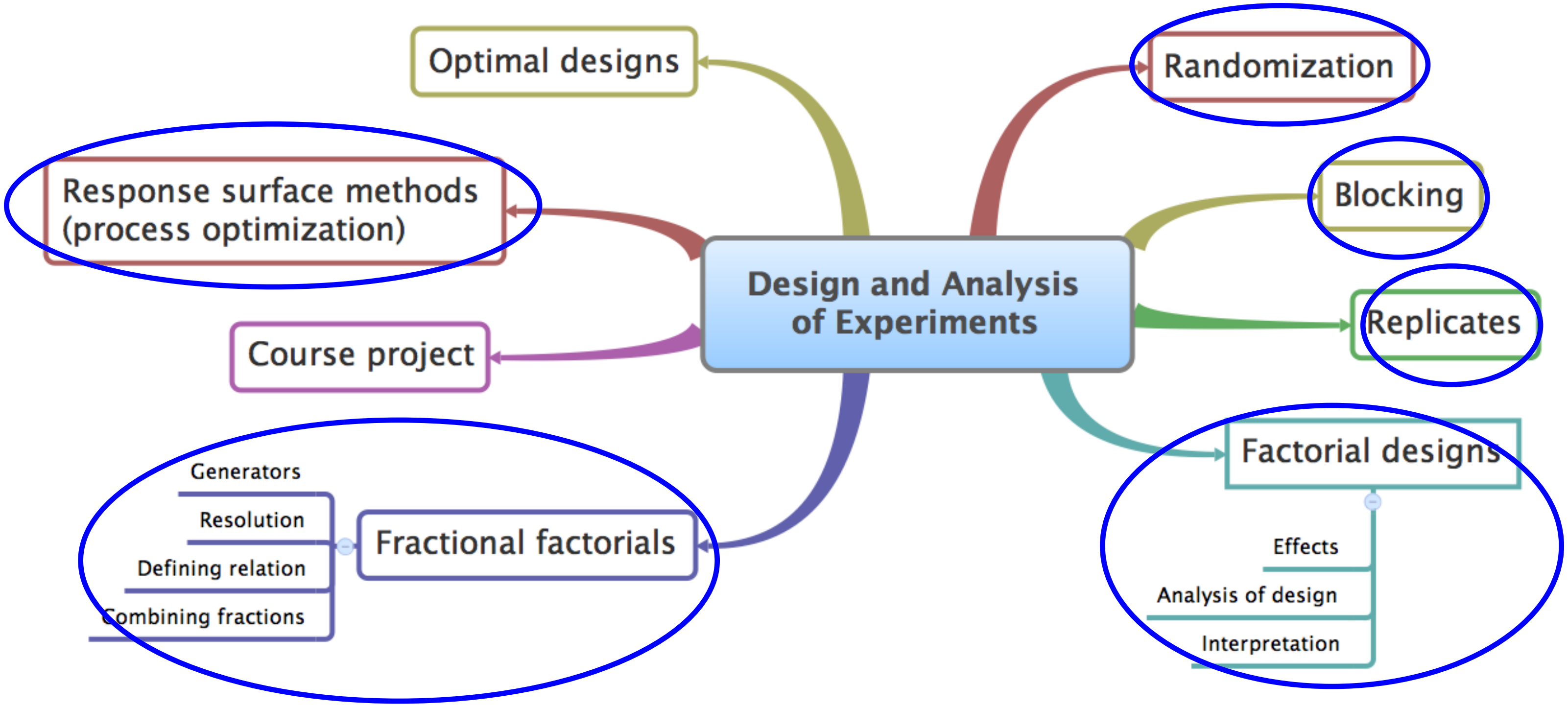
Jay Liu

Dept. Chemical Engineering

PKNU

Design of Experiments

- What we will cover



Reading:
http://www.chemometrics.se/index.php?option=com_content&task=view&id=18&Itemid=27

Fathers of Modern Experimental Design

Sir Ronald A. Fisher (1890-1962).

- Statistician, evolutionary biologist, eugenicist and geneticist.
- Credited with ANOVA and DOE
- *The Design of Experiments* (1935)

George E.P. Box, Professor Emeritus (1919~)

- Founder of Stat Department @ Univ. of Wisconsin, Madison
- DOE and RSM
- Famous quote: "Essentially, all models are wrong, but some are useful“.

Usage examples

- Colleague: 8 process variables seem to affect melt index. How to narrow them down? Which one has most effect on y?
- Engineer: 3 manipulated variables of interest; how to run the experiments?
- Manager: how do we analyze experimental data to optimize our process?
- Colleague: small changes in the flowrate lead to unsafe operation. Where can we operate to get similar results, but more safely?

Why design?

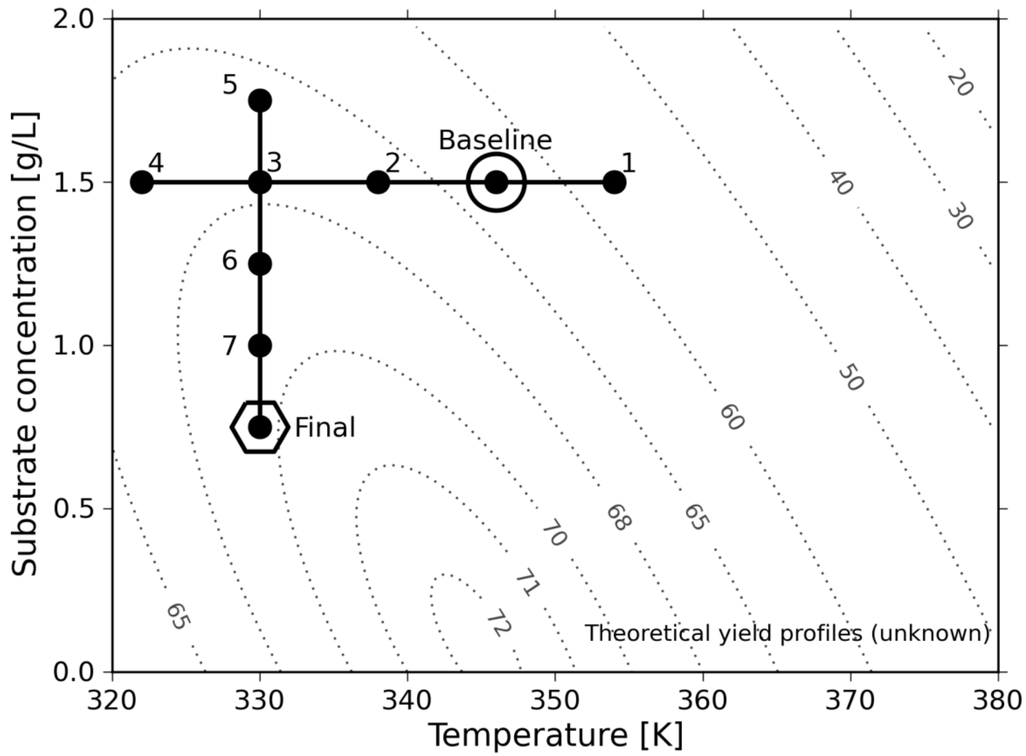
1. Ensure adequate variability in all key variables.
   - Variable *x* may have very important effect on process performance.
   - But if variation in it is small relative to noise level, then may
     - Accept H_0: effect of $x = 0$
     - Obtain confidence interval on effect of *x* to include zero.
   - This does not necessarily mean that effect of *x* is not important – only that it isn't large enough in this particular data set to detect significance.
   - Design of experiments provides a form of guarantee that accepting H_0 implies that the effect is not important.

Why design?

2. Ensure identifiability of all important effects & *interactions*
   - DOE helps ensure that all important main effects and interaction can be identified – minimizes confounding
   - Our bad experimental habits arise from the nature of university laboratories:
     - These undergrad labs aimed at demonstrating theoretical principles, not a building models, exploring for unknown effects, or optimizing processes.
     - Ex. Demonstrate the effect of temp. on reaction equilibrium – changing temp. holding all other variables constant!
   - COST approach is not good when searching for effects, building models, or optimizing processes.

[FYI]Changing One variable at a Single Time (COST)

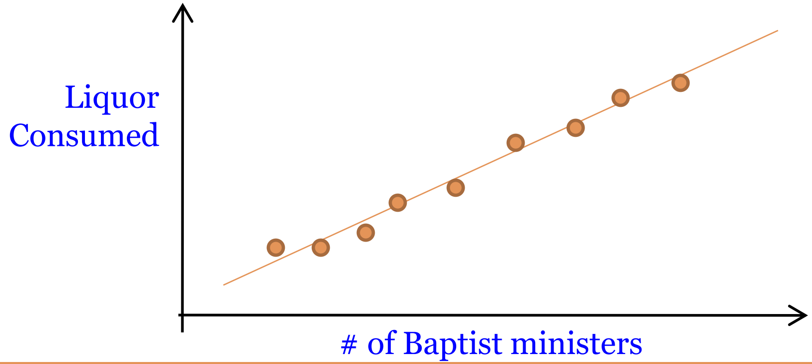
- We can hardly find values of conc. & temp. for max. yield using COST approach



- DOE: efficient ways of changing many variables at once

Why design?

3. Maximize the information obtained in fewest number of experiments
   - Examples of industrial screening experiment
     - Problem: in a new plant the cycle time in the filtration section was unacceptably long.
     - Need to de-bottleneck
     - Many factors suggested that might be responsible.
     - How to screen out important ones in fewest runs possible?
4. Distinguish between causality and correlation
   - Data from Australia over many years on
     - # of Baptist minister vs. amount of liquor consumed
     - Strong correlation? Causal effect?



Why design? (in plain words)

- The objective of experimental design is **to relate independent variables to dependent variables as efficiently as possible** (i.e., fewest number of experiments).

- Two general types of experimental design:
  - ***Screening***–Define the important variables or “Main effects”. (through Factorial design, fractional factorial design, …)
  - ***Empirical modeling***–Develop approximate models of true systems for further use. (Response surface method, …)

Analysis of effects of a single variable at two levels

- Simplest case:
  - catalyst A vs catalyst B
  - low RPM vs high RPM
  - Etc
- Measure n_A value from setup A
- Measure n_B values from setup B
- Hold all other variables constant (control disturbances)

➔ Two ways to answer this:

- Comparing means of X and Y
- Least squares

Using confidence interval of $\bar{X} - \bar{Y}$

- Test for difference

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A - 1 + n_B - 1} \qquad \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}}} \sim t_{n_A + n_B - 2}$$

- Confidence interval

$$\left[(\bar{X} - \bar{Y}) - t_{n_A + n_B - 2, \alpha/2} \sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}}, (\bar{X} - \bar{Y}) + t_{n_A + n_B - 2, \alpha/2} \sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}} \right]$$

-

Using least squares

- The same result can be achieved using least squares: $y_i = a_o + a_1 d_i$
  - d_i = 0 for A; d_i = 1 for B; y_i : the response variable

EXAMPLE : Etch rate of solutions 1 & 2

Solution 1	Solution 2
9.9	10.2
9.4	10.6
9.3	10.7
9.6	10.4
10.2	10.5
10.6	10.0
10.3	10.2
10.0	10.7
10.3	10.4
10.1	10.3

Engineers @ a semiconductor manufacturing plant want to know which solution has higher etching rate.

Using least squares (cont.)

- C. I approach

$$(\bar{x}_1-\bar{x}_2)-t_{\alpha/2,n_1+n_2-2}(s_p)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\le \mu_1-\mu_2\le(\bar{x}_1-\bar{x}_2)+t_{\alpha/2,n_1+n_2-2}(s_p)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$(9.97-10.4)-2.101(.340)\sqrt{\frac{1}{10}+\frac{1}{10}}\le \mu_1-\mu_2\le(9.97-10.4)+2.101(.340)\sqrt{\frac{1}{10}+\frac{1}{10}}$$

$$-0.749\le \mu_1-\mu_2\le -0.111$$

Zero included?

- LS approach
  - Find a LS solution in the model: $y = a_0 + a_1 d$
  - d_i = 0 for 1; d_i = 1 for 2; y_i : etching rate

Confidence intervals of a_0 & a_1

	value	S.E	t statistic	P-value	L.B 95%	U.B 95%
a0	9.97	0.107523	92.72469	1.41E-25	9.744103	10.1959
a1	0.43	0.15206	2.827832	0.011151	0.110534	0.749466

Zero included?

  - Same result and more (**significance test + prediction model**)

Several concepts in DOE

- Randomization and blocking
  - Comparative experiment: effect of two methods on strength of rubber strip
    - Run experiments Run order →

A	A	A	...	A	B	B	B	...	B

and do significance test (C.I of $\bar{X}_A - \bar{X}_B$) or least squares($y_i = a_o + a_1 d_i$)

..... Any problem with this?

- What if strip of rubber had variation along its length?

Then, $\bar{X}_A - \bar{X}_B$ might just be reflecting this difference.

- One solution → randomize allocation of rubber portion to methods (A&B)

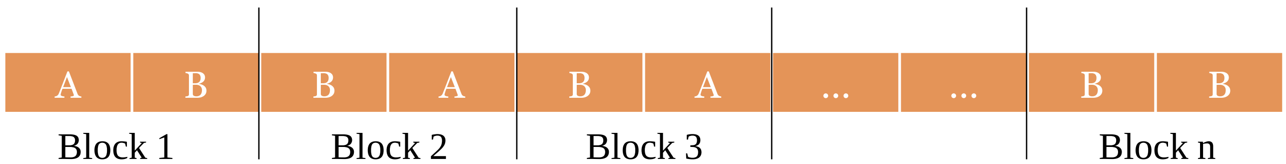
A	B	B	A	B	A	...	A	B	B

...... No problem with this?

Concepts in DOE - Randomization and blocking

- Suppose we expect variation in rubber to be progressive along length of the strip! Then, two different adjacent portion will be much more similar than two distant ones.

  → **block into pairs** of adjacent pieces. Assign methods (A&B) randomly within block .



(Randomized block design)

And only compare within block

block	A B	$D = X_A - X_B$
1	X_{A1} X_{B1}	$d_1 = X_{A1} - X_{B1}$
2	X_{A2} X_{B2}	d_2
…	…	…
n	X_{An} X_{Bn}	d_n

Blocking can remove effect of possible uncontrolled variations along the length of strip (**remember advantage of paring**)

Designs for experimental studies

※Objectives

- Screening studies

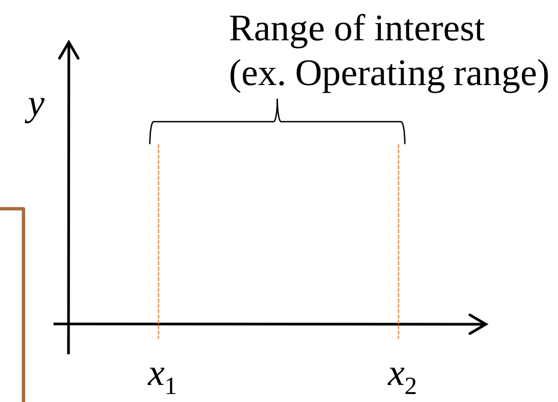
  : discovering which of a large number of variations affect response

- Empirical model building studies

  : true model unknown. Use approximate models, $y = f(x_1, x_2, \ldots, x_k)$

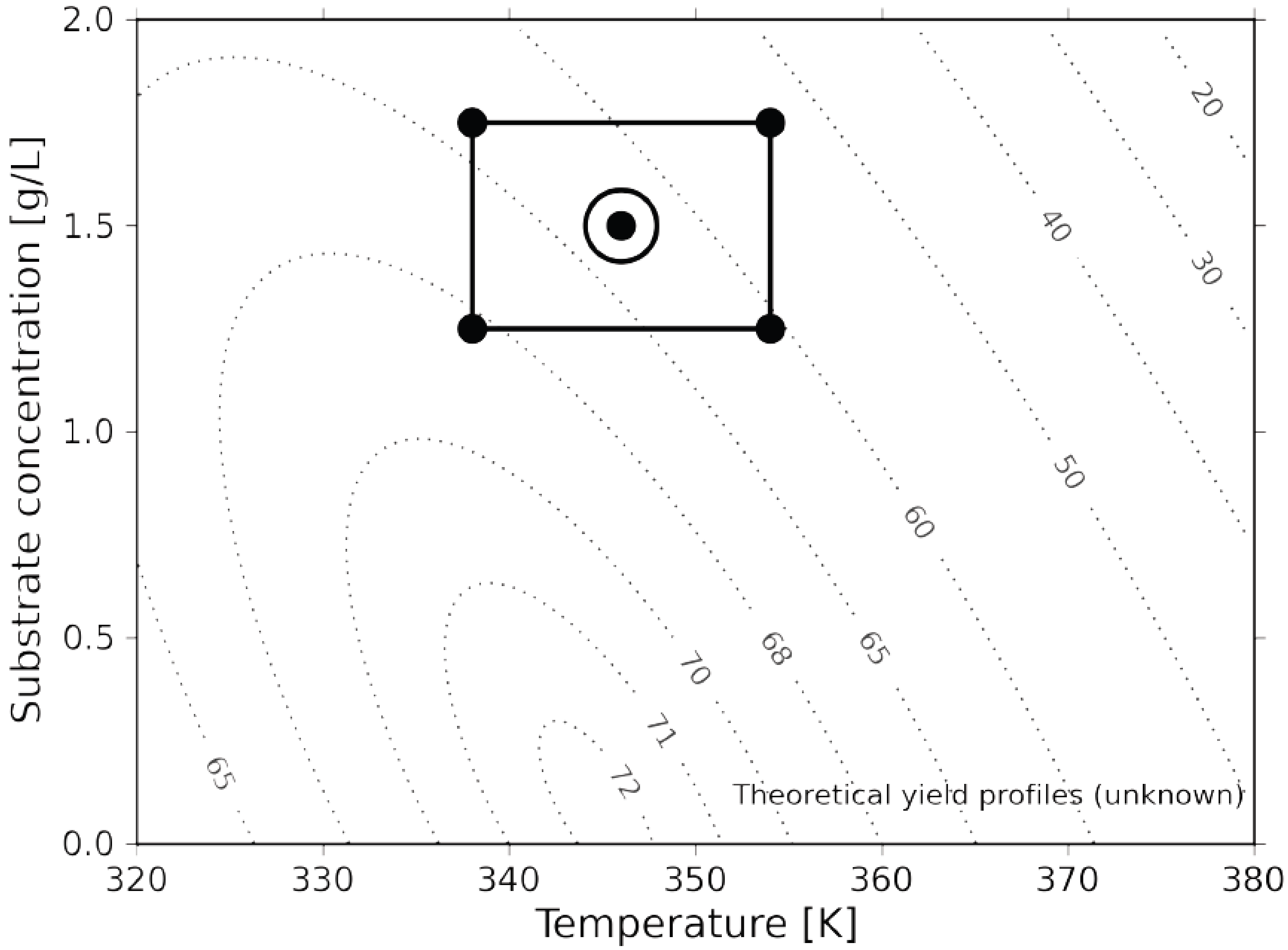
2^k factorial designs

- Want to estimate of *linear* effect of x on y.
- Best 2 experiments?



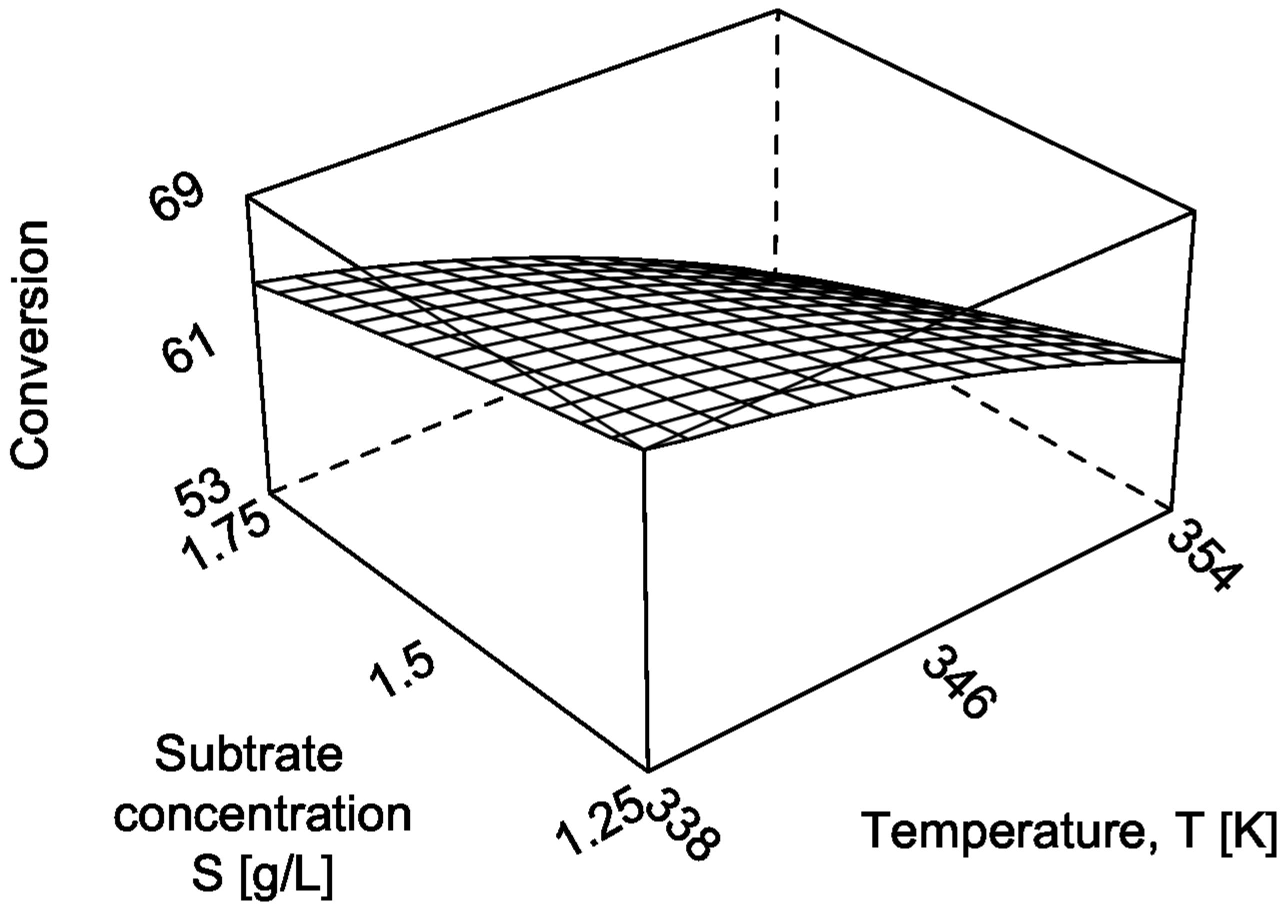
2^2 factorial design

We will use this system for our example.



2^2 factorial design (Cont.)

- This is the true surface plot



2^2 factorial design (Cont.)

- Two independent variables:

	range
Temperature (T, K)	338 ~ 354
Concentration (S, g/L)	1.25 ~ 1.75

  - Study effect of T & S on conversion y (%).

Two variables

- Design: 2^2 factorial in $2^2 = 4$ runs

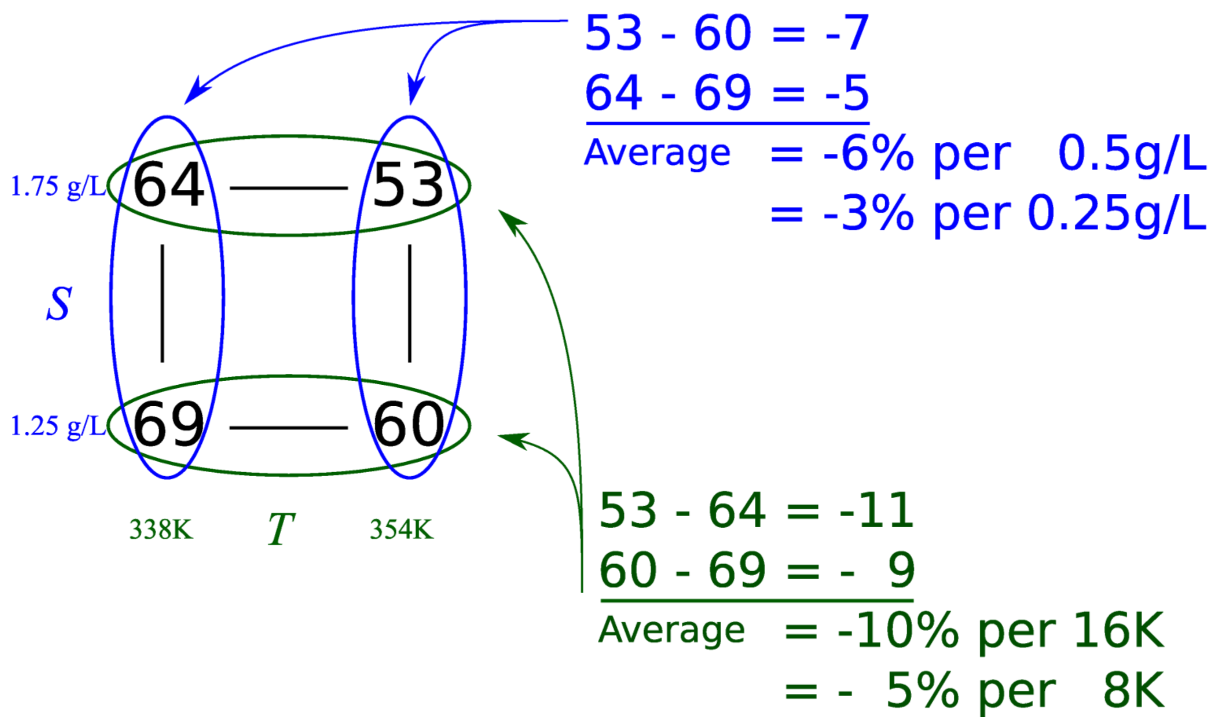
All possible combination
of two levels of two variables

Two levels

- & run the experiments:

2^2 factorial design (Cont.)

- Main effects of T & S



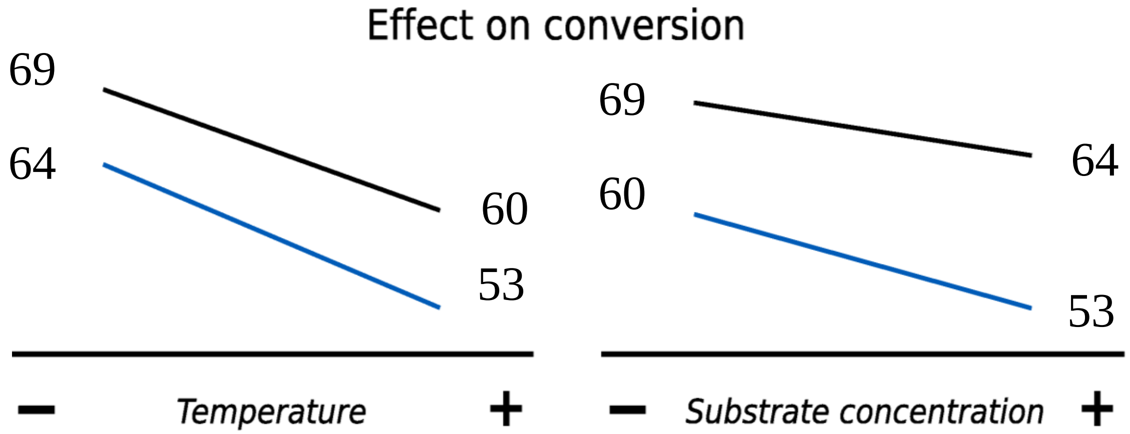
- Almost no difference between the values within each main effect (see interaction plot)

BTW, where would you run your next experiment(s) to improve yield?

2^2 factorial design (Cont.)

- Interaction between T & S
  - Do variables T & S act independent on y?
  - Or, is effect of T (or S) same at both levels of S (or T)?
  - If effect is different → T x S interaction.

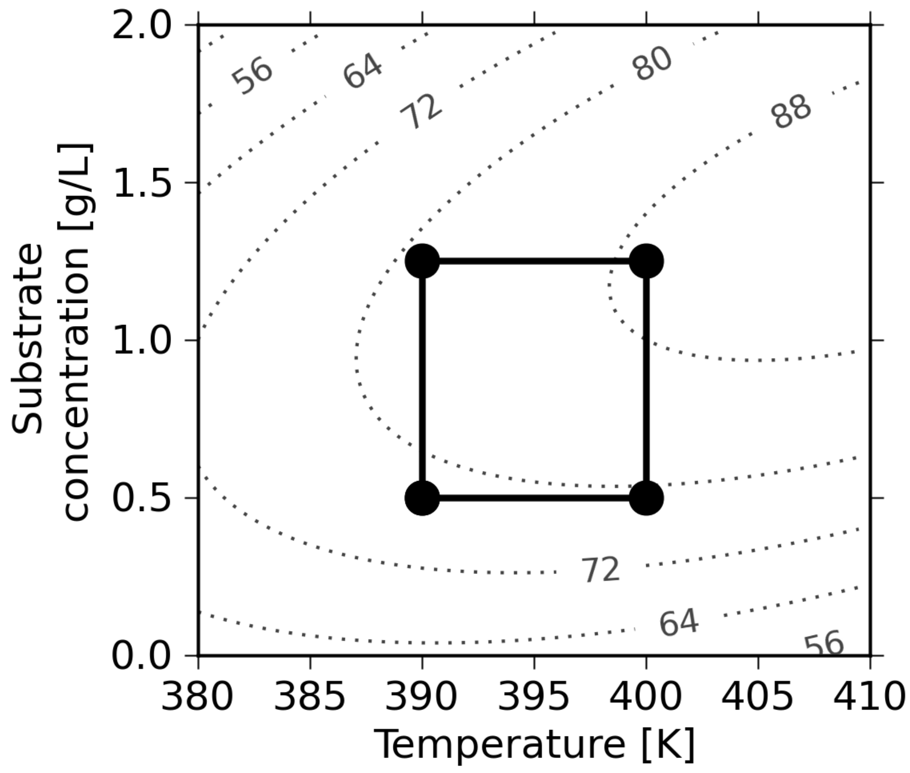
Visualize this with an interaction plot.



Lines are roughly parallel.

2^2 factorial design (Cont.)

- Consider another case



2^2 factorial design (Cont.)

- Main effects of T & S

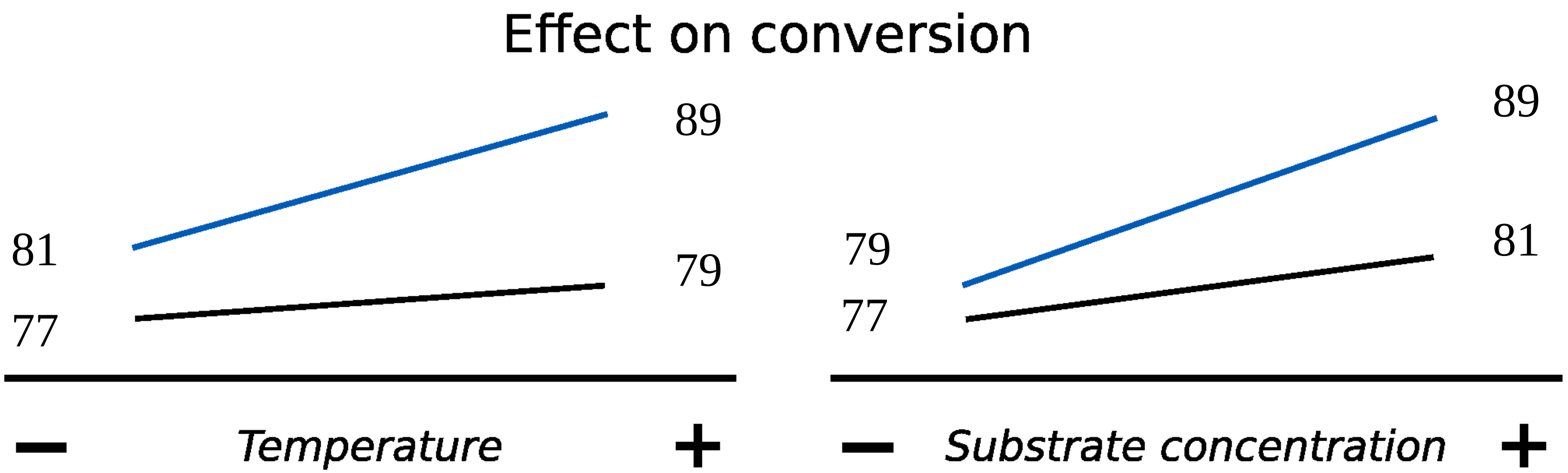
Experiment	T [K]	S [g/L]	y [%]
1	− (390K)	− (0.5 g/L)	77
2	+ (400K)	− (0.5 g/L)	79
3	− (390K)	+ (1.25 g/L)	81
4	+ (400K)	+ (1.25 g/L)	89

- Main effect of T: **5% per 10K**
  - $\Delta T_{S+} = 8\%$ per 10K
  - $\Delta T_{S-} = 2\%$ per 10K
- Main effect of S: **7% per 0.75g/L**
  - $\Delta S_{T+} = 10\%$ per 0.75g/L
  - $\Delta S_{T-} = 4\%$ per 0.75g/L

- There was an important phenomenon that we did not capture with the main effects alone.

2^2 factorial design (Cont.)

- Interaction between T & S



- Lines not parallel
- Implies there is an interaction
  - In this case, interaction between T and S
  - called the $T \times S$ interaction
  - it is a 2-factor interaction

Analysis by least squares (Cont.)

- Back to the 1st example (little interaction)
  - Design matrix (condition) & experimental results

T	S	y
338	1.25	69
354	1.25	60
338	1.75	64
354	1.75	53

※Center: usually current condition

  - Transform *x* variable (T & S) to scaled variables

※why?: remove scale effect

$$x_i = \frac{\text{variable} - \text{centerpoint}}{\text{Range} / 2}$$

$$x_1 = \frac{\text{T} - 346}{8}$$

$$x_2 = \frac{\text{C} - 1.5}{0.25}$$

Range of x_i's

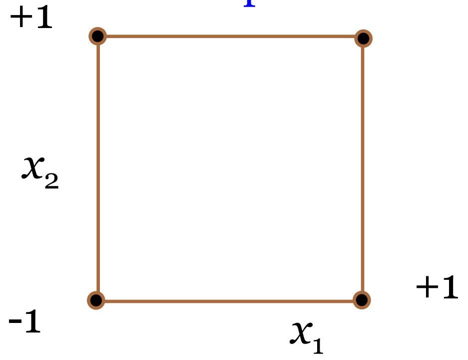
-1	to	+1
-1	to	+1

Analysis by least squares (Cont.)

- Design matrix becomes

x_1	x_2
-1	-1
+1	-1
-1	+1
+1	+1

"Cube plot"



- Fit model: $y = a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2$ Interaction term
  - four parameters & four data points

  → Zero D.O.F (no C.I possible)

Analysis by least squares (Cont.)

x_1	x_2	y
-1	-1	69
+1	-1	60
-1	+1	64
+1	+1	53

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2$$

In matrix-vector notation,

$$\mathbf{y} = \mathbf{X}\mathbf{a}$$

$1 \quad x_1 \quad x_2 \quad x_1 x_2$

$$\mathbf{y} = \begin{bmatrix} 60 \\ 72 \\ 54 \\ 68 \end{bmatrix} \qquad \mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_{12} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \quad \end{bmatrix}$$

- Regression coefficients (usually from S/W)

$$\mathbf{a} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

Columns of $\mathbf{X}$: orthogonal (i.e., $\mathbf{x}_i \cdot \mathbf{x}_j = \mathbf{x}_i^T \mathbf{x}_j = 0$)

$$\rightarrow \sum x_0 x_1 = \sum x_0 x_2 = \sum x_0 (x_1 x_2) = \sum x_1 x_2 \sum x_1 (x_1 x_2) = \sum x_2 (x_1 x_2) = 0$$

Analysis by least squares (Cont.)

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} \quad \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix} \qquad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} \quad \end{bmatrix}$$

$$\mathbf{a} = \begin{bmatrix} 1/\sum x_0^2 & 0 & 0 & 0 \\ 0 & 1/\sum x_1^2 & 0 & 0 \\ 0 & 0 & 1/\sum x_2^2 & 0 \\ 0 & 0 & 0 & 1/\sum (x_1 x_2)^2 \end{bmatrix} \begin{bmatrix} \sum x_0 y \\ \sum x_1 y \\ \sum x_2 y \\ \sum (x_1 x_2) y \end{bmatrix}$$

- i.e., $a_i = \dfrac{\sum x_i y_i}{\sum x_i^2}$

Each a_i can be calculated independently.

e.g., $a_0 = \dfrac{y_1 + y_2 + y_3 + y_4}{4}$

a_i = effect of changing variable x_i from 0 to +1.

Analysis by least squares (Cont.)

- Confidence interval of a_i
  - Four data points & four parameters: D.O.F is zero
  - Can't calculate C.I unless
    - σ is known
    - S can be calculated from replicates (or historical database)

$$\mathrm{var}(\mathbf{a}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2 \Rightarrow \mathrm{var}(a_i) = \frac{\sigma^2}{\sum x_i^2}$$

a_i are uncorrelated due to orthogonality of design

  - σ is known

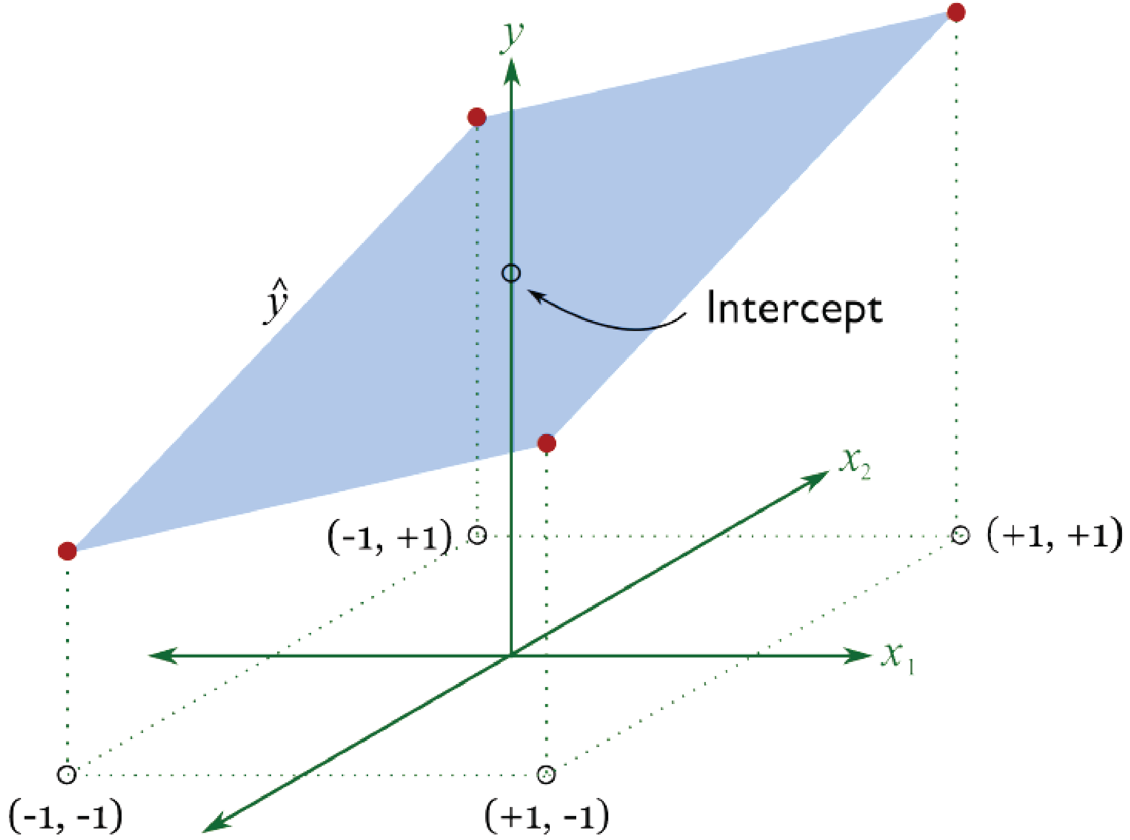
95% C.I $\quad a_i \pm z_{0.025}\sqrt{\sigma^2 / \sum x_i^2}$

  - S is known

95% C.I $\quad a_i \pm t_{\nu,0.025}\sqrt{s_{y/x}^2 / \sum x_i^2}$ (when σ unknown)

Analysis by least squares (Cont.)

- Least squares model for DOE in 2 factors



- Interaction term is small: blue plane is flat
- Interaction term is large: plane has curvature

Analysis by least squares (Cont.)

- Calculation by hand: 1st example (little interaction)

$$\mathbf{y} = \begin{bmatrix} 60 \\ 72 \\ 54 \\ 68 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & +1 & +1 \end{bmatrix} \quad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} 246 \\ -20 \\ -12 \\ -2 \end{bmatrix} \quad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

$$\mathbf{a} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{bmatrix} \begin{bmatrix} 246 \\ -20 \\ -12 \\ -2 \end{bmatrix} = \begin{bmatrix} 61.5 \\ -5 \\ -3 \\ -0.5 \end{bmatrix}$$

$$\begin{aligned} y &= a_0 + a_1x_1 + a_2x_2 + a_{12}x_1x_2 \\ &= 61.5 - 5x_1 - 3x_2 - 0.5x_1x_2 \end{aligned}$$

Analysis by least squares (Cont.)

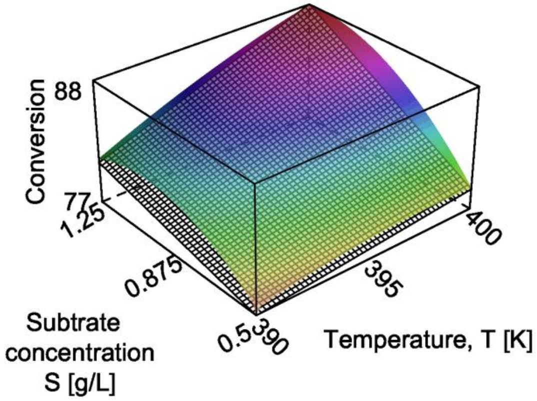
- $\mathbf{X}^T\mathbf{X}$: zeros on off-diagonals
  - orthogonal matrix
  - each column is varied independently of the others
- Interpret $a_1 = -5$?
  - x_1 (T) is changed in normalized temperature by 1 unit
  - Changing x_1 from 0 to 1 implies actual changes in T from 346K to 354K
  - -5% decrease in conversion for every 8K increase in temperature
- Interpret a_2 (S) $= -3$?

Analysis by least squares (Cont.)

- Calculation by hand: 1st example (strong interaction)

$$\mathbf{y} = \begin{bmatrix} 77 \\ 79 \\ 81 \\ 89 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & +1 & +1 \end{bmatrix}$$

  - Verify this yourself $y = 81.5 + 2.5x_1 + 3.5x_2 + 1.5x_1x_2$
    - Large interaction is verified.



Any stat. S/W can generate this.

2^3 factorial design

2^3 → 3 variables
→ 2 levels

Qualitative variable

- Three variables: T, C, and catalyst type (A and B)
  - Denote: x_3 = -1 for catalyst A

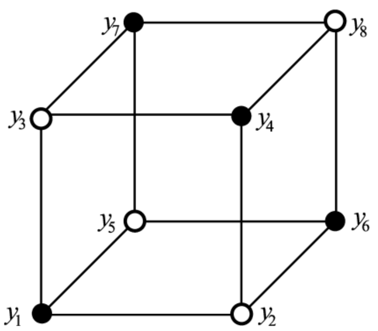
    = +1 for catalyst B
  - 2^3 factorial (= 8 runs): all combination of the 2 levels of the 3 variables.

Design

Matrix, **X**

x_0	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
+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

"Cube plot"



2^3 factorial design (cont.)

- Analysis by least squares
  - Fit model:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3 + a_{123}x_1x_2x_3$$

    In matrix-vector notation,

$$\mathbf{y} = \mathbf{X}\mathbf{a}$$

  - Again, by least squares

$$\mathbf{a} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \Rightarrow a_i = \frac{\sum x_i y}{\sum x_i^2}$$

C.I of a_i $\quad \mathrm{var}(\mathbf{a}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2 \Rightarrow \mathrm{var}(a_i) = \frac{\sigma^2}{\sum x_i^2}$

95% C.I $\quad a_i \pm z_{0.025}\sqrt{\sigma^2 / \sum x_i^2}$

95% C.I $\quad a_i \pm t_{\nu,0.025}\sqrt{s_{y/x}^2 / \sum x_i^2}$ (when σ unknown)

2^3 factorial design example

Plastics molding factory; waste treatment.

- Factor 1: C: chemical compound added (A or B)
- Factor 2: T: treatment temperature (72 F or 100F)
- Factor 3: S: stirring speed (200 rpm or 400 rpm)
- y = amount of pollutant discharged [lb]

Experiment	Order	C	T [°F]	S [rpm]	y [lb]
1	5	A	72	200	5
2	6	B	72	200	30
3	1	A	100	200	6
4	4	B	100	200	33
5	2	A	72	400	4
6	7	B	72	400	3
7	3	A	100	400	5
8	8	B	100	400	4

2^3 factorial design example

1. Calculate main effects, C, T and S
2. Calculate the 3 two-factor interactions:
   1. CT, CS and TS
3. and the single 3 factor interaction
   1. CTS
4. Main effects and interactions using least squares (by-hand)
5. S/W verification:

$$y = 11.25 + 6.25x_C + 0.75x_T - 7.25x_S + 0.25x_Cx_T - 6.75x_Cx_S - 0.25x_Tx_S - 0.25x_Cx_Tx_S$$

2^k factorial design

- Desirable features of factorial designs
  - Othogonal → easy calculations

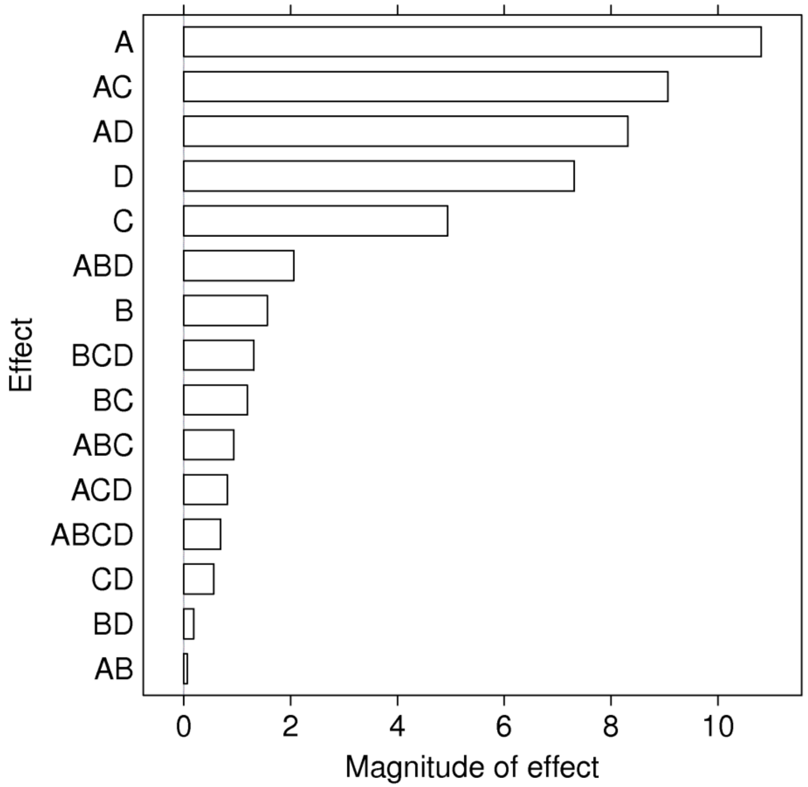
    → uncorrelated estimates a_i
  - Good variation in all variables
  - Efficient use of all data points
  - The only way to discover interactions between variables
  - Allows experiments to be performed in blocks
  - Allows designs of increasing order to be build up sequentially

Significance of effects

- For a 2^k factorial
  - 2^k parameters in the least squares model
  - 2^k data points collected
  - implies $S_r = 0$
  - Zero degrees of freedom
- How to assess if an effect is significant? Consider 2 approaches.

Significance of effects (cont.)

- Significant? : Pareto-plot (or normal probability plot)
  - 2^4 factorial: 15 parameters + intercept
  - Bar plot: any stat. S/W can do this.



Significance of effects (cont.)

- Caution: if an interaction is significant (e.g. BC), then no need to test the main effects, B and C
  - these main effects are ”automatically” significant
  - even if they have small numeric coefficients
  - since B and C act together to affect response y
  - so never exclude main effects whose interactions are significant

Significant effect?

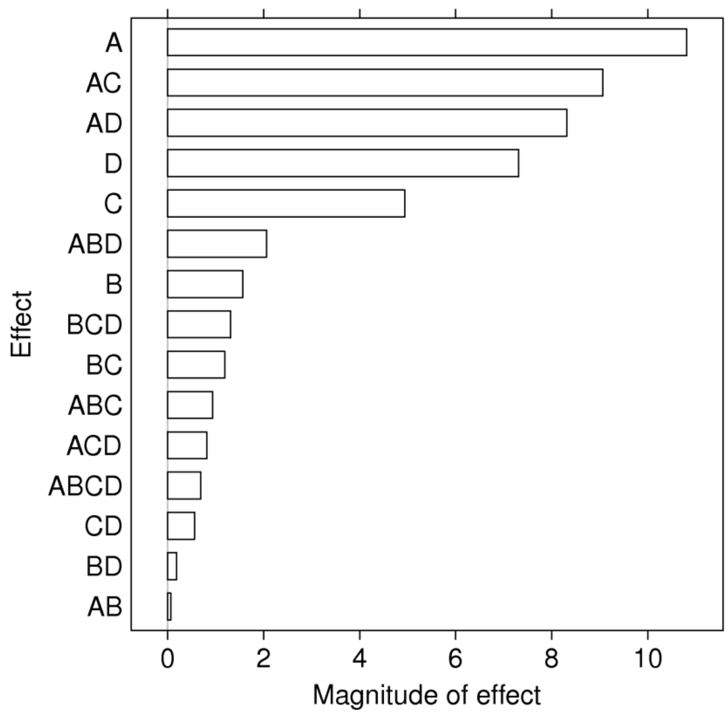
- We require degrees of freedom to construct confidence intervals. Two ways to get DoF:
  1. Replicate experiments
     - Easy (to calculate), but not doable when # of factor 4 ~
  2. Drop out a factor from a full factorial
     - Will five factor interaction $x_1 x_2 x_3 x_4 x_5$ be significant?
     - Or, drop smallest effects first.
  - In either case, delete non-significant effects (parameters) and re-fit
  - Now least squares model has new residuals and DOF.
  - Use all previous tools from least squares to check model
  - Use confidence interval of a_i to verify the effects are significant

Significant effect? (cont.)

- Replicate runs
  - replicated 2^3 factorial: 8 + 8 runs
  - $y_{i,1}$ & $y_{i,2}$ at condition i (i = 1, 2, . . . , 8)
  - $\bar{y}_i = 0.5(y_{i,1} + y_{i,2}), \quad d_i = y_{i,2} - y_{i,1}$
  - $s_i^2 = \frac{1}{2-1}\left(\left(y_{i,1} - \bar{y}_i\right)^2 + \left(y_{i,2} - \bar{y}_i\right)^2\right) = \frac{1}{2} d_i^2$
  - Pool variances for all 2^k levels
  - $s_{y/x}^2 = \frac{1}{2}\sum_i^{2^k} d_i^2$
  - Errors are t-distributed with 2^k degrees of freedom
  - $\therefore$ 95% C.I $\quad a_i \pm t_{\nu,0.025}\sqrt{s_{y/x}^2 / \sum x_i^2}$
  - determine if a main effect or interaction is significant

Significant effect? (cont.)

- No replicates
  - 2^4 factorial: 15 parameters + intercept → DOF (#data - #parameters) = 0
  - AB seems insignificant → set $a_{AB} = 0$ → now, DOF = 1

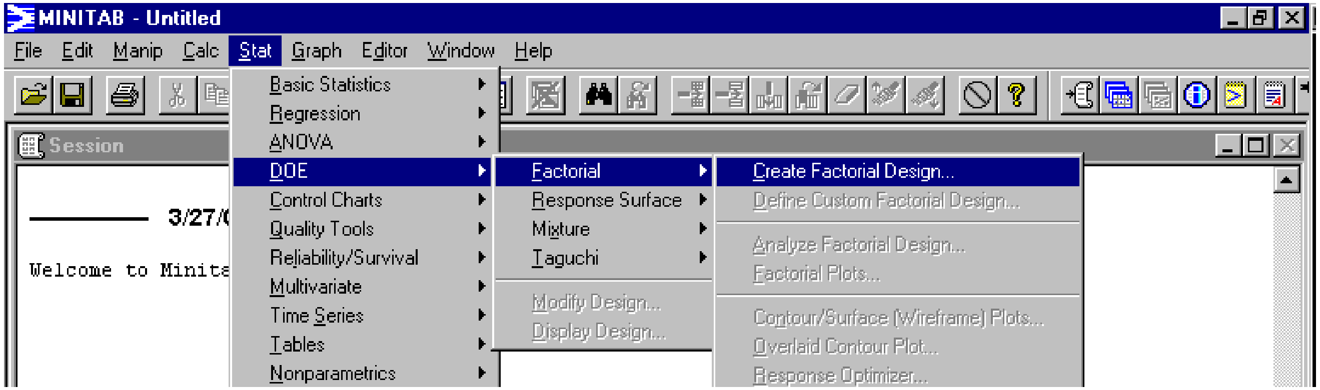


Exercise

- When you have replicates.
  - You're a process engineer @ a semiconductor plant who wants to determine factors affecting thickness of epitaxial layer on silicon wafer. The main factors (or input variables) you think are (deposition) time and (arsenic) flowrate. Assume only linear relationship.
  - Solution

1. 2^2 factorial design with 4 replicates @ corners

   stat>DOE>Factorial>create Factorial Design



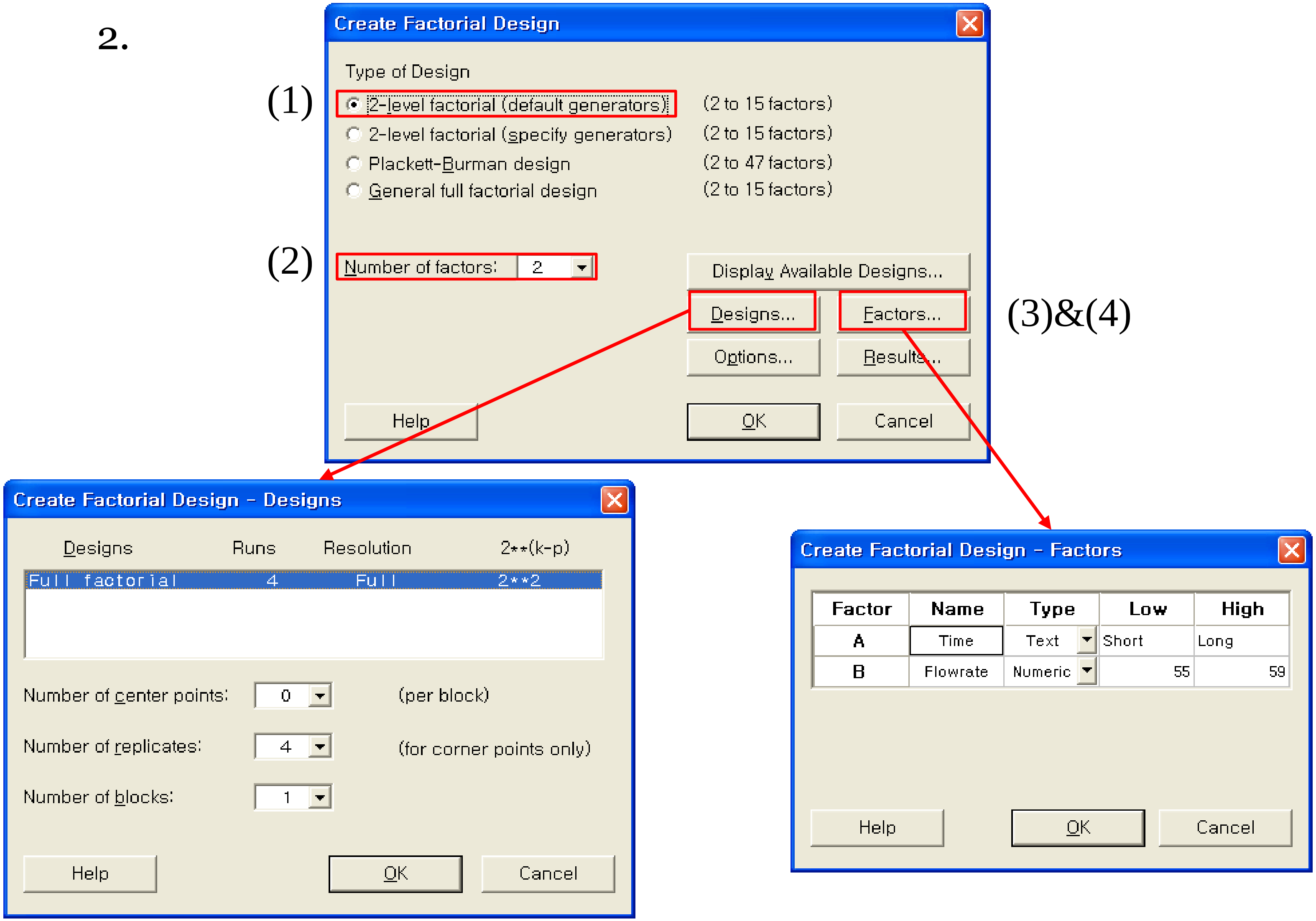
Exercise (cont.)

2.

(1)

(2)

(3)&(4)



Exercise (cont.)

3. Run experiments according to design matrix

Worksheet 2 ***

↓	C1	C2	C3	C4	C5-T	C6
	StdOrder	RunOrder	CenterPt	Blocks	Time	Flowrate
1	11	1	1	1	Short	59
2	15	2	1	1	Short	59
3	3	3	1	1	Short	59
4	2	4	1	1	Long	55
5	9	5	1	1	Short	55
6	8	6	1	1	Long	59
7	7	7	1	1	Short	59
8	10	8	1	1	Long	55
9	1	9	1	1	Short	55
10	4	10	1	1	Long	59
11	12	11	1	1	Long	59
12	5	12	1	1	Short	55
13	16	13	1	1	Long	59
14	6	14	1	1	Long	55
15	13	15	1	1	Short	55
16	14	16	1	1	Long	55

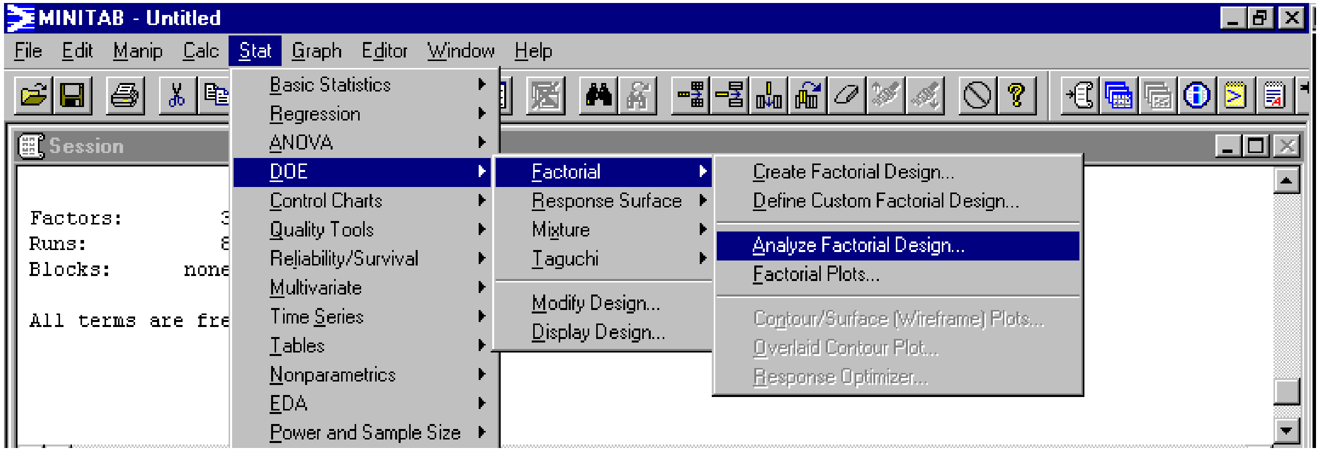
Why?

Exercise (cont.)

4. Analysis of experimental results

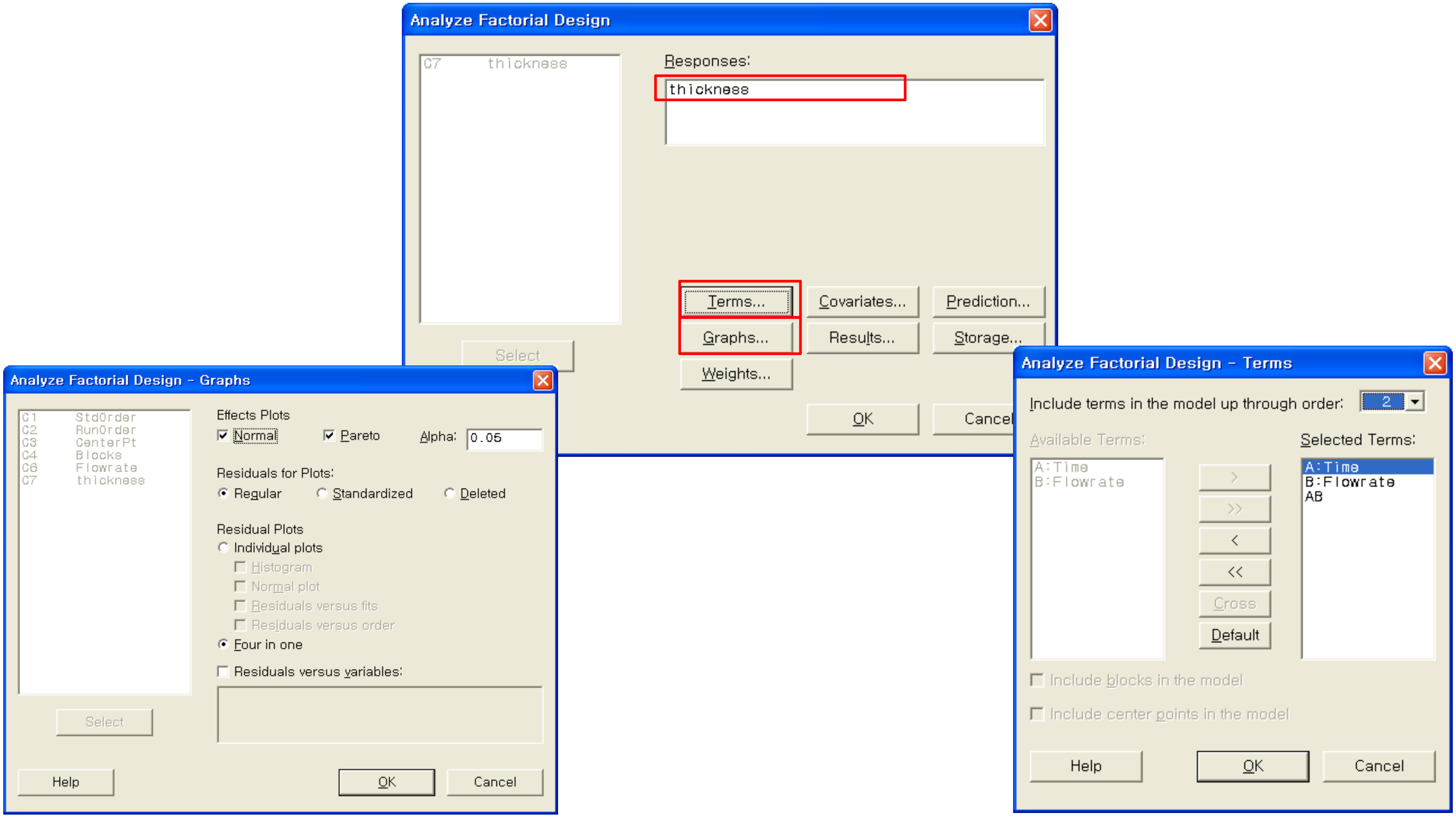
Using all analysis tools from least squares & main/interaction plots

DOE>Factorial>Analyze Factorial Design



Exercise (cont.)

4. Analysis of experimental results (cont.)



Exercise (cont.)

(a) ANOVA table (∵ we have replicates)

Estimated Effects and Coefficients for thickness (coded units)

Term	Effect	Coef	SE Coef	T	P
Constant		14.3884	0.03606	399.05	0.000
Time	0.8369	0.4184	0.03606	11.60	0.000
Flowrate	-0.0681	-0.0341	0.03606	-0.94	0.363
Time*Flowrate	0.0324	0.0162	0.03606	0.45	0.661

S = 0.144228   R-Sq = 91.88%   R-Sq(adj) = 89.85%

Analysis of Variance for thickness (coded units)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Main Effects	2	2.82000	2.82000	1.41000	67.78	0.000
2-Way Interactions	1	0.00419	0.00419	0.00419	0.20	0.661
Residual Error	12	0.24962	0.24962	0.02080		
Pure Error	12	0.24962	0.24962	0.02080		
Total	15	3.07382				

Unusual Observations for thickness

Obs	StdOrder	thickness	Fit	SE Fit	Residual	St Resid
11	12	14.4150	14.7890	0.0721	-0.3740	-2.99R

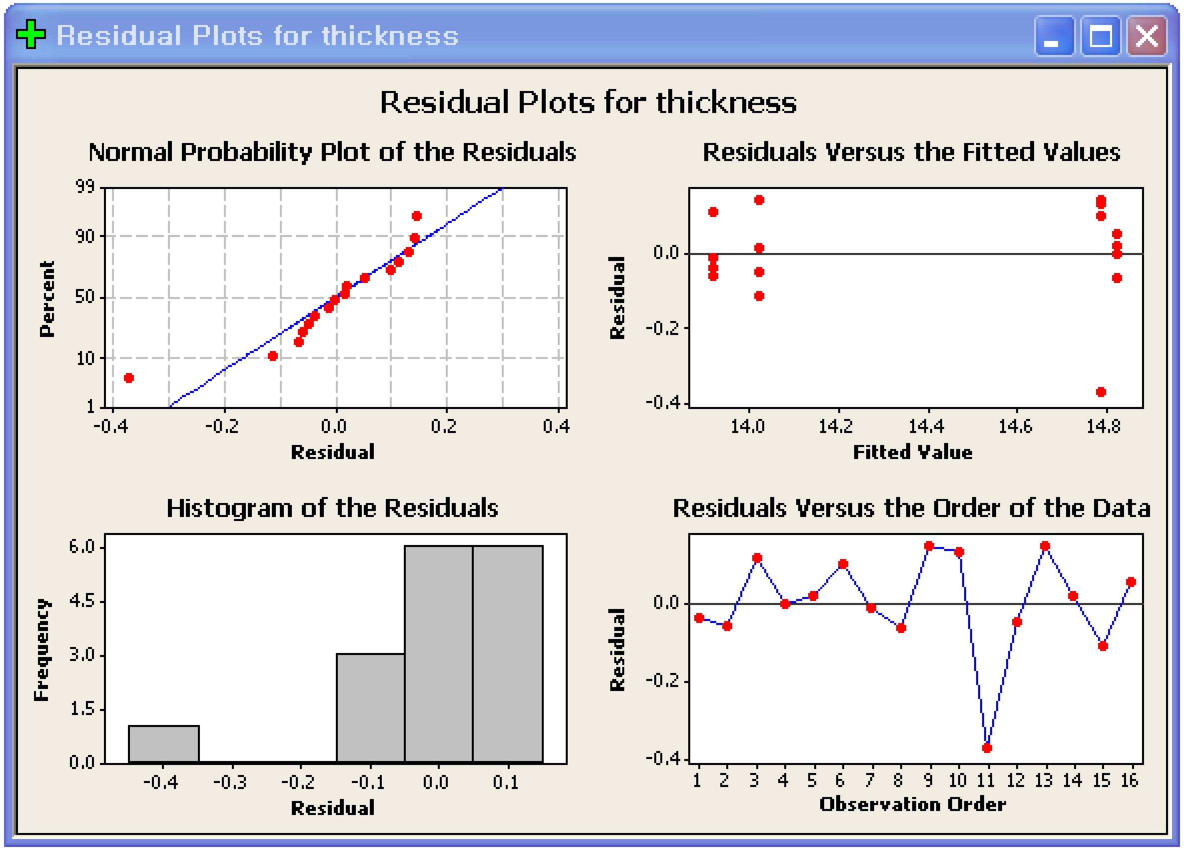
R denotes an observation with a large standardized residual.

Estimated Coefficients for thickness using data in uncoded units

Term	Coef
Constant	15.3592
Time	-0.04291
Flowrate	-0.0170313
Time*Flowrate	0.0080937

Exercise (cont.)

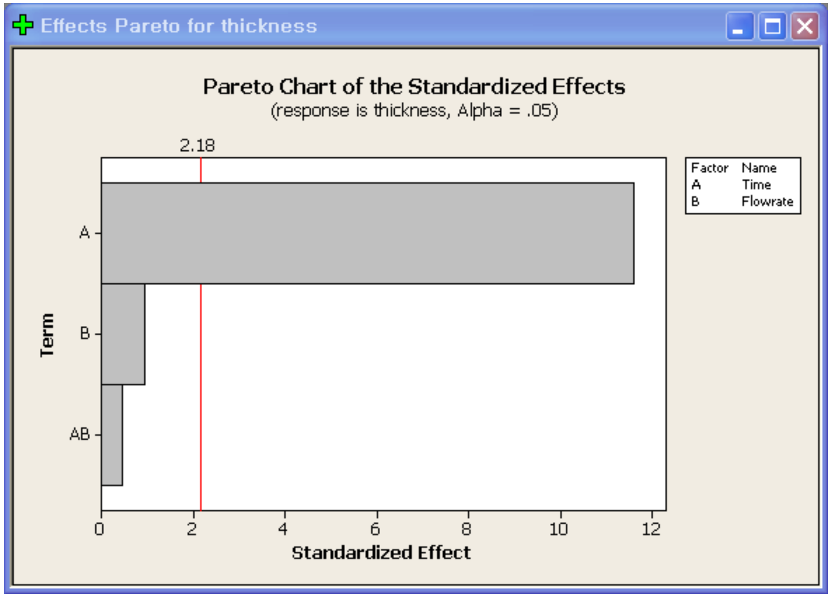
(b) Residual plots

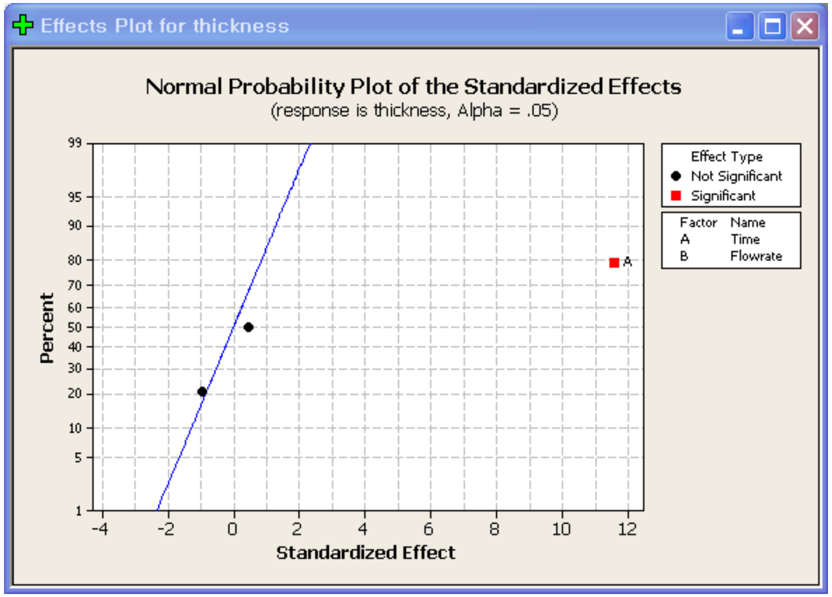


Exercise (cont.)

(c) Plots for effects

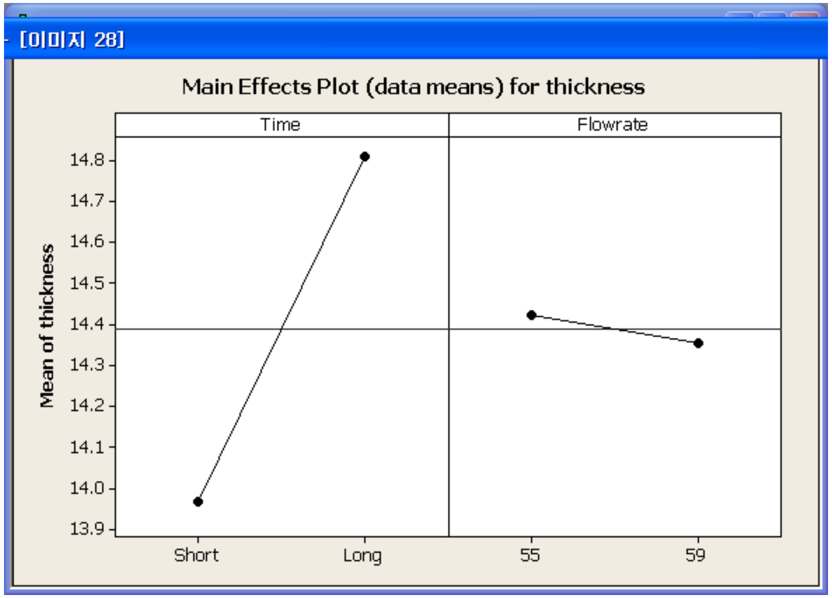
You can also determine which factors have significant effects.

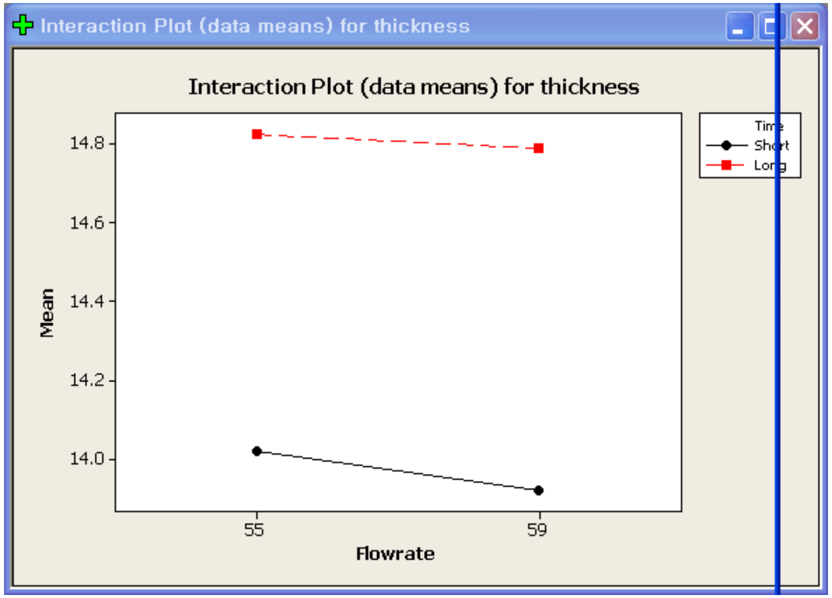




Exercise (cont.)

Alternatively, main/interaction plot





Exercise (cont.)

- Depending on your goal, you can refine a prediction model by selecting significant factors (variables) only.

  → less # of coefficients

  → more degree of freedom

  → more accurate estimate of C.I ($S_{y/x}$ can decrease)

This is very useful even when you have ***many factors*** and ***no replicates***.

Principle of sparsity of effects: the system (process) is usually dominated by the main effects and low-order interactions. That is, the three factor and higher-order interactions are usually negligible.

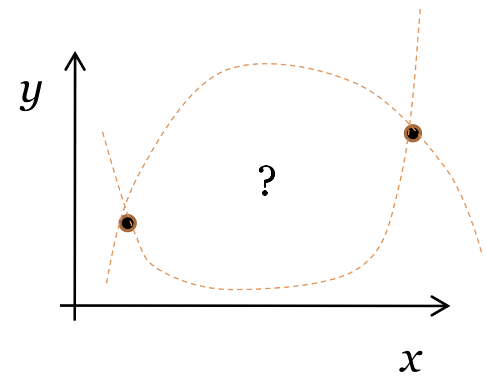
Design for 2nd order models

- If 1st order + interaction model exhibits "Lack of fit"

  → Include $x_1^2, x_2^2, \cdots$ terms

  But we need more than 2 level designs.

  → Central composite design or 3 level factorials



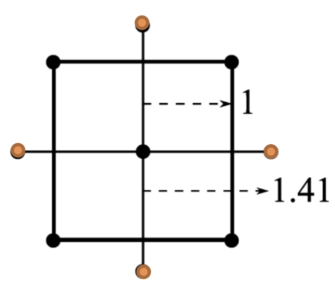
- Central composite design (k = 2)

  (1) Start with 2^k design with center points

  (2) Add vertices of star (for k=2, $\alpha = \sqrt{2}$)

  (3) Run experiments & analysis

"Cube plot"



	x_1	x_2
(1)	-1	-1
	+1	-1
	-1	+1
	+1	+1
	0	0
(2)	$-\alpha$	0
	$+\alpha$	0
	0	$-\alpha$
	0	$+\alpha$

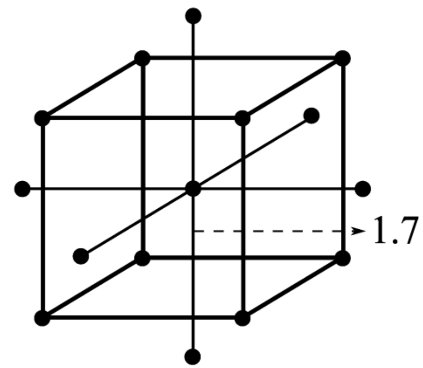
9 runs
For central composite design (k = 2)

Design for 2nd order models (cont.)

- Values of α

k	design	α
2	2^2	$\sqrt{2}$
3	2^3	$\sqrt{3}$
4	2^4	$\sqrt{4}$

Cube plot for 3 variables (factors)

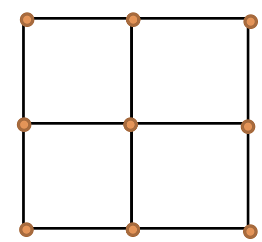


15 runs
For central
composite
design (k = 3)

- 3 level factorial

3^2 2 variables at all combinations of 3 levels

3^3 **27** runs for 3 variables



※Full quadratic model (assume 123 interaction is negligible.)

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_{12} x_1 x_2 + a_{13} x_1 x_3 + a_{23} x_2 x_3 + a_{11} x_1^2 + a_{22} x_2^2 + a_{33} x_3^2$$

Allows for approximation of many response.

Design for 2nd order models (cont.)

※A t-statistic for curvature

Average y of corner points

Average y of center points

$$t_{curvature} = \frac{\bar{y}_F - \bar{y}_C}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_F} + \frac{1}{n_C}\right)}}$$

Pure error calculated from center points

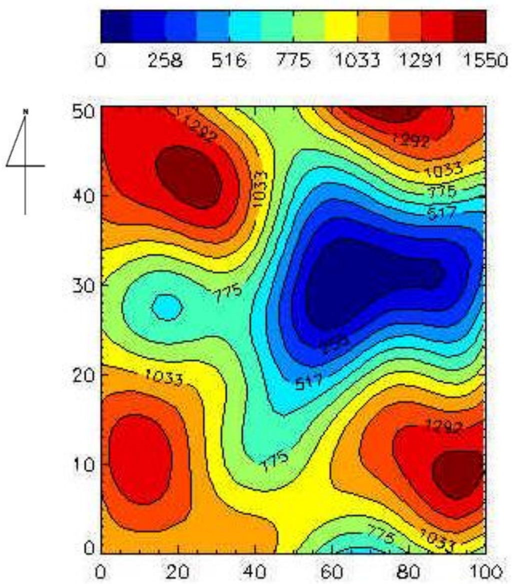
# of center points

# of corner points

Minitab uses ANOVA for testing curvature when center point replicates exist.

Response Surface Methods (RSM)

- Imagine you MUST climb a mountain,

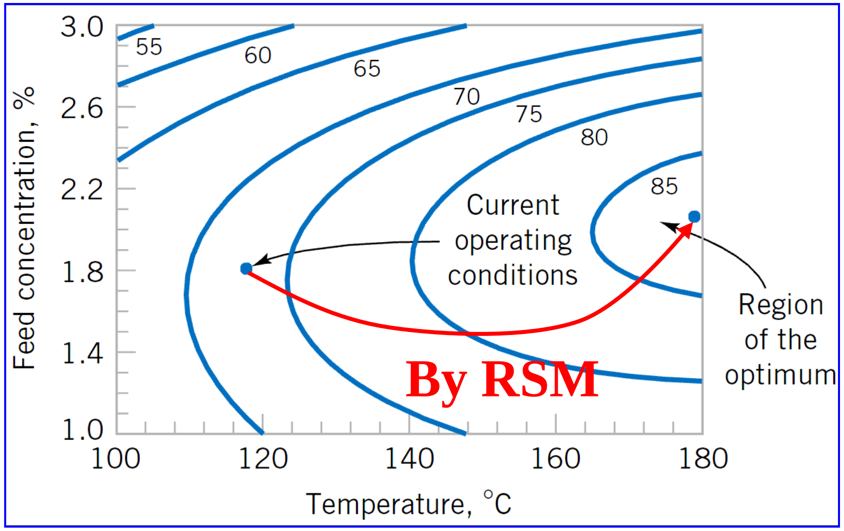


?

  - What you would do & how?
    - If you have GPSs and altimeters.
- Same situation: you want to increase a reactor's yield but don't know the process at all.

Response Surface Methods (RSM)

- RSM
  - Objective: optimize a process (or system) using mathematical & statistical techniques.
  - But, the process is usually unknown. (i.e., relationships between x & y variables are unknown.)



(1) The First step of RSM is to ***find a (approximate) model*** of the process using least squares & DOE.

(2) Next step is to ***improve process operation*** by moving to a better operating point using the model.

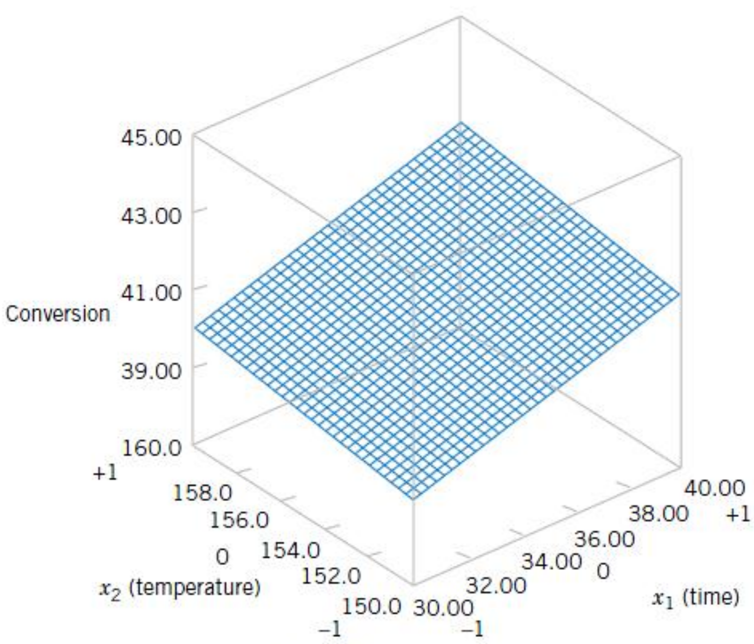
(3) Repeat this until optimum is reached.

FYI (For Your Information)

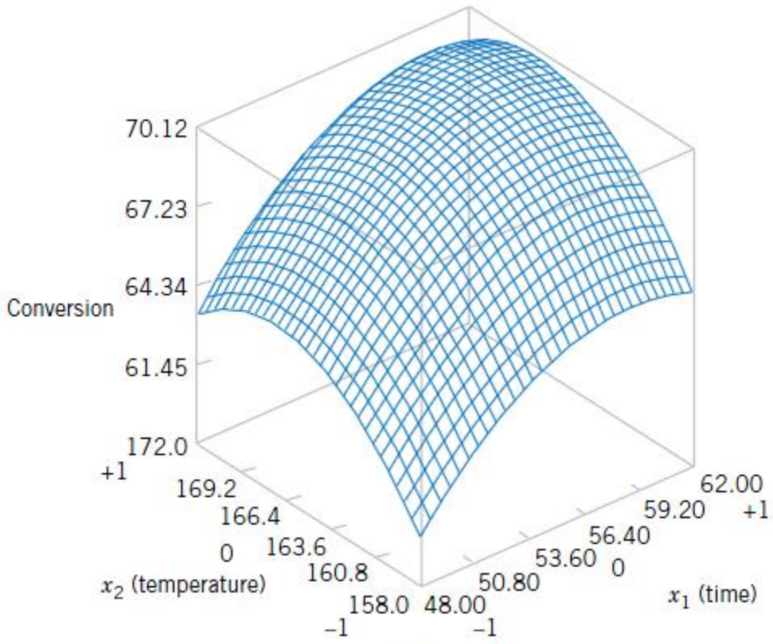
- Response surface?

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2$$

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2 + a_{11} x_1^2 + a_{22} x_2^2$$



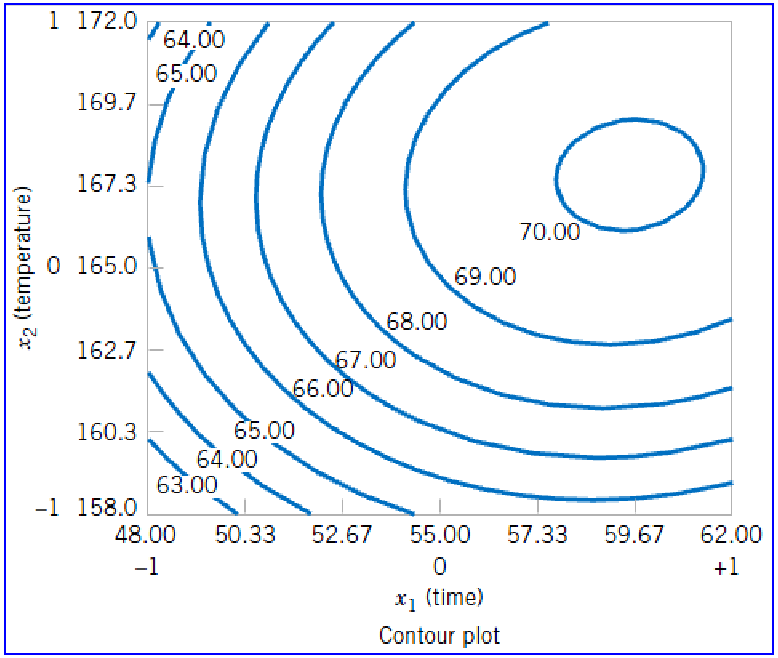
Three-dimensional surface plot

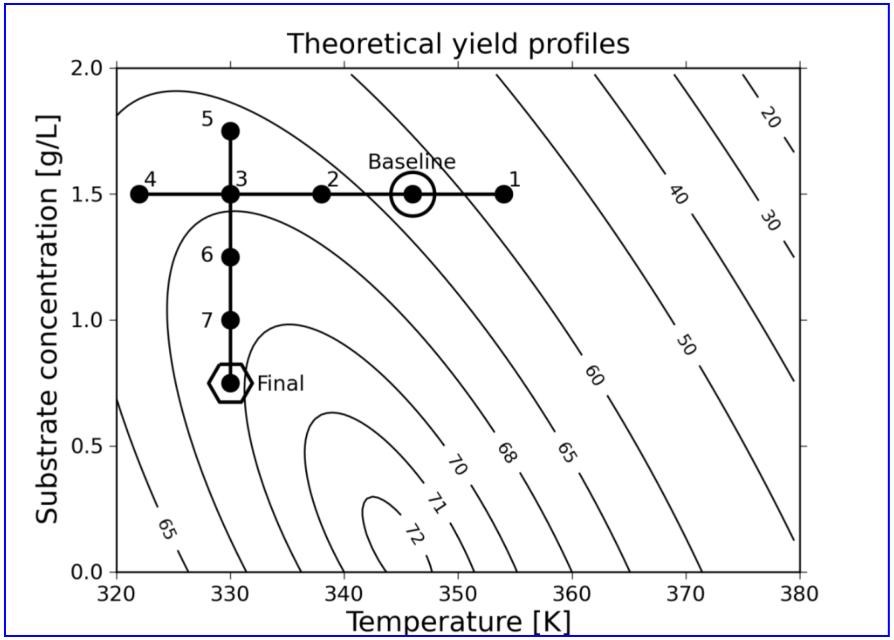


Surface plot

FYI (For Your Information)

- COST **costs too much** to find optimum when interaction exists.
  - Compare two cases

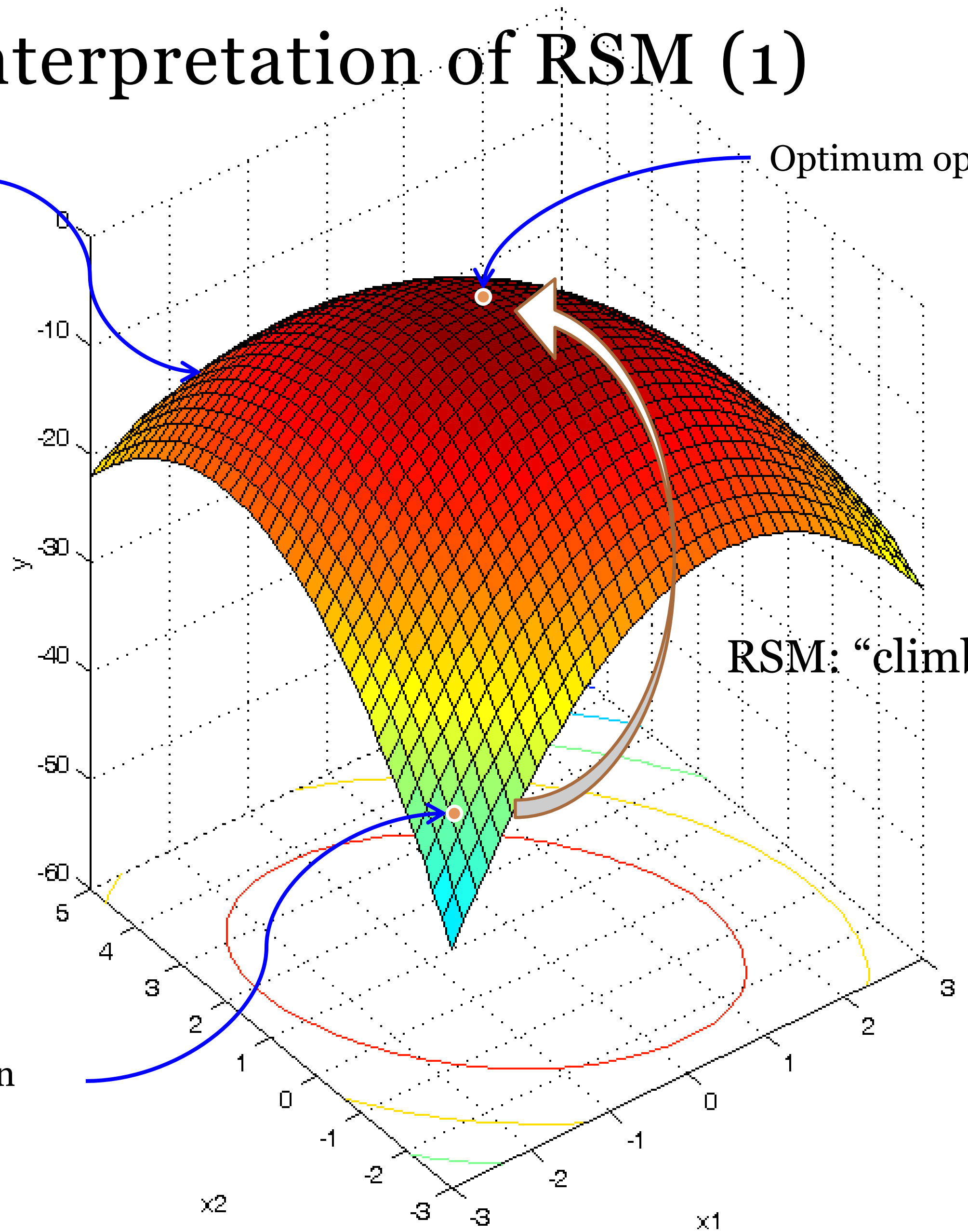




Graphical interpretation of RSM (1)

Unknown true process
$y = f(x_1, x_2)$

Optimum operating condition



RSM: “climbing a hill”

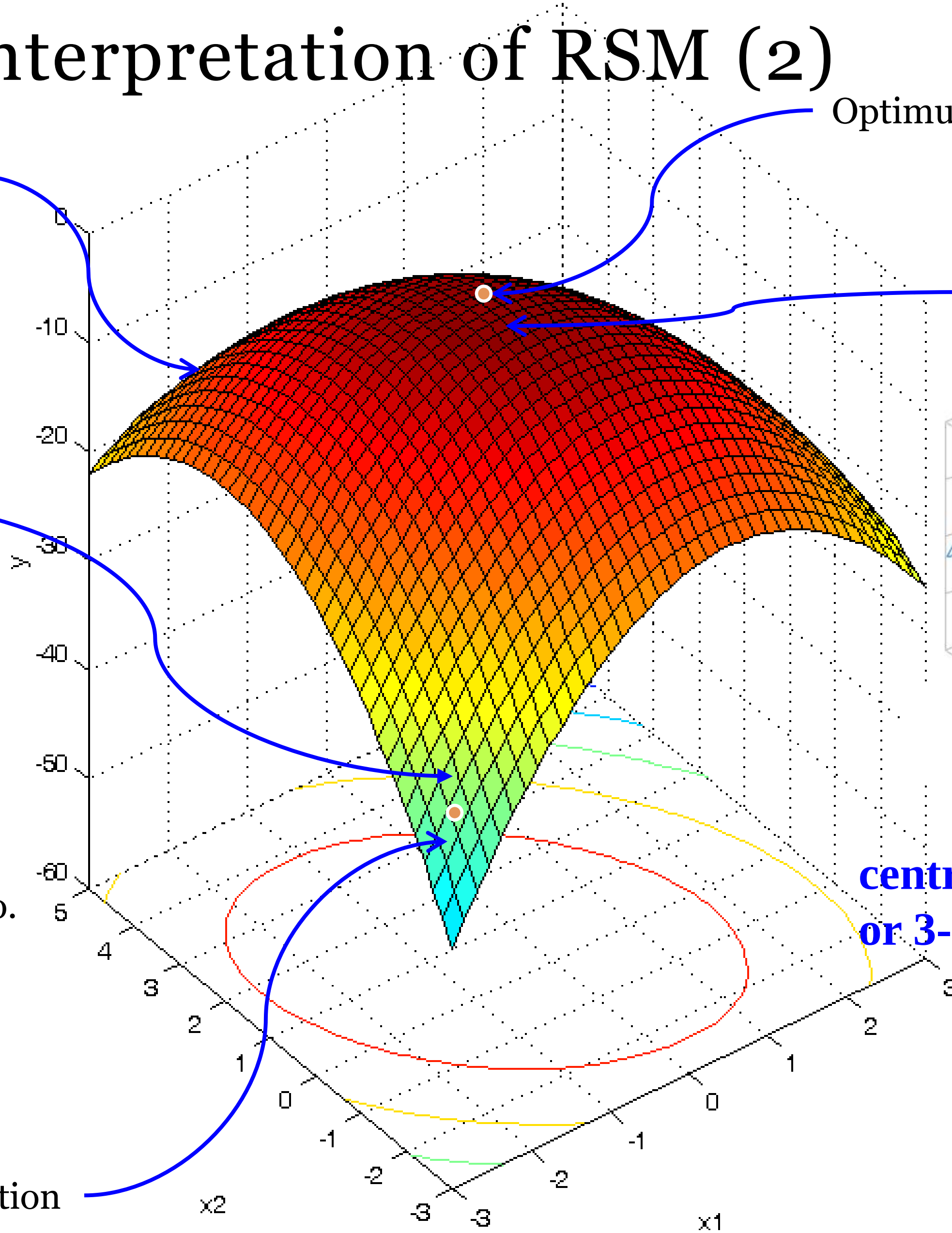
Current operating condition

Graphical interpretation of RSM (2)

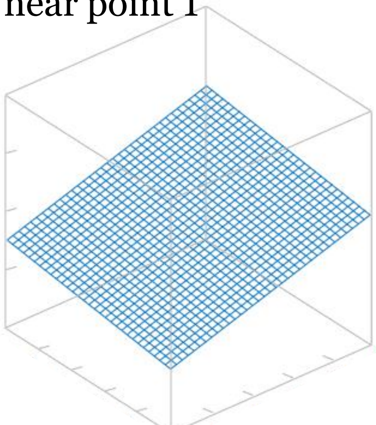
Unknown true process
$y = f(x_1, x_2)$

Optimum operating condition

Approximate model
near point 2

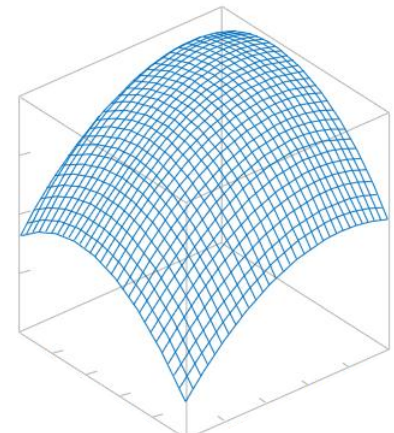


Approximate model
near point 1



A linear model form will do.

(full) factorial design



Quadratic model form
needed

**central composite design
or 3-level design**

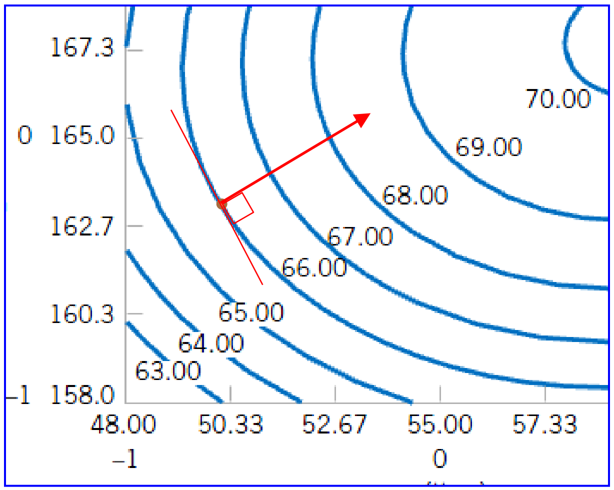
Current operating condition

Graphical interpretation of RSM (3)

Unknown true process
$y = f(x_1, x_2)$

Optimum operating condition

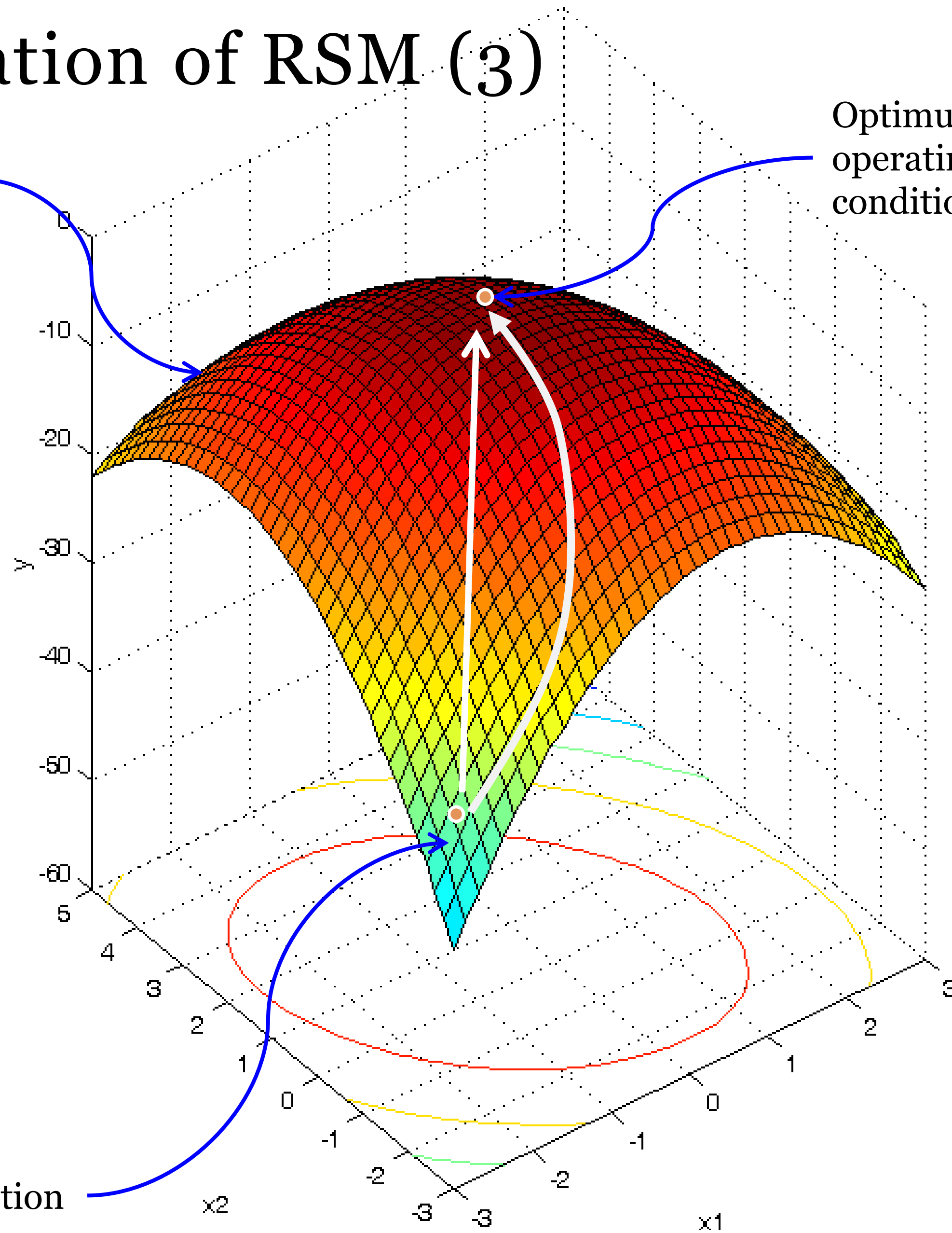
The fastest way to climb a hill
Methods of steepest ascent



Direction of steepest ascent

$$= \left(\frac{\partial y}{\partial x_1} \quad \frac{\partial y}{\partial x_2} \right)$$

$$\cong \left(a_1 \quad a_2 \right)$$ when interaction is smaller than main effects



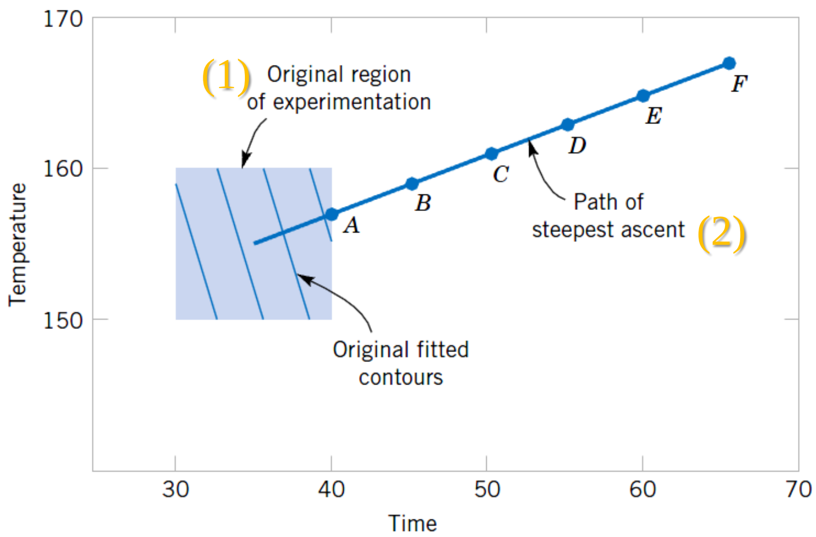
Current operating condition

RSM (cont.)

- General procedure
  1. Perform (fractional) factorial design around current operating conditions & fit a linear model form

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3 + a_{123}x_1x_2x_3$$

  2. Calculate direction of S.A. & perform experiments along this direction until response doesn't improve. (step size to be determined carefully)



Point A: 40 minutes, 157°F, y = 40.5
Point B: 45 minutes, 159°F, y = 51.3
Point C: 50 minutes, 161°F, y = 59.6
Point D: 55 minutes, 163°F, y = 67.1
Point E: 60 minutes, 165°F, y = 63.6
Point F: 65 minutes, 167°F, y = 60.7

RSM (cont.)

3. Lay down a new factorial design.
4. Repeat steps 1 ~ 3 until linear model is insufficient.
   - Curvature shows up.
   - 2-factor interaction dominate main effects.
5. Estimate a quadratic model if curvature and/or interaction is large relative to main effects.
   - Add star points → central composite design
   - Or three-level design

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3 + a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2$$

6. Plot response contour and move towards to best conditions

(most statistical software will do this)

RSM Exercise

Yield $y = f(T, S)$

Current operating conditions

- **T** = 325 K
- **S** = 0.75 g/L
- Profit = $407

Step 1

Experiment	T	S	Profit
1	−	−	193
2	+	−	310
3	−	+	468
4	+	+	571

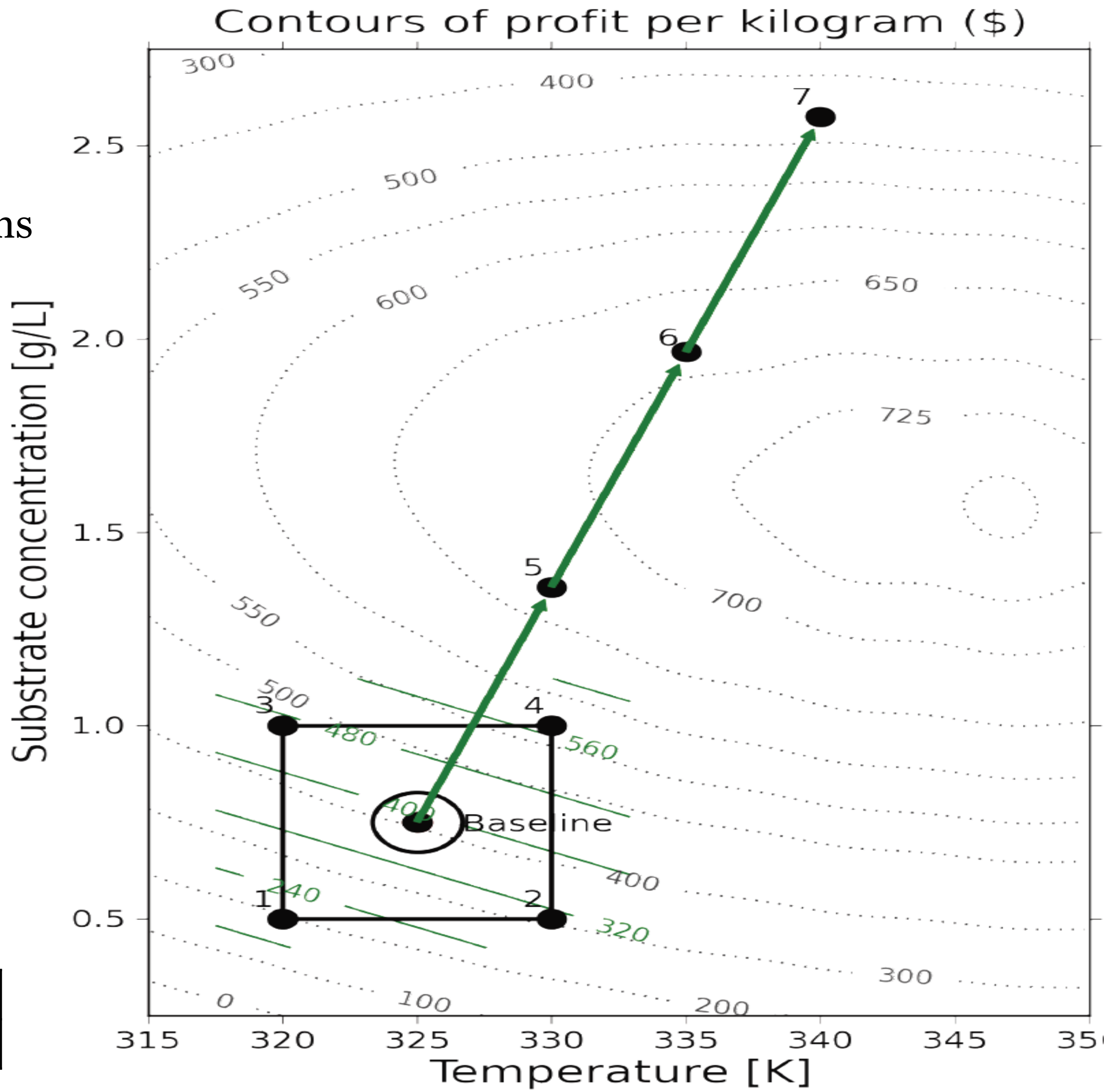
$$\hat{y} = 385.6 + 55x_T + 134x_S - 3.75x_Tx_S$$

Step 2

Derection of S.A

$$= \begin{pmatrix} \frac{\partial y}{\partial x_T} & \frac{\partial y}{\partial x_S} \end{pmatrix} \cong \begin{pmatrix} 55 & 134 \end{pmatrix}$$

experiment	5	6	7
profit	$669	$688	$463



RSM Exercise

Step 3

Experiment	T	S	Profit
8	−	−	694
9	+	−	725
10	−	+	620
11	+	+	642
6	0 (335 K)	0 (1.97 g/L)	688

$$\hat{y} = 670 + 13x_T - 39x_S - 2.4x_T x_S$$

Derection of S.A

$$= \left(\frac{\partial y}{\partial x_T} \quad \frac{\partial y}{\partial x_S} \right) \cong (13 \quad -39)$$

Profit (12) = 716 < profit (9)

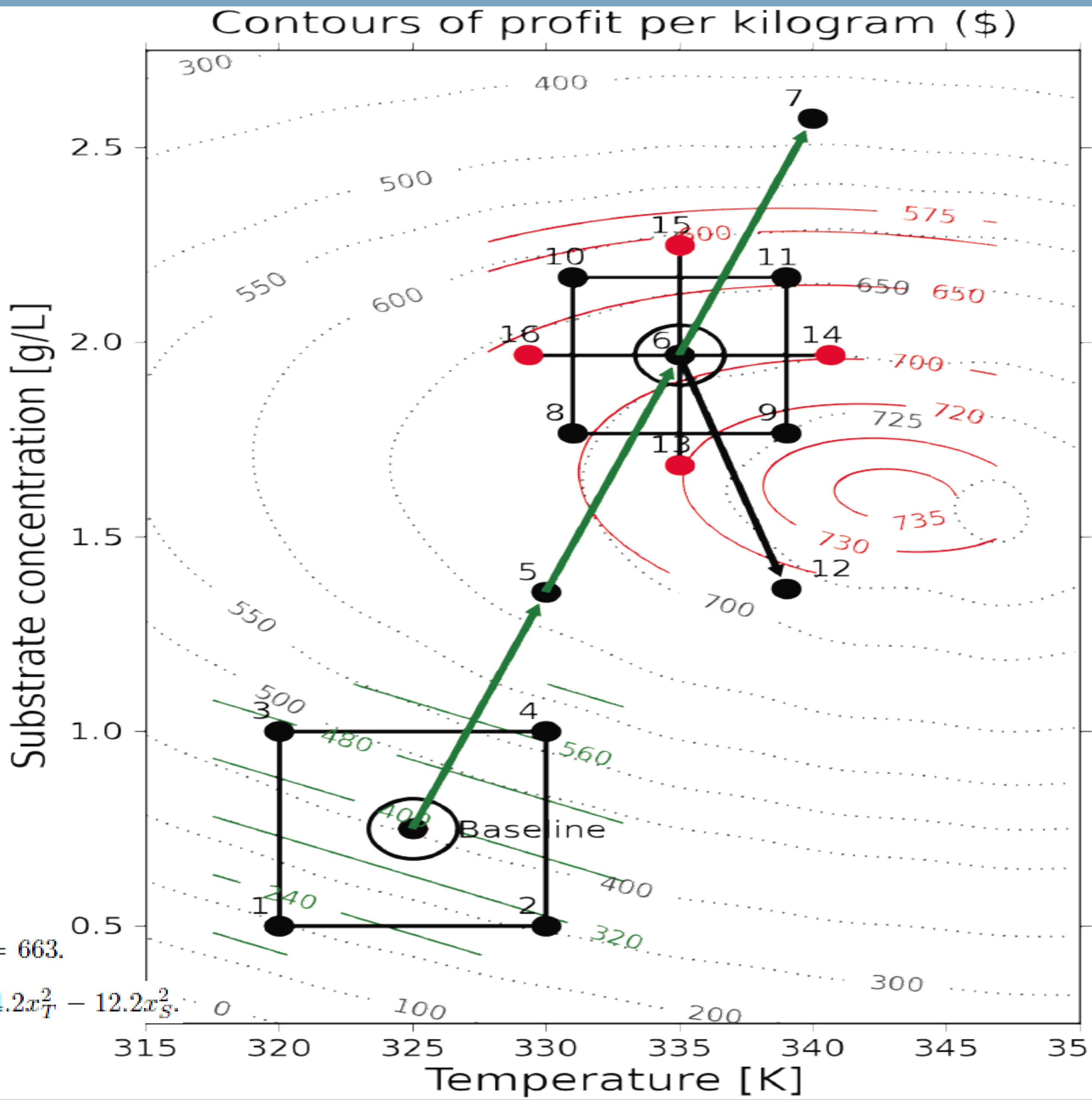
→ Strong interaction

Step 5

Star points

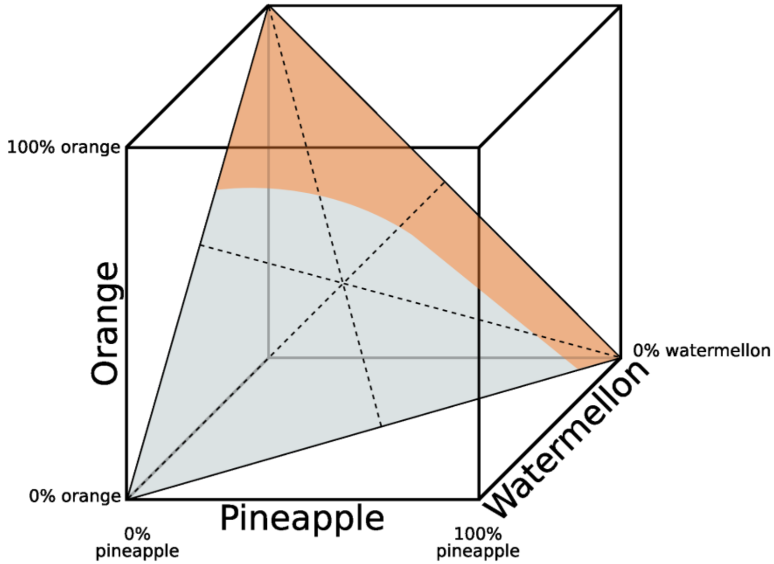
$y_{13} = 720$, $y_{14} = 699$, $y_{15} = 610$, and $y_{16} = 663$.

$$y = 688 + 12.9x_T - 39.1x_S - 2.4x_T x_S - 4.2x_T^2 - 12.2x_S^2.$$

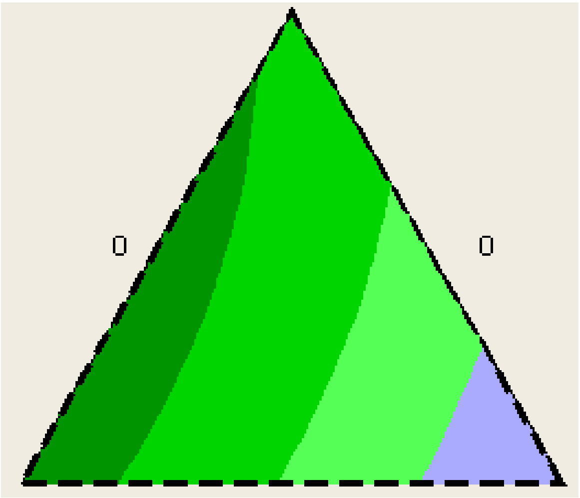


Mixture design

- Mixture design
  - Ordinary factorial design with a constraint
    - $0 \le x_A, x_B, x_C \le 1,\ \ x_A + x_B + x_C = 1$
  - Of course, RSM can be used to determine best mixture.

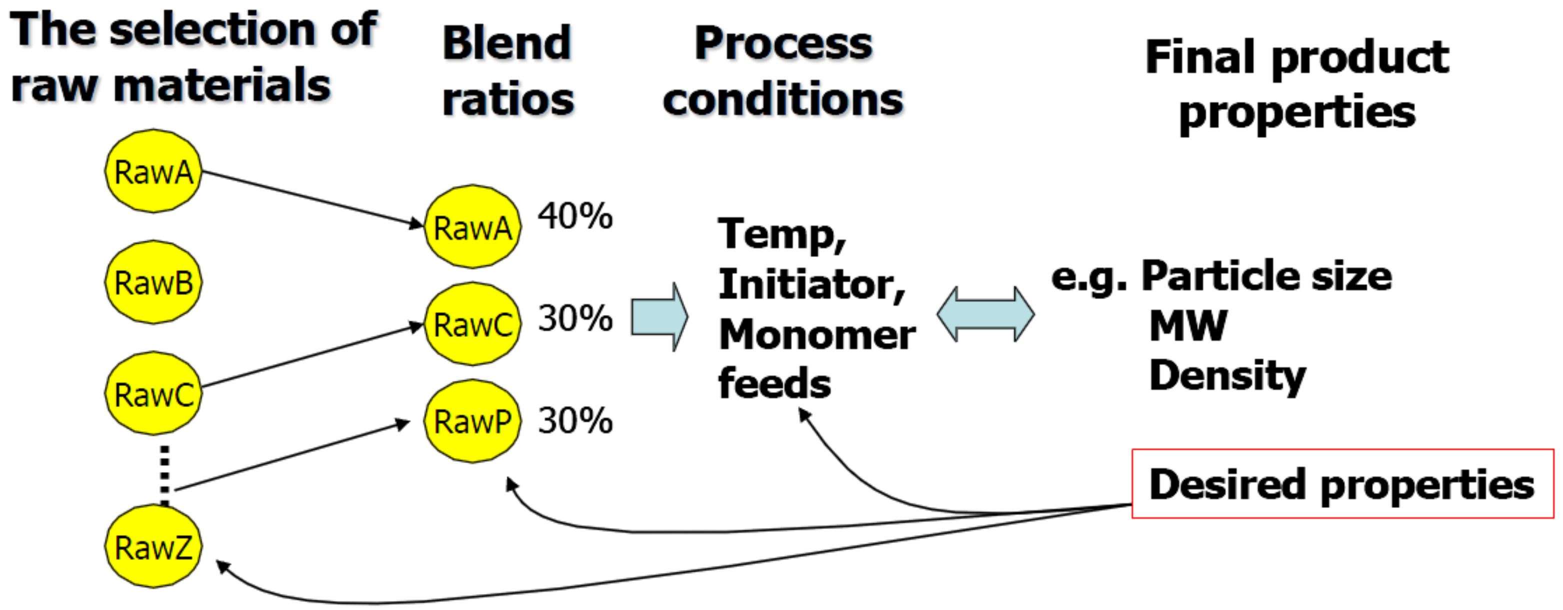


Mixture contour plot



Mixture design (cont.)

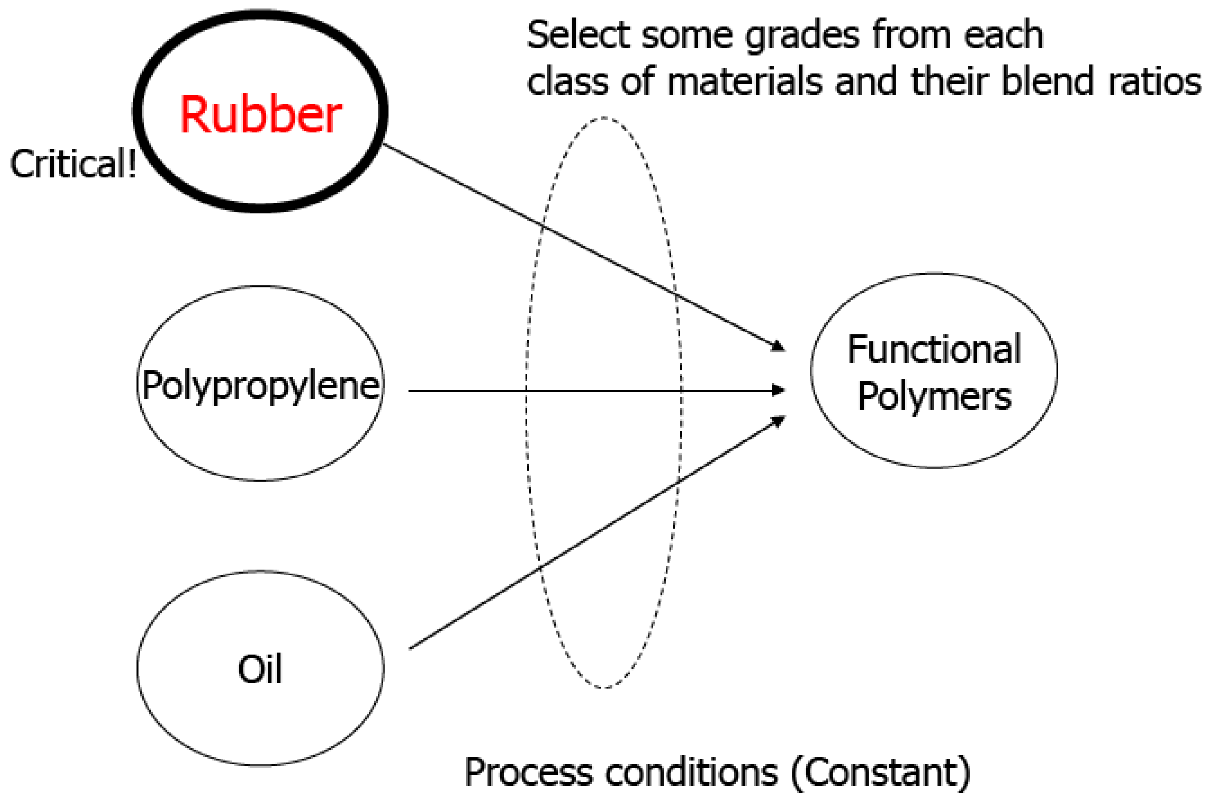
- Example: Product design (development)



Mixture design (cont.)

Example: Functional Polymer Development

Mitsubishi Chemicals



Mixture design (cont.)

- (Advanced) Mixture design example

Our approach has increased the resilience 1.7 times compared to previous products