

# 01NAEX - Lecture 05

## Introduction to $2^k$ Factorial Design

Jiri Franc

Czech Technical University  
Faculty of Nuclear Sciences and Physical Engineering  
Department of Mathematics

## Introduction to $2^k$ Factorial Design

By a factorial design, we mean that in each complete replication of the experiment all possible combinations of the levels of the factors are investigated.

- ▶  $2^k$  Factorial Design is special case of the general factorial design with  $k$  factors, all at 2 levels.
- ▶ The two levels are usually called low (-) and high (+).
- ▶  $2^k$  Factorial Design is widely used in industrial experimentation, especially in the early stages of experimental work, when many factors are likely to be investigated.

## Introduction to $2^k$ Factorial Design

Factors could be

- ▶ **quantitative**  
(such as two values of temperature, pressure, time, ...)
- ▶ **qualitative**  
(two operators, male and female, presence and absence, ...)

and it is assumed that **response is approximately linear over the range chosen.**

**We assume that**

- ▶ factors are fixed,
- ▶ the design is completely randomized,
- ▶ the usual normality assumptions are satisfied.

## Introduction to $2^k$ Factorial Design

The analysis procedure for a Factorial Design includes:

- ▶ Estimation of factor effects.
- ▶ Model formulation.
- ▶ Statistical testing (ANOVA).
- ▶ Refine the model.
- ▶ Analyze residuals (graphical).
- ▶ Interpret results.

## The 2<sup>2</sup> Design - Chemical process example

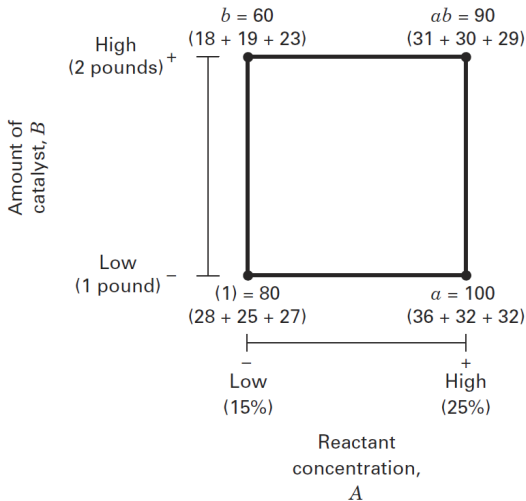
The 2<sup>2</sup> Design is the simplest case. In this example we consider an investigation into the effect of the concentration of the reactant and the amount of the catalyst on conversion in a chemical process.

	Factor		Treatment Combination	Replicate			Total
	A	B		I	II	III	
(1)	–	–	A low, B low	28	25	27	80
<i>a</i>	+	–	A high, B low	36	32	32	100
<i>b</i>	–	+	A low, B high	18	19	23	60
<i>ab</i>	+	+	A high, B high	31	30	29	90

- ▶ factor A - reactant concentration
- ▶ factor B - catalyst amount
- ▶ response *y* - recovery

The goal is to determine if adjustments to these factors would increase yield.

## Treatment combinations in the Chemical process example of $2^2$ design



- ▶ Low (-) and high (+) levels of a factor are arbitrary terms.
- ▶ The high level of any factor in the treatment combination is denoted by the corresponding lowercase letter.
- ▶ (1) is used to denote both factors at the low level.

## The 2<sup>2</sup> Design - Estimation of factor effects

The symbols (1),  $a$ ,  $b$ , and  $ab$  represents the total of all  $n$  replicates taken at the treatment combinations.

$$\begin{array}{lll} \text{main effect of A:} & A & = \frac{1}{2n} (ab + a - b - (1)) = \bar{y}_{A+} - \bar{y}_{A-} \\ \text{main effect of B:} & B & = \frac{1}{2n} (ab - a + b - (1)) = \bar{y}_{B+} - \bar{y}_{B-} \\ \text{interaction effect of AB:} & AB & = \frac{1}{2n} (ab - a - b + (1)) \end{array}$$

Where the **interaction effect AB** is defined as the average difference between the effect of A at high level of B and the effect of A at the low level of B.

## The 2<sup>2</sup> Design - Chemical process example

The estimation of average effects in chemical process example is:

<b>main effect of A</b>	$A$	$= \frac{1}{2 \times 3} (90 + 100 - 60 - 80) = 8.33$
<b>main effect of B</b>	$B$	$= \frac{1}{2 \times 3} (90 + 60 - 100 - 80) = 5.00$
<b>interaction effect of AB</b>	$AB$	$= \frac{1}{2 \times 3} (90 + 80 - 100 - 60) = 1.67$

Faster manually calculating of sum of squares:  $SS_A = \frac{[Contrast_A]^2}{N}$ .

$$SS_A = \frac{1}{12} (90 + 100 - 60 - 80)^2 = 208.33$$

$$SS_B = \frac{1}{12} (90 + 60 - 100 - 80)^2 = 75.00$$

$$SS_{AB} = \frac{1}{12} (90 + 80 - 100 - 60)^2 = 8.33$$

$$SS_T = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^n y_{ijk}^2 - \frac{y_{...}^2}{N} = 323.00$$

$$SS_E = SS_T - SS_A - SS_B - SS_{AB} = 31.33$$



## The 2<sup>2</sup> Design - Chemical process example

The Analysis of Variance for the Chemical process example, model with interaction effect:

```
> summary(yield.aov1 <- aov(yield~reactant*catalyst))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
reactant	1	208.33	208.33	53.191	8.44e-05	***
catalyst	1	75.00	75.00	19.149	0.00236	**
reactant:catalyst	1	8.33	8.33	2.128	0.18278	
Residuals	8	31.33	3.92			

-----

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 1

Notice (Notation in R):

$Y \sim X$  response variable  $Y$  is modeled as a function of predictor  $X$ .

$+X$  add (include) another explanatory variable  $X$ .

$-X$  delete (exclude) variable  $X$ .

$X:Z$  include the interaction between explanatory variables  $X$  and  $Z$ .

\* include the explanatory variables  $X$ ,  $Z$ , and the interactions between them.

## The 2<sup>2</sup> Design - Chemical process example

The Analysis of Variance for the Chemical process example, model without interaction effect:

```
> summary(yield.aov2 <- aov(yield~reactant+catalyst))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
reactant	1	208.33	208.33	47.27	7.27e-05	***
catalyst	1	75.00	75.00	17.02	0.00258	**
Residuals	9	39.67	4.41			

```
> anova(yield.aov2,yield.aov1)
```

Analysis of Variance Table

Model 1: yield ~ reactant + catalyst

Model 2: yield ~ reactant \* catalyst

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	9	39.667				
2	8	31.333	1	8.3333	2.1277	0.1828

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## The 2<sup>2</sup> Design - Chemical process example

The summary of linear regression model with coded variables  $x_i = \pm 1$ .

In our example  $x_1 = +1$ , if concentration is at the high level and  $-1$  if concentration is at the low level. Other variables in the same way.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Call: `lm(formula = yield ~ reactant.num * catalyst.num)`

Coefficients:

	Estimate	Std.Error	t value	Pr(> t )
(Intercept)	27.5000	0.5713	48.135	3.84e-11***
reactant.num	4.1667	0.5713	7.293	8.44e-05***
catalyst.num	-2.5000	0.5713	-4.376	0.00236 **
reactant.num:catalyst.num	0.8333	0.5713	1.459	0.18278

Residual standard error: 1.979 on 8 degrees of freedom

Multiple R-squared: 0.903, Adjusted R-squared: 0.8666

F-statistic: 24.82 on 3 and 8 DF, p-value: 0.0002093

$$y = 27.5 + 4.2x_1 - 2.5x_2 + 0.8x_3$$

Note:  $x_3 = x_1 \cdot x_2$ .

## The 2<sup>2</sup> Design - Chemical process example

Since the influence of the interaction effect is not significant, we remove it from the model:

```
Call: lm(formula = yield ~ reactant.num + catalyst.num)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	27.500	0.606	45.377	6.13e-12	***
reactant.num	4.167	0.606	6.875	7.27e-05	***
catalyst.num	-2.500	0.606	-4.125	0.00258	**

Residual standard error: 2.099 on 9 degrees of freedom

Multiple R-squared: 0.8772, Adjusted R-squared: 0.8499

F-statistic: 32.14 on 2 and 9 DF, p-value: 7.971e-05

```
> anova(yield.lm_coded2, yield.lm_coded1)
```

Analysis of Variance Table

Model 1: yield ~ reactant.num + catalyst.num

Model 2: yield ~ reactant.num \* catalyst.num

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	9	39.667				
2	8	31.333	1	8.3333	2.1277	0.1828

## The 2<sup>2</sup> Design - Chemical process example

The linear regression model in natural factors levels:

```
Call: lm(formula = yield ~ reactant.num + catalyst.num)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	18.3333	3.0902	5.933	0.00022	***
reactant.num	0.8333	0.1212	6.875	7.27e-05	***
catalyst.num	-5.0000	1.2121	-4.125	0.00258	**

Residual standard error: 2.099 on 9 degrees of freedom

Multiple R-squared: 0.8772, Adjusted R-squared: 0.8499

F-statistic: 32.14 on 2 and 9 DF, p-value: 7.971e-05

Conversion between Coded and Natural Variables:

$$\begin{aligned}\hat{y} &= 27.5 + 4.2 \left( \frac{Con - (Con_l + Con_h)/2}{(Con_h - Con_l)/2} \right) - 2.5 \left( \frac{Cat - (Cat_l + Cat_h)/2}{(Cat_h - Cat_l)/2} \right) \\ &= 18.3 + 0.83 \cdot Concentration - 5.00 \cdot Catalyst\end{aligned}$$

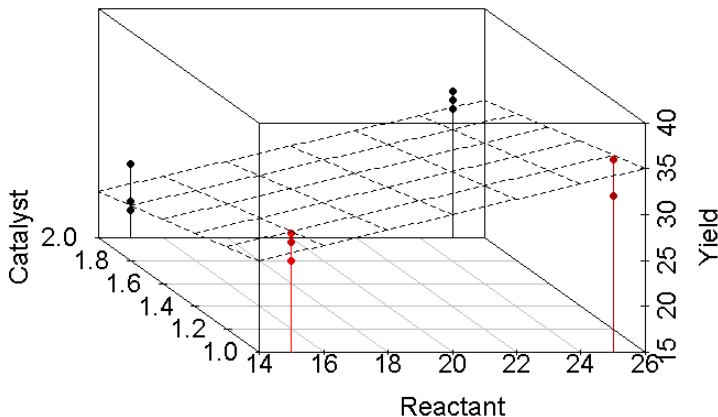
Concentration: Low = 15 and High = 25

Catalyst: Low = 1 and High = 2

## The $2^2$ Design - Chemical process example

The linear regression model in natural factors levels can be used to generate response surface plots.

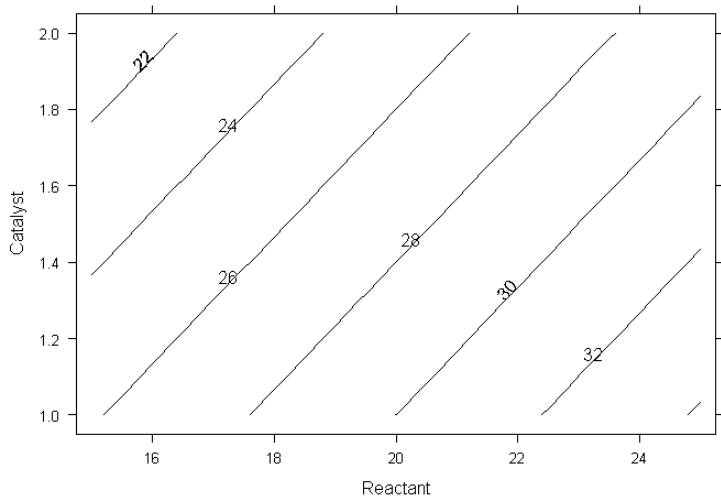
**3D Scatter Plot with Vertical Lines and Regression Planes**



## The 2<sup>2</sup> Design - Chemical process example

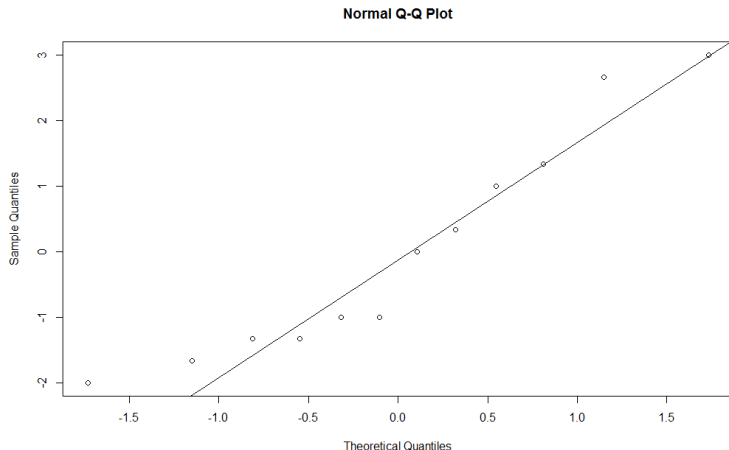
The linear regression model in natural factors levels can be used to generate response contour plots.

**Contour plot of Chemical process model with new predicted dataset**



## The 2<sup>2</sup> Design - Chemical process example

The model adequacy checking - normal plot of residuals.

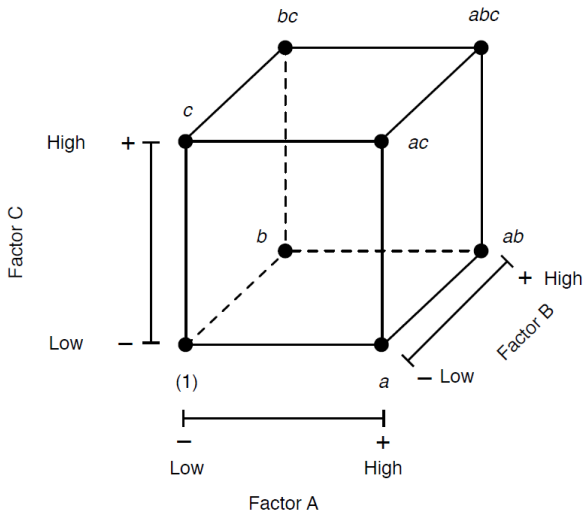


+ perform all normality and homoscedasticity tests.



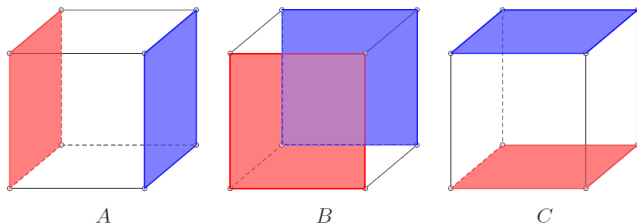
## The $2^3$ Design

If we now consider three factors, A, B, and C, we obtain a  $2^3$  factorial design with eight treatment combinations that can be displayed as a cube.

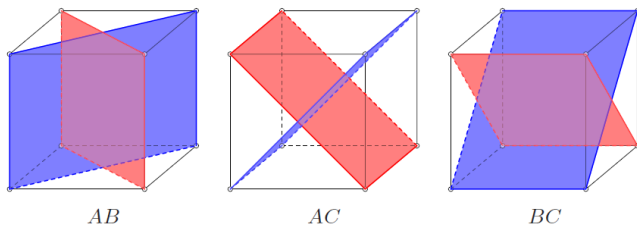


## The $2^3$ Design

Geometric presentation of contrast corresponding to the main effects and interactions in the  $2^3$  factorial design.



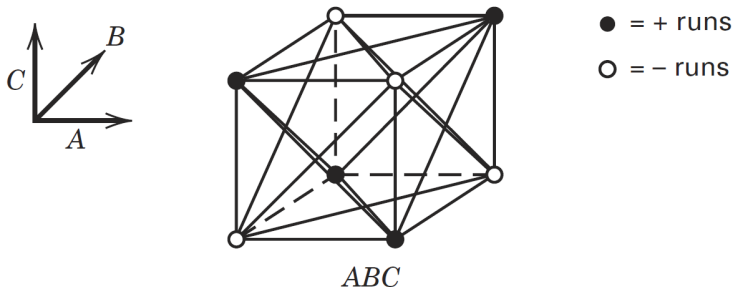
(a) Main effects



(b) Two-factor interaction

## The $2^3$ Design

Geometric presentation of contrast corresponding to the main effects and interactions in the  $2^3$  factorial design.



## The $2^3$ Design

Three different notations used for the runs in the  $2^k$  design:

Run	A	B	C	Labels	A	B	C
1	-	-	-	(1)	0	0	0
2	+	-	-	a	1	0	0
3	-	+	-	b	0	1	0
4	+	+	-	ab	1	1	0
5	-	-	+	c	0	0	1
6	+	-	+	ac	1	0	1
7	-	+	+	bc	0	1	1
8	+	+	+	abc	1	1	1

- ▶  $\pm$  notation is called Yates's order.
- ▶ Lowercase letter label to identify the treatment combinations.
- ▶ 1 and 0 denote high and low factor levels.

## Algebraic Signs for Calculating Effects in the $2^3$ Design:

Treatment Combination	Factorial Effect							
	<i>I</i>	<i>A</i>	<i>B</i>	<i>AB</i>	<i>C</i>	<i>AC</i>	<i>BC</i>	<i>ABC</i>
(1)	+	−	−	+	−	+	+	−
<i>a</i>	+	+	−	−	−	−	+	+
<i>b</i>	+	−	+	−	−	+	−	+
<i>ab</i>	+	+	+	+	−	−	−	−
<i>c</i>	+	−	−	+	+	−	−	+
<i>ac</i>	+	+	−	−	+	+	−	−
<i>bc</i>	+	−	+	−	+	−	+	−
<i>abc</i>	+	+	+	+	+	+	+	+

Note that except for column *I*, every column has an equal number of + and − signs.

The product of any two columns yields a column in the table (i.e.  $A \times B = AB$ ,  $AB \times BC = AB^2C = AC$ ).

Orthogonality is an important property shared by all factorial designs.

The estimation of average effects:

<b>main effect of A:</b>	$\begin{aligned} A &= \bar{y}_{A+} - \bar{y}_{A-} \\ &= \frac{1}{4n} (a + ab + ac + abc - (1) - b - c - bc) \end{aligned}$
<b>main effect of B:</b>	$\begin{aligned} B &= \bar{y}_{B+} - \bar{y}_{B-} \\ &= \frac{1}{4n} (b + ab + bc + abc - (1) - a - c - ac) \end{aligned}$
<b>main effect of C:</b>	$\begin{aligned} C &= \bar{y}_{C+} - \bar{y}_{C-} \\ &= \frac{1}{4n} (c + ac + bc + abc - (1) - a - b - ab) \end{aligned}$
<b>interaction of AB:</b>	$AB = \frac{1}{4n} (abc + ab + c + (1) - a - b - ac - bc)$
<b>interaction of AC:</b>	$AC = \frac{1}{4n} (abc + ac + b + (1) - a - c - ab - bc)$
<b>interaction of BC:</b>	$BC = \frac{1}{4n} (abc + bc + a + (1) - b - c - ab - ac)$
<b>interaction of ABC:</b>	$ABC = \frac{1}{4n} (abc + a + b + c - ab - ac - bc - (1))$

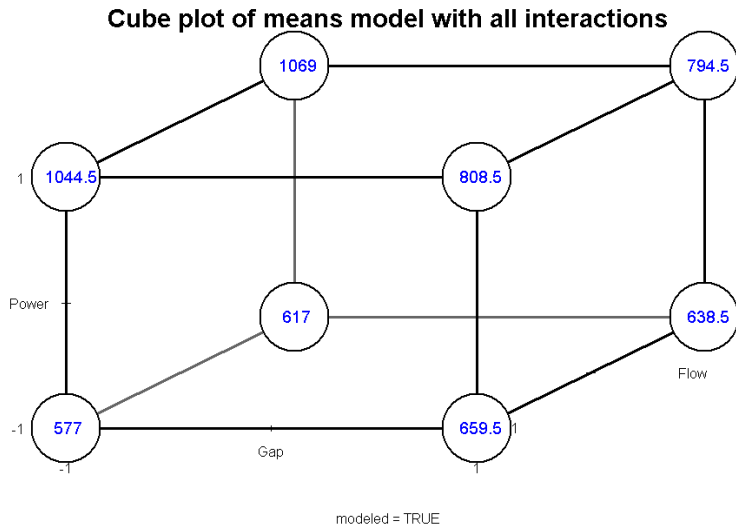
## The 2<sup>3</sup> Design - The Etch Plasma Experiment

The design factors are

- ▶ A: gap between electrodes in cm  
Low (-1) 0.8 and High (+1) 1.2
- ▶ B: gas  $C_2F_6$  flow in SCCM  
Low (-1) 125 and High (+1) 200
- ▶ C: RF power applied to the cathode in W  
Low (-1) 275 and High (+1) 325

Run	Coded Factors			Etch Rate		Total
	A	B	C	Replicate 1	Replicate 2	
1	-1	-1	-1	550	604	(1) = 1154
2	1	-1	-1	669	650	<i>a</i> = 1319
3	-1	1	-1	633	601	<i>b</i> = 1234
4	1	1	-1	642	635	<i>ab</i> = 1277
5	-1	-1	1	1037	1052	<i>c</i> = 2089
6	1	-1	1	749	868	<i>ac</i> = 1617
7	-1	1	1	1075	1063	<i>bc</i> = 2178
8	1	1	1	729	860	<i>abc</i> = 1589

## The $2^3$ Design - The Etch Plasma Experiment





## The 2<sup>3</sup> Design - The Etch Plasma Experiment

Analysis of Variance results:

```
> summary(aov(etch~A*B*C))
```

Df	Sum Sq	Mean Sq	F value	Pr(>F)	
A	1	41311	41311	18.339	0.002679 **
B	1	218	218	0.097	0.763911
C	1	374850	374850	166.411	1.23e-06 ***
A:B	1	2475	2475	1.099	0.325168
A:C	1	94403	94403	41.909	0.000193 ***
B:C	1	18	18	0.008	0.930849
A:B:C	1	127	127	0.056	0.818586
Residuals	8	18020	2253		

Gap (A), Power (C) and its interaction are highly significant.

We remove other variables and compare new reduced model with this one.

## The 2<sup>3</sup> Design - The Etch Plasma Experiment

Analysis of Variance results of reduced model and computed Lack of fit.

```
> summary(etch~A*C)
              Df Sum Sq Mean Sq F value    Pr(>F)
A              1  41311    41311    23.77 0.000382 ***
C              1 374850   374850   215.66 4.95e-09 ***
A:C            1  94403    94403    54.31 8.62e-06 ***
Residuals     12  20858         1738
```

Analysis of Variance Table

Model 1: etch ~ A \* C

Model 2: etch ~ A \* B \* C

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	12	20858				
2	8	18021	4	2837.2	0.3149	0.8604

Lack of fit is the difference between Residuals in reduced models and pure error (residuals) in old model with all variables and interactions.

## The 2<sup>3</sup> Design - The Etch Plasma Experiment

Linear regression model results for coded ( $\pm 1$ ) variables.

Estimated effects are double of mentioned values.

```
Call: lm(formula = etch ~ A.num * B.num * C.num)
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
```

(Intercept)	776.062	11.865	65.406	3.32e-12	***
A.num	-50.812	11.865	-4.282	0.002679	**
B.num	3.688	11.865	0.311	0.763911	
C.num	153.062	11.865	12.900	1.23e-06	***
A.num:B.num	-12.437	11.865	-1.048	0.325168	
A.num:C.num	-76.812	11.865	-6.474	0.000193	***
B.num:C.num	-1.062	11.865	-0.090	0.930849	
A.num:B.num:C.num	2.813	11.865	0.237	0.818586	

Residual standard error: 47.46 on 8 degrees of freedom

Multiple R-squared: 0.9661, Adjusted R-squared: 0.9364

F-statistic: 32.56 on 7 and 8 DF, p-value: 2.896e-05

## The 2<sup>3</sup> Design - The Etch Plasma Experiment

Linear regression results for reduced model with coded ( $\pm 1$ ) variables.

```
Call: lm(formula = etch ~ A.num * C.num)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	776.06	10.42	74.458	< 2e-16	***
A.num	-50.81	10.42	-4.875	0.000382	***
C.num	153.06	10.42	14.685	4.95e-09	***
A.num:C.num	-76.81	10.42	-7.370	8.62e-06	***

Residual standard error: 41.69 on 12 degrees of freedom

Multiple R-squared: 0.9608, Adjusted R-squared: 0.9509

F-statistic: 97.91 on 3 and 12 DF, p-value: 1.054e-08

The proportion of total variability in etch rate that is explained by this reduced model is

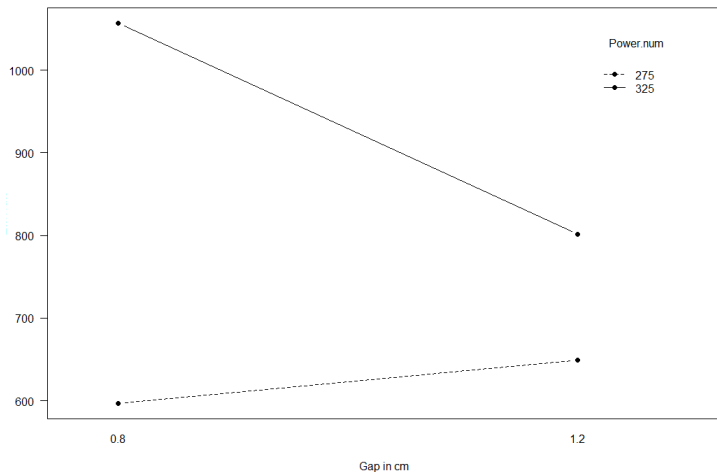
$$R^2 = 1 - \frac{SS_{Res}}{SS_{Total}} = \frac{SS_{Model}}{SS_{Total}} = \frac{SS_A + SS_C + SS_{AC}}{SS_{Total}} = \frac{510564}{531422} = 0.961$$

It is smaller than for the full model, but the removing of the nonsignificant terms from the full model is good for reducing the length of confidence intervals and for predicting.

## The 2<sup>3</sup> Design - Plasma Etch Rate Experiment

The interaction plot.

```
interaction.plot (Gap.num, Power.num, etch, type="b", pch=19,  
                 fixed=T, xlab="Gap in cm", ylab="Etch rate")
```

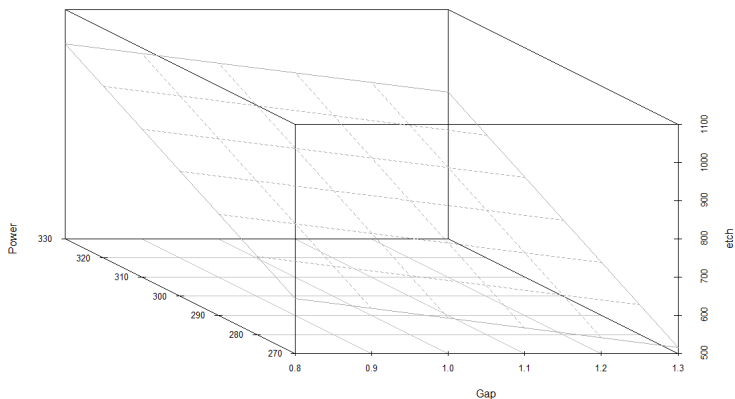


## The $2^3$ Design - Plasma Etch Rate Experiment

The linear regression model in natural factors levels can be used to generate response surface plots.

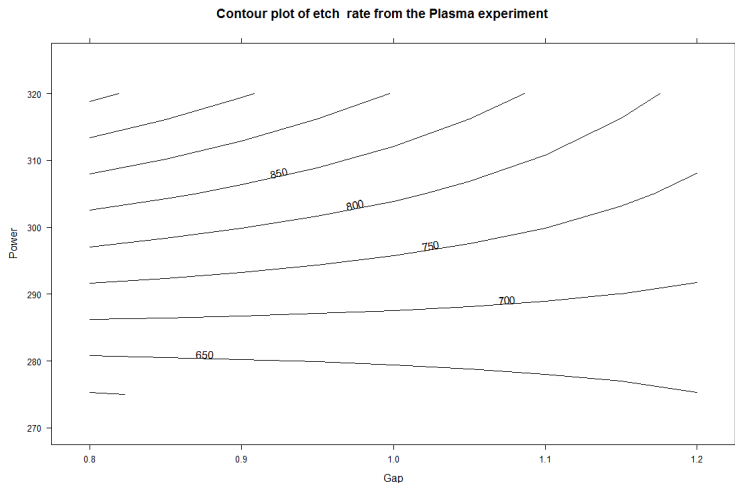
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3 + \hat{\beta}_{13} x_1 x_3 = 776.06 - 50.81 x_1 + 153.06 x_3 - 76.81 x_1 x_3$$

Response surface plot of etch rate from the Plasma experiment



## The $2^3$ Design - Plasma Etch Rate Experiment

The linear regression model in natural factors levels can be used to generate response contour plots.



## The 2<sup>3</sup> Design - Plasma Etch Rate Experiment

Model Summary Statistics:

Standard error of model coefficients:

$$se(\hat{\beta}) = \sqrt{\frac{\sigma^2}{2n^k}} = \sqrt{\frac{MS_E}{2n^k}} = \sqrt{\frac{1738}{16}} = \frac{41.69}{4} = 10.42.$$

The standard errors of all model coefficients are equal because the design is orthogonal.

Confidence interval on model coefficients:

$$\hat{\beta} - t_{\alpha/2, df_E} se(\hat{\beta}) \leq \beta \leq \hat{\beta} + t_{\alpha/2, df_E} se(\hat{\beta}),$$

where the degrees of freedom  $df_E$  on  $t$  are the number of DF for Error.

$df_E = N - p = 12$ , where  $N = 16$  is the total number of runs in the experiment and  $p = 4$  is the number of parameters in the regression model.



## The General $2^k$ Factorial Design

In a general  $2^k$  Factorial Design will be

- ▶  $k$  main effects
- ▶  $\binom{k}{2}$  two-factor interactions
- ▶  $\binom{k}{3}$  three-factor interactions
- ▶  $\vdots$
- ▶ 1  $k$ -factor interaction

## The General $2^k$ Factorial Design

```
> k=3
```

```
> plan2a <- FrF2(2^k, k, replications = 2, randomize = FALSE,  
  factor.names = c("A", "B", "C"))
```

```
> plan2a
```

run.no	run.no.	std.rp	A	B	C
1	1	1.1	-1	-1	-1
2	2	2.1	1	-1	-1
3	3	3.1	-1	1	-1
4	4	4.1	1	1	-1
5	5	5.1	-1	-1	1
6	6	6.1	1	-1	1
7	7	7.1	-1	1	1
8	8	8.1	1	1	1
9	9	1.2	-1	-1	-1
10	10	2.2	1	-1	-1
11	11	3.2	-1	1	-1
12	12	4.2	1	1	-1
13	13	5.2	-1	-1	1
14	14	6.2	1	-1	1
15	15	7.2	-1	1	1
16	16	8.2	1	1	1

```
class=design, type= full factorial
```

NOTE: columns run.no and run.no.std.rp are annotation, not part of the

```
result <- rnorm(16)
```

```
plan2a <- add.response(plan2a,result)
```

## Unreplicated $2^k$ Factorial Design

- ▶ These are  $2^k$  factorial designs with one observation at each corner of the "cube" and sometimes is called single replicated design.
- ▶ Very widely used type of design, especially in first planning and testing.
- ▶ If the factors are spaced too closely, it increases the chances that the noise will overwhelm the signal in the data.
- ▶ More aggressive spacing is usually best.
- ▶ Lack of replication causes potential problems in statistical testing:
  - ▶ With no replication, fitting the full model results in zero degrees of freedom for error.
  - ▶ Pooling high-order interactions to estimate error.
  - ▶ Normal probability plotting of effects (Daniels, 1959).

Examples and further discussion in the next lesson.

## Today Exercise

Solve problems 6.1, 6.2.

(from D. C. Montgomery DAoE ed.8 - chapter 6)

**6.1.** An engineer is interested in the effects of cutting speed ( $A$ ), tool geometry ( $B$ ), and cutting angle ( $C$ ) on the life (in hours) of a machine tool. Two levels of each factor are chosen, and three replicates of a  $2^3$  factorial design are run. The results are as follows:

$A$	$B$	$C$	Treatment Combination	Replicate		
				I	II	III
–	–	–	(1)	22	31	25
+	–	–	$a$	32	43	29
–	+	–	$b$	35	34	50
+	+	–	$ab$	55	47	46
–	–	+	$c$	44	45	38
+	–	+	$ac$	40	37	36
–	+	+	$bc$	60	50	54
+	+	+	$abc$	39	41	47

- (a) Estimate the factor effects. Which effects appear to be large?
- (b) Use the analysis of variance to confirm your conclusions for part (a).
- (c) Write down a regression model for predicting tool life (in hours) based on the results of this experiment.
- (d) Analyze the residuals. Are there any obvious problems?
- (e) On the basis of an analysis of main effect and interaction plots, what coded factor levels of  $A$ ,  $B$ , and  $C$  would you recommend using?

**6.2.** Reconsider part (c) of Problem 6.1. Use the regression model to generate response surface and contour plots of the tool life response. Interpret these plots. Do they provide insight regarding the desirable operating conditions for this process?