Design of experiments

Mikko Mäkelä

Aalto University, School of Chemical Engineering
Department of Bioproducts and Biosystems
Espoo, Finland

Session 2

Diagnostics

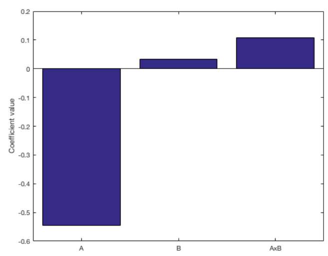
- Coefficients
- Random error assumption
- Statistical distributions
- ANOVA
- R^2
- Residuals

Research problem

An engineer is interested on the effect of temperature (A) and catalyst concentration (B) on the molecular weight of produced bio-oil. He performed a replicated 2^2 factorial design

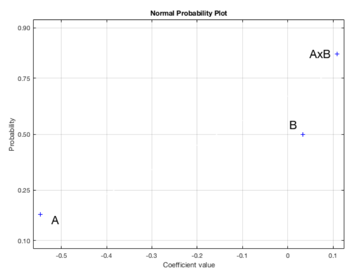
Exp.	A (°C)	B (%)	MW (kg mol^{-1})
1	160	0.2	2.0, 2.2
2	320	0.2	0.85, 0.73
3	160	0.8	1.8, 2.1
4	320	0.8	1.0, 1.15

Coefficients



Bar plot

- Large coefficient → important effect
- Small coefficient → negligible effect



Probability plot

- ”Outlying” effects not random
- Only for factorial designs

Random error assumption

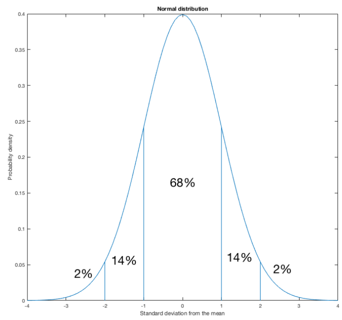
Observations that differ from the model due to random experimental error

→ Residuals approach the normal distribution

- Mean μ, variance $\hat{\sigma}^2$

Normalised residuals

- < |3| form 99.7%



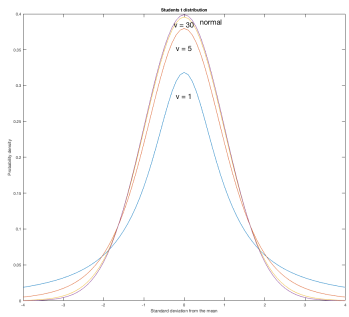
Student's t distribution

In practice population $\hat{\sigma}^2$ is unknown

- An estimate of $\hat{\sigma}^2$ depends on the number of observations

Student's t distribution

- Depends on dfs
- Two-sided



Student's t distribution

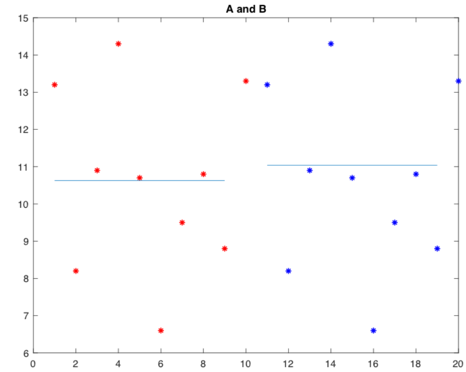
Randomized pair comparison

A

13.2	6.6
8.2	9.5
10.9	10.8
14.3	8.8
10.7	13.3

B

14.0	6.4
8.8	9.8
11.2	11.3
14.2	9.3
11.8	13.6



Box et al., Statistics for Experimenters; 2005:81.

Student's t distribution

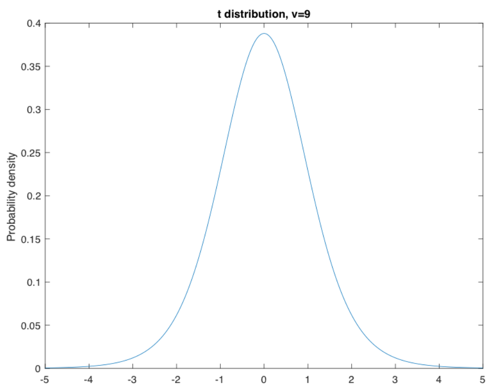
Random sampling of differences

$$s_d^2 = \sum \frac{(d-d)^2}{n-1} = 0.15$$

$$s_d = \sqrt{\frac{s_d^2}{n}} = 0.12$$

$$\rightarrow \; t_0 = \frac{0.41-0}{0.12} = 3.35$$

→ One- or two-sided?



Student’s t distribution

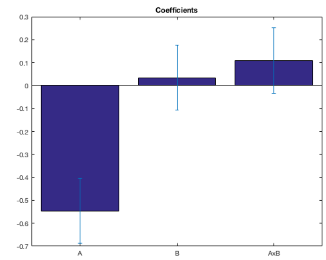
For regression coefficients

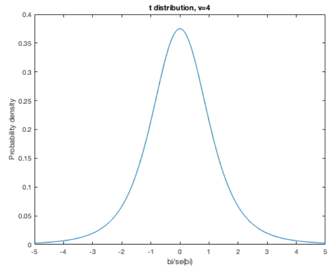
- Compare with zero
- H_0: $\beta_i = 0$ and H_1: $\beta_i \neq 0$

$\rightarrow H_1$ if $\left|\frac{\beta_i}{se(\beta_i)}\right| > t_{\frac{\alpha}{2},n-p}$

Different ways to calculate se(β_i)

- Residuals
- Replicates

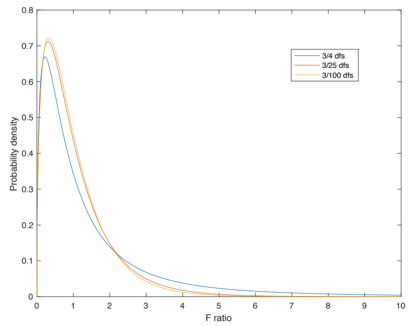




F distribution

For the ratio of sample variances

- Distribution for every combination of dfs
- One-sided with low dfs



ANOVA

Blood sugar from different diets

A	B	C	D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
63	65	68	63
59	66	68	64

Box et al., Statistics for Experimenters; 2005:134.

ANOVA

Blood sugar from different diets

	A	B	C	D
	62	63	68	56
	60	67	66	62
	63	71	71	60
	59	64	67	61
	63	65	68	63
	59	66	68	64
Diet mean	**61**	**66**	**68**	**61**
From grand mean	**-3**	**2**	**4**	**-3**

ANOVA

Blood sugar from different diets

$y_i - \overline{y}$

A	B	C	D
-2	-1	4	-8
-4	3	2	-2
-1	7	7	-4
-5	0	3	-3
-1	1	4	-1
-5	2	4	0

=

$\overline{y}_t - \overline{y}$

A	B	C	D
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3

+

$y_i - \overline{y}_t$

A	B	C	D
1	-3	0	-5
-1	1	-2	1
2	5	3	-1
-2	-2	-1	0
2	-1	0	2
-2	0	0	3

ANOVA

Blood sugar from different diets

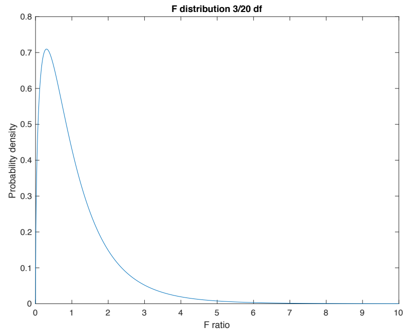
Parameter	df	Sum of squares (SS)	Mean square (MS)	F ratio
Total corrected	23	340		
Diets	3	228	76	14
Residual	20	112	5.6	

F distribution

Testing model significance

- F = Model variance / residual variance
- H_0: $\beta_1 = \ldots = \beta_i = 0$
- H_1: at least one $\beta \neq 0$

$\rightarrow H_1$ if $\dfrac{MS_{\text{mod}}}{MS_{res}} > F_{\alpha,k,n-p}$



ANOVA

For the original example, main effects model

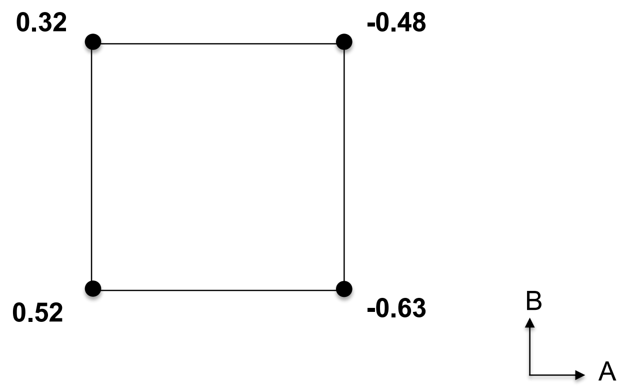
Parameter	df	Sum of squares (SS)	Mean square (MS)	F-value	p-value
Total corrected	7	2.6			
Model	2	2.4	1.2	34	<0.01
Residual	5	0.2	0.04		

ANOVA

For the original example

- **y** around its mean

$\rightarrow SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$



ANOVA

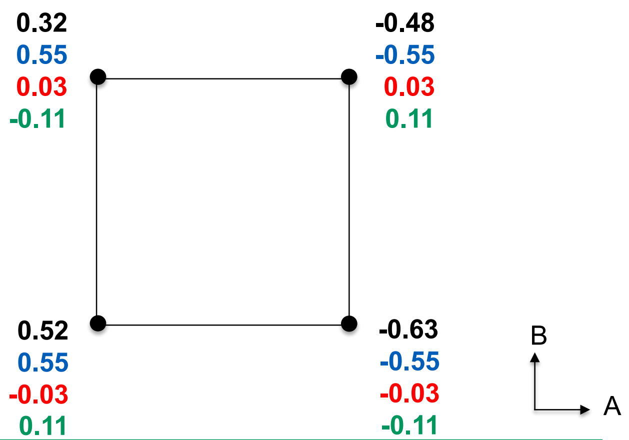
For the original example

- **y** around its mean

$\rightarrow SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$

- Value predicted by a variable

$\rightarrow SS_k = \sum_{i=1}^{n} x_{k,i} b_{k,i}{}^2$



ANOVA

Parameter	df	Sum of squares (SS)	Mean square (MS)	F-value	p-value
Total corrected	7	2.57			
Model	3	2.49	0.83	39.8	<0.01
A	1	2.39	2.39	114	
B	1	0.01	0.01	0.44	
AB	1	0.09	0.09	4.54	
Residual	4	0.08	0.04		

R^2 value

The variation explained by the model

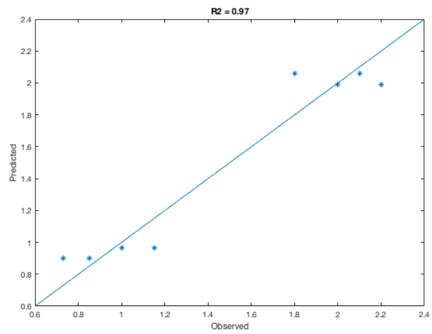
$$R^2 = \frac{SS_{\text{mod}}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- Sum of squares are additive
- Always increases with more terms

→ Easy to overfit

With $R^2 = 0.50$ the model equals noise

→ No model

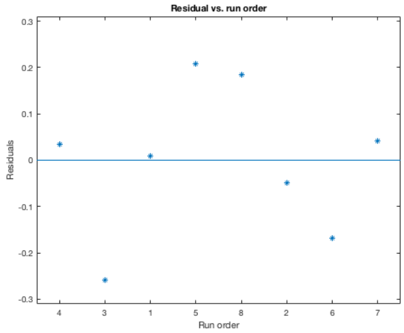


Residuals

Easier to identify abnormalities

- Raw residuals not very useful

→ Different normalisation norms are used

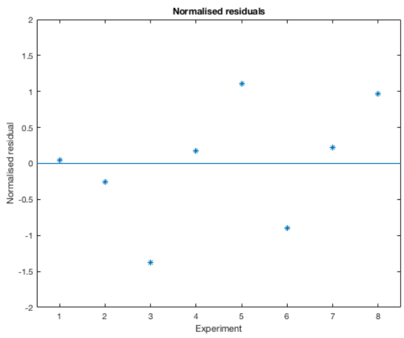


Residuals

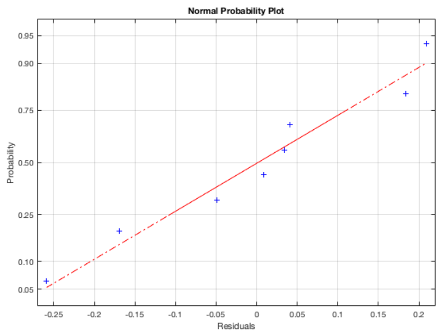
Normalised residuals

→ Residuals due to pure random error

- 4.6% > |2|
- 0.3% > |3|



Residuals



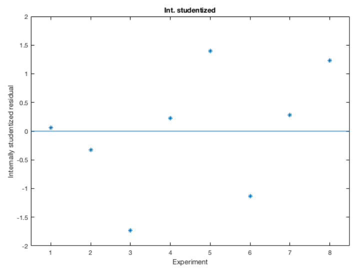
Normal probability plot

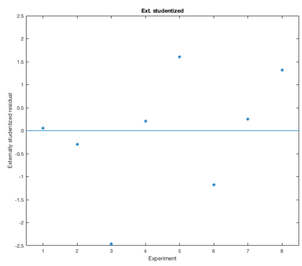
- Normally distributed should lie on a straight line
- Same with coefficients?

Aalto University
School of Chemical
Technology

25.3.2019
23

Residuals





Studentized residuals

- Take into consideration leverage

Aalto University
School of Chemical
Technology

25.3.2019
24

Session 2

Diagnostics

- Coefficients
- Random error assumption
- Statistical distributions
- ANOVA
- R^2
- Residuals

Nomenclature

Sum of squares
Overfitting
Bar plot
Probability plot
Random error
Degrees of freedom
Variance
Normalisation
Studentized residuals
Leverage

Thank you!