

Introductory Course: Using LS-OPT[®] on the TRACC Cluster

1.2b -Introduction to Response Surface Methodology; Numerical Example

By: Cezary Bojanowski, PhD

Problem Description

- Data from: M. Sosada “Optimal Conditions for Fractionation of Rapeseed Lecthin with Alcohols”, Journal of the American Oil Chemists' Society, Volume 70, Number 4 / April, 1993
- <http://www.springerlink.com/content/217t0t4573k54242/fulltext.pdf>
- Influence of four variables on the yield and phosphatidylcholine enrichment (PCE) was investigated in the experiment. Full factorial point selection method was used in the experiment.
- Variables:
 - t - extraction time
 - V - solvent volume
 - C - ethanol concentration
 - T - temperature

Problem Description

- Given:
 - set of data for 16 points
 - Problem has 4 variables
 - Points are selected using 2^4 full factorial method
 - Two responses are investigated
- Goals:
 - Construct linear metamodel
 - Perform Model Adequacy check

Coding of Variables

Natural Variables				Coded Variables				Response	
t	V	C	T	A	B	C	D	Yield	PCE
15	10	98	25	1	1	1	1	27.6	43.8
5	5	98	25	-1	-1	1	1	16.6	27.2
...

- In order to simplify the calculation, it is advisable to use coded variables
- Independent variables are described in (-1,1) interval
- 0 is in the middle of the design and
- +/- 1 are the furthest distances from the center

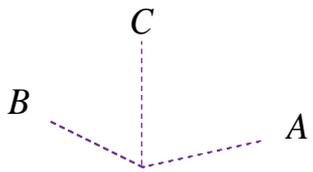
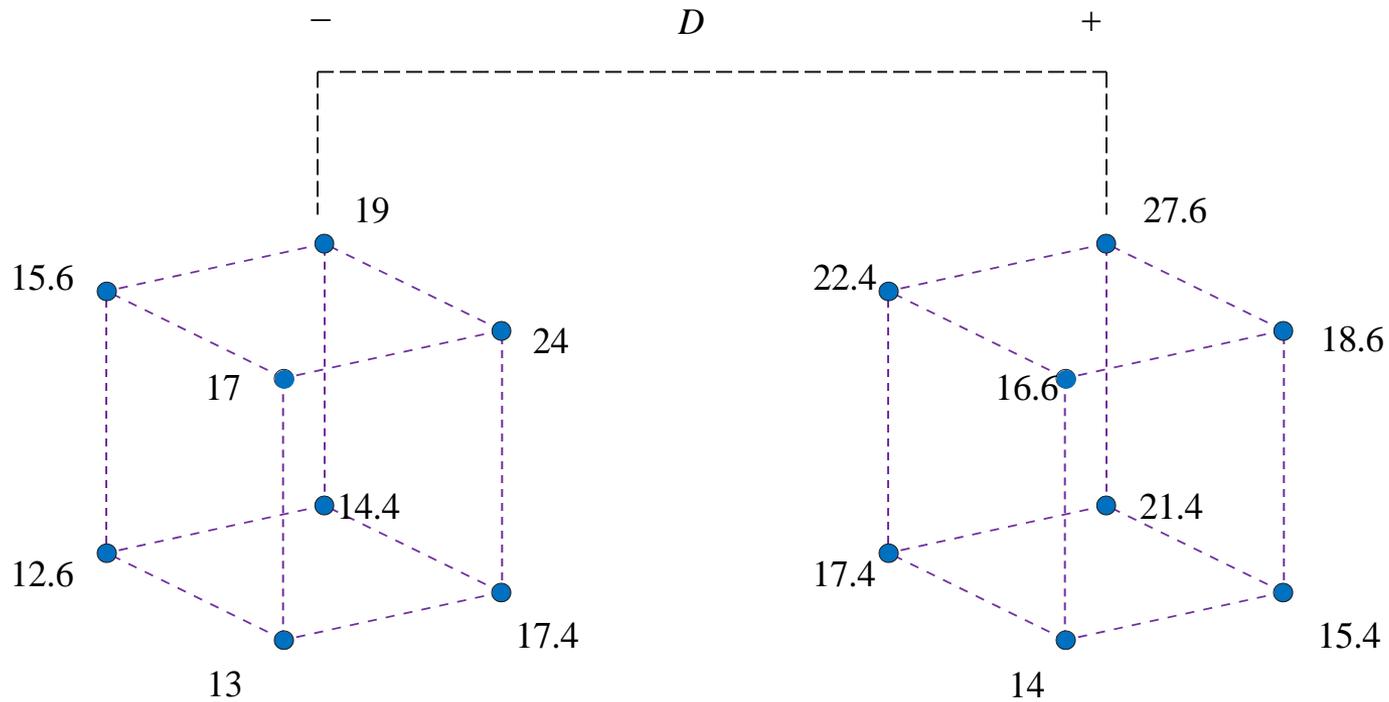
$$A = \frac{t - 10}{5} \quad B = \frac{V - 7.5}{2.5} \quad C = \frac{Con - 95}{3} \quad D = \frac{T - 20}{5}$$

Data for Fitting the First Order Model

Natural Variables				Coded Variables				Response	
t	V	C	T	A	B	C	D	Yield	PCE
15	10	98	25	1	1	1	1	27.6	43.8
5	5	98	25	-1	-1	1	1	16.6	27.2
15	5	92	25	1	-1	-1	1	15.4	23.6
5	10	92	25	-1	1	-1	1	17.4	26.2
15	5	98	15	1	-1	1	-1	17.0	27.8
5	10	98	15	-1	1	1	-1	19.0	30.2
15	10	92	15	1	1	-1	-1	17.4	25.2
5	5	92	15	-1	-1	-1	-1	12.6	18.8
15	5	98	25	1	-1	1	1	18.6	28.8
5	10	98	25	-1	1	1	1	22.4	36.8
15	10	92	25	1	1	-1	1	21.4	33.4
5	5	92	25	-1	-1	-1	1	14.0	21.0
15	10	98	15	1	1	1	-1	24.0	38.0
5	5	98	15	-1	-1	1	-1	15.6	23.6
15	5	92	15	1	-1	-1	-1	13.0	20.2
5	10	92	15	-1	1	-1	-1	14.4	22.6



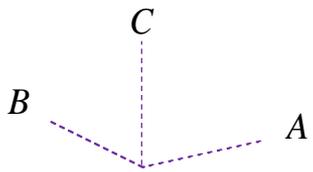
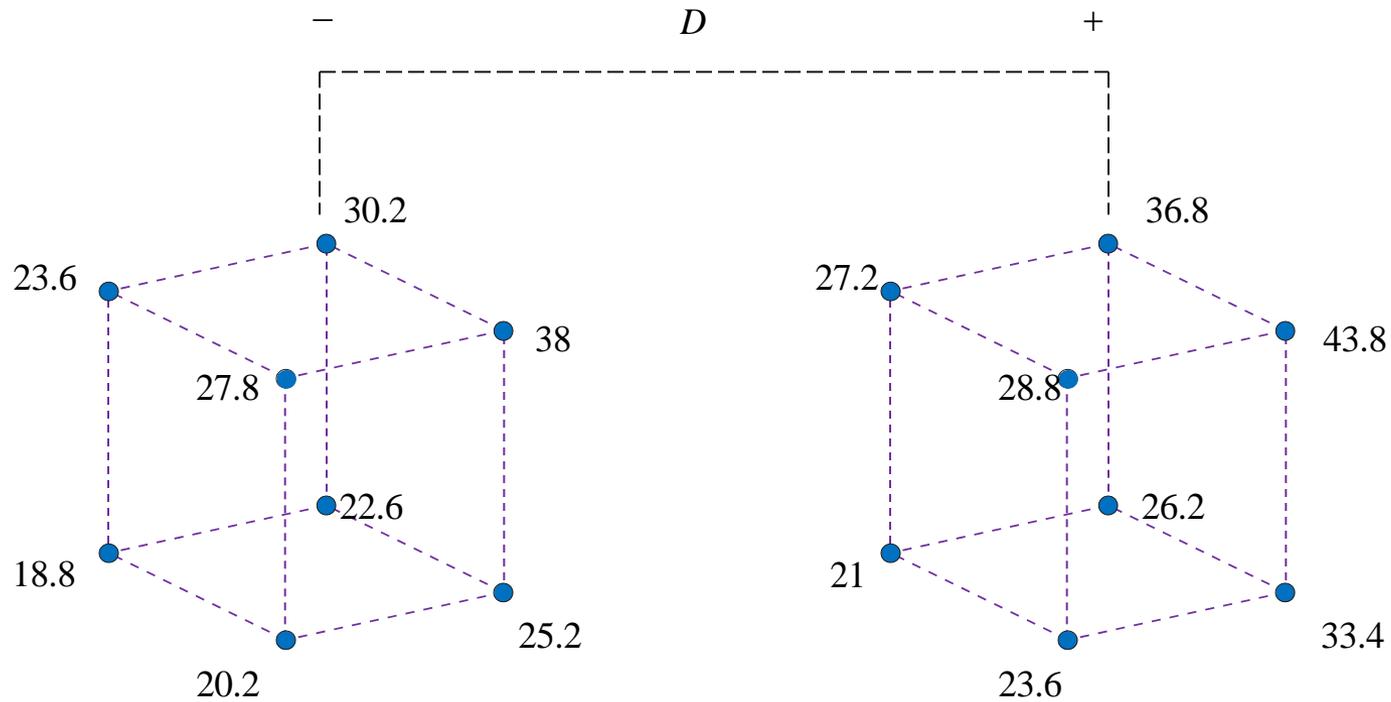
Geometric View of the Response Variable Yield



2^3

2^3

Geometric View of the Response Variable PCE



2^3

2^3

Multiple Linear Model

$$\hat{y} = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D$$

$$\hat{\beta} = (X' X)^{-1} X' y$$

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 \end{bmatrix}$$

$$X' X = \begin{bmatrix} 16 & 0 & 0 & 0 & 0 \\ 0 & 16 & 0 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 \\ 0 & 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 0 & 16 \end{bmatrix}$$

$$(X' X)^{-1} = \begin{bmatrix} 0.0625 & 0 & 0 & 0 & 0 \\ 0 & 0.0625 & 0 & 0 & 0 \\ 0 & 0 & 0.0625 & 0 & 0 \\ 0 & 0 & 0 & 0.0625 & 0 \\ 0 & 0 & 0 & 0 & 0.0625 \end{bmatrix}$$

Multiple Linear Model - PCE Response

$$\hat{y} = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D$$

$$\hat{\beta} = (X' X)^{-1} X' y$$

$$X' y = \begin{bmatrix} 447.2 \\ 34.4 \\ 65.2 \\ 65.2 \\ 34.4 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} 27.950 \\ 2.150 \\ 4.075 \\ 4.075 \\ 2.50 \end{bmatrix}$$

$$y_{PCE} = \begin{bmatrix} 43.8 \\ 27.2 \\ 23.6 \\ 26.2 \\ 27.8 \\ 30.2 \\ 25.2 \\ 18.8 \\ 28.8 \\ 36.8 \\ 33.4 \\ 21.0 \\ 38.0 \\ 23.6 \\ 20.2 \\ 22.6 \end{bmatrix}$$

$$\hat{y}_{PCE} = 27.95 + 2.15A + 4.075B + 4.075C + 2.5D$$

Multiple Linear Model - Yield Response

$$\hat{y} = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D$$

$$\hat{\beta} = (X' X)^{-1} X' y$$

$$X' y = \begin{bmatrix} 286.0 \\ 22.8 \\ 40.4 \\ 35.6 \\ 20.8 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} 17.9 \\ 1.4 \\ 2.55 \\ 2.2 \\ 1.275 \end{bmatrix}$$

$$y_{Yield} = \begin{bmatrix} 27.6 \\ 16.6 \\ 15.4 \\ 17.4 \\ 17.0 \\ 19.0 \\ 17.4 \\ 12.6 \\ 18.6 \\ 22.4 \\ 21.4 \\ 14.0 \\ 24.0 \\ 15.6 \\ 13.0 \\ 14.4 \end{bmatrix}$$

$$\hat{y}_{Yield} = 17.9 + 1.4A + 2.55B + 2.2C + 1.275D$$

Degrees of Freedom

- Estimates of statistical parameters can be based upon different amounts of information or data.
- The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom.
- Geometrically, the degrees of freedom can be interpreted as the dimension of certain vectors.

Model Adequacy Checking

Variation	Sum of Squares	Degrees of Freedom	Mean Square	F ₀
Regression	SS_R	q	MS_R	MS_R / MS_E
Error of Residuals	SS_E	$N-q-1$	MS_E	
Total	SS_y	$N-1$		

N – number of observations

q – number of independent variables

$$SS_y = SS_R + SS_E$$

$$S_{yy} = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n}$$

$$SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

Sums of Squares for PCE Response

$$y' y = 13235.04$$

$$\hat{\beta}' X' y = 13178.54$$

$$\frac{\left(\sum_{i=1}^n y_i \right)^2}{n} = 12499.24$$

$$SS_R = 13178.54 - 12499.24 = 679.30$$

$$SS_E = 13235.04 - 13178.54 = 56.50$$

$$SS_{yy} = 13235.04 - 12499.24 = 735.80$$

Mean Squares for PCE Response

$$MS_R = \frac{SS_R}{q} = \frac{679.30}{4} = 169.825$$

$$MS_E = \frac{SS_E}{N - q - 1} = \frac{56.5}{16 - 4 - 1} = 5.136$$

$$F = \frac{SS_R / q}{SS_E / N - q - 1} = \frac{MS_R}{MS_E} = 33.063$$

Sums of Squares for Yield Response

$$y' y = 5386.16$$

$$\hat{\beta}' X' y = 5365.41$$

$$\frac{\left(\sum_{i=1}^n y_i \right)^2}{n} = 5126.56$$

$$SS_R = 5365.41 - 5126.56 = 238.85$$

$$SS_E = 5386.16 - 5365.41 = 20.75$$

$$SS_{yy} = 5386.16 - 5126.56 = 259.60$$

Mean Squares for Yield Response

$$MS_R = \frac{SS_R}{q} = \frac{238.85}{4} = 59.712$$

$$MS_E = \frac{SS_E}{N - q - 1} = \frac{20.75}{16 - 4 - 1} = 1.886$$

$$F = \frac{SS_R / q}{SS_E / N - q - 1} = \frac{MS_R}{MS_E} = 31.655$$

Statistical Tests

- Statistical hypotheses form the basis of statements and conclusions we can make about sets of data
- Two sets of hypotheses are formulated
 - Null hypothesis $H_0 : \mu = \mu_0$
 - Alternate hypothesis $H_1 : \mu > \mu_0$
- The hypotheses are designed to be proven or disproven, such as “the sample means of two sets of data are statistically the same and the samples come from the same overall population”

F-test

- The test for significance of regression is a test to determine if there is a linear relationship between the response y and a subset of the regressor variables x_i

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \beta_j \neq 0$$

- The test procedure for H_0 is to compute:

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)}$$

$$F_0 = \frac{MS_R}{MS_E}$$

explained variance

random variance

- If this value exceeds the statistical one then H_0 is rejected.
- Meaning at least one variable x_i contributes to the model significantly.

t-test

- We are frequently interested in the importance of the distinct variable to the response.
- Perhaps the model would be more effective with the inclusion of additional variable? Or maybe with the deletion of one or more?
- The hypotheses for t-test are:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

- If the null hypothesis is not rejected then it indicates that *j-th* variable can be deleted from the model

t-test

- The expected value of the SS_E error:

$$SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

- ... can be proven to be:

$$E(SS_E) = \sigma^2(n - p)$$

- Estimator of variance is given by:

$$\hat{\sigma}^2 = \frac{SS_E}{(n - p)}$$

t-test

- The test procedure for this hypothesis is to calculate:

$$t_0 = \frac{b_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \quad j = 1, 2, \dots, k$$

- where C_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to b_j
- $\hat{\sigma}^2 C_{jj}$ is called standard error of the regression coefficient b_j
- The null hypothesis is rejected if t_0 is greater than statistical value of t
- meaning the variable is important for the model

$$|t_0| > t_{\alpha/2, n-k-1}$$

see tables for that value

F-test

- The test for significance of regression is a test to determine if there is a linear relationship between the response y and a subset of the regressor variables x_i

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \beta_j \neq 0$$

$$F_0^{PCE} = \frac{SS_R / q}{SS_E / N - q - 1} = \frac{MS_R}{MS_E} = 33.063$$

$$F_0^{Yield} = \frac{SS_R / q}{SS_E / N - q - 1} = \frac{MS_R}{MS_E} = 31.655$$

$$F_{\alpha, q, N-q-1} = F_{0.05, 4, 11} = 3.36$$

$$F_{0.05, 4, 11} < F_0$$

- If this value exceeds the statistical one then H_0 is rejected.
- Meaning at least one variable x_i contributes to the model significantly.

Coefficient of Multiple Determination

- represents ability of the model to capture the variability of the real response

$$0 \leq R^2 \leq 1$$

- PCE response:

$$R^2 = \frac{SS_R}{SS_y} = 1 - \frac{SS_E}{SS_y} = \frac{679.30}{735.80} = 0.923$$

- Yield response:

$$R^2 = \frac{SS_R}{SS_y} = 1 - \frac{SS_E}{SS_y} = \frac{238.85}{259.60} = 0.92$$

t-test for PCE Response

- The hypotheses for t-test are:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

- If the null hypothesis is not rejected then it indicates that *j-th* variable can be deleted from the model

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

$$\hat{\sigma}^2 = \frac{SS_E}{(n-q)} = \frac{56.5}{16-4} = 4.71$$

t-test for PCE Response

$$t_A = \frac{2.15}{\sqrt{4.71 \cdot 0.0625}} = 3.96$$

$$t_B = \frac{4.075}{\sqrt{4.71 \cdot 0.0625}} = 2.88$$

$$t_C = \frac{4.075}{\sqrt{4.71 \cdot 0.0625}} = 2.88$$

$$t_D = \frac{2.15}{\sqrt{4.71 \cdot 0.0625}} = 3.96$$

$$t_{\frac{\alpha}{2}, N-q-1} = t_{0.025, 11} = 2.201 < t_A, t_B, t_C, t_D$$

t-test for Yield Response

- The hypotheses for t-test are:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

- If the null hypothesis is not rejected then it indicates that *j-th* variable can be deleted from the model

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

$$\hat{\sigma}^2 = \frac{SS_E}{(n-q)} = \frac{20.75}{16-4} = 1.72$$

t-test for Yield Response

$$t_A = \frac{1.40}{\sqrt{1.72 \cdot 0.0625}} = 4.27$$

$$t_B = \frac{2.55}{\sqrt{1.72 \cdot 0.0625}} = 7.78$$

$$t_C = \frac{2.20}{\sqrt{1.72 \cdot 0.0625}} = 6.71$$

$$t_D = \frac{1.275}{\sqrt{1.72 \cdot 0.0625}} = 3.89$$

$$t_{\frac{\alpha}{2}, N-q-1} = t_{0.025, 11} = 2.201 < t_A, t_B, t_C, t_D$$

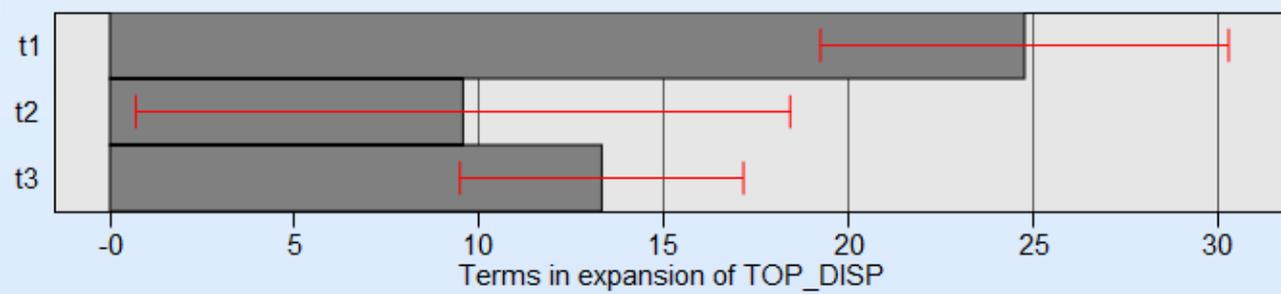
Confidence Intervals in Multiple Regression - PCE

$$\hat{\beta}_j - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{jj}}$$

$$t_{\frac{\alpha}{2}, N-q-1} = t_{0.025, 11} = 2.201 < t_A, t_B, t_C, t_D$$

$$\beta_0 = 27.95 \pm 1.19 \quad \beta_1 = 2.15 \pm \sqrt{4.71 \cdot 0.0625} \cdot 2.201 = 2.15 \pm 1.19$$

$$\beta_2 = 4.075 \pm 1.19 \quad \beta_3 = 4.075 \pm 1.19 \quad \beta_4 = 2.5 \pm 1.19$$



Confidence Intervals in Multiple Regression - Yield

$$\hat{\beta}_j - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{jj}}$$

$$t_{\frac{\alpha}{2}, N-q-1} = t_{0.025, 11} = 2.201 < t_A, t_B, t_C, t_D$$

$$\beta_0 = 17.9 \pm 0.33 \quad \beta_1 = 1.4 \pm \sqrt{1.72 \cdot 0.0625} \cdot 2.201 = 1.4 \pm 0.33$$

$$\beta_2 = 2.55 \pm 0.33 \quad \beta_3 = 2.2 \pm 0.33 \quad \beta_4 = 1.275 \pm 0.33$$