**Problem statement**

IJ Barnabas, JR Dean, WR Tomlinson and SP Owen (1995) used a central composite design (CCD) to study the effects of four factors, Pressure (PRES) , Temperature (TEMP), Extraction Time (ET) and Methanol Content (MC) on total recovery of Polycyclic Aromatic Hydrocarbons (PAHs) when extracted from soil.

Total design was composed of nf  = 24 = 16 factorial points and na = 2p = 8 axial points.

The experimenters chose to use noa = 2, axial centre points and nof = 4, factorial centre points.

RUN B1 B2 PRES TEMP ET MC Y

== First Factorial Block ===

1 1 0 250 55 47.5 15 391.8

2 1 0 150 85 47.5 15 413.6

3 1 0 250 55 22.5 5 68.7

4 1 0 250 85 47.5 5 143.0

5 1 0 150 85 22.5 5 104.0

6 1 0 150 55 22.5 15 309.1

7 1 0 200 70 35.0 10 400.6

8 1 0 250 85 22.5 15 402.5

9 1 0 150 55 47.5 5 77.7

10 1 0 200 70 35.0 10 426.5

== Second Factorial Block ===

11 0 1 250 85 47.5 15 457.5

12 0 1 150 55 22.5 5 56.9

13 0 1 250 85 22.5 5 94.1

14 0 1 250 55 22.5 15 409.7

15 0 1 150 55 47.5 15 410.9

16 0 1 150 85 22.5 15 375.8

17 0 1 150 85 47.5 5 110.5

18 0 1 200 70 35.0 10 387.8

19 0 1 250 55 47.5 5 103.0

20 0 1 200 70 35.0 10 399.1

== Axial Factorial Block ===

21 -1 -1 200 70 35.0 10 416.9

22 -1 -1 200 40 35.0 10 359.8

23 -1 -1 200 70 10.0 10 276.1

24 -1 -1 200 70 60.0 10 462.3

25 -1 -1 100 70 35.0 10 311.5

26 -1 -1 200 70 35.0 10 346.5

27 -1 -1 200 70 35.0 0 46.8

28 -1 -1 200 70 35.0 20 418.7

29 -1 -1 200 100 35.0 10 413.9

30 -1 -1 300 70 35.0 10 429.4

**Terminology**

|  |  |
| --- | --- |
| **Lack of fit tests** | In statistics, a sum of squares due to lack of fit, or more tersely a lack-of-fit sum of squares, is one of the components of a partition of the sum of squares of residuals in an analysis of variance, used in the numerator in an F-test of the null hypothesis that says that a proposed model fits well.  In RSM, p-value of lack-of-fit, if >0.05 (not significant) means that the model fits well. and there is significant effect on parameters on output response.  A regression model exhibits lack-of-fit when it fails to adequately describe the functional relationship between the experimental factors and the response variable. Lack-of-fit can occur if important terms from the model such as interactions or quadratic terms are not included. It can also occur if several, unusually large residuals result from fitting the model.  *Ref: https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/regression-models/lack-of-fit-and-lack-of-fit-tests/* |
| **Pure error** | Pure error reflects the variability of the observations within each treatment. The sum of squares for the pure error is the sum of the squared deviations of the responses from the mean response in each set of replicates.  A lack-of-fit error significantly larger than the pure error indicates that something remains in the residuals that can be removed by a more appropriate model.  As replicate measurements are available, a test indicating the significance of the replicate error in comparison to the model dependent error is performed. This test splits the residual or error sum of squares into two portions, one which is due to pure error which is based on the replicate measurements and the other due to lack-of-fit based on the model performance.  *Ref: https://core.ac.uk/download/pdf/11776921.pdf* |
| **FO** | FO() in the model formula in rsm indicates a first-order response surface (i.e., a linear function) in its arguments. |
| **TWI** | TWI() indicates two-way interactions such as ET \* PRES |
| **PQ** | PQ() indicates pure quadratic terms (squares of the FO() terms) such as PRES \* PRES |
| **Saddle point** | A saddle point yields neither a maximum or minimum for the fitted model. The surface close to a saddle point is reminiscent of a horse saddle - raising up from front to back but sloping down from side to side. Instead, these will be found at the edge of the local design region.  A saddle point is indicative that the polynomial is not fitting the data points very well.  R*ef: https://ieeexplore.ieee.org/document/8736803* |
| **Stationary point** | The point for which the response, y^, is optimized is the point at which the partial derivatives, ∂y^/∂x1, ∂y^/∂x2, ...∂y^/∂xk, are all equal to zero. This point is called the stationary point. The stationary point may be a point of maximum response, minimum response or a saddle point. |
| **CCD** | Central Composite Designs (CCD) were designed by Box and Wilson in 1951, and they are nowadays the most popular second-order designs. Each design consists of standard first order design with nf orthogonal factorial points and no center points, augmented by na axial points.  The CCD can be made orthogonal (no correlation between parameter estimates) or rotatable (the variance of the estimated response is a function only of the distance from the design center and not the direction) by the location of the axial treatments. In spite of the reduced number of treatments compared with a full 3k factorial, the CCD is relatively efficient in estimation of the quadratic and interaction terms. Blocking of experimental units is often desirable and is sometimes required. |

**Solution**:

A Response Surface Methodology (RSM) was used to solve this problem. A rotatable central composite design (CCD) with orthogonal blocking was used to study the effects of four factors - pressure (PRES), temperature (TEMP), extraction time (ET) and methanol content (MC) on the total recovery of polycyclic aromatic hydrocarbons (Y) when extracted from soil.

*RSM has many advantages when compared to classical methods. It needs fewer experiments to study the effects of all the factors and the optimum combination of all the variables can be revealed. The interaction (the behavior of one factor may be dependent on the level of another factor) between factors can be determined.*

In RSM methods, appropriate coding of data is important not only for numerical stability, but for proper scaling of results.

Coding is performed on the factors (aka variables) so that +1 and -1 represent the extreme levels of each factor:

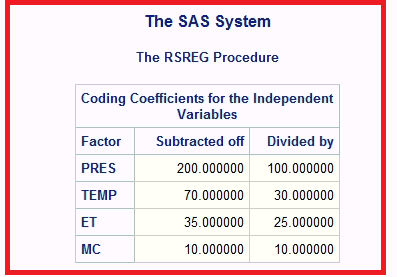
PRES = (PRES - 200) / 100

TEMP = (TEMP - 70) / 30

ET = (ET - 35) / 25

MC = (MC - 10) / 10

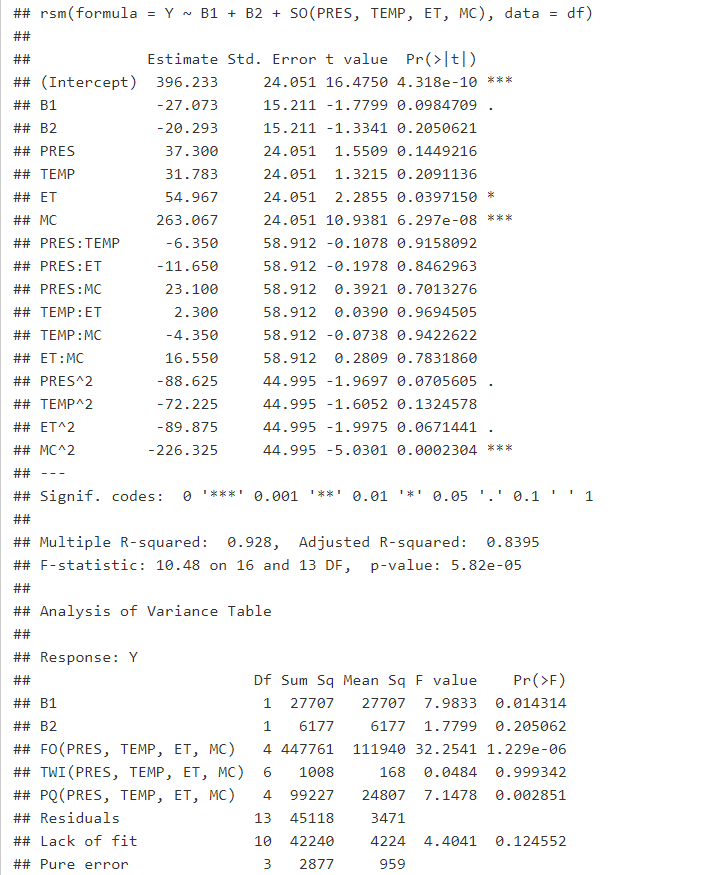
It is done automatically in SAS. Please refer below:

****

**Results**

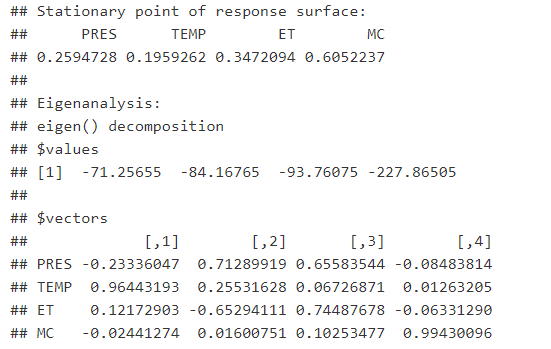
**Using R package, rsm()**

Analysis of Variance including least square parameter estimates, lack of fit test and the corresponding t tests:



**Table 1: R output – Analysis of Variance, lack-of-fit test, and parameter estimates for coded factor levels**

There is a unique treatment combination called stationary point. Here coded variables are used.



**Table 2: R output – Canonical Analysis of Response Surface**

PRES = 0.2594728; TEMP = 0.1959262; ET = 0.3472094; MC = 0.6052237

For uncoded, the stationary point:

PRES = 225.947278; TEMP = 75.87779; ET = 43.68023; MC = 16.052237

Predicted value at stationary point = 493.3356

## Eigen analysis:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| Eigen value | -71.25655 | -84.16765 | -93.76075 | ***-227.86505*** |
| **Eigen vectors** | | | | |
| PRES | -0.23336047 | 0.71289919 | 0.65583544 | -0.08483814 |
| TEMP | 0.96443193 | 0.25531628 | 0.06726871 | 0.01263205 |
| ET | 0.12172903 | -0.65294111 | 0.74487678 | -0.06331290 |
| MC | -0.02441274 | 0.01600751 | 0.10253477 | 0.99430096 |

* If the eigenvalues are all negative, the stable point is a maximum.
* If the eigenvalues are all positive, the stable point is a minimum.
* If the eigenvalues are of mixed sign, the stable point is a saddle.

For this experiment, all eigen values (canonical coefficients) are negative, so the stationary point is maximum.

The eigen values are each scaled to be of length one. The last eigen value, with value = -227.86505 is the largest in magnitude. From the corresponding eigen vector, the primary component is that of MC with value, 0.99430096.

So, the fitted model has greatest curvature at the stationary point when moving in either direction determined by this fourth eigen vector, which is nearly parallel to the MC axis.

R Code

### R commands and output:

library(rsm)

### -------------------------

### FUNCTION, y\_hat\_calcuate

### -------------------------

y\_hat\_calcuate <- function(PRES, TEMP, ET, MC) {

y\_hat <- 396.233 + (37.3 \* PRES) +

(31.783 \* TEMP) + (54.967 \* ET ) + ( 263.067 \* MC) -

(6.350 \* PRES \* TEMP) - (11.650 \* PRES \* ET) +

(23.1 \* PRES \* MC) + (2.3 \* TEMP \* ET) - (4.35 \* TEMP \* MC) +

(16.55 \* ET \* MC) - (88.625 \* PRES \* PRES) -

(72.225 \* TEMP \* TEMP) - (89.875 \* ET \* ET) - (226.325 \* MC \* MC)

return(y\_hat)

}

###

### END OF FUNCTION, y\_hat\_calcuate

### -------------------------------

###

## Read data and save relevant variables.

fname = "D:/MDRF-FNDR/RSM/data/PAH-RECOVERY.DAT"

m = matrix(scan(fname, skip = 1),ncol = 8,byrow = T)

print(m)

run = m[,1]

B1 = m[,2]

B2 = m[,3]

PRES = m[,4]

TEMP = m[,5]

ET = m[,6]

MC = m[,7]

Y = m[,8]

###

cat("\n PRES")

print(max(PRES))

print(min(PRES))

###

cat("\n TEMP")

print(max(TEMP))

print(min(TEMP))

###

cat("\n TEMP")

print(max(TEMP))

print(min(TEMP))

###

cat("\n ET")

print(max(ET))

print(min(ET))

###

### CODING COEFFICIENTS FOR THE INDEPENDENT VARIABLES

###

cat("\n MC")

print(max(MC))

print(min(MC))

###

PRES = (PRES - 200) / 100

TEMP = (TEMP - 70) / 30

ET = (ET - 35) / 25

MC = (MC - 10) / 10

## Create data frame.

df = data.frame(run,B1, B2, PRES, TEMP, ET, MC, Y)

CR2.rsm <- rsm(Y ~ B1 + B2 + SO(PRES, TEMP, ET, MC), data = df)

CR2\_summary = summary(CR2.rsm)

print(CR2\_summary)

PRES\_VAL = 0.2594728

TEMP\_VAL = 0.1959262

ET\_VAL = 0.3472094

MC\_VAL = 0.6052237

RES = y\_hat\_calcuate(PRES\_VAL, TEMP\_VAL, ET\_VAL, MC\_VAL)

cat("\n=======================================================================")

cat("\nThere is a unique treatment combination called stationary point")

cat("\nPRES = 0.2594728; TEMP = 0.1959262; ET = 0.3472094; MC = 0.6052237\n")

cat("\nPredicted value at stationary point ", RES)

cat("\n=======================================================================")

SAS code

DATA PAH;

INPUT RUN B1 B2 PRES TEMP ET MC Y;

LINES;

1 1 0 250 55 47.5 15 391.8

2 1 0 150 85 47.5 15 413.6

3 1 0 250 55 22.5 5 68.7

4 1 0 250 85 47.5 5 143.0

5 1 0 150 85 22.5 5 104.0

6 1 0 150 55 22.5 15 309.1

7 1 0 200 70 35.0 10 400.6

8 1 0 250 85 22.5 15 402.5

9 1 0 150 55 47.5 5 77.7

10 1 0 200 70 35.0 10 426.5

11 0 1 250 85 47.5 15 457.5

12 0 1 150 55 22.5 5 56.9

13 0 1 250 85 22.5 5 94.1

14 0 1 250 55 22.5 15 409.7

15 0 1 150 55 47.5 15 410.9

16 0 1 150 85 22.5 15 375.8

17 0 1 150 85 47.5 5 110.5

18 0 1 200 70 35.0 10 387.8

19 0 1 250 55 47.5 5 103.0

20 0 1 200 70 35.0 10 399.1

21 -1 -1 200 70 35.0 10 416.9

22 -1 -1 200 40 35.0 10 359.8

23 -1 -1 200 70 10.0 10 276.1

24 -1 -1 200 70 60.0 10 462.3

25 -1 -1 100 70 35.0 10 311.5

26 -1 -1 200 70 35.0 10 346.5

27 -1 -1 200 70 35.0 0 46.8

28 -1 -1 200 70 35.0 20 418.7

29 -1 -1 200 100 35.0 10 413.9

30 -1 -1 300 70 35.0 10 429.4

;

\* SORT by independent variables for lack of fit test;

PROC SORT;

BY B1 B2 PRES TEMP ET MC;

;

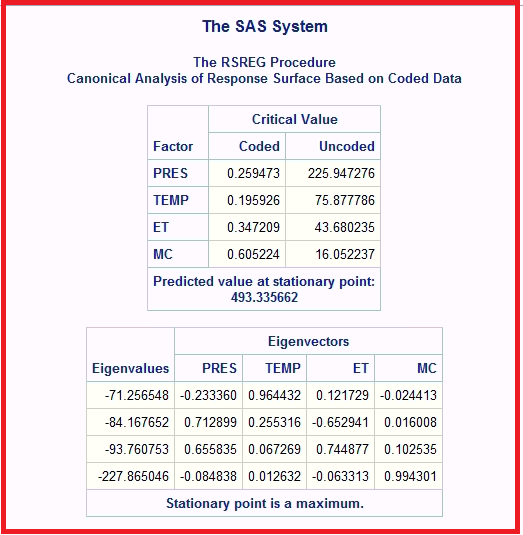
\* RESPONSE SURFACE REGRESSION;

PROC RSREG;

MODEL Y = B1 B2 PRES TEMP ET MC / COVAR = 2 LACKFIT;

;

**Output from SAS**



**Table 3: SAS output – Canonical Analysis of Response Surface**