Bartlett Test

Note that for calculating the confidence intervals, we assumed that the true variance σ^2 is the same for all observations and that the observations are independent.

- How can we check if this assumption is valid? **Bartlett Test**

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2 = \sigma_y^2$$

$$H_1 : \text{at least one } \sigma_i^2 \neq \sigma_j^2, i \neq j$$

$$\chi^2_{cal} = \frac{M}{C}$$

where

$$M = (N-m)\ln(s_p^2) - \sum_{i=1}^{m}(n_i - 1)\ln(s_i^2)$$

$$C = 1 + \frac{1}{3(m-1)}\left[\left(\sum_{i=1}^{m}\frac{1}{n_i - 1}\right) - \frac{1}{N-m}\right]$$

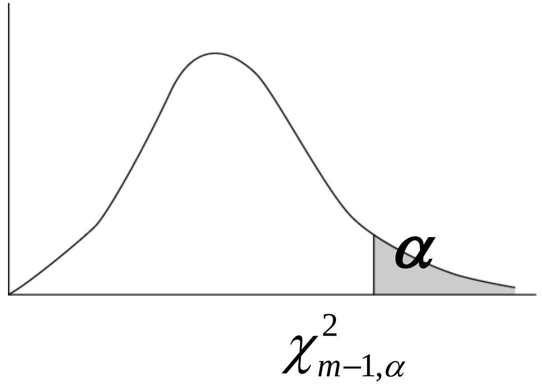
$m = 2^k$ *(Total experiments)*

Sample size = n $N = n_1 + n_2 + \ldots + n_m$

The value of M will be large if the sample variances s_i^2 differ greatly in magnitude, and will be zero if all the sample variances are exactly equal.

We will reject H_0 if χ^2_{cal} is too large, i.e.,

$$\chi^2_{\text{cal}} > \chi^2_{m-1,\alpha}$$



Bartlett Test

Example: Bartlett Test

$$\chi^2_{v=m-1} = \frac{M}{C} \quad \text{where, } M = (N-m)\ln(s_p^2) - \sum_{i=1}^{m}(n_i - 1)\ln(s_i^2), \quad \textit{and} \quad C = 1 + \frac{1}{3(m-1)}\left[\left(\sum_{i=1}^{m}\frac{1}{n_i-1}\right) - \frac{1}{N-m}\right]$$

Here, N =   m =

The value of M will be large if the sample variances s_i^2 differ greatly in magnitude, and will be zero if all the sample variances are exactly equal.

$S_1^2 = 24.5$, $S_2^2=21.78$, $S_3^2=134.48$, $S_4^2=242.0$,

$S_5^2=3.92$, $S_6^2=8.82$, $S_7^2=33.62$, $S_8^2=72.00$

We will reject H_0 if χ^2_{cal} is too large, i.e., $\chi^2_{cal} > \chi^2_{m-1,\alpha}$

$$S_p{}^2 = \frac{[(y_{a1}-\bar{y}_1)^2+(y_{b1}-\bar{y}_1)^2+\ldots\ldots\ldots\ldots+(y_{a8}-\bar{y}_8)^2+(y_{b8}-\bar{y}_8)^2]}{(2-1)+\ldots\ldots\ldots+(2-1)} = 67.64$$

$$\begin{aligned} M &= (16-8)\ln 67.64 - [(2-1)\ln 24.5 + (2-1)\ln 21.78 \\ &\quad +(2-1)\ln 134.48 + (2-1)\ln 242 + (2-1)\ln 3.92 \\ &\quad +(2-1)\ln 8.82 + (2-1)33.62 + (2-1)\ln 72.0] \\ &= 5.713 \end{aligned}$$

$$\begin{aligned} C &= 1 + \frac{1}{3(8-1)}\left[\left(\sum_{i=1}^{8}\frac{1}{2-1}\right) - \frac{1}{16-8}\right] \\ &= 1 + \frac{1}{21}[8-5] = 1.357 \end{aligned}$$

$$\chi^2_{cal} = \frac{5.713}{1.357} = 4.21 \qquad \chi^2_{7,\alpha=0.05} = 14.1$$

Test #	X1	X2	X3	Y_{ai} (kpsi)	Y_{bi} (kpsi)	Average (kpsi)
1	-1	-1	-1	84	91	87.5
2	1	-1	-1	90.6	84	87.3
3	-1	1	-1	69.6	86	77.8
4	1	1	-1	76	98	87
5	-1	-1	1	77.7	80.5	79.1
6	1	-1	1	99.7	95.5	97.6
7	-1	1	1	82.7	74.5	78.6
8	1	1	1	93.7	81.7	87.7

Example

The yield form, a certain chemical depends on

either the **chemical formulation of the input materials** or **the mixer speed**, **or both**.

A 2-level factorial design was run with three replicates and the yield data are shown below.

Find Main and Interaction Effects and their Confidence Intervals, and Significance using ANOVA

x_1	x_2	y_a	y_b	y_c	$\bar{y}$
-1	-1	10	20	30	20
1	-1	40	30	50	40
-1	1	60	30	60	50
1	1	40	45	50	45

x_1	x_2	y_a	y_b	y_c	$\bar{y}$
-1	-1	10	20	30	20
1	-1	40	30	50	40
-1	1	60	30	60	50
1	1	40	45	50	45

$x_1 x_2$

$$E_1 = \frac{1}{2}\left(-20+40+(-50)+45\right)$$

$$E_1 = 7.5$$

$$E_2 = \frac{1}{2}\left(-20-40+50+45\right) = 17.5$$

$$E_{12} = \frac{1}{2}\left(20-40-50+45\right) = -12.5$$

What are sample variances? How to find confidence intervals for E_1, E_2, E_{12}?

$$S_1^2 = \sum \frac{(y_i-\bar{y})^2}{n-1} = \frac{(10-20)^2+(20-20)^2+(30-20)^2}{2} = 100$$

$$S_2^2 = \frac{(40-40)^2+(30-40)^2+(50-40)^2}{2} = 100$$

$$S_3^2 = 300$$

$$S_4^2 = 25$$

pooled variance

$$S_p^2 = \frac{\sum DOF \times S_i^2}{TDOF} = \frac{2(100)+2(100)+2(300)+2(25)}{2+2+2+2}$$

$$= 131.25$$

x_1	x_2	y_a	y_b	y_c	$\bar{y}$
-1	-1	10	20	30	20
1	-1	40	30	50	40
-1	1	60	30	60	50
1	1	40	45	50	45

Confidence interval for E_1 $\quad V(cy) = c^2 V(\bar{y})$

$$V(E_1) = V\left(\frac{\bar{y}_2 - \bar{y}_1 + \bar{y}_4 - \bar{y}_3}{2}\right)$$

$$= \frac{1}{4} V\left(\frac{y_{2a}+y_{2b}+y_{2c}}{3} - \frac{y_{1a}+y_{1b}+y_{1c}}{3} + \frac{y_{4a}+y_{4b}+y_{4c}}{3} - \frac{y_{3a}+y_{3b}+y_{3c}}{3}\right)$$

$$= \frac{1}{36}\, 12\, V(y_{1a}) = \frac{12}{36}\sigma^2 = \frac{1}{3}\sigma^2$$

$$V(E_1) = \sigma^2/3$$

$$\underline{V(E_2)} = \sigma^2/3 = V(E_{12})$$

Confidence interval $\quad E_i \pm t_{\nu,\alpha}\sqrt{\frac{S_p^2}{3}}$

at 95% confidence $\quad t_{\nu,\alpha} = t_{8,0.025} = 2.306$ from table

$$E_i \pm 2.306\sqrt{131.25/3} = E_i \pm 15.25$$

78

Statistical Design of Experiments

ME 794

Instructors: Prof. Suhas Joshi, Prof. Soham Mujumdar (sohammujumdar@iitb.ac.in)

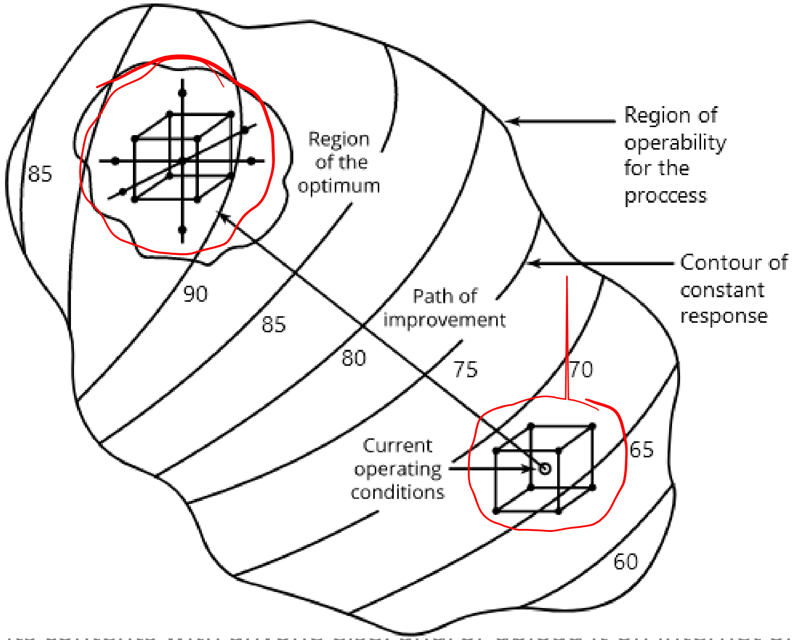
Response Surface Methodology

Acknowledgement: Design and Analysis of Experiments by Montegamory. Some of the course material has been adopted from similar courses taught previously by Prof. Shiv Kapoor (UofI), and Prof. Suhas Joshi (IITB).

Department of Mechanical Engineering
Indian Institute of Technology Bombay

Goal of RSM

- So far, the focus of the design of experiments was ***'factor screening'*** – which factors strongly affect the process, which factors are less important, how the factors interact ..
- After screening, we now shift our focus to ***optimization*** – which factor level combinations give us maximum (e.g. yield) or minimum (e.g. cost), or target result.
- *The objective of Response Surface Methods (RSM) is optimization, finding the best set of factor levels to achieve some goal.*



Example

Suppose, yield (y) of a chemical process depends on temperature (x_1) and pressure (x_2). The chemical engineer would like to find out which levels of temperature and pressure give the maximum yield.

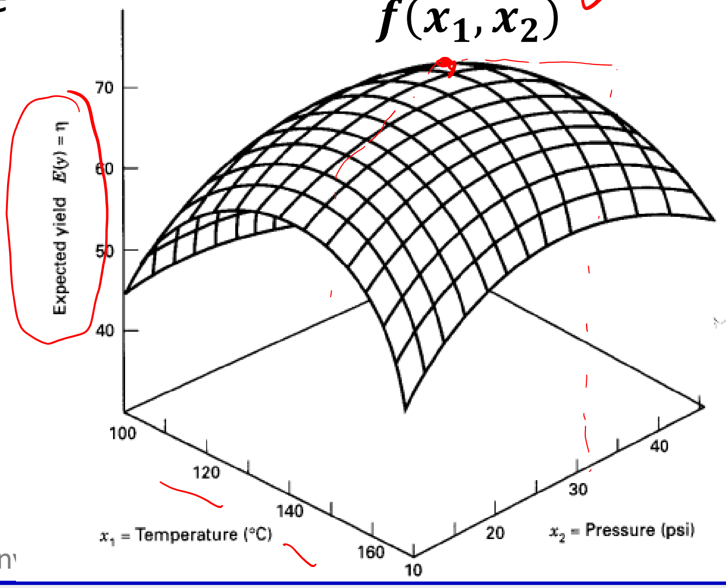
One may write, $$\boldsymbol{y = f(x_1, x_2) + \epsilon}$$

Where $'\epsilon'$ is the error/noise observed in response 'y'

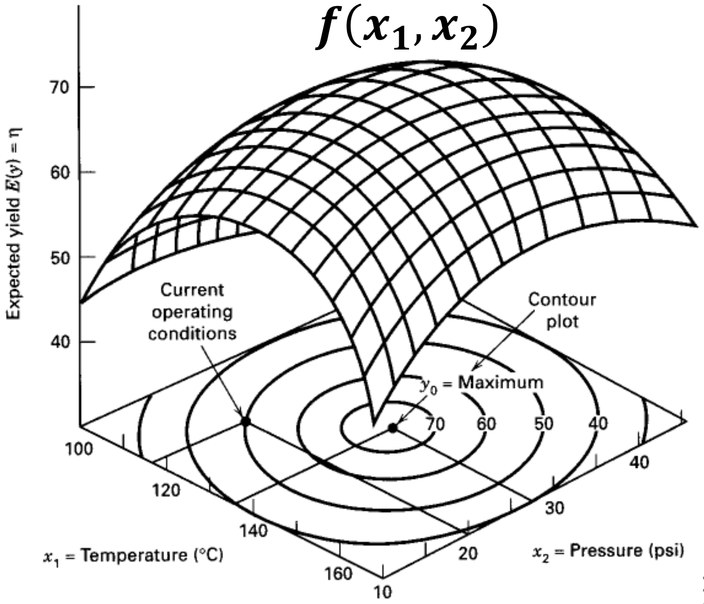
The expected value of the response 'y' will be **E(y) = $\boldsymbol{f(x_1, x_2)}$**

One could show this graphically,

$\boldsymbol{f(x_1, x_2)}$



$\boldsymbol{f(x_1, x_2)}$



Sequential Process

'RSM' is sequential procedure

- In most problems, *the exact relationship between the response variable and the independent variables is unknown*
- Therefore, the first step in RSM is to *find a suitable approximation* of the true functional relationship between response and independent variables.
- Typically, the approximations are in the *form of low-order polynomials* in some region of independent variables

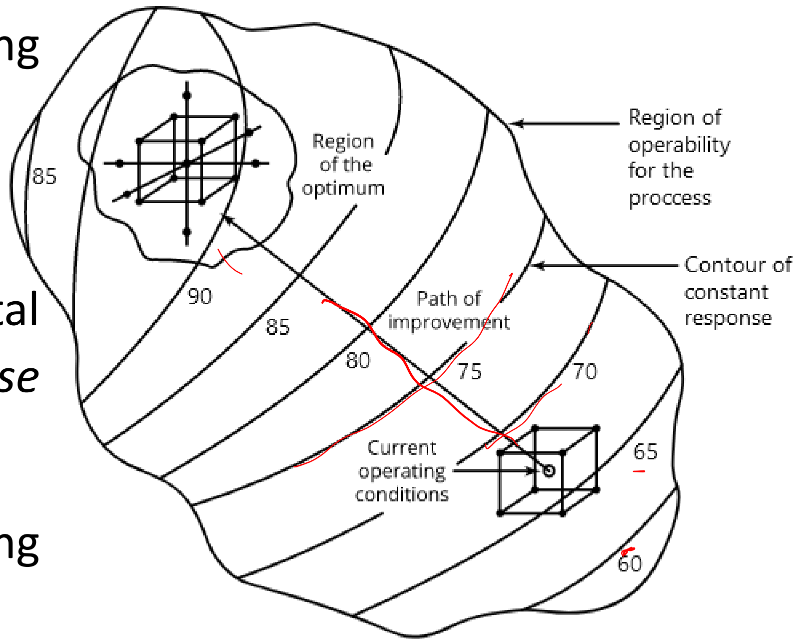
For example, if response (y) is well modeled by linear function of independent variables $(x_1, x_2, x_3, \dots, x_k)$, then we can write the approximate function as **'first order model'**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots . + \beta_k x_k + \epsilon$$

If there is curvature/non-linearity in the system, we must use polynomial of 2nd or higher degree,

For example, **second degree model** : $y = \beta_0 + \sum_{i=0}^{k} \beta_i x_i + \sum_{i=0}^{k} \beta_{ii} x_i^2 + \Sigma\Sigma\, \beta_{ij} x_i x_j + \epsilon$

- In real-problems, it is unlikely that these polynomials will provide reasonable approximation of the true functional relationship over the ENTIRE range of independent variables, but they work quite well for a relatively small region
- The coefficients in the RSM models (model parameters) are estimated using least square method (least square fitting)
- The response surface analysis is then performed on the fitted surface
- The model parameters can be obtained more effectively if proper experimental designs are used to collect the data (responses). *Designs for fitting the response surfaces are called response surface designs.*
- Often we start at a point that is far from optimum such as the existing operating conditions. If the region is linear, we use first order model.
- We then take the shortest and most efficient path towards the optimum
- As we near the optimum, there may be non-linearities, so we can employ higher order models



Method of Steepest Ascent

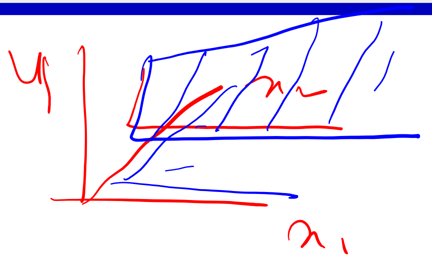
- If we want to find maximum response, then we will be 'climbing the hill', if we want to minimize the response, we will be 'descending into a valley'
- We then take the shortest and most efficient path towards the optimum
- 'Method of steepest ascent' is a procedure of moving sequentially alor path of steepest ascent, i.e., direction of the maximum increase i response.
- If minimization is desired, we follow the 'method steepest descent'
- If we use first order model,

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i$$

Then, the contours of y will be a set of parallel lines

So the path of steepest ascent will be along a line perpendicular to the cor from center of the region

The actual step-size will be dependent on other practical considerations



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

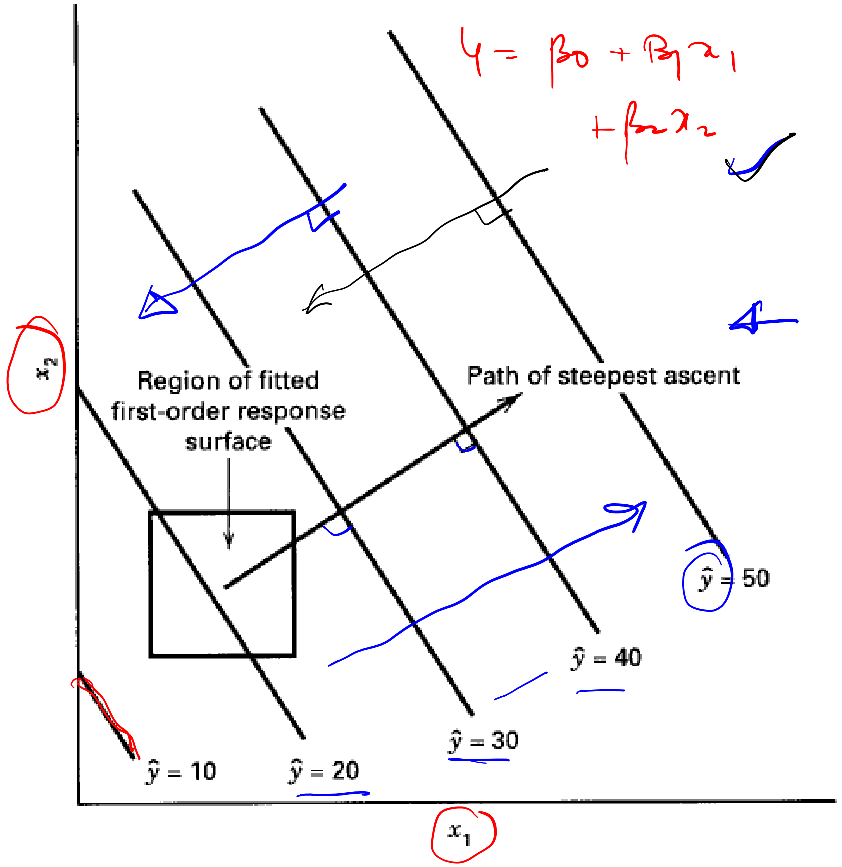
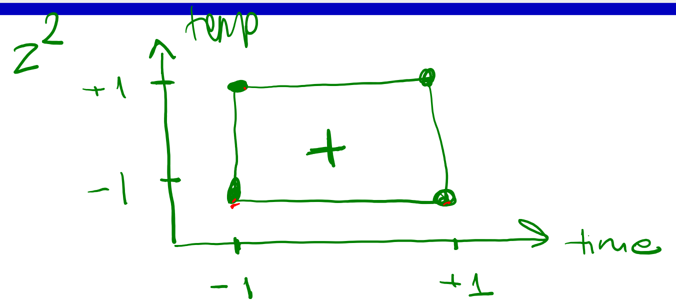


Figure 11-4 First-order response surface and path of steepest ascent.

Example

A chemical engineer is interested in determining the operating conditions that maximize the yield of a process. Two controllable variables influence process yield: reaction time and reaction temperature. The engineer is currently operating the process with a reaction time of 35 minutes and a temperature of 155°F, which result in yields of around 40 percent.

ξ_1 reaction time ξ_2 temperature



Region of (30, 40) minutes of time, and (150, 160) F temperature was explored and responses were collected.

Note the experimental design is 2^2 factorial design augmented by five center points. 5 replications at the center point [35, 155] allow estimation of error as well as help us determine adequacy of linear (first-order) model

$$x_1 = \frac{\xi_1 - 35}{5} \quad \text{and} \quad x_2 = \frac{\xi_2 - 155}{5}$$

Time, Temp

2^2

Natural Variables		Coded Variables		Response
ξ_1	ξ_2	x_1	x_2	y
30	150	−1	−1	39.3
30	160	−1	1	40.0
40	150	1	−1	40.9
40	160	1	1	41.5
35	155	0	0	40.3
35	155	0	0	40.5
35	155	0	0	40.7
35	155	0	0	40.2
35	155	0	0	40.6

2^2 Factorial Analysis

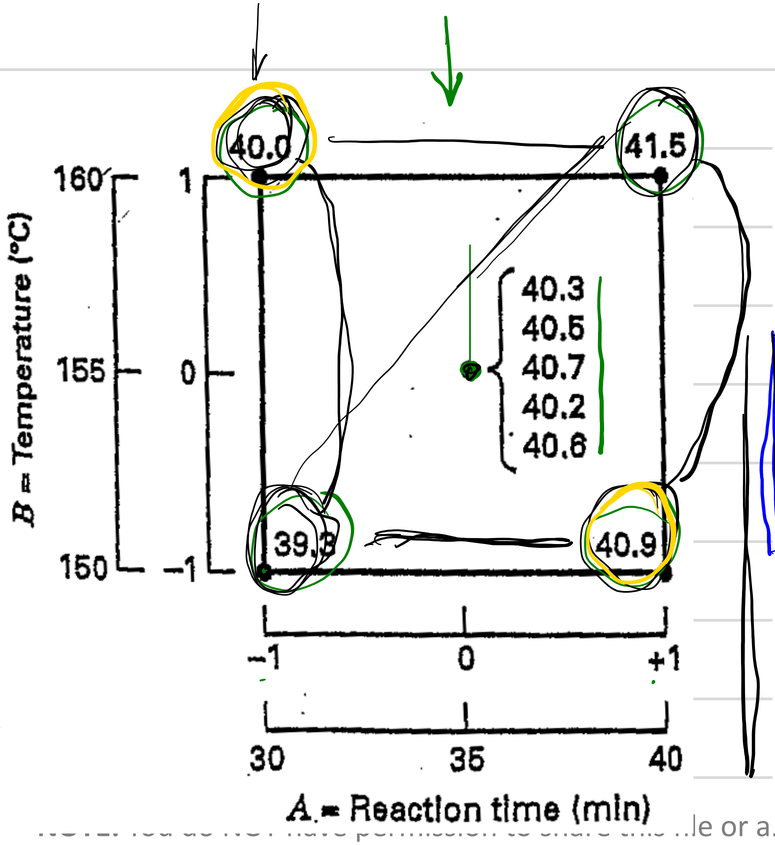
Can we find which terms are important?

What will be the first-order model?

Will a first-order model be appropriate?

First order model, $\underset{\text{Obs}}{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$ 0

predicted, $\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2}$



Natural Variables		Coded Variables		Response
ξ_1	ξ_2	x_1	x_2	y
30	150	−1	−1	39.3
30	160	−1	1	40.0
40	150	1	−1	40.9
40	160	1	1	41.5
35	155	0	0	40.3
35	155	0	0	40.5
35	155	0	0	40.7
35	155	0	0	40.2
35	155	0	0	40.6

First let's find out (DIY)

$\beta_0, \beta_1, \beta_2, \beta_{12}$

$\beta_0 = \left(\frac{39.3 + 40 + 40.9 + 41.5}{4}\right)$

$= 40.425$

$\beta_1 = \frac{1}{2}(-39.3 - 40.0 + 40.9 + 41.5)$

$= 1.55$

$\beta_2 = \frac{1}{2}(-39.3 + 40 - 40.9 + 41.5)$

$= 0.65$

$\beta_{12} = -0.05$

ANOVA

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \quad ?$$

can we reduce this to a 1-order model

$$+ (\beta_{11} x_1^2 + \beta_{22} x_2^2)$$

SS_T , SS_{mean}

ANOVA TABLE

	DF	SS	MS	F_0
Total	9			
x_1	1	2.4025		55.87
x_2	1	0.4225		9.83
$x_1 x_2$	1	0.0025		0.06
mean	1			
ε	5-1	0.172		
and	1	?		

$$SS_{x_1} = 2\left((39.65 - 40.425)^2 + (41.2 - 40.425)^2\right)$$

$$SS_{x_2} = 2\left((40.1 - 40.425)^2 + (40.75 - 40.425)^2\right)$$

$$SS_{x_1 x_2} = 2\left((40.45 - 40.425)^2 + (40.9 - 40.2425)^2\right)$$

$$= 0.0025$$

$$\varepsilon = (y_1 - \bar{y}_c)^2 + (y_2 - \bar{y}_c)^2 + \cdots + (y_5 - \bar{y}_c)^2$$

$$= (40.3 - 40.46)^2 + (\quad) + \cdots + (40.6 - 40.46)^2$$

$$=$$

ANOVA

Find $SS_{quad} = \frac{n_F n_c (\bar{y}_F - \bar{y}_c)^2}{n_F + n_c} = \frac{4 \times 5 (40.425 - 40.46)^2}{9}$

Another check of the adequacy of the straight-line model is obtained by applying the check for pure quadratic curvature effect described in Section 6-6. Recall that this consists of comparing the average response at the four points in the factorial portion of the design, say $\bar{y}_F = 40.425$, with the average response at the design center, say $\bar{y}_C = 40.46$. If there is quadratic curvature in the true response function, then $\bar{y}_F - \bar{y}_C$ is a measure of this curvature. If β_{11} and β_{22} are the coefficients of the "pure quadratic" terms x_1^2 and x_2^2, then $\bar{y}_F - \bar{y}_C$ is an estimate of $\beta_{11} + \beta_{22}$. In our example, an estimate of the pure quadratic term is

$$\begin{aligned}\hat{\beta}_{11} + \hat{\beta}_{22} &= \bar{y}_F - \bar{y}_C \\ &= 40.425 - 40.46 \\ &= -0.035\end{aligned}$$

‘Climbing the hill’

$\frac{0.325}{0.775} = 0.42$

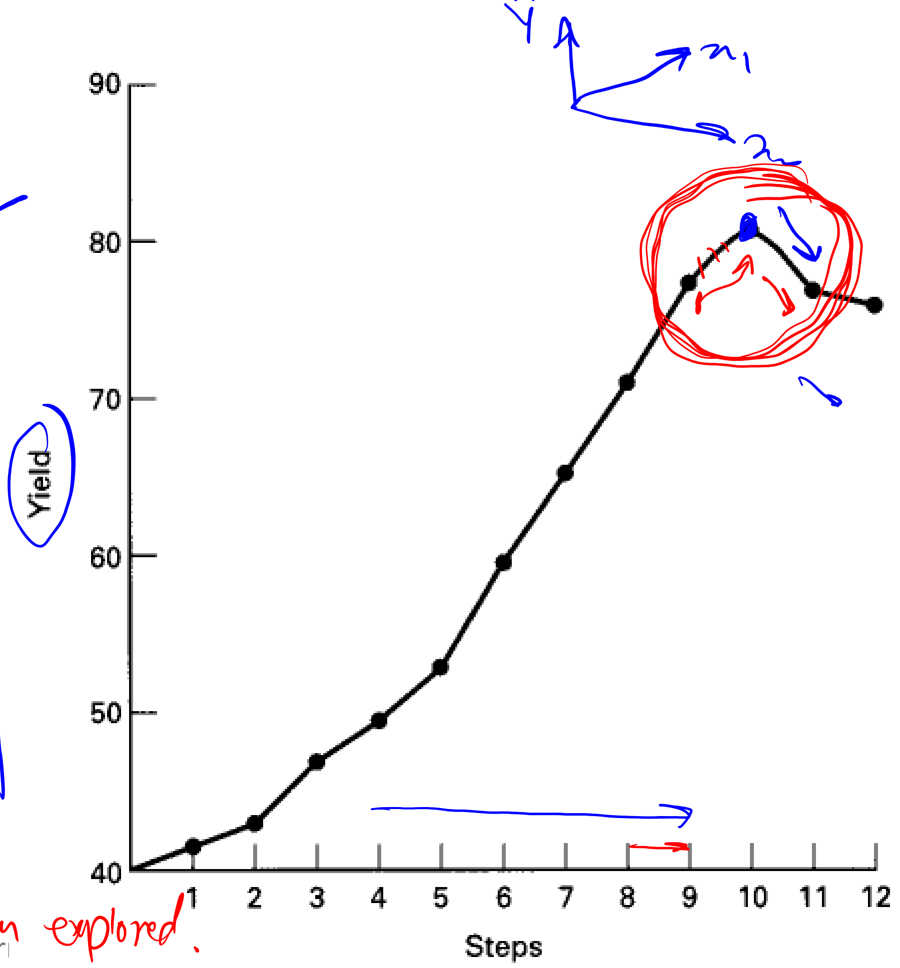
$$\hat{y} = 40.44 + 0.775x_1 + 0.325x_2$$

Only applicable in region explored.

note that $\hat{y} = f(x_1, x_2)$ is a plane

Table 11-3 Steepest Ascent Experiment for Example 11-1

Steps	Coded Variables x_1	Coded Variables x_2	Natural Variables ξ_1	Natural Variables ξ_2	Response y
Origin	0	0	35	155	40.44
Δ	1.00	0.42	5	2	
Origin + Δ	1.00	0.42	40	157	41.0
Origin + 2Δ	2.00	0.84	45	159	42.9
Origin + 3Δ	3.00	1.26	50	161	47.1
Origin + 4Δ	4.00	1.68	55	163	49.7
Origin + 5Δ	5.00	2.10	60	165	53.8
Origin + 6Δ	6.00	2.52	65	167	59.9
Origin + 7Δ	7.00	2.94	70	169	65.0
Origin + 8Δ	8.00	3.36	75	171	70.4
Origin + 9Δ	9.00	3.78	80	173	77.6
Origin + 10Δ	10.00	4.20	85	175	80.3
Origin + 11Δ	11.00	4.62	90	179	76.2
Origin + 12Δ	12.00	5.04	95	181	75.1



New model needs to be employed around [85, 175]

these are outside the region you explored.