Sample Mean($\bar{y}$)

A sample mean($\bar{y}$) of a random sample of n observations $y_1, y_2, y_3, \cdots, y_n$ is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Note:

1) $\bar{y} \neq \mu$ (Population mean).

2) But $E(\bar{y}) = \mu$.

Variance(σ^2)

$$\text{Var}(y) = \sigma^2 = E(y-\mu)^2$$
$$= E(y^2) - \mu^2.$$

Sample variance

$$S^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1}$$

NOTE: 1) $S^2 \neq \sigma^2$.

2) $E(S^2) = \sigma^2$.

Sample Co-variance:

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$$

person	Height(x)	weight (y in pounds)	$x_i y_i$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	69	153	10557	6.00	136.89
2	74	175	12950	6.50	106.09
3	68	155	10540	11.90	94.09
4	72	135	9450	2.10	882.09
5	72	172	12384	0.30	53.29
6	67	150	10050	19.80	216.09
7	66	115	7590	29.70	2470.09
8	70	137	9570	2.90	767.29
9	76	200	15200	20.70	1246.09
10	68	130	8840	11.90	1204.09
11	72	140	10080	0.30	610.09
12	79	265	20935	57.00	10062.09
13	74	185	13690	6.50	412.09
14	67	112	7504	19.80	2777.29
15	66	140	9240	29.70	610.09
16	71	150	10650	0.20	216.09
17	74	165	12210	6.50	0.09
18	75	185	13875	12.60	412.09
19	75	210	15750	12.60	2052.09
20	76	220	16720	20.70	3058.09
	$\sum x = 1429$	$\sum y = 3294$	$\sum x_i y_i = 237805$	$\sum (x_i - \bar{x})^2 = 276.9$	$\sum (y_i - \bar{y})^2 = 27,384$
	$\bar{x} = 71.64$	$\bar{y} = 164.7$			

Covariance :

$$\delta_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = 128.87894736842$$

NOTE : Covariance depends on units of variables.
i.e if units are changes from inches to feet and pounds to kgs in the previous example, covariance (δ_{xy}) will also change

Correlation Co-efficient for Sample.

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$S_{xy} = \sqrt{S_{xx}}, \quad S_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Correlation

Population Co-efficient of x & y:

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Note:

① r_{xy} is independent of units and is a constant $-1 \le r_{xy} \le 1$.

② If $r_{xy} = 1 \Rightarrow$ The variable x & y are perfectly co-related.

– If $r_{xy} = -1 \Rightarrow$ Then also x & y are perfectly correlated but in opposite direction.
i.e.. one ↑ other ↓.

– If $r_{xy} = 0 \Rightarrow$ There is no relation b/w x & y i.e.. orthogonal.

For the adj example:

$$S_x^2 = \frac{276.9}{19} = 14.573.$$

$$S_x = 3.8175.$$

$$S_y^2 = 1441.2631.$$

$$S_y = 37.9639.$$

$$r_{xy} = \frac{128.88}{(3.8175)(37.9639)} = 0.889276135243 8.$$

Covariance Matrix:

The Sample Covariance matrix $S = (S_{jk}) = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$ is the matrix of Sample variations and Covariances of P-Variables.

NOTE: It's a Symmetric Matrix Since $S_{jk} = S_{kj}$.

Ex:

Location No.	y_1 (Available Soil Ca)	y_2 (exchanged soil Ca)	y_3 (Turnip green Ca)
1.	35	3.5	2.8
2.	35	4.9	2.7
3.	40	30	4.38
4.	10	2.8	3.21
5.	6	2.7	2.73
6.	20	2.8	2.81
7.	35	4.6	2.88
8.	35	10.9	2.9
9.	35	8	3.28
10.	30	1.6	3.2
	$\sum y_1 = 281$	$\sum y_2 = 71.8$	$\sum y_3 = 30.89$

Find $S = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ \cdot & S_{22} & S_{23} \\ \cdot & \cdot & S_{33} \end{bmatrix}$

$\bar{y}_1 = 28.1$

$\bar{y}_2 = 7.18$

$\bar{y}_3 = 3.089$

$$S_{11} = S_1^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_{1i}^2 - n\bar{y}_1^2\right)$$

$$= \frac{1}{9}\left(9161 - 10(789.61)\right) = \frac{1}{9}(1264.9) = 140.545.$$

$$S_{12} = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_{1i}y_{2i} - n\bar{y}_1\bar{y}_2\right)$$

$= 49.68.$

$S_{13} = 1.941$

$S_{22} = 72.25$

$S_{23} = 3.68$

$S_{33} = 0.25.$

$$S = \begin{bmatrix} 140.54 & 49.68 & 1.941 \\ 49.68 & 72.25 & 3.68 \\ 1.941 & 3.68 & 0.25 \end{bmatrix}$$

Correlation Matrix:

The random Correlation matrix b/w the j^{th} & k^{th} variables is defined as

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} S_{kk}}} = \frac{S_{jk}}{S_j \cdot S_k}$$

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

where $r_{jj} = 1$; $j = 1, 2, 3, \ldots, p$.

$$\equiv R = \begin{bmatrix} 1 & r_{12} & \cdots \\ r_{21} & 1 & \cdots \\ \vdots & & \end{bmatrix}$$

→ Conversion of Correlation matrix to Covariance:

define diagonal matrix, $D_S = \text{diag}(\sqrt{S_{11}}, \sqrt{S_{22}}, \ldots, \sqrt{S_{pp}})$

$$= \text{diag}(S_1, S_2, \ldots, S_p)$$

$$= \begin{bmatrix} S_1 & 0 & \cdots & 0 \\ 0 & S_2 & \cdots & 0 \\ \vdots & 0 & & \vdots \\ 0 & 0 & \cdots & S_p \end{bmatrix}$$

Note: $\boxed{R = D_S^{-1} S D_S^{-1}}$ (or)

$\boxed{S = D_S R D_S}$ where D_S^{-1} = inverse of D_S.

For Previous example,

$$R = \begin{bmatrix} 1 & 0.596 & 0.8 \\ 0.596 & 1 & 0.86 \\ 0.8 & 0.86 & 1 \end{bmatrix}$$

Q. Find D_S given $S = \begin{bmatrix} 140.56 & 69.68 & 1.941 \\ 69.68 & 72.25 & 3.68 \\ 1.941 & 3.68 & 0.25 \end{bmatrix}$.

$$D_S = \begin{bmatrix} 11.8551 & 0 & 0 \\ 0 & 8.4999 & 0 \\ 0 & 0 & 0.5001 \end{bmatrix}$$

$$|D_S| = 11.8551 (8.4999 \times 0.5001)$$
$$= 50.3936589761449.$$

$$\text{adj}(D_S) = \begin{bmatrix} 4.25 & 0 & 0 \\ 0 & 5.92873 & 0 \\ 0 & 0 & 100.7671 \end{bmatrix}$$

$$D_S^{-1} = \frac{\text{adj}(D_S)}{|D_S|} = \begin{bmatrix} 0.084 & 0 & 0 \\ 0 & 0.1176 & 0 \\ 0 & 0 & 1.9996 \end{bmatrix}$$

$$\text{adj}(D_S) = \begin{bmatrix} 4.25 & 0 & 0 \\ 0 & 5.92873 & 0 \\ 0 & 0 & 100.7671 \end{bmatrix}$$

$$R = D_S^{-1} S D_S^{-1} = \begin{bmatrix} 0.084 & 0 & 0 \\ 0 & 0.1176 & 0 \\ 0 & 0 & 1.9996 \end{bmatrix} \begin{bmatrix} 140.54 & 49.68 & 1.961 \\ 49.68 & 72.25 & 3.68 \\ 1.961 & 3.68 & 0.25 \end{bmatrix} \begin{bmatrix} 0.084 & 0 & 0 \\ 0 & 0.1176 & 0 \\ 0 & 0 & 1.9996 \end{bmatrix}$$

$$= \begin{bmatrix} 11.80536 & 4.17312 & 0.163044 \\ 5.842368 & 8.4966 & 0.632768 \\ 3.8812236 & 7.358528 & 0.4999 \end{bmatrix} \begin{bmatrix} 0.084 & 0 & 0 \\ 0 & 0.1176 & 0 \\ 0 & 0 & 1.9996 \end{bmatrix}$$

$$\simeq \begin{bmatrix} 0.9916 & 0.490758912 & 0.3260227824 \\ 0.490758912 & 0.9992006 & 0.8653628928 \\ 0.3260227824 & 0.8653628928 & 0.999600064 \end{bmatrix}$$

$$\simeq \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}$$

Note:

#Mean Vectors and Co-Variance matrices for subsets of Variables:

Sometimes a ~~resthe~~ researcher is intrested in two diff kinds of variables of variables, both measured on same sampling unit.

Let us denote two subvectors by y & x, with p variables in y & q variables in x.

Thus each observation vector in a sample is partitioned as

$$\begin{pmatrix} y_i \\ x_i \end{pmatrix} = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{ip} \\ x_{i1} \\ \vdots \\ x_{iq} \end{pmatrix}, \quad i = 1, 2, \cdots, n.$$

Hence, there are $p+q$ variables in each of n observations vectors.

For the sample of n observation vectors, the mean vector and the covariance matrix have the form.

$$\begin{pmatrix} \bar{y} \\ \bar{x} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_p \\ \bar{x}_1 \\ \vdots \\ \bar{x}_q \end{pmatrix} ; S = \begin{pmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{pmatrix}$$

where $[S_{yy}]_{p\times p}$, $[S_{yx}]_{p\times q}$, $[S_{xy}]_{q\times p}$, $[S_{xx}]_{q\times q}$

$\rightarrow S'_{yx} = S_{xy}$, where S'_{yx} is the transpose of S_{yx}.

Let, $p=2$ & $q=3$. Then,

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} ; x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$S_{yy} = \begin{pmatrix} s_1^2 & s_{12} & s_{13} \\ s_{21} & s_2^2 & s_{23} \\ s_{31} & s_{32} & s_3^2 \end{pmatrix}.$$

$$S_{yy} = \begin{pmatrix} s_{y_1}^2 & s_{y_1y_2} \\ s_{y_2y_1} & s_{y_2}^2 \end{pmatrix}_{2\times 2} ; S_{yx} = \begin{pmatrix} s_{y_1x_1} & s_{y_1x_2} & s_{y_1x_3} \\ s_{y_2x_1} & s_{y_2x_2} & s_{y_2x_3} \end{pmatrix}_{2\times 3}$$

$$= S'_{xy}.$$

$$S_{xx} = \begin{pmatrix} s_{x_1}^2 & s_{x_1x_2} & s_{x_1x_3} \\ s_{x_2x_1} & s_{x_2}^2 & s_{x_2x_3} \\ s_{x_3x_1} & s_{x_3x_2} & s_{x_3}^2 \end{pmatrix}_{3\times 3}$$

Ex: Research of Miller measured five variables in a comparision of normal Patients and diabetics.

x_1 = glucose tolerance, x_2 = insulin response to oral glucose,

x_3 = insulin resistance, y_1 = relative weight, y_2 = fasting plasma glucose.

$$S = \left[\begin{array}{cc|ccc} 0.0162 & 0.216 & 0.7872 & -0.2138 & 2.189 \\ 0.216 & 70.56 & 26.23 & -23.96 & -20.84 \\ \hline 0.7872 & 26.23 & 1106 & 396.7 & 108.4 \\ -0.2138 & -23.96 & 396.7 & 2382 & 1143 \\ 2.189 & -20.84 & 108.4 & 1143 & 2136 \end{array}\right]_{5\times5}$$

$$S_{xx} = \begin{bmatrix} 1106 & 396.7 & 108.4 \\ 396.7 & 2382 & 1143 \\ 108.4 & 1143 & 2136 \end{bmatrix}_{3\times3} \qquad S_{yx} = \begin{bmatrix} 0.7872 & -0.2138 & 2.189 \\ 26.23 & -23.96 & -20.84 \end{bmatrix}_{2\times3}$$

$$S_{yy} = \begin{bmatrix} 0.0162 & 0.216 \\ 0.216 & 70.56 \end{bmatrix}_{2\times2} \qquad S_{xy} = \begin{bmatrix} 0.7872 & 26.23 \\ -0.2138 & -23.96 \\ 2.189 & -20.84 \end{bmatrix}_{3\times2}$$

Dependent Variable (D.V) & Independent Variable (I.V):

Covariates:

Q) Patient no.

	y_1	y_2	x_1	x_2	x_3	y_1^2	y_2^2	y_1y_2	y_2x_1	x_1^2	x_2^2	x_3^2	x_1x_2
1	0.81	80	356	124	55								
2	0.95	97	289	117	76								
3	0.94	105	319	143	105								
4	1.04	90	356	199	108								
5	1	90	323	240	143								

$$S_{y_1}^2 = \frac{1}{n-1}\left(\Sigma y_1^2 - n\bar{y}_1^2\right)$$

$$\begin{bmatrix} S_{y_1}^2 & S_{y_1y_2} & S_{y_1x_1} & S_{y_1x_2} & S_{y_1x_3} \\ S_{y_2y_1} & S_{y_2}^2 & S_{y_2x_1} & S_{y_2x_2} & S_{y_2x_3} \\ S_{y_3y_1} & S_{y_3y_2} & S_{y_3x_1} & S_{y_3x_2} & S_{y_3x_3} \\ S_{y_4y_1} & S_{y_4y_2} & S_{y_4x_1} & S_{y_4x_2} & S_{y_4x_3} \\ S_{y_5y_1} & S_{y_5y_2} & S_{y_5x_1} & S_{y_5x_2} & S_{y_5x_3} \end{bmatrix}$$

Ex: Research of Miller measured five variables in a comparision of normal Patients and diabetics.

x_1 = glucose tolerance, x_2 = insulin response to oral glucose,

x_3 = insulin resistance, y_1 = relative weight, y_2 = fasting plasma glucose.

$$S = \left[\begin{array}{cc|ccc} 0.0162 & 0.216 & 0.7872 & -0.2138 & 2.189 \\ 0.216 & 70.56 & 26.23 & -23.96 & -20.84 \\ \hline 0.7872 & 26.23 & 1106 & 396.7 & 108.4 \\ -0.2138 & -23.96 & 396.7 & 2382 & 1143 \\ 2.189 & -20.84 & 108.4 & 1143 & 2136 \end{array}\right]_{5\times5}$$

$$S_{xx} = \begin{bmatrix} 1106 & 396.7 & 108.4 \\ 396.7 & 2382 & 1143 \\ 108.4 & 1143 & 2136 \end{bmatrix}_{3\times3} \qquad S_{yx} = \begin{bmatrix} 0.7872 & -0.2138 & 2.189 \\ 26.23 & -23.96 & -20.84 \end{bmatrix}_{2\times3}$$

$$S_{yy} = \begin{bmatrix} 0.0162 & 0.216 \\ 0.216 & 70.56 \end{bmatrix}_{2\times2} \qquad S_{xy} = \begin{bmatrix} 0.7872 & 26.23 \\ -0.2138 & -23.96 \\ 2.189 & -20.84 \end{bmatrix}_{3\times2}$$

Dependent Variable (D.V) & Independent Variable (I.V):

Covariates:

Patient no.	y_1	y_2	x_1	x_2	x_3	y_1^2	y_2^2	y_1y_2	y_2x_1	x_1^2	x_2^2	x_3^2	x_1x_2
1	0.81	80	356	124	55								
2	0.95	97	289	117	76								
3	0.94	105	319	143	105								
4	1.04	90	356	199	108								
5	1	90	323	240	143								

$$S_{y_1}^2 = \frac{1}{n-1}\left(\sum y_1^2 - n\bar{y}_1^2\right)$$

$$\begin{bmatrix} S_{y_1}^2 & S_{y_1y_2} & S_{y_1x_1} & S_{y_1x_2} & S_{y_1x_3} \\ S_{y_2y_1} & S_{y_2}^2 & S_{y_2x_1} & S_{y_2x_2} & S_{y_2x_3} \\ S_{y_3y_1} & S_{y_3y_2} & S_{y_3x_1} & S_{y_3x_2} & S_{y_3x_3} \\ S_{y_4y_1} & S_{y_4y_2} & S_{y_4x_1} & S_{y_4x_2} & S_{y_4x_3} \\ S_{y_5y_1} & S_{y_5y_2} & S_{y_5x_1} & S_{y_5x_2} & S_{y_5x_3} \end{bmatrix}$$

Linear Combination of Variables:

Let, z be the linear combination of the variables $y_1, y_2, y_3, \cdots, y_p$.

$$z = \sum_{i=1}^{p} a_i y_i = a_1 y_1 + a_2 y_2 + \cdots + a_p y_p.$$

Let, $a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} \Rightarrow a' = (a_1, a_2, \cdots, a_p)$.

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}$$

$$\Rightarrow \boxed{z = a'y}.$$

If the Sample Co-efficient vector a is applied to each y_i in a sample, we have-

$$z_i = a_1 y_{i1} + a_2 y_{i2} + a_3 y_{i3} + \cdots + a_p y_{ip}.$$

$$= a' y_i, \; i = 1, 2, \cdots, n.$$

$$z_1 = a_1 y_{11} + a_2 y_{12} + \cdots + a_p y_{1p} = a' \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1p} \end{pmatrix} = a' y_1.$$

$$z_2 = a_1 y_{21} + a_2 y_{22} + \cdots + a_p y_{2p} = a' y_2.$$

$$z_3 = a_1 y_{31} + a_2 y_{32} + \cdots + a_p y_{3p} = a' y_3.$$

$$\vdots$$

$$z_n = a_1 y_{n1} + \cdots + a_p y_{np} = a' y_n.$$

Sample mean of z will be:

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i = \frac{1}{n} \sum_{i=1}^{n} a' y_i = a' \frac{1}{n} \sum_{i=1}^{n} y_i = a' \bar{y}.$$

$$\boxed{\bar{z} = a' \bar{y}}.$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

Similarly, the Sample Variance of $z_i = a' y_i, \; i = 1, 2, \cdots, n$, can be found as the Sample Variance of $z_1, z_2, \cdots, z_p$ or directly from a and S, where S is the Sample Covariance matrix of $y_1, y_2, y_3, \cdots, y_n$.

$$S_z^2 = \frac{\sum_{i=1}^{n} (z_i - \bar{z})^2}{n-1} = a' S a$$

$$\boxed{S_z^2 = a' S a}.$$

Note: Since a Variable is always non-negative, $S_z^2 \geq 0$ and therefore $a'Sa \geq 0$, for every a. Hence S is atleast positive semidefinite.

Let, $w \neq z$, is a new linear combination of y_i

$$w = b_1 y_1 + b_2 y_2 + \cdots + b_p y_p.$$

$\boxed{w = b'y}$, where $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$, $b' = (b_1, b_2, \cdots, b_p)$.

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}$$

Then Sample Co-Variance of z and w is given by

$$S_{zw} = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(w_i - \bar{w})}{n-1} = \boxed{a'Sb = S_{zw}}$$

The Sample Correlation between z and w is

$$r_{zw} = \frac{S_{zw}}{\sqrt{S_z^2 S_w^2}} = \boxed{\frac{a'Sb}{\sqrt{(a'Sa)(b'Sb)}} = r_{zw}}$$

Ex: Timm reported the results of an experiment where subjects responded to "Probe words" at five positions in a sentence. The variables are response times for the j^{th} probe word, y_j, $j = 1, 2, \cdots, 5$.

Subject No.	y_1	y_2	y_3	y_4	y_5
1	51	36	50	35	42
2	27	20	26	17	27
3	37	22	41	37	30
4	42	36	32	34	27
5	27	18	33	14	29
6	43	32	43	35	40
7	41	22	36	25	38
8	38	21	31	20	16
9	36	23	27	25	28
10	26	31	31	32	36
11	29	20	25	26	25
Total:	397	281	375	300	338

$$z = 3y_1 - 2y_2 + 4y_3 - y_4 + y_5$$

Sol: Method-1:

$z = a'y$ $\quad a' = (3, -2, 4, -1, 1)$

$\bar{z} = a'\bar{y}$

$$\bar{y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \bar{y}_4 \\ \bar{y}_5 \end{pmatrix} = \begin{pmatrix} 36.09 \\ 25.54 \\ 36.09 \\ 27.27 \\ 30.73 \end{pmatrix}$$

$$\bar{z} = (3, -2, 4, -1, 1)_{1\times5} \begin{pmatrix} 36.09 \\ 25.54 \\ 36.09 \\ 27.27 \\ 30.73 \end{pmatrix}_{5\times1}$$

$\bar{z} = (108.27 - 51.08 + 136.36 - 27.27 + 30.73)$

$\bar{z} = 197.01$

Method-2:

$z_1 = 3(51) + (-2)(36) + 4(50) - 35 + 42 = 288$;

$z_2 = 3(27) + (-2)(20) + 4(26) - 17 + 27 = 155$.

$z_3 = 224$, $z_4 = 175$, $z_5 = 192$, $z_6 = 242$, $z_7 = 236$, $z_8 = 192$, $z_9 = 173$,

$z_{10} = 164$, $z_{11} = 146$.

$$\bar{z} = \frac{z_1 + z_2 + \cdots + z_{11}}{11} = \frac{2167}{11}$$

$\bar{z} \approx 197.01$

$$S_z^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (z_i - \bar{z})^2$$

$$S_z^2 = a' S a$$

$$S = \begin{pmatrix} 65.09 & 33.65 & 47.59 & 36.77 & 25.43 \\ 33.65 & 46.07 & 28.95 & 40.34 & 28.36 \\ 47.59 & 28.95 & 60.69 & 37.37 & 41.13 \\ 36.77 & 40.34 & 37.37 & 62.82 & 31.68 \\ 25.43 & 28.36 & 41.13 & 31.68 & 58.22 \end{pmatrix}$$

$$W = y_1 + 3y_2 - y_3 + y_4 - 2y_5$$

$$\bar{w} = b'\bar{y} = 44.45$$

$$S_w^2 = b'Sb = 605.67$$

Covariance: $S_{zw} = a'Sb$

$$\frac{1}{(n-1)} \sum_{i=1}^{n} (z_i - \bar{z})(w_i - \bar{w})$$

Correlation: $\dfrac{S_{zw}}{\sqrt{S_w^2 S_z^2}} = 0.0358.$

let, $a = a_1$, $b = a_2$

Define: $A = \begin{pmatrix} a_1' \\ a_2' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1p} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2p} \end{pmatrix}_{2 \times p}$

and $z = \begin{pmatrix} a_1' y \\ a_2' y \end{pmatrix} = \begin{pmatrix} a_1' \\ a_2' \end{pmatrix} y = Ay.$

$$\boxed{z = Ay.}$$

$z_i = Ay_i \; ; \; i = 1, 2, \cdots, n$

$$\bar{z} = \begin{pmatrix} \bar{z}_1 \\ \bar{z}_2 \end{pmatrix} = \begin{pmatrix} a_1' \bar{y} \\ a_2' \bar{y} \end{pmatrix} = \begin{pmatrix} a_1' \\ a_2' \end{pmatrix} \bar{y}$$

$$\boxed{\bar{z} = A\bar{y}}$$

Sample Covariance matrix for z

$$S_z = \begin{pmatrix} S_{z_1}^2 & S_{z_1 z_2} \\ S_{z_2 z_1} & S_{z_2}^2 \end{pmatrix}$$

$$S_z = \begin{pmatrix} a_1' S a_1 & a_1' S a_2 \\ a_2' S a_1 & a_2' S a_2 \end{pmatrix}$$

$$S_z = \begin{pmatrix} a_1' \\ a_2' \end{pmatrix} S \begin{pmatrix} a_1 & a_2 \end{pmatrix}$$

$$\boxed{S_z = ASA'.}$$

let, k linear transformations/combinations are expressed as —

$$z_1 = a_{11} y_1 + a_{12} y_2 + a_{13} y_3 + \cdots + a_{1p} y_p = a_1' y$$

$$z_2 = a_{21} y_1 + a_{22} y_2 + \cdots + a_{2p} y_p = a_2' y$$

$$\vdots$$

$$z_k = a_{k1} y_1 + a_{k2} y_2 + \cdots + a_{kp} y_p = a_k' y$$

Or in Matrix:

$$\underset{z}{\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{pmatrix}_{k\times 1}} = \underset{A}{\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & & & \\ a_{k1} & a_{k2} & \cdots & a_{kp} \end{pmatrix}_{k\times p}} \underset{y}{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}_{p\times 1}}$$

$$= \begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_k' \end{pmatrix} y = a'y$$

$$\boxed{z = Ay.}$$

If $z_i = Ay_i$, $i = 1, 2, \ldots, n$.

The sample mean vector $\bar{z}$ is

$$\bar{z} = \begin{pmatrix} a_1'\bar{y} \\ a_2'\bar{y} \\ \vdots \\ a_k'\bar{y} \end{pmatrix} = \begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_k' \end{pmatrix} \bar{y}$$

$$\boxed{\bar{z} = A\bar{y}.}$$

Sample Covariance matrix

$$S_z = \begin{pmatrix} a_1'Sa_1 & a_1'Sa_2 & \cdots & a_1'Sa_k \\ a_2'Sa_1 & a_2'Sa_2 & \cdots & a_2'Sa_k \\ \vdots & \vdots & & \vdots \\ a_k'Sa_1 & a_k'Sa_2 & \cdots & a_k'Sa_k \end{pmatrix}_{k\times k}$$

$$S_z = \begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_k' \end{pmatrix} S \begin{pmatrix} a_1 & a_2 & \cdots & a_k \end{pmatrix}$$

$$\boxed{S_z = ASA'.}$$

NOTE:

① $tr(ASA') = \sum_{i=1}^{k} a_i' S a_i$

② A more general linear transformation

$$z_i = Ay_i + b \quad \text{and} \quad S_z = ASA'$$

$$\boxed{\bar{z} = A\bar{y} + b}$$

a) $z_1 = y_1 + y_2 + y_3 + y_4 + y_5$

$z_2 = 2y_1 - 3y_2 + y_3 - 2y_4 - y_5$

$z_3 = -y_1 - 2y_2 + y_3 - 2y_4 + 3y_5$

These linear combinations in matrix form can be written as.

$$z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & -3 & 1 & -2 & -1 \\ -1 & -2 & 1 & -2 & 3 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}}_{y}$$

$$\bar{y} = \begin{pmatrix} 36.09 \\ 25.55 \\ 36.09 \\ 27.27 \\ 30.73 \end{pmatrix} \quad \text{(Previous ex)}$$

$$\bar{z} = A\bar{y}$$

$$\bar{z} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & -3 & 1 & -2 & -1 \\ -1 & -2 & 1 & -2 & 3 \end{pmatrix} \begin{pmatrix} 36.09 \\ 25.55 \\ 36.09 \\ 27.27 \\ 30.73 \end{pmatrix} = \begin{pmatrix} 153.73 \\ -55.65 \\ -5.45 \end{pmatrix}$$

NOTE:

③ Generated Sample Variance $= |S|$.

④ Total Sample Variance $= tr(S) = S_{11} + S_{22} + \cdots + S_{pp}$

Exercise-4

$$S_z = A S A'$$

$$R_z = D_z^{-1} S_z D_z^{-1} \qquad \text{where } D_z = \left(\right.$$

1Q) y_1: available soil Calcium.

y_2: exchangeable soil Calcium.

y_3: turnip green Calcium.

Loc. no	y_1	y_2	y_3
1	35	3.5	2.8
2	35	4.9	2.7
3	40	30	4.38
⋮	⋮	⋮	⋮
10	30	1.6	3.2

Data given in one of the recent examples.

Define: $z = 3y_1 - y_2 + 2y_3$

Find $\bar{z}$ and s_z^2 in two ways.

(a) Direct.

(b) Use $\bar{z} = a'\bar{y}$.

and $s_z^2 = a'Sa$.

2Q) Define $w = -2y_1 + 3y_2 + y_3$

Find r_{zw} in two ways.

(a) Direct.

(b) Matrix.

3Q) $z_1 = y_1 + y_2 + y_3$

$z_2 = 2y_1 - 3y_2 + 2y_3$

$z_3 = -y_1 - 2y_2 - 3y_3$

(a) Find $\bar{z}$ and S_z.

(b) Find R_z.

Regression Analysis

Marks (Y) = α Hours (X_1) + β Lectures (X_2)

- Marks (Y): D.V, Response Var
- Hours (X_1), Lectures (X_2): Independent Var, Input/Predictor var.

#Linear regression:

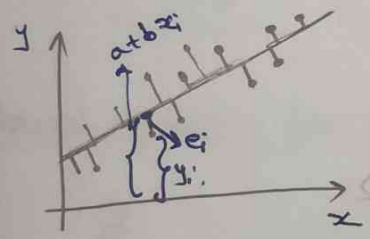
(x_i, y_i) , $i = 1, \dots, n$.

x = I.V, Predictor var, Input var.

y = D.V, Response Var.

Ex:

x	0	1	2	3	4	4	5	6
y	25	20	30	40	45	50	60	50



$y = a + bx$

Residuals: $e_i = |y_i - (a + bx_i)|$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - a - bx_i]^2.$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

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

Note: In above notations, there's no '$n-1$'.

The least square line (regression line).

$\hat{y} = \hat{\alpha} + \hat{\beta} x$ (Eq^n of regression line y on x).

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Residual Sum of squares / Error Sum of Squares:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

Ex①: $\hat{y} = \hat{\alpha} + \hat{\beta} x = 22 + 6x$

$SSE = 270.$

Ex②: x = width, y = height

$n = 50, \bar{x} = 88.36, \bar{y} = 305.58$

$S_{xx} = 7239.22, \quad S_{xy} = 17840.1, \quad S_{yy} = 66976.2.$

(a) Find the least square line for predicting height from width.

(b) " " " " width from height.

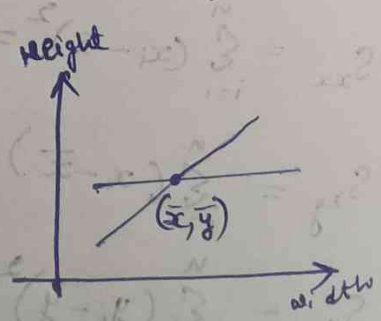
(c) Make a scatter plot & show both lines.

(a) $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{17840.1}{7239.22} = 2.464$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 305.58 - 2.464 \times 88.36$

$\hat{\alpha} = 88.26$

$\hat{y} = \hat{\alpha} + \hat{\beta} x = 88.26 + 2.464 x$

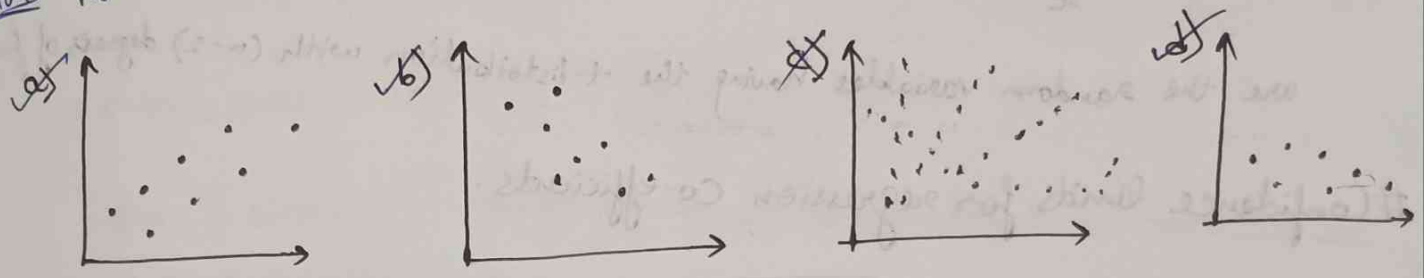


Method of Least Square

The procedure of finding the equation of regression line that best fits the set of paired data is called "method of least square".

Q Can we always get best fitting line for a set of paired data?

Ans No.



Standard error of the estimate (S_e)

$$S_e^2 = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{(n-2)} = \frac{SSE}{n-2}$$

Estimate of σ^2

$$E(\sigma^2) = S_e^2$$

Normal equation for Least Squares estimates

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Differentiating partially w.r.t 'a' & 'b' and equating to zero, we get

w.r.t a

$$2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i$$

w.r.t b

$$2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

$$y_i = a + bx_i$$

$a \Rightarrow \hat{\alpha}$

$b \Rightarrow \hat{\beta}$

$$\Rightarrow \hat{y} = \hat{\alpha} + \hat{\beta}x.$$

t-statistics

$$t = \frac{(\hat{\alpha} - \alpha)}{Se} \sqrt{\frac{n S_{xx}}{S_{xx} + n(\bar{x})^2}}$$

and

$$t = \frac{(\hat{\beta} - \beta)}{Se} \sqrt{S_{xx}}$$

are the random variables having the t-distribution with $(n-2)$ degrees of freedom

Confidence limits for regression co-efficients.

$$\alpha = \hat{\alpha} \pm t_{\alpha/2} \cdot Se \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}} \qquad -t_{\alpha/2} < t < t_{\alpha/2}$$

and

$$\beta = \hat{\beta} \pm t_{\alpha/2} \, Se \frac{1}{\sqrt{S_{xx}}}$$

Ex: $n = 10$, $S_{xx} = 132000$, $S_{xy} = 505.4$, $\bar{x} = 200$, $S_{yy} = 2.13745$, $\bar{y} = 0.835$.

Construct a 95% confidence interval for the regression coefficient 'α'.

Sol^{y} $$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$= 2.13745 - \frac{(505.4)^2}{132000}$$

$$= 2.13745 - 1.93506 = 0.20239.$$

For 95% Confidence Intervals, the level of significance is $\alpha = 0.05$.

$$t_{\alpha/2, 8} = t_{0.025, 8} = 2.306.$$

$$Se^2 = \frac{SSE}{n-2} = \frac{0.20239}{8} = 0.0252$$

$$Se = 0.1587;$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{505.4}{132000} = 0.0038$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 0.835 - 0.0038 \times 200$$

$$= 0.835 - 0.765$$

$$= 0.07.$$

$$\alpha = \hat{\alpha} \pm t_{\alpha/2}\, Se \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

$$= 0.07 \pm 0.05 \times 0.1587 \sqrt{\frac{1}{10} + \frac{(200)^2}{132006}}$$

$$= 0.07 \pm 0.0079 \sqrt{0.1 + 0.30}$$

$$= 0.07 \pm 0.0079 \times 0.634$$

$$= 0.07 \pm 0.05$$

$\beta =$

Q) Test the null hypothesis $\beta = 0$ against the alternate hypothesis $\beta \neq 0$ at the 0.05 ~~boot~~ level of significance.

Given, $\hat{\beta} = 0.003829$

$Se = 0.1531$

Sol

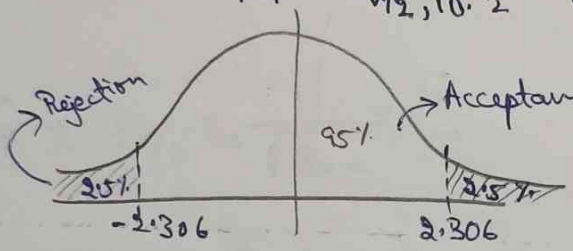
① Null hypothesis : $\beta = 0$

Alternate hypothesis : $\beta \neq 0$.

② Level of Significance : $\alpha = 0.05$.

③ Criterion : Reject the null hypothesis if

$|t| > t_{\alpha/2, 10.2} = 2.306$.



④ Calculation:

$$t = \frac{(\hat{\beta} - \beta)}{Se} \sqrt{S_{xx}}$$

$$= \frac{(0.03829 - 0)}{0.1591} \sqrt{132000} = 8.746.$$

$\because 8.744 > t_{\alpha/2, 8} = 2.306$

⑤ Decision: Null hypothesis must be rejected.

Given (x_i, y_i)

Transformation,

$y_i \longrightarrow \log y_i, \sqrt{y_i}, \frac{1}{y_i}$

$x_i \longrightarrow \log x_i$

1) Exponential type

$y = \alpha \beta^x$

$\boxed{\log y = \log \alpha + x \log \beta}$ linear in x & $\log y$

Ex: y = battery capacity (amp - hre)
x = rate of discharge (amp)

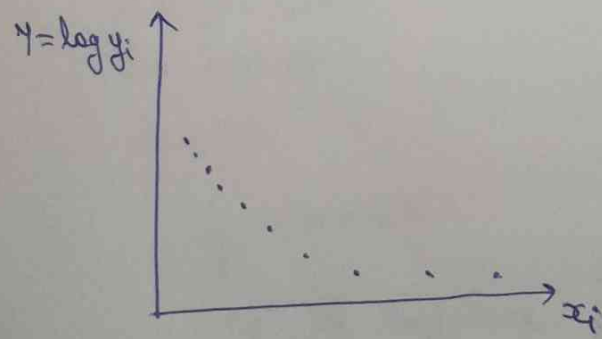
Rate of discharge (A) x	Capacity (Ah) y	$Y = \log y$
2	164.7	2.21
3	156.1	2.19
6	142.5	2.15
10	133.8	2.12
15	116.6	2.05
20	107.1	2.02

$\Sigma Y = 12.74$

OR

$\Sigma Y = \ln y = 29.62$

a) Graph $(x_i, \log y_i)$



b) Fit an exponential curve by applying the method of least square to the data points $(x_i, \log y_i)$.

$$\bar{x} = \frac{56}{6} = 9.3333$$

$$\bar{Y} = \overline{\ln y} = \frac{29.42552}{6} = 4.9043$$

$$S_{xy} = S_{x \log y} = 268.6762 - 29.42552\left(\frac{56}{6}\right)$$

$$= -5.961987.$$

$$S_{xx} = 774 - \frac{(56)^2}{6} = 251.3333$$

Then the estimated slope

$$\frac{S_{xy}}{S_{xx}} = -0.02372$$

and the estimated intercept is

$$4.9043 + 9.3333(0.02372) = 5.1257$$

$$Y = \text{Intercept} + \text{Slope}.x$$

$$\ln y = 5.1257 - 0.02372x$$

$$\Rightarrow \text{predicted } y = \exp(5.1257 - 0.02372x)$$

$$\hat{y} = e^{5.1257} \cdot e^{-0.2372x}$$

$$\hat{y} = 169 \cdot e^{-0.2372x}$$

Q) solve normal eqn and find the values for the least square to estimate.

$8.35 = 10a + 2000b$
$21754 = 2000a + 5320000b$ } $\hat{\alpha} = ?$, $\hat{\beta} = ?$

$y_i = a + bx_i$

$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i$ ——— ①

$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$.

$\hat{\alpha} = 0.069$
$\hat{\beta} = 0.08383$

Q)

x	0	1	2	2	4	4	5	6
y_i	1/25	1/20	1/30	1/40	1/80	1/50	1/60	1/50
$1/y_i$	25	20	30	40	45	50	60	50
$x_i y_i$	0	20	60	80	180	200	300	300
x_i^2	0	1	4	4	16	16	25	36

$24/8 = 3$.
$320/8 = 40$.
1140.
102.

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$

Eqn of regression line y on x.

$\hat{y} = \hat{\alpha} + \hat{\beta}x$

$S_{xx} = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 = 102 - 8(9) \Rightarrow 102 - 72 = \underline{30}$.

$S_{xy} = \sum_{i=1}^{n} x_i y_i - n(\bar{x})(\bar{y}) = 1140 - 960 \Rightarrow \underline{180}$.

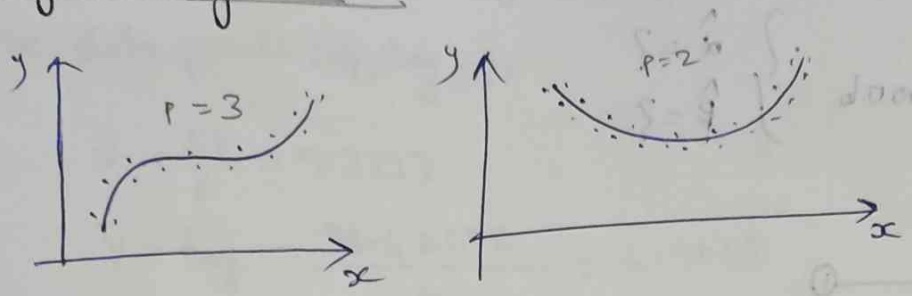
$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{180}{30} \Rightarrow \underline{6}$.

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
$\hat{\alpha} = 40 - 6(3)$
$\hat{\alpha} = 40 - 18$
$\hat{\alpha} = 22$

$\frac{1}{y} = a + bx$

eqn of reg. line: $\boxed{\frac{1}{y} = 22 + 6x}$.

#Polynomial Regression:



The general equation of path degree polynomial.

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_p x^p \quad \text{—①}$$

Given a set of data consisting of n-points (x_i, y_i), we estimate the co-efficients $b_0, b_1, b_2, \cdots, b_p$ of the p^{th} degree polynomial by minimizing

$$\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_p x_i^p)]^2$$

Taking the partial derivatives w.r.t $b_0, b_1, b_2, \cdots, b_p$ equating these partial derivatives to zero.

from ①,

$$\sum_{i=1}^{n} y = nb_0 + b_1 \Sigma x_i + b \cdots + b_p \Sigma x_i^p$$

$$\Sigma xy = b_0 \Sigma x_i + b_1 \Sigma x_i^2 + \cdots + b_p \Sigma x_i^{p+1}$$

$$\vdots$$

$$\Sigma x^p y = b_0 \Sigma x_i^p + b_1 \Sigma x_i^{p+1} + \cdots + b_p \Sigma x_i^{2p}$$

NOTE:

$$\Sigma = \sum_{i=1}^{n}, \quad \Sigma xy = \sum_{i=1}^{n} x_i y_i, \quad \Sigma x^p = \sum_{i=1}^{n} x_i^p$$

$$\Sigma x^p y = \sum_{i=1}^{n} x_i^p y_i$$

The eqⁿ of best fitting curve will be given by

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \cdots + \hat{\beta}_p x^p$$

where $\hat{\beta}_0 = b_0$ (Best estimated)
$\hat{\beta}_1 = b_1$
$\vdots \quad \vdots$
$\hat{\beta}_p = b_p$

Ex: The following are data on the drying time of a certain varnish and the amount of an additive that is intended to reduce the drying time.

Amount of varnish additive x (grams)	Drying time y (hours)
0	12
1	10.5
2	10
3	8
4	7
5	8
6	7.5
7	8.5
8	9

(a) Find a second degree polynomial.

(b) Predict the drying time of varnish when 6.5 g of the addive is added

$y = b_0 + b_1 x + b_2 x^2$ [General 2nd-degree polynomial].

The Normal eq's are:

$$\Sigma y = n b_0 + b_1 \Sigma x + b_2 \Sigma x^2$$

$$\Sigma xy = b_0 \Sigma x + b_1 \Sigma x^2 + b_2 \Sigma x^3$$

$$\Sigma x^2 y = b_0 \Sigma x^2 + b_1 \Sigma x^3 + b_2 \Sigma x^4$$

(concatinating columns to above table):

x^2	x^3	x^4	xy	x^2y
0	0	0	0	0
1	1	1	10.5	10.5
4	8	16	20	40
9	27	81	24	72
16	64	256	28	112
25	125	625	40	200
36	216	1296	45	270
49	343	2401	59.5	416.5
64	512	4096	72	576

Now, the normal eq^{n}s will be

$80.5 = 9b_0 + 36b_1 + 204b_2$

$299 = 36b_0 + 204b_1 + 1296b_2$

$1697 = 204b_0 + 1296b_1 + 8772b_2$

$$D = \begin{bmatrix} 9 & 36 & 204 \\ 36 & 204 & 1296 \\ 204 & 1296 & 8772 \end{bmatrix}_{3\times3}$$

$$D_1 = \begin{bmatrix} 80.5 & 36 & 204 \\ 299 & 204 & 1296 \\ 1697 & 1296 & 8772 \end{bmatrix} \Rightarrow b_0 = \frac{|D_1|}{|D|}$$

$$D_2 = \begin{bmatrix} 9 & 80.5 & 204 \\ 36 & 299 & 1296 \\ 204 & 1697 & 8772 \end{bmatrix} \Rightarrow b_1 = \frac{|D_2|}{|D|}$$

$$D_3 = \begin{bmatrix} 9 & 36 & 80.5 \\ 36 & 204 & 299 \\ 204 & 1296 & 1697 \end{bmatrix} \Rightarrow b_2 = \frac{|D_3|}{|D|}$$

$$B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}_{3\times1}$$

$$C = \begin{bmatrix} 80.5 \\ 299 \\ 1697 \end{bmatrix}_{3\times1}$$

$DB = C$

$\Rightarrow \boxed{B = D^{-1}C}$

Augmented Matrix:

$$\left[\begin{array}{ccc|c} 9 & 36 & 204 & 80.5 \\ 36 & 204 & 1296 & 299 \\ 204 & 1296 & 8772 & 1697 \end{array}\right]$$

Alternate:

$R_2 \to 4R_1 - R_2$, $R_3 \to R_3 - \frac{204}{9} R_1$

$$\sim \left[\begin{array}{ccc|c} 9 & 36 & 204 & 80.5 \\ 0 & -60 & -680 & 23 \\ 0 & 480 & 4148 & -127.67 \end{array}\right]$$

$R_3 \to R_3 + 8R_2$, $R_2 \to -\frac{1}{60} R_2$

$$\sim \left[\begin{array}{ccc|c} 9 & 36 & 204 & 80.5 \\ 0 & 1 & 8 & -23/60 \\ 0 & 0 & 308 & 56.33 \end{array}\right]$$

$308 b_2 = 56.33$

$\Rightarrow \boxed{b_2 = 0.1828}$

$b_1 + 8b_2 = -23/60$

$\Rightarrow \boxed{b_1 = -1.845}$

$9b_0 + 36b_1 + 204 b_2 = 80.5$

$\Rightarrow \boxed{b_0 = 12.1848}$

∴ The eqⁿ of the least square polynomial is:

$$\boxed{\hat{y} = 12.1848 - 1.845x + 0.183x^2}$$

(b) $\hat{y}\,(x = 6\text{ o'clock}) = \boxed{7.9 \text{ hrs.}}$

Multiple Regression (Linear):

	y	x_1	x_2	. . . . .	x_r
1	y_1	x_{11}	x_{12}	. . . . .	x_{1r}
2	y_2	x_{21}	x_{22}		x_{2r}
.	.	.	.		.
.	.	.	.		.
n	y_n	x_{n1}	x_{n2}		x_{nr}

$n \times (r+1)$ tuples.

$i = 1, 2, 3, \ldots, r.$

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_r x_r$$

Sum of squares of the vertical distances from the observations y_i

$$\sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_1 + \ldots + b_r x_r) \right]^2$$

Let $r = 2$,

$$\sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_1 + b_2 x_2) \right]^2$$

$$y = b_0 + b_1 x_1 + b_2 x_2.$$

Normal equations

$$\Sigma y = n b_0 + b_1 \Sigma x_1 + b_2 \Sigma x_2.$$

$$\Sigma x_1 y = b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2.$$

$$\Sigma x_2 y = b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2.$$

$$\Sigma x_1 = \sum_{i=1}^{n} x_{i1}$$

$$\Sigma x_1 x_2 = \sum_{i=1}^{n} x_{i1} x_{i2}$$

$$\Sigma x_1 y = \sum_{i=1}^{n} x_{i1} y_i$$

The eqn of regression curve is

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

$\hat{\beta}_0 = b_0$, $\hat{\beta}_1 = b_1$, $\hat{\beta}_2 = b_2$

Ex: The following are data on the number of twists required to break a certain kind of forged alloy bar and the % of two alloying elements present in the metal:

no of twists y	% of element A x_1	% of element B x_2
41	1	5
49	2	5
69	3	5
65	4	5
40	1	10
50	2	10
58	3	10
57	4	10
31	1	15
36	2	15
44	3	15
57	4	15
19	1	20
31	2	20
33	3	20
43	4	20

Fit a least squares regression plane and use its equation to estimate the number of twists required to break one of bars when $x_1 = 2.5$ & $x_2 = 12$.

Sol^{n}: Using Computer. Software:

THE REGRESSION EQUATION IS.

$Y = 46.4 + 7.78 X1 - 1.65 X2$ ①

PREDICTOR	COEFF ②	STD.DEV	③ T-RATIO	P ⑦
CONSTANT	46.438	3.517	13.20	0.000
X_1	7.775	0.9485	8.20	0.000
X_2	-1.655	0.1897	-8.72	0.000

$S = 4.242$ ⑥ $R-SQ = 91.7\%$ ⑤

ANALYSIS OF VARIANCE

SOURCE	DF	④ SS	MS
REGRESSION	2	25.785	1289.3
ERROR	13	233.9	18.0
TOTAL	15	2812.4	

① The least squares regression plane $\hat{y} = 46.4 + 7.78 x_1 - 1.65 x_2$ This eq^{n} estimates that the average number of twists required to break a bar, increases by 7.78 if the % of element A is increased by 1% and x_2 remains fixed.

② The least squares estimates and their corresponding estimated error are:

$\hat{\beta}_0 = 46.438$ with estimated standard error 3.517.

$\hat{\beta}_1 = 7.775$ - - - - - - - 0.9485.

$\hat{\beta}_2 = -1.655$ - - - - - - - 0.1897.

③ The t-ratio 13.2, 8.2, -8.72 are all highly significant so all the terms are needed in the model.

④ $\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$

total sum of squares | error sum of squares | regression sum of squares

$2812.4 = 233.9 + 2578.5$

⑤ $R^2 = \frac{\text{regression SS}}{\text{total SS}} = \frac{2578.5}{2812.4} = 0.917.$

The estimate of σ^2 is

$s_e^2 = \frac{233.9}{13} = 18$

or $\boxed{s_e = 4.242}$

The ⑦ p-values confirm the significance of the t-ratios and thus the fact that all the terms are required in the model.

Alternate method:

Normal equations,

$723 = 16b_0 + 40b_1 + 200b_2$
$1963 = 40b_0 + 120b_1 + 500b_2$
$8210 = 200b_0 + 500b_1 + 3000b_2$

$\hat{\beta} = B^{-1}C.$

$\rightarrow \hat{\beta}_0 = 46.4,\ \hat{\beta}_1 = 7.78,\ \hat{\beta}_2 = -1.65$

$\therefore \hat{y} = 46.4 + 7.78x_1 - 1.65x_2$

Multiple Linear Regression (Matrix Notation):

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}_{n\times 3}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \text{ and } b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}_{3\times 1}$$

$\hat{\beta}$, the least squares estimates of the multiple regression coefficients.

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

where X' is the transpose of matrix X.

$(XX)^{-1}$ is the inverse of matrix $X'X$.

$$X'X = \begin{bmatrix} n & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_2 & \Sigma x_2 x_1 & \Sigma x_2^2 \end{bmatrix}_{3\times 3}$$

$$X'Xb = \begin{bmatrix} b_0 n + b_1 \Sigma x_1 + b_2 \Sigma x_2 \\ b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 \\ b_0 \Sigma x_2 + b_1 \Sigma x_2 x_1 + b_2 \Sigma x_2^2 \end{bmatrix}_{3\times 1}$$

$$X'y = \begin{bmatrix} \Sigma y \\ \Sigma x_1 y \\ \Sigma x_2 y \end{bmatrix}_{3\times 1}$$

$$\Rightarrow X'Xb = X'y$$

$$\underbrace{(X'X)^{-1}(X'X)}_{I} b = (X'X)^{-1} X'y$$

$$\Rightarrow \boxed{b = (X'X)^{-1} X'y} \quad \hat{\beta}$$

Ex: $\Sigma x_1 = 40$, $\Sigma x_2 = 200$, $\Sigma x_1^2 = 120$,

$\Sigma x_1 x_2 = 500$, $\Sigma x_2^2 = 3000$, $n = 16$

$\Sigma y = 723$, $\Sigma x_1 y = 1963$ & $\Sigma x_2 y = 8210$

Sol: Here,

$$X'X = \begin{bmatrix} 16 & 40 & 200 \\ 40 & 120 & 500 \\ 200 & 500 & 3000 \end{bmatrix}$$

$$X'X = \begin{bmatrix} n & \Sigma x_1 & \Sigma x_2 & \Sigma x_3 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 & \Sigma x_1 x_3 \\ \Sigma x_2 & \Sigma x_2 x_1 & \Sigma x_2^2 & \Sigma x_2 x_3 \\ \Sigma x_3 & \Sigma x_3 x_1 & \Sigma x_3 x_2 & \Sigma x_3^2 \end{bmatrix}$$

$$\begin{bmatrix} 110000 & -20000 & -400 \\ -20000 & 8000 & 0 \\ -4000 & 0 & 320 \end{bmatrix}$$

where $|X'X| = 1,60,000$

$$\beta = (x'x)^{-1} x'y = \begin{bmatrix} 46.4375 \\ 7.7750 \\ -1.6550 \end{bmatrix}$$

Residual sum of squares:

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}_{3\times3} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}_{3\times1}$$

or $\boxed{\hat{y} = X\hat{\beta}}_{3\times1}$

then residual sum of squares

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

$$= (y - X\hat{\beta})'_{1\times3} (y - X\hat{\beta})_{3\times1}$$

the estimate s_e^2 of σ^2 can be expressed as –

$$s_e^2 = \frac{1}{n-3} (y - X\hat{\beta})'(y - X\hat{\beta})$$

\# $X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & & & \\ \vdots & \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & \cdots & x_{nk} \end{bmatrix}$ $\beta = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$

$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ $\boxed{\hat{\beta} = (x'x)^{-1} x'y}$

$$s_e^2 = \frac{1}{(n-k-1)} (y - X\hat{\beta})'(y - X\hat{\beta})$$

df = n – no of β's = $n-(k+1)$