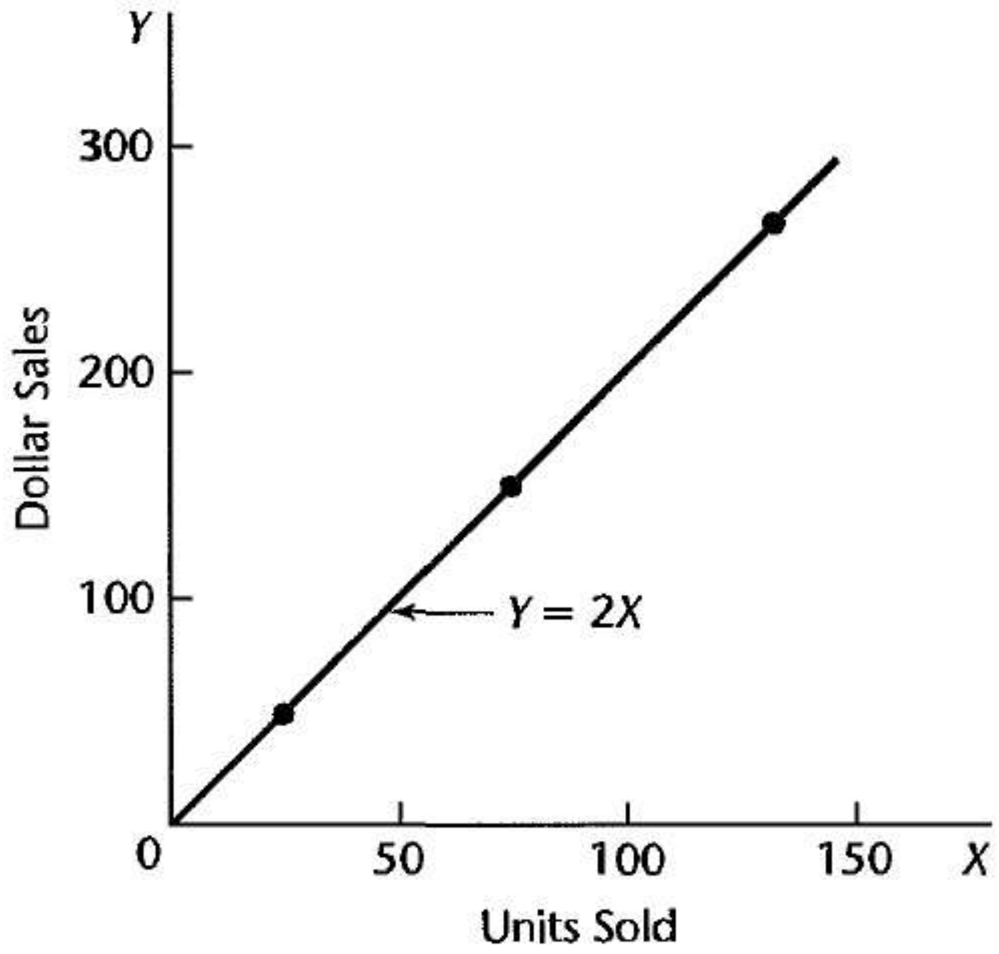
Machine Learning

Linear Regression

Samatrix.io

Functional Relationship – 2 Variables



$$Y = f(X)$$

Unit price $2

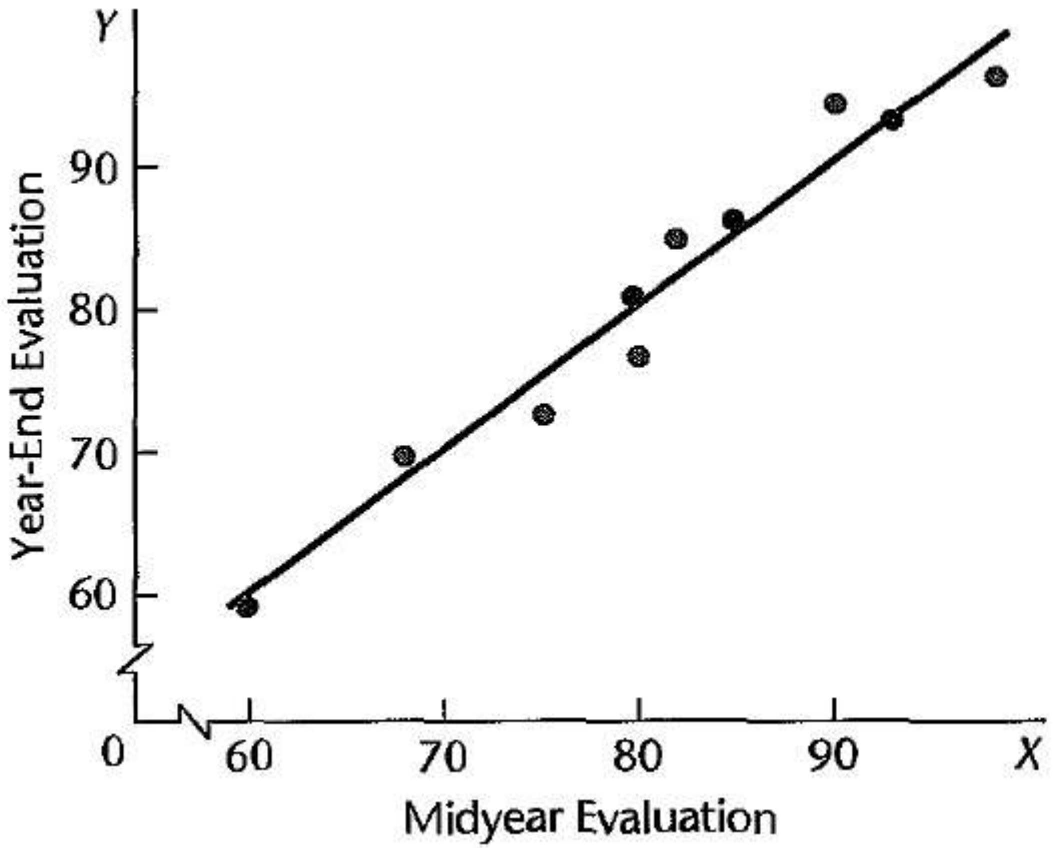
Period	# Units (X)	Sales $ (Y)
1	75	150
2	25	50
3	130	260

$$(X_1, Y_1) = (75, 150)$$
$$(X_2, Y_2) = (25, 50)$$
$$(X_3, Y_3) = (130, 260)$$

Function Relationship is perfect

Samatrix.io

Statistical Relationship – 2 Variables



Performance Evaluation of 10 employees at Mid Year and Year-End

There is a relationship but not perfect

For 2 Employees Mid Year Evaluation is same at $X = 80$ But different Year End Evaluation

Indicates general tendency by which Year End Tendency vary with Midyear Evaluation

Samatrix.io

Formal Statement of Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Y_i is the value of the response variable in the ith trial

β_0 and β_1 are parameters

X_i the value of the predictor variable in the ith trial

ϵ_i random error with mean $E\{\epsilon_i\} = 0$ and Variance $\sigma^2\{\epsilon_i\} = \sigma^2$

Project Statement

A certain spare part is manufactured by a company once a month in lots which vary in size as demand fluctuates. Data on lot size and number of man hours of labour for 10 production run performed under similar production conditions

First Trial $(X_1, Y_1) = (30, 73)$

$ith\ Trial\ (X_i, Y_i)\ where\ i = 1, \dots, n$

Production Run i	Lot Size X_i	Man-Hour Y_i
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

Analysis of Data

$n = 10, \bar{X} = \frac{500}{10} = 50, \bar{Y} = \frac{1100}{10} = 110$

$Run\ (i)$	$Lot\ Size\ (X_i)$	$ManHour\ (Y_i)$	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	30	73	-20	-37	400	1369
2	20	50	-30	-60	900	3600
3	60	128	10	18	100	324
4	80	170	30	60	900	3600
5	40	87	-10	-23	100	529
6	50	108	0	-2	0	4
7	60	135	10	25	100	625
8	30	69	-20	-41	400	1681
9	70	148	20	38	400	1444
10	60	132	10	22	100	484
Total	500	1100	0	0	3400	13660
	$\sum X_i$	$\sum Y_i$			$\sum (X_i - \bar{X})^2$	$\sum (Y_i - \bar{Y})^2$

Samatrix.io

Mean and Standard Deviation Lot Size X_i

Sample Size: $n = 10$

Degree of Freedom: $DF = n - 1 = 9$

Mean: $\bar{X} = \frac{\sum X_i}{n} = \frac{500}{10} = 50$

Variance: $Var = \frac{\sum(\boldsymbol{X_i} - \bar{\boldsymbol{X}})^2}{n-1} = \frac{3400}{9} = 377.77$

Standard Deviation :$s_X = \sqrt{Var} = \sqrt{377.77} = 19.436$

Samatrix.io

Mean and Standard Deviation Man-Hours Y_i

Sample Size: $n = 10$

Degree of Freedom: $DF = n - 1 = 9$

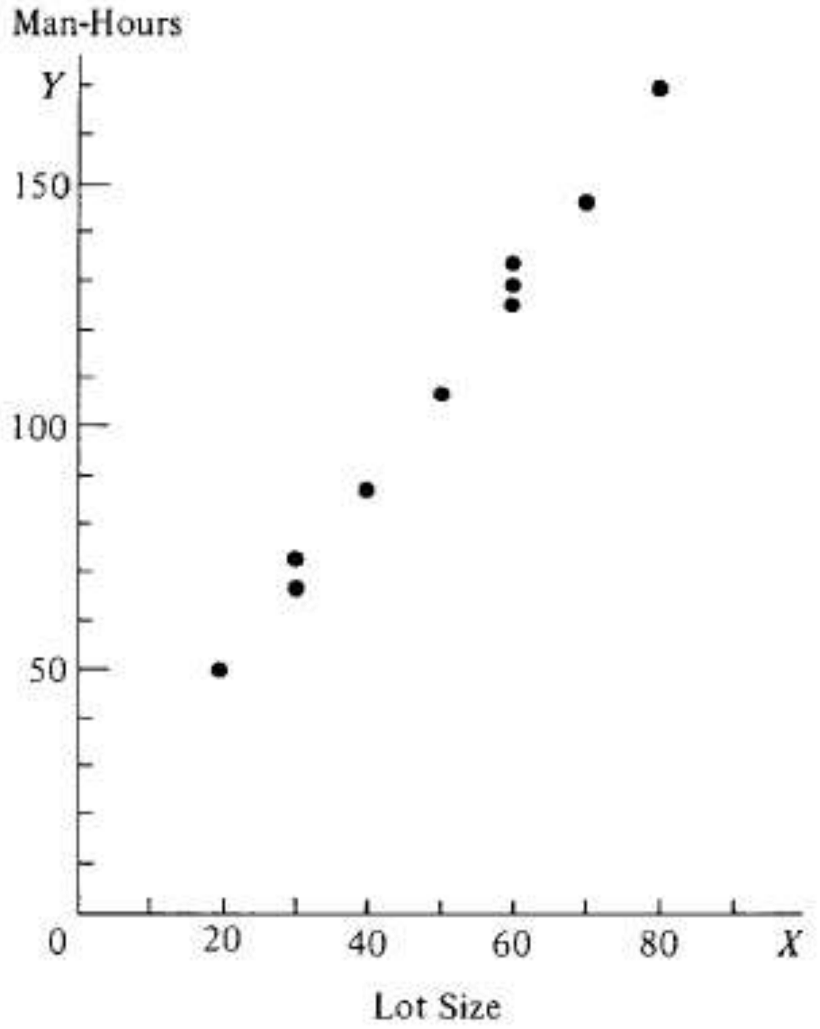
Mean: $\bar{Y} = \frac{\sum Y_i}{n} = \frac{1100}{10} = 110$

Variance: $Var = \frac{\sum(\boldsymbol{Y_i} - \bar{\boldsymbol{Y}})^2}{n-1} = \frac{13660}{9} = 1517.778$

Standard Deviation : $s_Y = \sqrt{Var} = \sqrt{1517.778} = 38.9587$

Samatrix.io

Scatter Diagram



Scatter Diagram or Scatter Plot suggests direct relationship between $lot\ size$ and $man - hour$

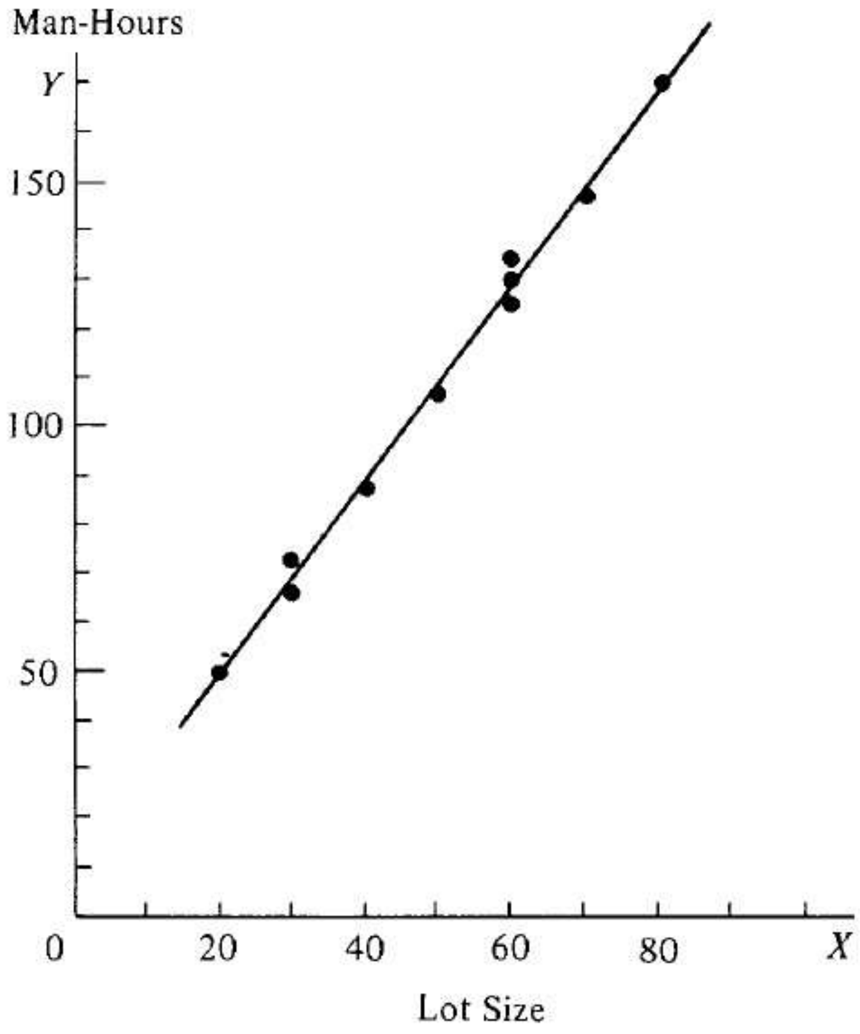
Larger $lot\ size$ needs more $man - hour$

Relation is not perfect

Production run 1 and 8 of 30 parts needs different $man - hour$

Each point in scatter diagram represents $Observation$ or $Trial$

Statistical Relationship



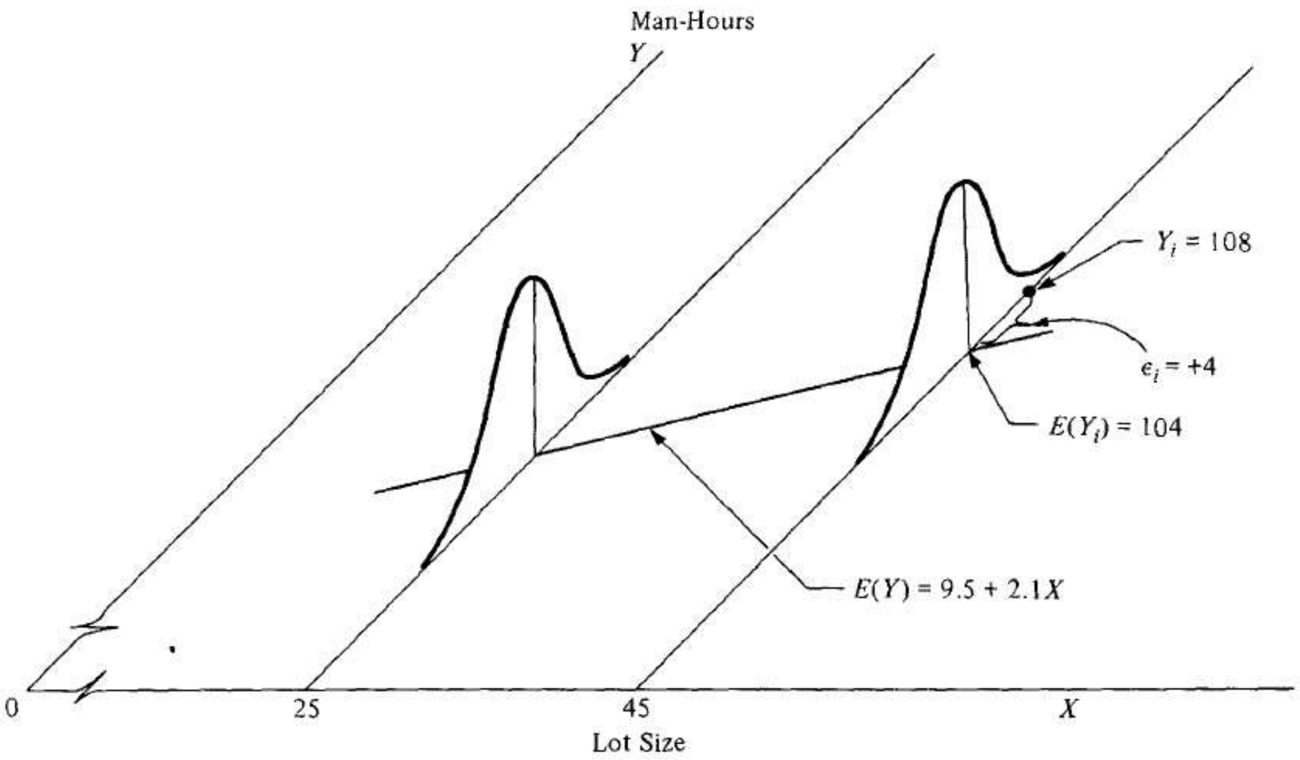
Line of Relationship describes statistical relationship between $lot\ size$ and $man - hour$

Shows general tendency by which $man - hour$ changes with $lot\ size$

Scattering of points around the line represents the variation in $man - hour$ which is not associated with $lot\ size$

Samatrix.io

Probability Distribution



Relation – Lot Size (X_i) and Required Man-Hours (Y_i)

$$Y_i = 9.5 + 2.1\,X_i + \epsilon_i$$

$$\hat{Y}_i = E\{Y_i\} = 9.5 + 2.1\,X_i$$

For $(X_i, Y_i) = (45, 108)$

$$E\{Y_i\} = 9.5 + 2.1 \times 45 = 104$$

$$Y_i = 104 + 4 = 108$$

Probability Distribution of Y when $X = 45$ indicates from where in the distribution $Y = 108$ comes

Samatrix.io

Method of Least Square

The objective of $method\ of\ Least\ Square$ is to find estimates $(\hat{\beta}_0\ and\ \hat{\beta}_1)$ for β_0 and β_1 for which the sum of n square deviations is minimum

For each sample observation (X_i, Y_i), the $method\ of\ Least\ Square$ consider the deviation of Y_i from its expected value

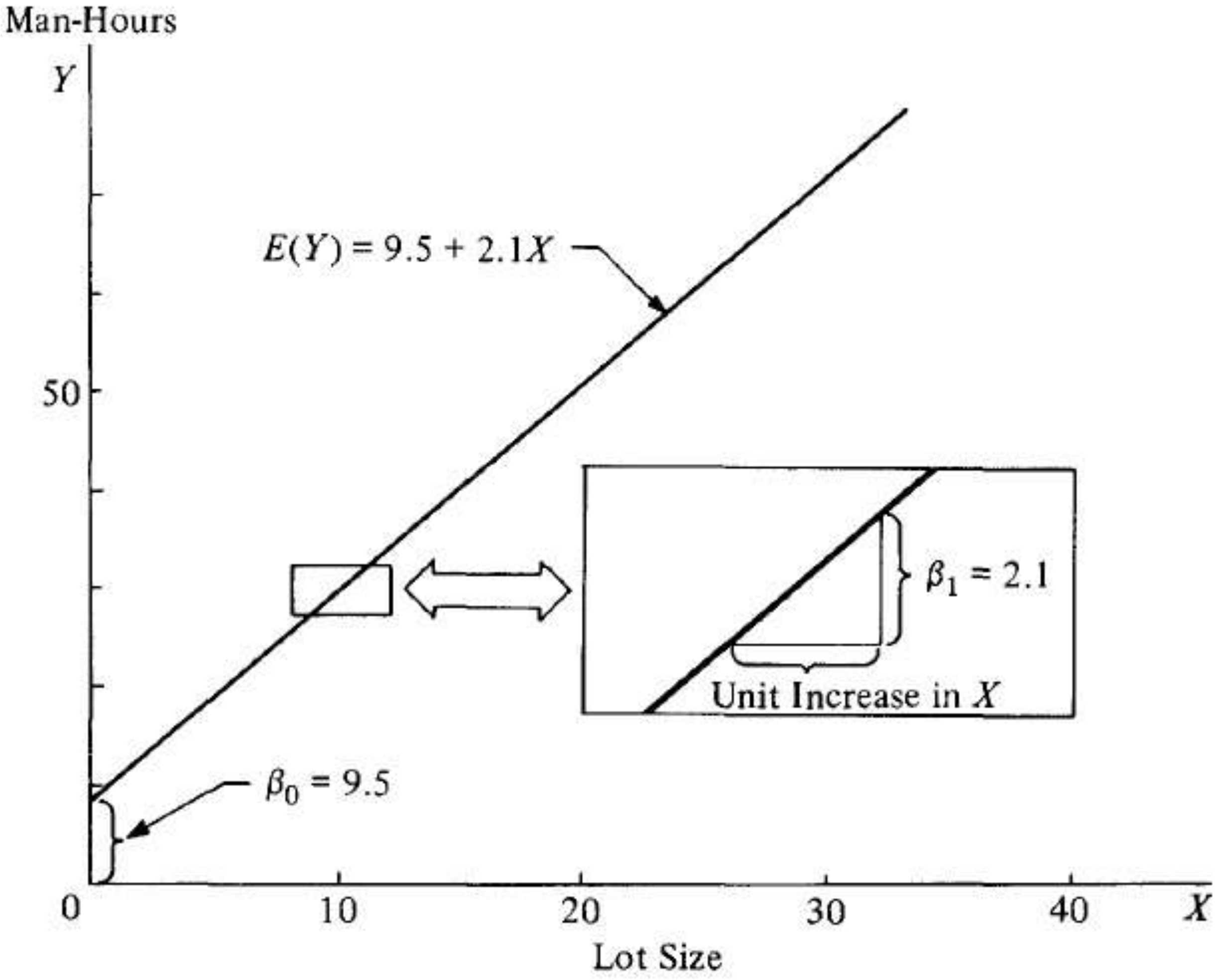
$$Y_i - (\beta_0 + \beta_1 X_i)$$

If Sum of Square Deviation is Q

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Samatrix.io

Linear Regression Coefficients



Parameters β_0 and β_1 are called *Regression Coefficients*

β_1 is slope of regression line

Indicates change in the mean of probability distribution of Y per unit increase in X

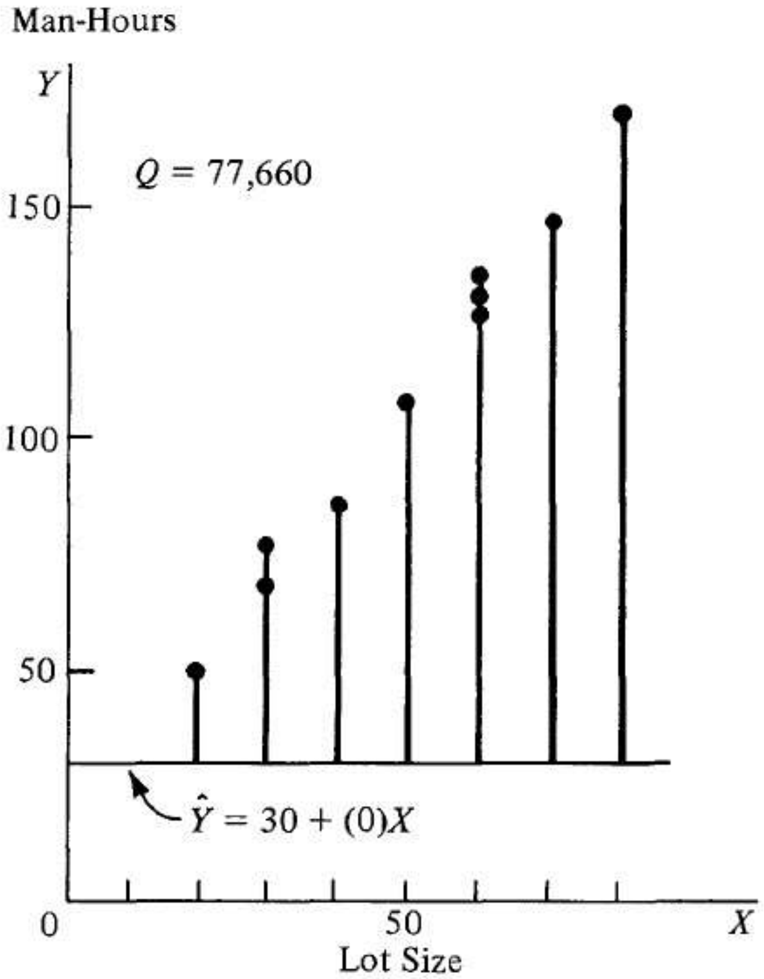
β_0 indicates mean of probability distribution of Y when $X = 0$

Least Square Estimators could be found by trial and error methods

Can also be found using *Normal Equations*

Samatrix.io

Trial and Error Method



Draw a random line

$$\hat{Y}_i = E\{Y_i\} = 30 + (0)X_i$$

$$\hat{\beta}_0 = 30 \; and \; \hat{\beta}_1 = 0$$

The sum of Squared Error is 77,660

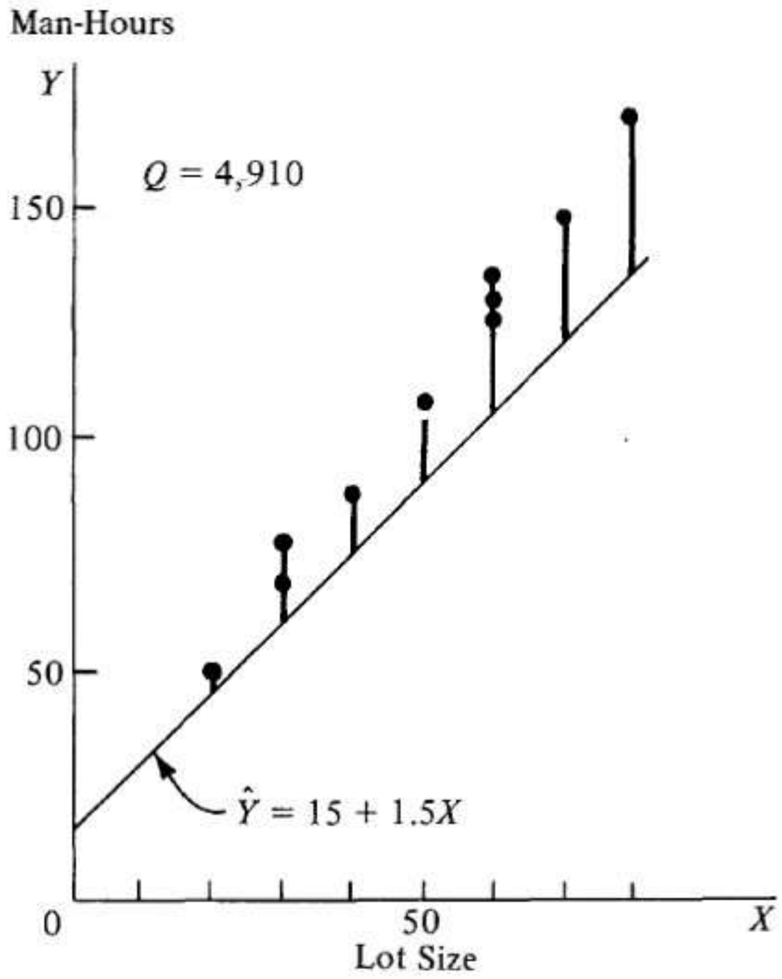
Deviation is large, so fit is poor

$$Q = \sum_{i=1}^{n} (Y_i - 30 - (0)X_i)^2$$

$Run\ (i)$	$Lot\ Size\ (X_i)$	$ManHour\ (Y_i)$	$\hat{Y}_i = 30$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
1	30	73	30	43	1849
2	20	50	30	20	400
3	60	128	30	98	9604
4	80	170	30	140	19600
5	40	87	30	57	3249
6	50	108	30	78	6084
7	60	135	30	105	11025
8	30	69	30	39	1521
9	70	148	30	118	13924
10	60	132	30	102	10404
					77660

Samatrix.io

Linear Regression Coefficients



Draw another line that is closer to the observations

$$\hat{Y}_i = E\{Y_i\} = 15 + (1.5)X_i$$

$$\hat{\beta}_0 = 15 \; and \; \hat{\beta}_1 = 1.5$$

The sum of Squared Error is 4,910

Deviation is lesser, so fit is some what better

Try drawing lines until you find the line with minimum error

$$Q = \sum_{i=1}^{n} (Y_i - 15 - (1.5)X_i)^2$$

$Run\ (i)$	$Lot\ Size\ (X_i)$	$ManHour\ (Y_i)$	$\hat{Y}_i = 15 + 1.5X_i$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
1	30	73	60	13	169
2	20	50	45	5	25
3	60	128	105	23	529
4	80	170	135	35	1225
5	40	87	75	12	144
6	50	108	90	18	324
7	60	135	105	30	900
8	30	69	60	9	81
9	70	148	120	28	784
10	60	132	105	27	729
					4910

Least Square Estimators

Least Square Estimators could be found by trial and error methods

Can also be found using *Normal Equations*

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

Samatrix.io

Least Square Estimators

$n = 10, \bar{X} = \frac{500}{10} = 50, \bar{Y} = \frac{1100}{10} = 110$

$Run\ (i)$	$Lot\ Size\ (X_i)$	$ManHour\ (Y_i)$	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
1	30	73	-20	-37	740	400
2	20	50	-30	-60	1800	900
3	60	128	10	18	180	100
4	80	170	30	60	1800	900
5	40	87	-10	-23	230	100
6	50	108	0	-2	0	0
7	60	135	10	25	250	100
8	30	69	-20	-41	820	400
9	70	148	20	38	760	400
10	60	132	10	22	220	100
Total	500	1100	0	0	6800	3400
	$\sum X_i$	$\sum Y_i$			$\sum (X_i - \bar{X})(Y_i - \bar{Y})$	$\sum (X_i - \bar{X})^2$

Samatrix.io

Least Square Estimators

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

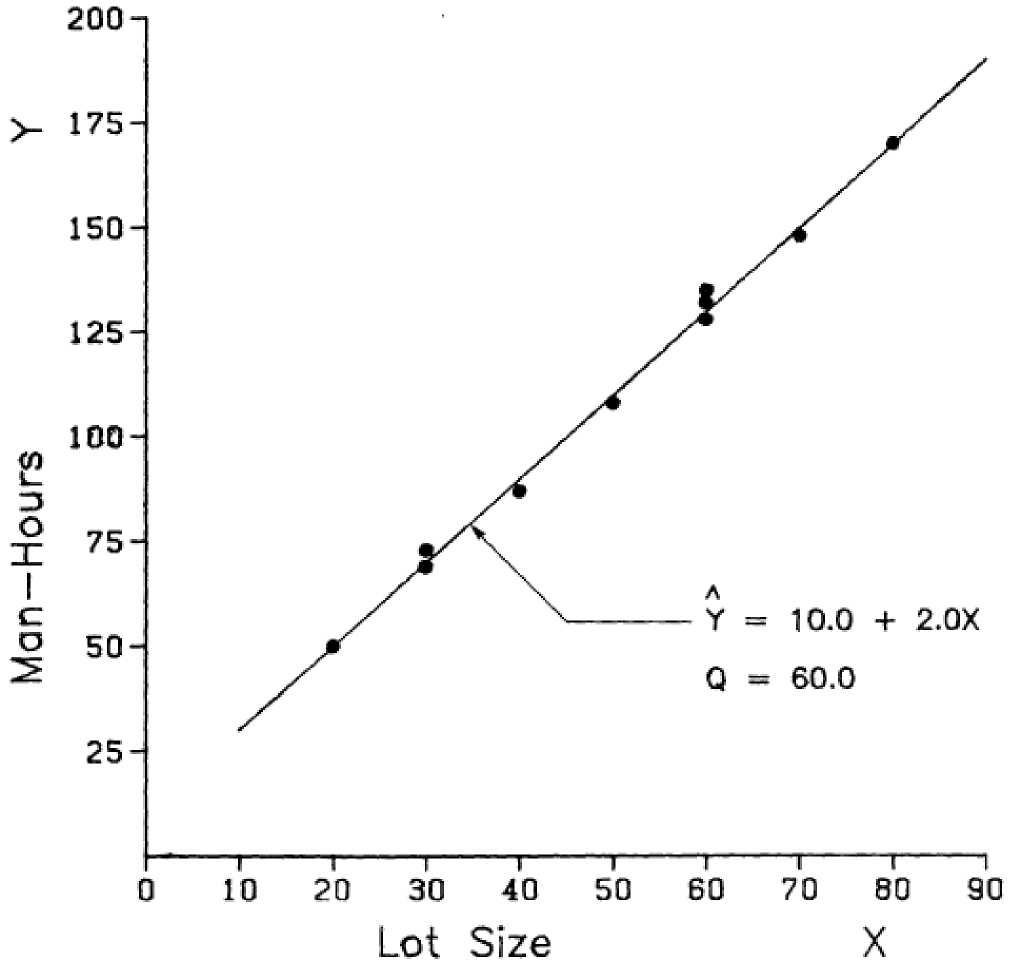
$\sum(X_i - \bar{X})(Y_i - \bar{Y})$ = 6800, $\sum(X_i - \bar{X})^2$= 3400, $n = 10, \bar{X} = \frac{500}{10} =$ 50, $\bar{Y} = \frac{1100}{10} =$ 110

$$\hat{\beta}_1 = \frac{6800}{3400} = 2.0$$

$$\hat{\beta}_0 = 110 - 2 \times 50 = 10.0$$

Samatrix.io

Linear Regression Coefficients



$$\hat{Y}_i = E\{Y_i\} = 10 + (2.0)X_i$$

$$\hat{\beta}_0 = 10 \; and \; \hat{\beta}_1 = 2.0$$

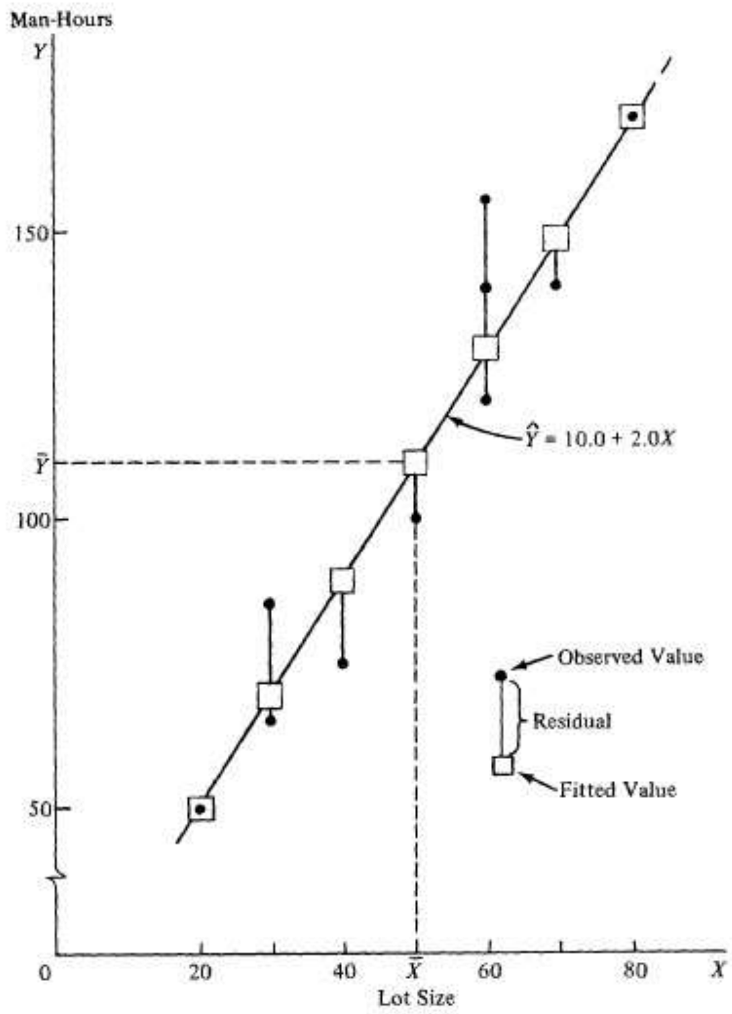
The sum of Squared Error is 60

Mean Number of Man-Hours when $X = 55$

$$\hat{Y} = 10 + 2 \times 55 = 120$$

Samatrix.io

Residual



ith *Residual* is the difference between the observed value Y_i and the corresponding fitted value $\hat{Y}_i$

$$e_i = Y_i - \hat{Y}_i$$

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

Samatrix.io

Residuals

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

$Run\ (i)$	$Lot\ Size\ (X_i)$	$ManHour\ (Y_i)$	$\hat{Y}_i = 10 + 2X_i$	$e_i = \hat{Y}_i - Y_i$	$e_i^2 = (\hat{Y}_i - Y_i)^2$
1	30	73	70	3	9
2	20	50	50	0	0
3	60	128	130	-2	4
4	80	170	170	0	0
5	40	87	90	-3	9
6	50	108	110	-2	4
7	60	135	130	5	25
8	30	69	70	-1	1
9	70	148	150	-2	4
10	60	132	130	2	4
Total	500	1100	1100	0	60

Samatrix.io

Properties of Fitted Regression Line

- The sum of Residuals is Zero

$$\sum_{i=1}^{n} e_i = 0$$

- The sum of Residual Square $\sum e_i^2$ is minimum
- The sum of Observed Value Y_i is equal to sum of Fitted Value $\hat{Y}_i$

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

Properties of Fitted Regression Line

- The sum of weighted residuals is Zero when the residual in the ith trial is weighted by the level of independent variable in ith trial

$$\sum_{i=1}^{n} X_i e_i$$

- The sum of weighted residuals is Zero when the residual in the ith trial is weighted by the fitted value of response variable in ith trial

$$\sum_{i=1}^{n} \widehat{Y}_i e_i$$

- The Regression Line Always go through $\bar{X}, \bar{Y}$

Regression Summary Output

MULTIPLE R	0.99780	← r
R SQUARE	0.99561	← r^2
STANDARD ERROR	2.73861	← $\sqrt{MSE}$

------------------ VARIABLES IN THE EQUATION ------------------

VARIABLE	B		STD ERROR B	F
SIZE	2.000000 ← b_1	$s(b_1)$ →	0.04697	1813.333 ← F^*
(CONSTANT)	10.00000 ← b_0			

VARIABLE	MEAN	STANDARD DEV	CASES
SIZE	$\bar{X}$ → 50.0000	s_X → 19.4365	10 ← n
HOURS	$\bar{Y}$ → 110.0000	s_Y → 38.9587	10 ← n

ANALYSIS OF VARIANCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	1.	SSR → 13600.00000	MSR → 13600.00000
RESIDUAL ← Error	8.	SSE → 60.00000	MSE → 7.50000

Samatrix.io

Regression Summary Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9978014
R Square	0.9956076
Adjusted R Square	0.9950586
Standard Error	2.7386128
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	13600	13600	1813.3333	1.01959E-10
Residual	8	60	7.5		
Total	9	13660			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	10	2.502939448	3.9953024	0.0039758	4.228211282	15.771789	4.2282113	15.771789
X Variable 1	2	0.046966822	42.583252	1.02E-10	1.891694315	2.1083057	1.8916943	2.1083057

Samatrix.io

$Error\ Sum\ of\ Square\ (SSE)$

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	SSR 13600	MSR 13600	1813.3333	1.01959E-10
Residual Error	8	SSE 60	MSE 7.5		
Total	9	SSTO 13660			

Y_i come from different probability distributions with different means, depending upon the level X_i

Deviation of an observation Y_i must be calculated around its estimated mean $\hat{Y}_i$

Deviation of residuals is

$$Y_i - \hat{Y}_i = e_i$$

$Error\ Sum\ of\ Square$ or $Residual\ Sum\ of\ Square\ (SSE)$ is

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = \sum_{i=1}^{n} e_i^2$$

$Error\ Mean\ Square\ (MSE)$

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	SSR 13600	MSR 13600	1813.3333	1.01959E-10
Residual Error	8	SSE 60	MSE 7.5		
Total	9	SSTO 13660			

The Error Sum of Square MSE has $n - 2$ degree of freedom associated with it

Two $degree\ of\ freedom$ are lost because both β_0 and β_1 had to be estimated in obtaining $\hat{Y}_i$

$Error\ Mean\ Square\ (MSE)$ is

$$MSE = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum(Y_i - b_0 - b_1 X_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

Samatrix.io

$Standard\ Error\ -\ \sigma$

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	SSR 13600	MSR 13600	1813.3333	1.01959E-10
Residual Error	8	SSE 60	MSE 7.5		
Total	9	SSTO 13660			

MSE is unbiased estimator of σ^2 for the regression model

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

So

$$Standard\ Error\ (\sigma) = \sqrt{E(MSE)}$$

Samatrix.io

Residuals

$e_i = Y_i - b_o - b_1 X_i$

$Run\ (i)$	$Lot\ Size\ (X_i)$	$ManHour\ (Y_i)$	$\widehat{Y}_i = 10 + 2X_i$	$e_i = \widehat{Y}_i - Y_i$	$e_i^2 = (\widehat{Y}_i - Y_i)^2$
1	30	73	70	3	9
2	20	50	50	0	0
3	60	128	130	-2	4
4	80	170	170	0	0
5	40	87	90	-3	9
6	50	108	110	-2	4
7	60	135	130	5	25
8	30	69	70	-1	1
9	70	148	150	-2	4
10	60	132	130	2	4
Total	500	1100	1100	0	60

SSE

Samatrix.io

Example

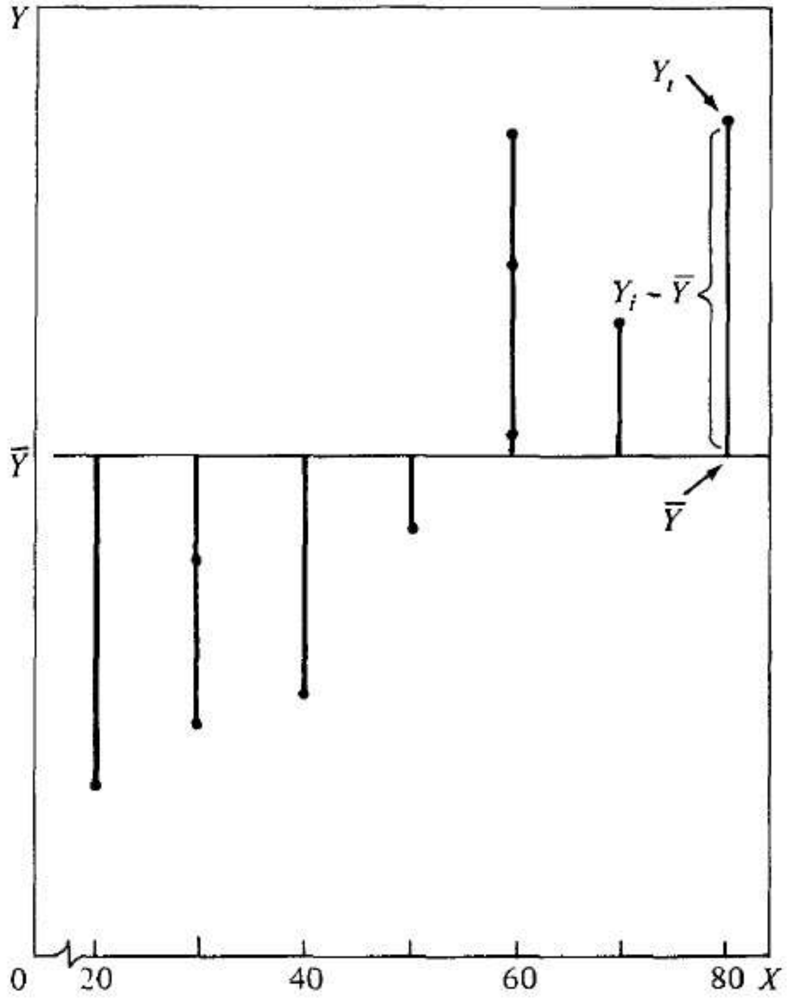
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	SSR 13600	MSR 13600	1813.3333	1.01959E-10
Residual Error	8	SSE 60	MSE 7.5		
Total	9	SSTO 13660			

$$SSE = \sum_{i=1}^{n} e_i^2 = 60$$

$$MSE = \frac{SSE}{n-2} = \frac{60}{10-2} = \frac{60}{8} = 7.5$$

$$Standard\ Error\ (\sigma) = \sqrt{E(MSE)} = \sqrt{7.5} = 2.7386$$

SSTO



$Total\ Sum\ of\ Squares$

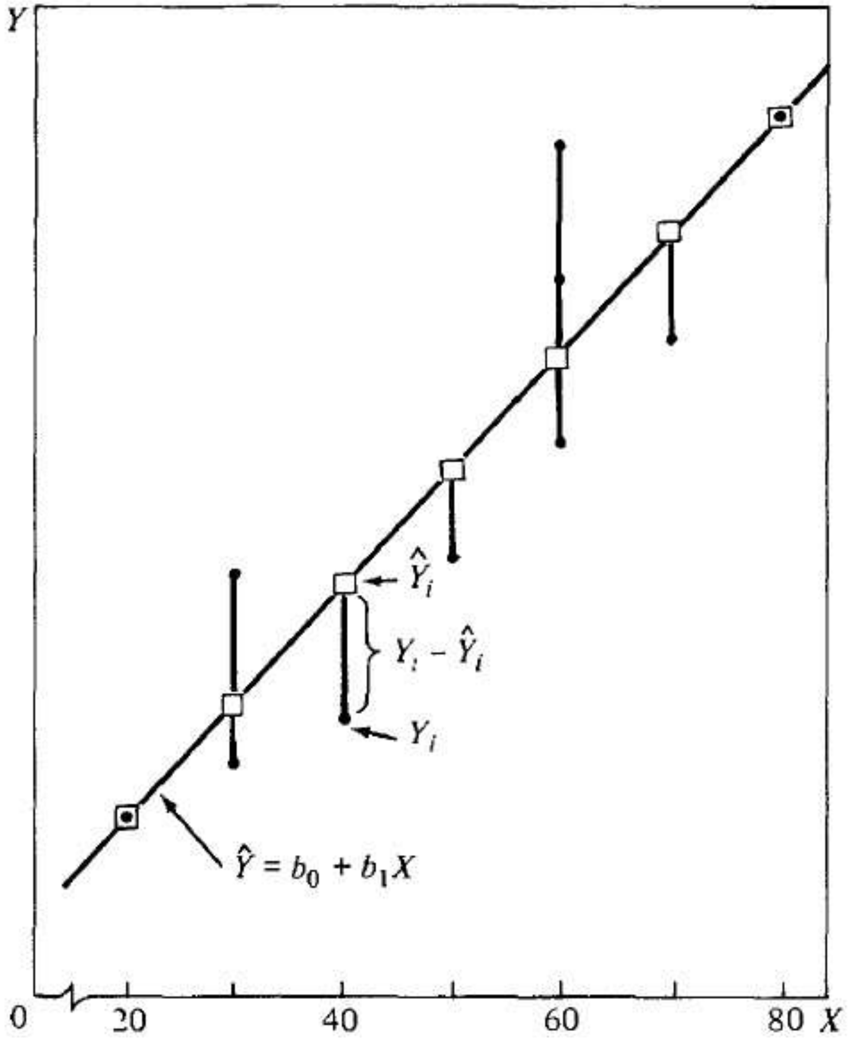
Measures Total Variation

Greater the SSTO, greater the variation among the Y observations

$$SSTO = \sum (Y_i - \bar{Y})^2$$

Samatrix.io

SSE



Error Sum of Squares

Uncertainty of data of Y observations around regression line

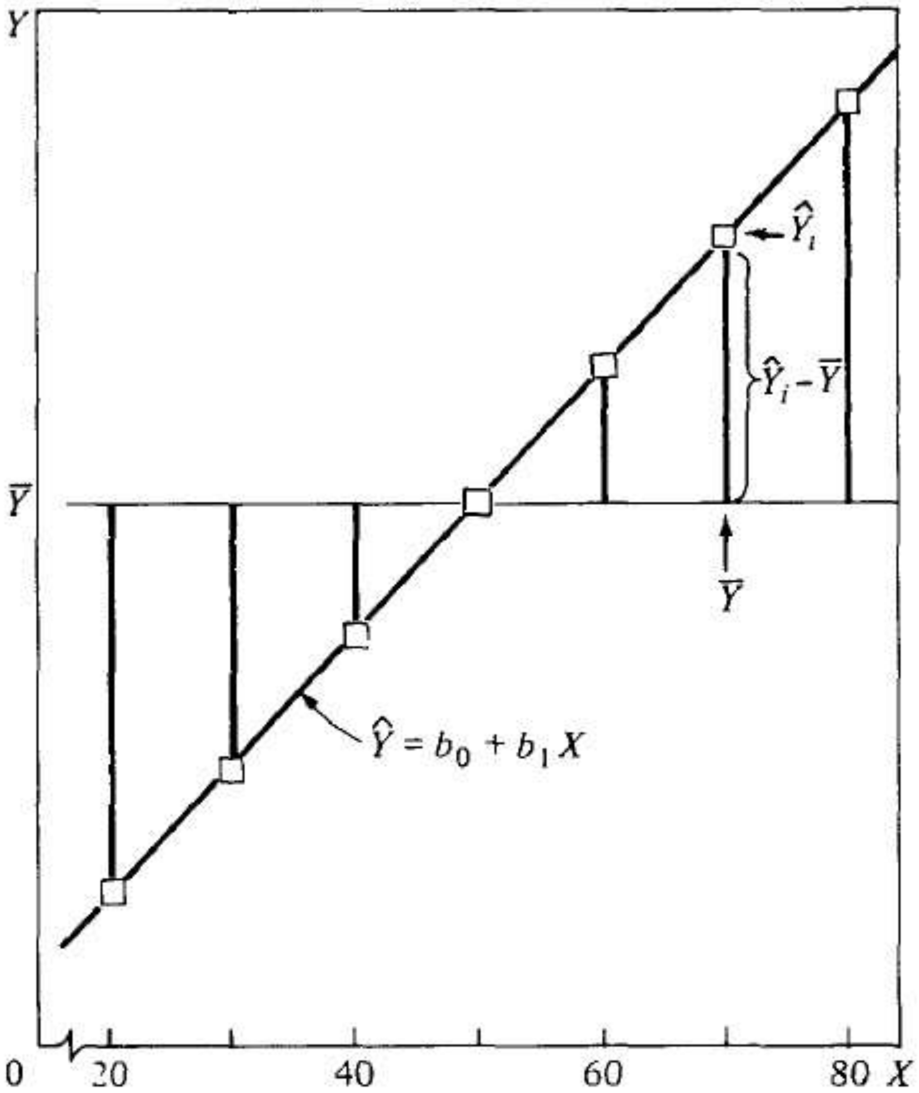
If SSE = 0, all observations fall on Regression Line

Larger the SSE, the greater is the variation of Y observation around Regression Line

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

Samatrix.io

SSR



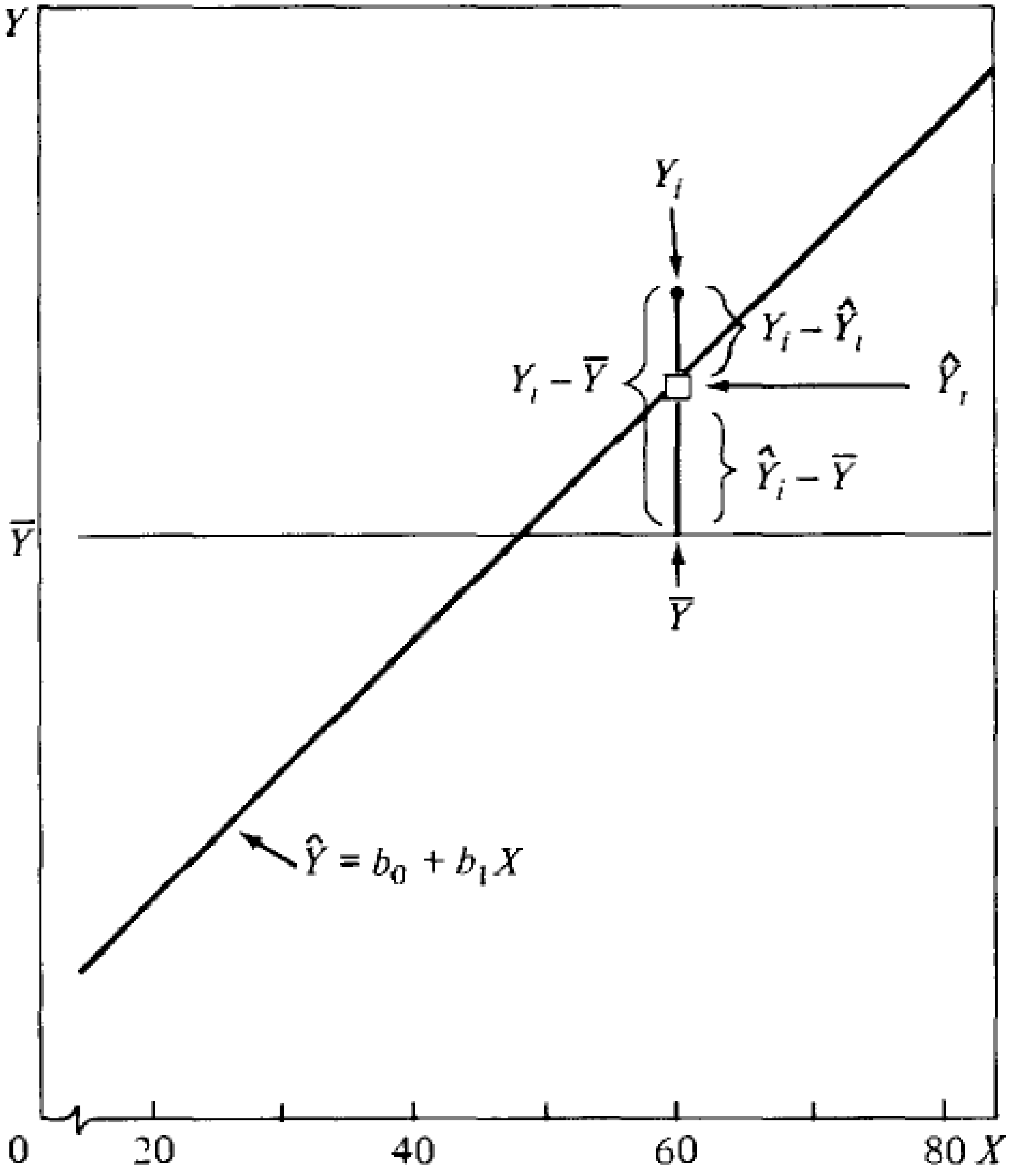
Regression Sum of Squares

Difference between fitted value on Regression line and the mean of fitted value

Measure of the variability of the Y's associated with Regression Line

Larger the SSR in relation to SSTO, greater the effect of regression relation in accounting for total variation in the Y observations

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2$$



Relationship SSTO, SSE, SSR

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Total Deviation = Deviation of Fitted Regression Line + Deviation around Regression Line.

Sum of Squared deviation have the same relationship

$$\sum (Y_i - \bar{Y})^2$$

$$= \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

$$SSTO = SSE + SSR$$

Samatrix.io

SSTO = SSE+SSR

$n = 10, \bar{X} = \frac{500}{10} = 50, \bar{Y} = \frac{1100}{10} = 110$

$\boldsymbol{Run\ (i)}$	$\boldsymbol{Lot\ Size\ (X_i)}$	$\boldsymbol{ManHour\ (Y_i)}$	$\hat{Y}_i$	$(Y_i - \bar{Y})^2$	$(Y_i - \hat{Y}_i)^2$	$(\hat{Y}_i - \bar{Y})^2$
1	30	73	70	1369	9	1600
2	20	50	50	3600	0	3600
3	60	128	130	324	4	400
4	80	170	170	3600	0	3600
5	40	87	90	529	9	400
6	50	108	110	4	4	0
7	60	135	130	625	25	400
8	30	69	70	1681	1	1600
9	70	148	150	1444	4	1600
10	60	132	130	484	4	400
Total	500	1100	0	13660	60	13600
	$\sum X_i$	$\sum Y_i$		$\sum (Y_i - \bar{Y})^2$	$= \sum (Y_i - \hat{Y}_i)^2 +$	$\sum (\hat{Y}_i - \bar{Y})^2$

Samatrix.io

Degree of Freedom df

$SSTO\ df = (n-1)$

1 df is lost because deviation $Y_i - \bar{Y}$ should sum to 0

$SSE\ df = (n-2)$

2 df are lost because two parameter β_0 and β_1 were used to calculate $\hat{Y}_i$

$SSR\ df = 1$

There are two parameters in regression equation (intercept and slope). One df is lost because $\boldsymbol{\hat{Y}_i - \bar{Y}}$ should sum to Zero. So one df is lost

$$Degree\ of\ Freedom\ are\ additive$$

$$(n-1) = (n-2) + 1$$

Samatrix.io

Mean Squares

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	SSR 13600	MSR 13600	1813.3333	1.01959E-10
Residual Error	8	SSE 60	MSE 7.5		
Total	9	SSTO 13660			

Sum of Squares divided by degree of Freedom is called $Mean\ Square\ (MS)$

Regression Mean Square $MSR = \frac{SSR}{1} = SSR$

Error Mean Square $MSE = \frac{SSE}{(n-2)}$

<u>In Our Example</u>

MSR = SSR = 13600

MSE=$\frac{60}{8}$=7.5

Samatrix.io

F Test

To Establish a Relationship between the Response and Predictors

We check whether $\beta_1 = 0$ using Hypothesis

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

$$F^* = \frac{MSR}{MSE}$$

$$If\ F^* \leq F(1-\alpha; 1; n-2), conclude\ H_0$$
$$If\ F^* \geq F(1-\alpha; 1; n-2), conclude\ H_a$$

Samatrix.io

F Test

$$MSR = 13600; MSE = 7.5$$

$$F^* = \frac{MSR}{MSE} = \frac{13600}{7.5} = 1813.333$$

For $\alpha = 0.05\ and\ n = 10$

$$F(1 - 0.05; 1; 8) = 5.32$$

$$Since\ F^* \geq 5.32, we\ conclude\ H_a$$

Calculate excel function `=F.INV(0.95,1,8)`

Hence, there is a linear association between lot-size and man-hours

Samatrix.io

Analysis Of Variance Table (ANOVA)

Source of Variation	SS	Df	MS	F
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$F^* = \frac{MSR}{MSE}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{(n-2)}$	
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	SSR 13600	MSR 13600	1813.3333	1.01959E-10
Residual Error	8	SSE 60	MSE 7.5		
Total	9	SSTO 13660			

Samatrix.io

Analysis Of Variance Table (ANOVA)

Source of Variation	SS	Df	MS	F
Regression	$SSR = 13600$	1	$MSR = \frac{13600}{1} = 13600$	$F^* = \frac{13600}{7.5} = 1813.333$
Error	$SSE = 60$	8	$MSE = \frac{60}{8} = 7.5$	
Total	$SSTO = 13660$	9		

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	SSR 13600	MSR 13600	1813.3333	1.01959E-10
Residual Error	8	SSE 60	MSE 7.5		
Total	9	SSTO 13660			

Samatrix.io

Assessing Model Accuracy

Quantify the extent to which model fits the data or measure of lack of fit

4 Methods

1. Multiple R
2. R^2 Statistics
3. Adjusted R^2 Statistics
4. Residual Standard Error

Regression Statistics	
Multiple R	0.9978014
R Square	0.9956076
Adjusted R Square	0.9950586
Standard Error	2.7386128
Observations	10

Residual Standard Error (RSE)

RSE - Average amount that response will deviate from true regression line

RSE is estimate of the standard deviation of ϵ

$$RSE = \sqrt{MSE} = \sqrt{\frac{SSE}{(n-2)}} = \sqrt{\frac{\sum(Y_i - \widehat{Y}_i)^2}{(n-2)}}$$

If the predictions using the model are very close to true outcome value, then we can conclude that model fits the data very well

If the predictions are very far from the true outcome value, RSE will may be large, we can conclude that model does not fit the data well

RSE gives the absolute value in terms of Y but we are not sure about what constitutes good RSE (eg 2.7386 in our example)

In this case $\text{RSE} = \sqrt{7.5} = 2.7386$

r^2– Coefficient of Determination

Regression Statistics	
Multiple R	0.9978014
R Square	0.9956076
Adjusted R Square	0.9950586
Standard Error	2.7386128
Observations	10

Alternate is r^2 Statistics – Proportion of Variance Explained.

Value Varies between 0 and 1 and independent of the scale of Y

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

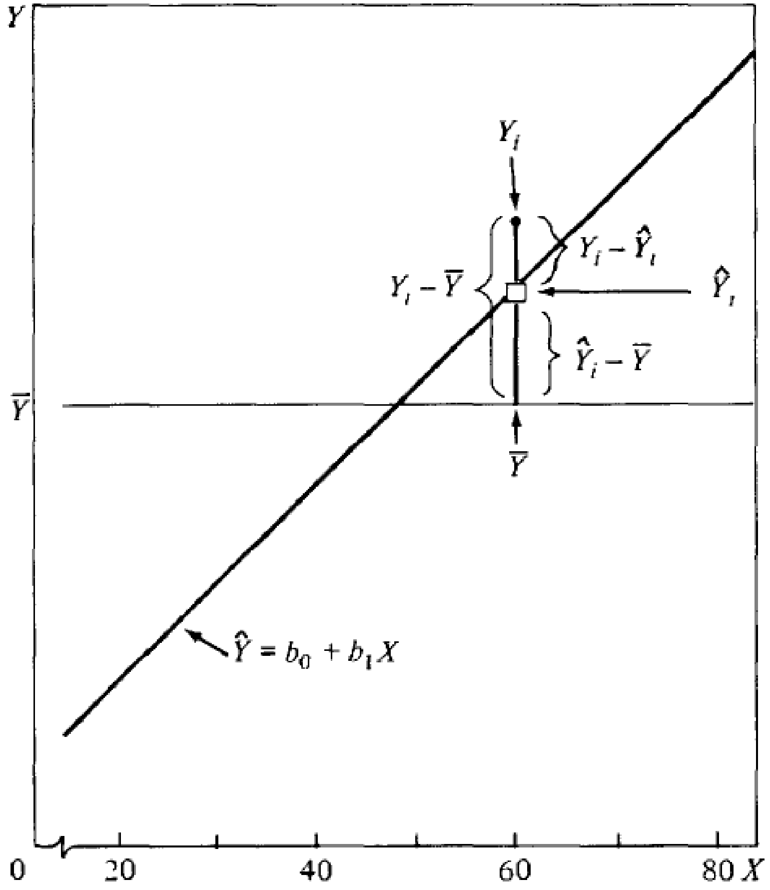
Since $0 \le SSE \le SSTO$ it follows

$$0 \le r^2 \le 1$$

SSE measures the amount of variability that is left unexplained after performing the Regression

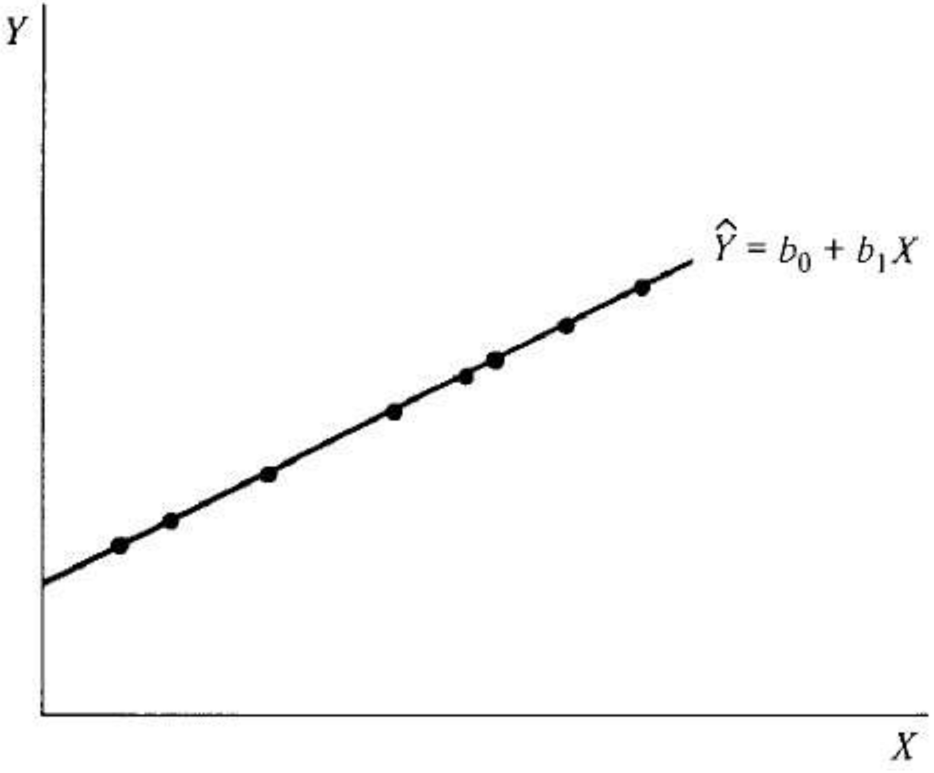
$SSTO - SSE$ measures the amount of variability that is explained (or Removed) after performing the Regression

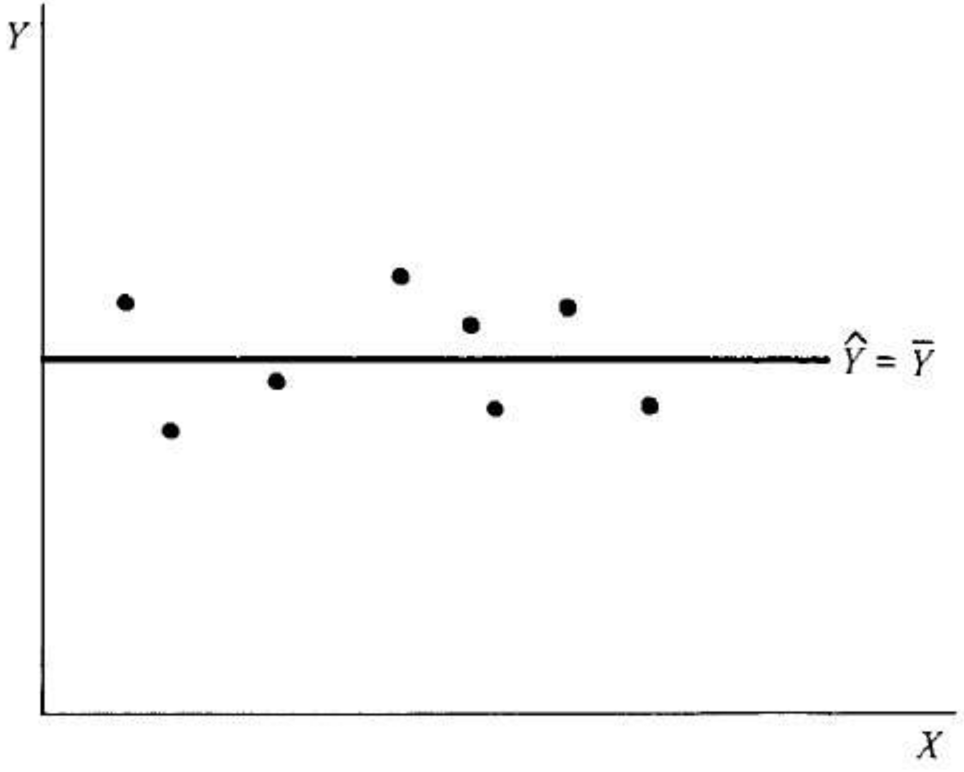
r^2Proportion of variability in Y that can be explained using X



Samatrix.io

r^2 Statistics





Fig(a) - $r^2 = 1$ & $SSE = 0$

X accounts for all variations in Y

Fig (b) - $r^2 = 0$ & $SSE = SSTO$

No linear relationship between X - Y

Samatrix.io

r^2 Statistics

Regression Statistics	
Multiple R	0.9978014
R Square	0.9956076
Adjusted R Square	0.9950586
Standard Error	2.7386128
Observations	10

r^2 Statistics is measure of Linear Relationship between X and Y

In practice r^2 is not equal to 0 or 1. It is some where between

Closer to 1 - greater degree of linear association between X and Y

Our Example

$$r^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{60}{13660} = 0.9956$$

<u>Means</u> - The variation in man-hours is reduced by 99.56% when lot-size is considered

Samatrix.io

Multiple r – Coefficient of Correlation

Regression Statistics	
Multiple R	0.9978014
R Square	0.9956076
Adjusted R Square	0.9950586
Standard Error	2.7386128
Observations	10

Multiple r (Coefficient of Correlation) is Square Root of r^2

$$r = \pm\sqrt{r^2}$$

A Plus/Minus sign is attached to measure according to whether the slope of fitted regression line is positive or negative

The Range of r is:

$$-1 \leq r \leq +1$$

Any r^2 other then 0 or 1, $r^2 < |r|$, r may give an impression of a closer relationship between X and Y than r^2.

Example : $r = \sqrt{r^2} = \sqrt{0.9956} = +\,0.9978$ (+ because b_1 is positive)

Samatrix.io

Adjusted r^2

Regression Statistics	
Multiple R	0.9978014
R Square	0.9956076
Adjusted R Square	0.9950586
Standard Error	2.7386128
Observations	10

It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable.

It penalizes you for adding independent variable that do not help in predicting the dependent variable.

Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.

$$Adjusted\ r^2 = 1 - \frac{\frac{SSE}{df}}{\frac{SSTO}{df}} = 1 - \frac{\frac{60}{8}}{\frac{13660}{9}} = 0.9950586$$

Samatrix.io

Adjusted r^2

Regression Statistics	
Multiple R	0.9978014
R Square	0.9956076
Adjusted R Square	0.9950586
Standard Error	2.7386128
Observations	10

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

$$Adjusted\ r^2 = 1 - \frac{(1 - r^2)(N - 1)}{(N - p - 1)} = 1 - \frac{(1 - 0.9956)(10 - 1)}{(10 - 1 - 1)}$$
$$= 0.9950586$$

Where r^2 is the value of r^2

p is the number of predictors

N is the Sample Size

Samatrix.io

What are the flaws in R-squared?

- There are two major flaws of R-squared:
- **Problem- 1:** As we are adding more and more predictors, R^2 always increases irrespective of the impact of the predictor on the model. As R^2 always increases and never decreases, it can always appear to be a better fit with the more independent variables(predictors) we add to the model. This can be completely misleading.
- **Problem- 2:** Similarly, if our model has too many independent variables and too many high-order polynomials, we can also face the problem of over-fitting the data. Whenever the data is over-fitted, it can lead to a misleadingly high R^2 value which eventually can lead to misleading predictions.

Thanks

Samatrix Consulting Pvt Ltd

Samatrix.io