Study the relationship between two or more variables using regression

Example: relationship between advertising expenditures and sales

As advertising expenditures increase

Sales increase

Example: relationship between number of hours practice and errors

As hours of practice increase

Errors decrease

Develop a model to show how the variables are related and to predict

Example: predict sales for a given level of advertising

Dependent Variable – the variable we are trying to predict

y Sales

Independent Variable – the variable we use to predict the dependent variable

x Advertising Expenditures

Simple Linear Regression:

Simple – one independent variable and one dependent variable

Linear – the relationship is approximated using a straight line

Multiple Regression – two or more independent variables

Simple Linear Regression Model:

$$\boxed{y = \beta_0 + \beta_1 x + \varepsilon}$$

where:

β_0 is the y-intercept of the regression line

β_1 is the slope of the regression line

ε is the error term.

β_0 and β_1 are the population parameters

b_0 and b_1 are the sample statistics used to estimate β_0 and β_1

Estimated Simple Linear Regression Equation:

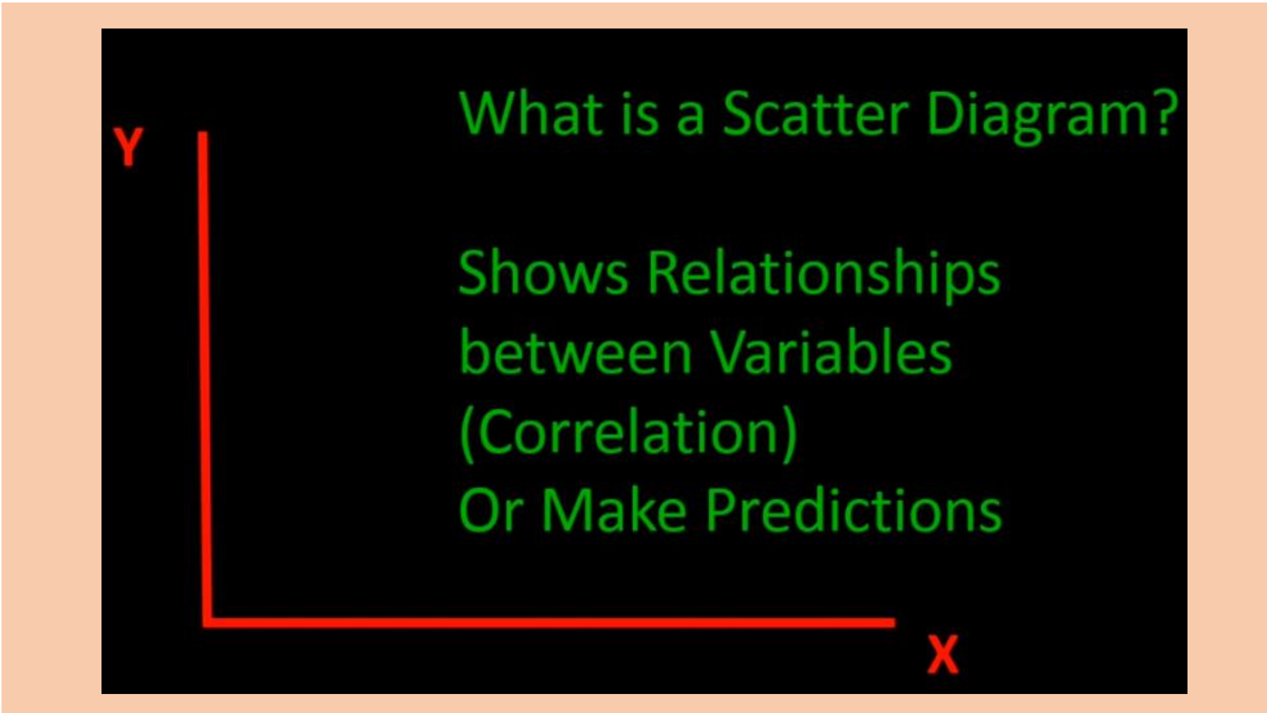
$$\boxed{\hat{y} = b_0 + b_1 x}$$

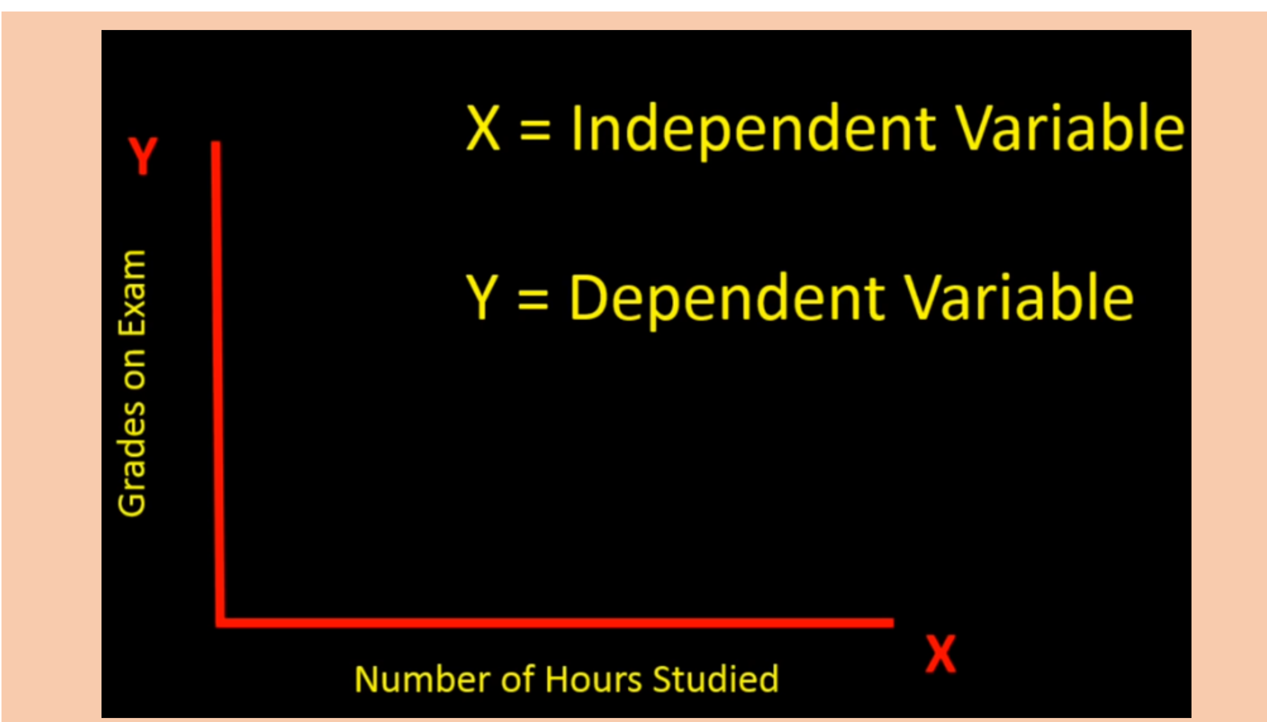
where:

$\hat{y}$ is the predicted value of y for a given x value.

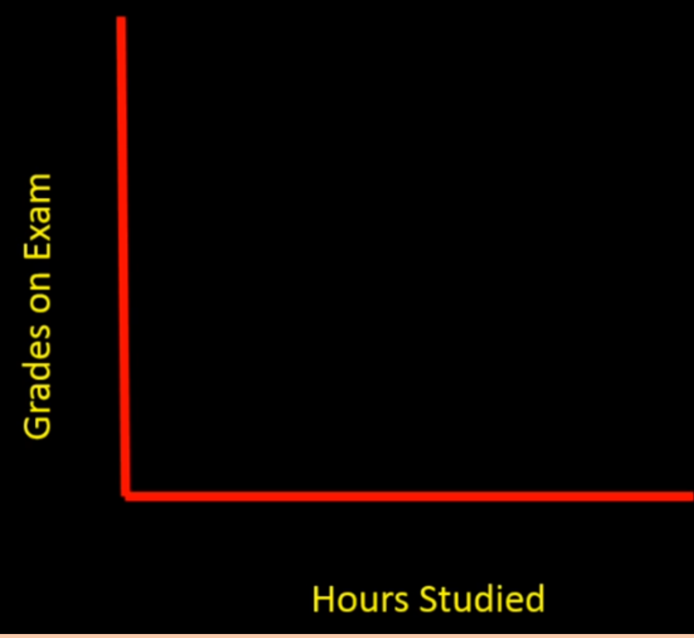
b_0 is the y intercept of the line.

b_1 is the slope of the line.

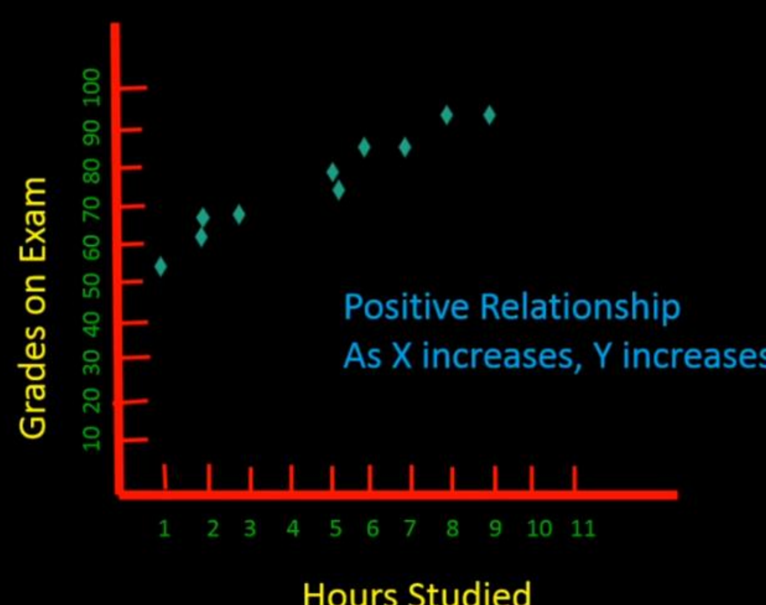

What is a Scatter Diagram?
Shows Relationships between Variables (Correlation)
Or Make Predictions
Y
X

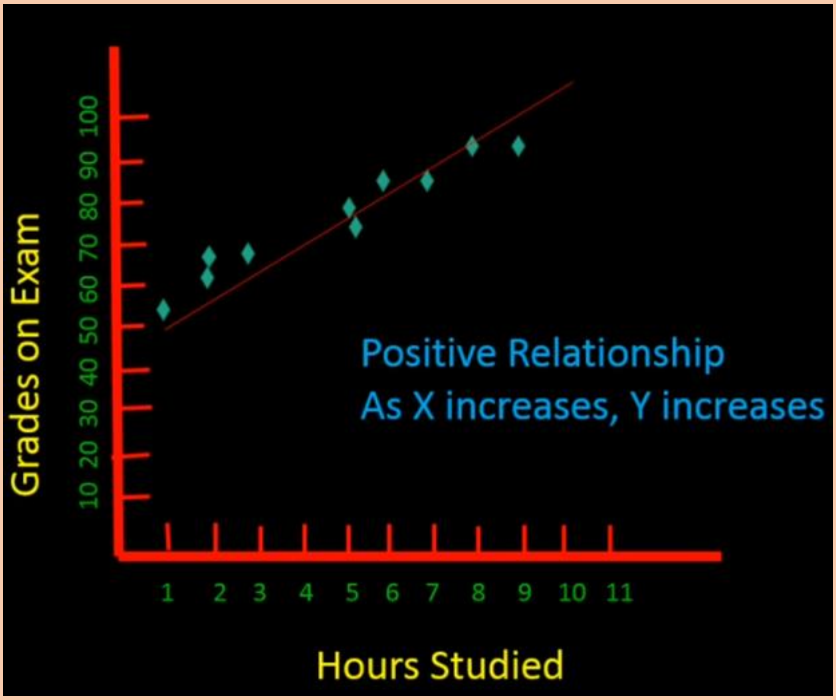

X = Independent Variable
Y = Dependent Variable
Y
Grades on Exam
X
Number of Hours Studied

Hours Studied	Grade on Exam
2.00	69.00
9.00	98.00
5.00	82.00
5.00	77.00
3.00	71.00
7.00	84.00
1.00	55.00
8.00	94.00
6.00	84.00
2.00	64.00



Hours Studied	Grade on Exam
2.00	69.00
9.00	98.00
5.00	82.00
5.00	77.00
3.00	71.00
7.00	84.00
1.00	55.00
8.00	94.00
6.00	84.00
2.00	64.00





Least Squares Method:

Use Sample data to find the line of regression

$$\boxed{\hat{y} = b_0 + b_1 x}$$

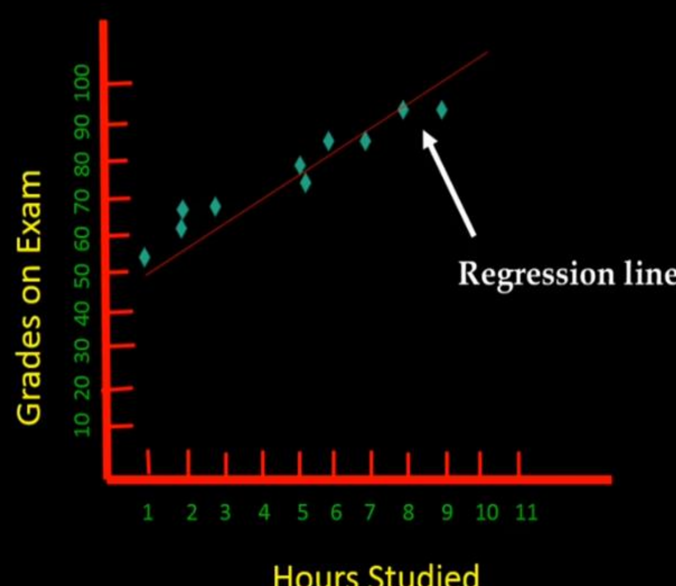
where:

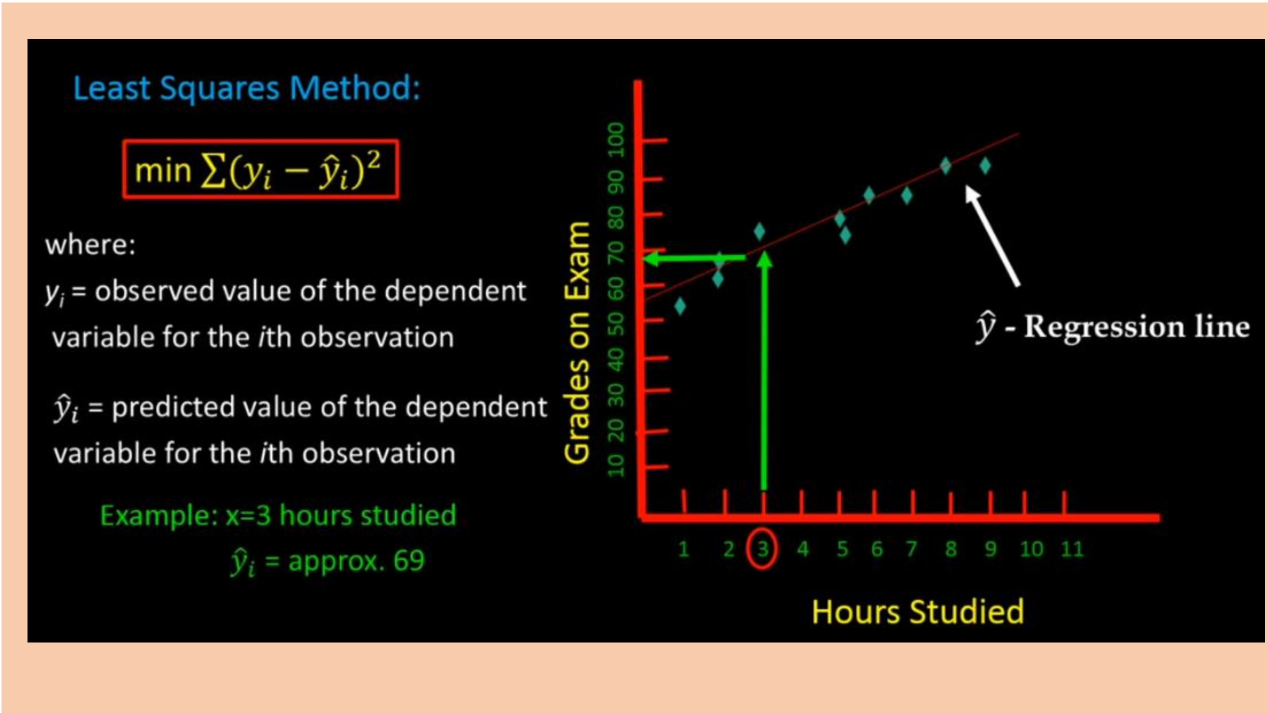
$\hat{y}$ is the predicted grade on exam

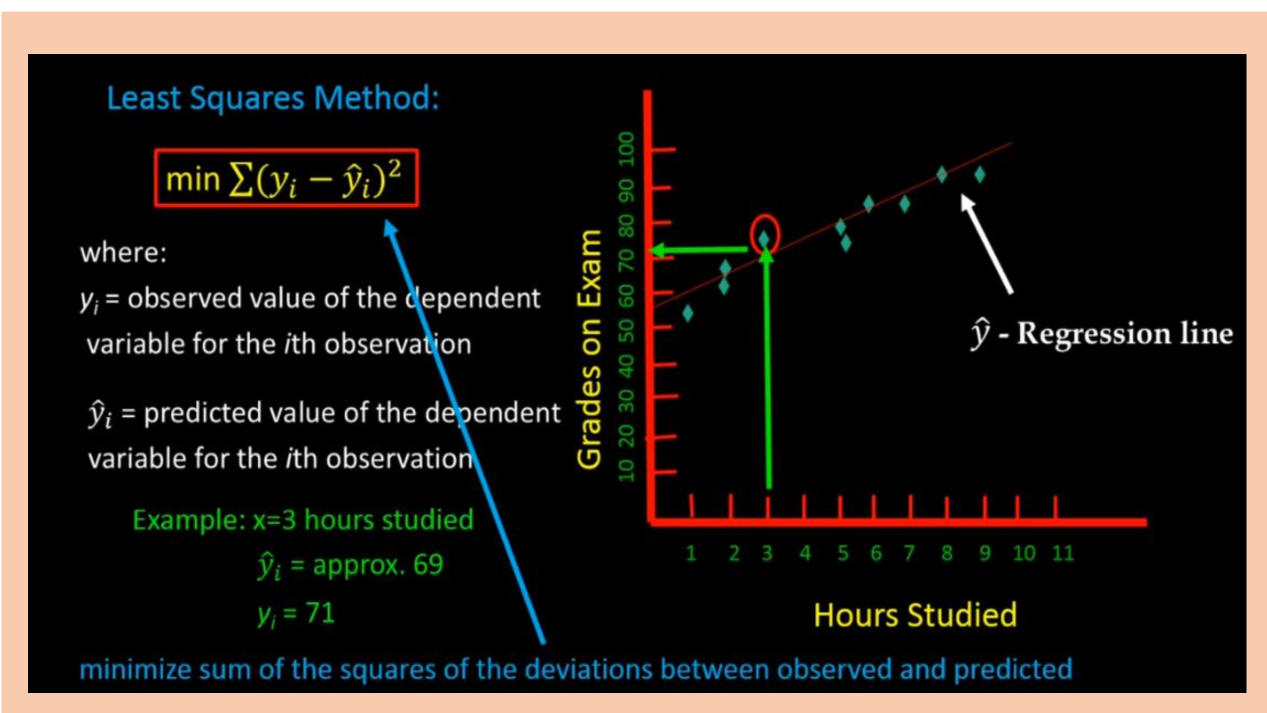
b_0 is the y intercept of the line.

b_1 is the slope of the line.

x is number of hours studied




Least Squares Method:
$\min \sum (y_i - \hat{y}_i)^2$
where:
y_i = observed value of the dependent variable for the ith observation
$\hat{y}_i$ = predicted value of the dependent variable for the ith observation
Example: x=3 hours studied
$\hat{y}_i$ = approx. 69
Grades on Exam
10 20 30 40 50 60 70 80 90 100
1 2 3 4 5 6 7 8 9 10 11
Hours Studied
$\hat{y}$ - Regression line


Least Squares Method:
$\min \sum (y_i - \hat{y}_i)^2$
where:
y_i = observed value of the dependent variable for the ith observation
$\hat{y}_i$ = predicted value of the dependent variable for the ith observation
Example: x=3 hours studied
$\hat{y}_i$ = approx. 69
y_i = 71
Grades on Exam
10 20 30 40 50 60 70 80 90 100
1 2 3 4 5 6 7 8 9 10 11
Hours Studied
$\hat{y}$ - Regression line
minimize sum of the squares of the deviations between observed and predicted

Calculating the Slope:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

where:

x_i = value of independent variable for ith observation

y_i = value of dependent variable for ith observation

$\bar{x}$ = mean value for independent variable

$\bar{y}$ = mean value for dependent variable

Calculating the y – intercept:

$$b_0 = \bar{y} - b_1\bar{x}$$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
2	69	-2.8	-8.8	24.64	7.84
9	98	4.2	20.2	84.84	17.64
5	82	.2	4.2	.84	.04
5	77	.2	-.8	-.16	.04
3	71	-1.8	-6.8	12.24	3.24
7	84	2.2	6.2	13.64	4.84
1	55	-3.8	-22.8	86.64	14.44
8	94	3.2	16.2	51.84	10.24
6	84	1.2	6.2	7.44	1.44
2	64	-2.8	-13.8	38.64	7.84
$\Sigma x_i = 48$	$\Sigma y_i = 778$			320.6	67.6
$\bar{x} = 48/10 = 4.8$	$\bar{y} = 778/10 = 77.8$			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

$$\boxed{b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}}$$

$$b_1 = \frac{320.6}{67.6} = 4.74$$

$$\boxed{b_0 = \bar{y} - b_1\bar{x}}$$

$\bar{x} = 48/10 = 4.8$ $\quad$ $\bar{y} = 778/10 = 77.8$

$$b_0 = 77.8 - 4.74(4.8)$$

$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
24.64	7.84
84.84	17.64
.84	.04
-.16	.04
12.24	3.24
13.64	4.84
86.64	14.44
51.84	10.24
7.44	1.44
38.64	7.84
320.6	67.6
$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

$$\boxed{b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}}$$

$$b_1 = \frac{320.6}{67.6} = 4.74$$

$$\boxed{b_0 = \bar{y} - b_1\bar{x}}$$

$\bar{x} = 48/10 = 4.8$ $\quad$ $\bar{y} = 778/10 = 77.8$

$$b_0 = 77.8 - 4.74(4.8)$$

$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
24.64	7.84
84.84	17.64
.84	.04
-.16	.04
12.24	3.24
13.64	4.84
86.64	14.44
51.84	10.24
7.44	1.44
38.64	7.84
320.6	67.6
$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

$$\boxed{b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}}$$

$$b_1 = \frac{320.6}{67.6} = 4.74$$

$$\boxed{b_0 = \bar{y} - b_1\bar{x}}$$

$\bar{x}=48/10 = 4.8$ $\bar{y}=778/10 = 77.8$

$$b_0 = 77.8 - 4.74(4.8) = 55.048$$

$$\boxed{\hat{y} = b_0 + b_1 x}$$

$$\hat{y} = 55.048 + 4.74x$$

$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$
24.64	7.84
84.84	17.64
.84	.04
-.16	.04
12.24	3.24
13.64	4.84
86.64	14.44
51.84	10.24
7.44	1.44
38.64	7.84
320.6	67.6
$\Sigma(x_i-\bar{x})(y_i-\bar{y})$	$\Sigma(x_i-\bar{x})^2$

Estimated Regression Line:

$$\hat{y} = 55.048 + 4.74x$$

Use regression line to predict the value of y for a given x

Suppose Number of hours studied = 3

What is the predicted grade on exam?

x=3 what is the predicted value of y?

$$\hat{y} = 55.048 + 4.74\,(3)$$

$$\hat{y} = 69.268$$

Coefficient of Determination:

How well does the regression line fit the data?

$$\boxed{r^2 = SSR/SST}$$

where:

SSR = sum of squares due to regression $= \sum(\hat{y}_i - \bar{y})^2$

SST = total sum of squares $= \sum(y_i - \bar{y})^2$

SSE = sum of squares due to error $= \sum(y_i - \hat{y}_i)^2$

SST = SSR + SSE

x_i	y_i	Predicted Grades $\hat{y}_i = 55.048 + 4.74x_i$
2	69	64.528
9	98	97.708
5	82	78.748
5	77	78.748
3	71	69.268
7	84	88.228
1	55	59.788
8	94	92.968
6	84	83.488
2	64	64.528

x_i	y_i	Predicted Grades $\hat{y}_i = 55.048 + 4.74x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$	Deviation $y_i - \bar{y}$
2	69	64.528	4.472	19.9988	-8.8
9	98	97.708	.292	.0852	20.2
5	82	78.748	3.252	10.5755	4.2
5	77	78.748	-1.748	3.0555	-.8
3	71	69.268	1.732	2.9998	-6.8
7	84	88.228	-4.228	17.8759	6.2
1	55	59.788	-4.788	22.9249	-22.8
8	94	92.968	1.032	1.0650	16.2
6	84	83.488	.512	.2621	6.2
2	64	64.528	-.528	.2788	-13.8

x_i	y_i	Predicted Grades $\hat{y}_i = 55.048 + 4.74x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$	Deviation $y_i - \bar{y}$	Squared Deviation $(y_i - \bar{y})^2$
2	69	64.528	4.472	19.9988	-8.8	77.44
9	98	97.708	.292	.0852	20.2	408.04
5	82	78.748	3.252	10.5755	4.2	17.64
5	77	78.748	-1.748	3.0555	-.8	.64
3	71	69.268	1.732	2.9998	-6.8	46.24
7	84	88.228	-4.228	17.8759	6.2	38.44
1	55	59.788	-4.788	22.9249	-22.8	519.84
8	94	92.968	1.032	1.0650	16.2	262.44
6	84	83.488	.512	.2621	6.2	38.44
2	64	64.528	-.528	.2788	-13.8	190.44
				SSE =79.1215		

SSE =79.1215

SST = 1599.6

Coefficient of Determination:

$\boxed{r^2 = SSR/SST}$ $= 1520.4785/1599.6$ $= .9505$

$SST = SSR + SSE$

$SSR = SST - SSE$

$SSR = 1599.6 - 79.1215$

$= 1520.4785$

r^2 = percent of variability in y can be explained by x

Correlation Coefficient: measures the strength of association between x and y

Values of Correlation Coefficient, r, are between -1 and +1

r = +1 means perfect positive linear relationship
r = -1 means perfect negative linear relationship

r = 0 means no linear relationship

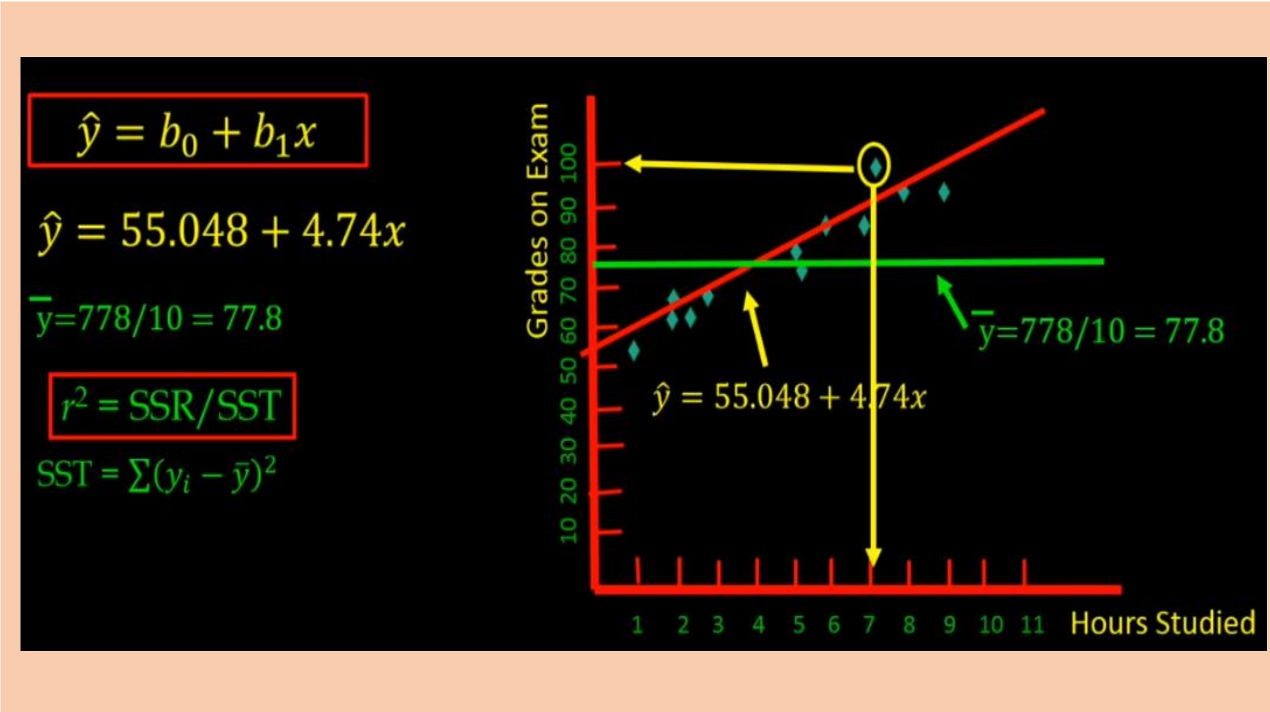
$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of Determination}} = (\text{sign of } b_1)\sqrt{r^2}$$

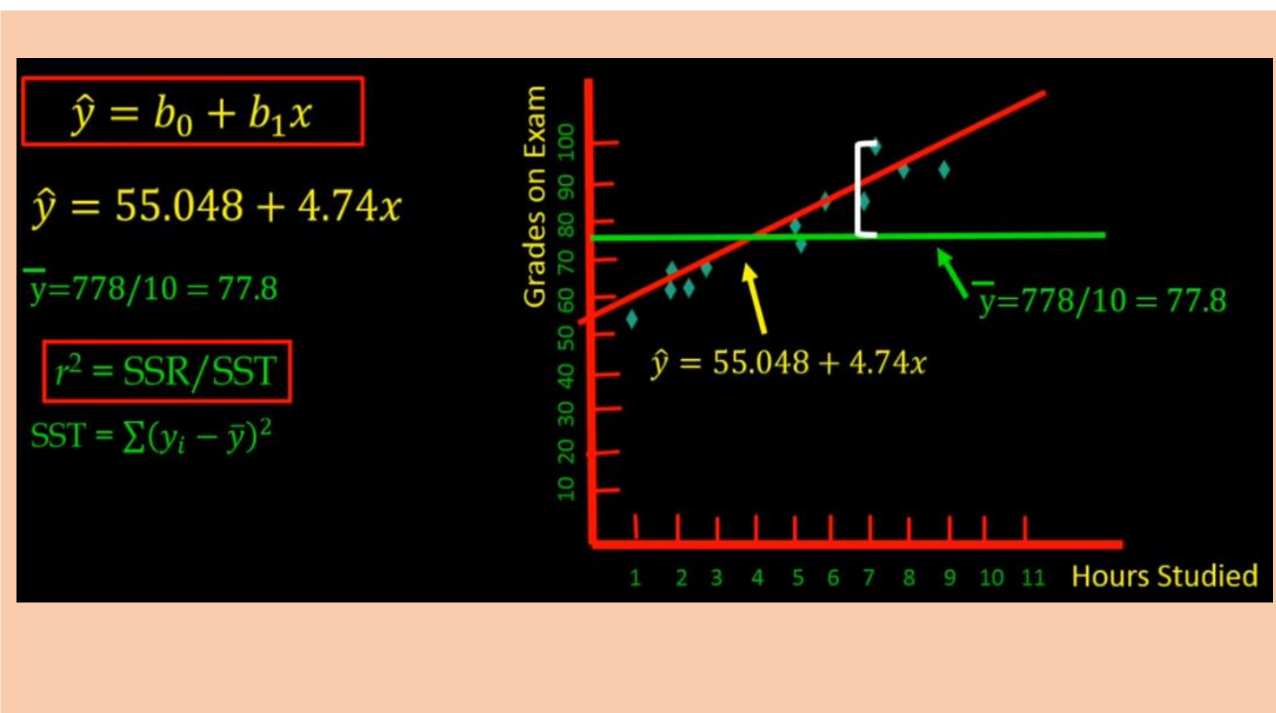
$$r_{xy} = +\sqrt{.9505}$$

$r^2 = .9505$

$$r_{xy} = +.9749$$

+.9749 indicates a very strong positive linear relationship between x and y


$\hat{y} = b_0 + b_1 x$
$\hat{y} = 55.048 + 4.74x$
$\bar{y}$=778/10 = 77.8
r^2 = SSR/SST
SST = $\sum(y_i - \bar{y})^2$
Grades on Exam
10 20 30 40 50 60 70 80 90 100
$\hat{y} = 55.048 + 4.74x$
$\bar{y}$=778/10 = 77.8
1 2 3 4 5 6 7 8 9 10 11
Hours Studied


$\hat{y} = b_0 + b_1 x$
$\hat{y} = 55.048 + 4.74x$
$\bar{y}$=778/10 = 77.8
r^2 = SSR/SST
SST = $\sum(y_i - \bar{y})^2$
Grades on Exam
10 20 30 40 50 60 70 80 90 100
$\hat{y} = 55.048 + 4.74x$
$\bar{y}$=778/10 = 77.8
1 2 3 4 5 6 7 8 9 10 11
Hours Studied

$$\boxed{\hat{y} = b_0 + b_1 x}$$

$$\hat{y} = 55.048 + 4.74x$$

$$\bar{y} = 778/10 = 77.8$$

$$\boxed{r^2 = SSR/SST}$$

$$SST = \sum(y_i - \bar{y})^2$$

$$\hat{y} = 55.048 + 4.74(7)$$

$$\hat{y} = 88.228$$

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

$$\boxed{\hat{y} = b_0 + b_1 x}$$

$$\hat{y} = 55.048 + 4.74x$$

$$\bar{y} = 778/10 = 77.8$$

$$\boxed{r^2 = SSR/SST}$$

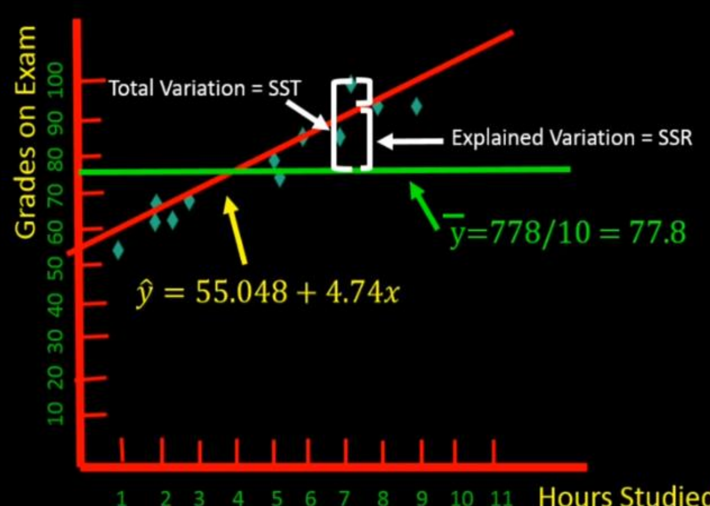
$$SST = \sum(y_i - \bar{y})^2$$

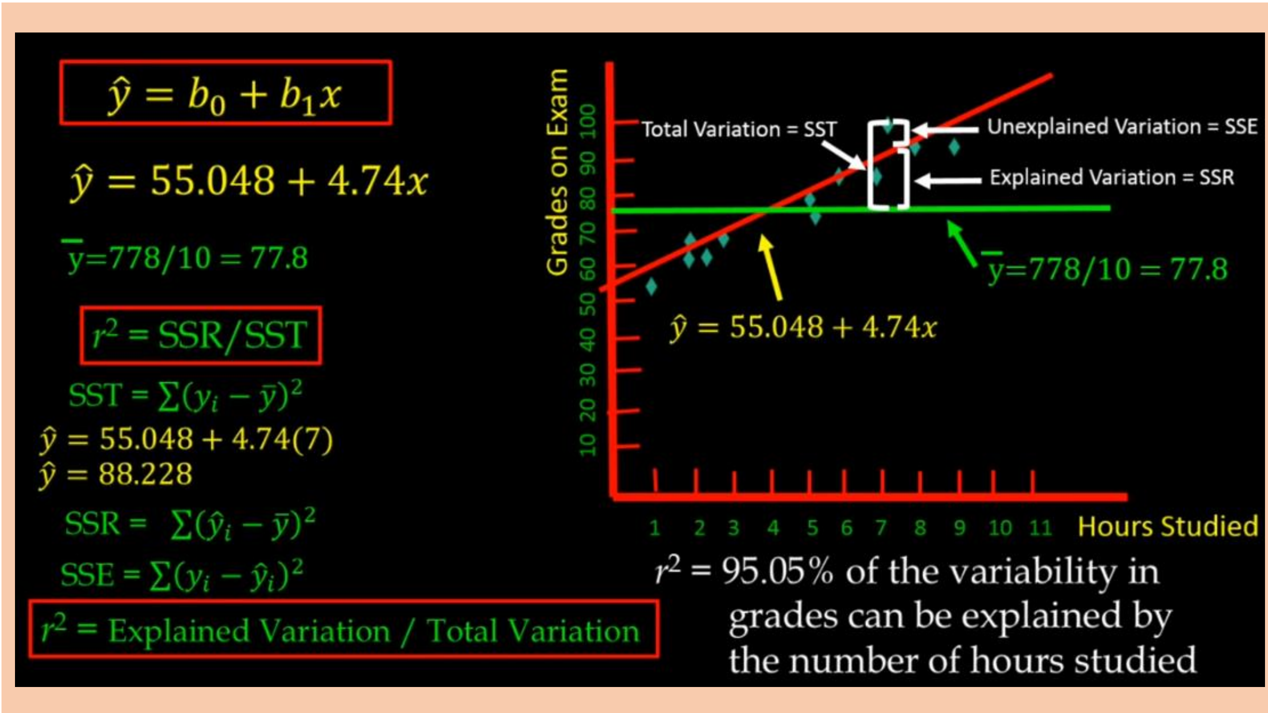
$$\hat{y} = 55.048 + 4.74(7)$$

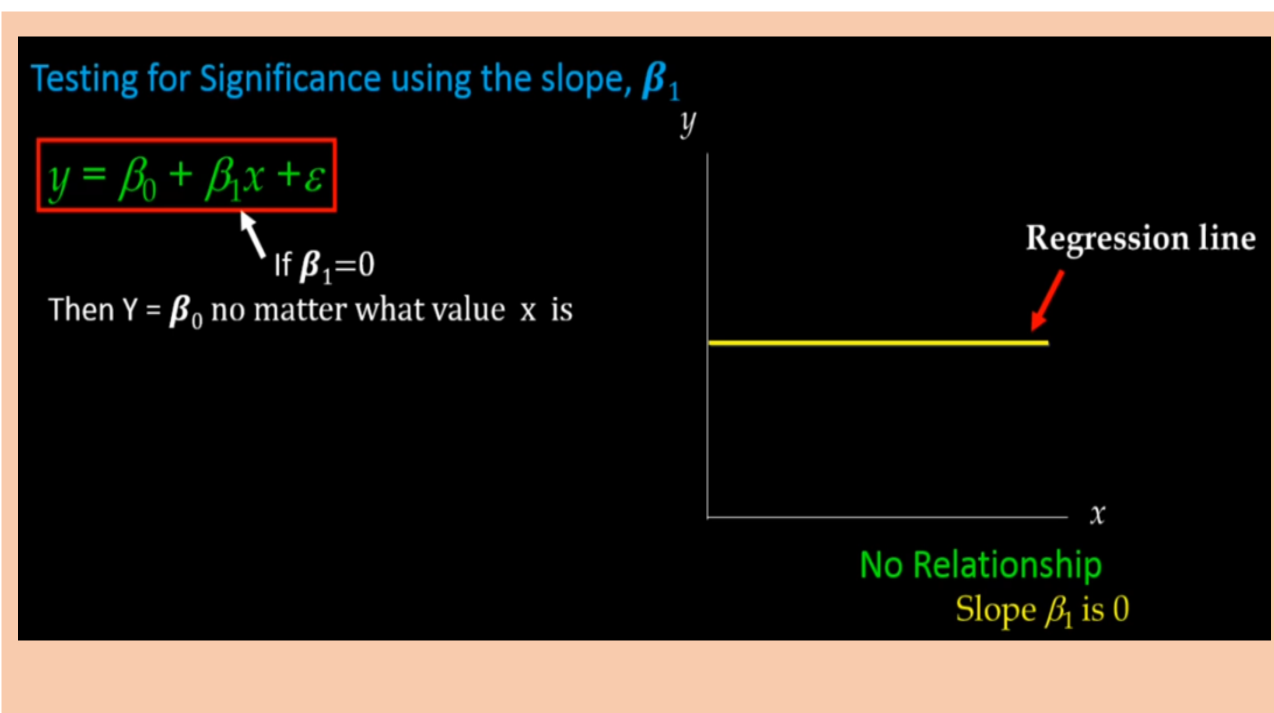
$$\hat{y} = 88.228$$

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

$$SSE = \sum(y_i - \hat{y}_i)^2$$




$\hat{y} = b_0 + b_1 x$
$\hat{y} = 55.048 + 4.74x$
$\bar{y}=778/10 = 77.8$
$r^2 = SSR/SST$
$SST = \sum(y_i - \bar{y})^2$
$\hat{y} = 55.048 + 4.74(7)$
$\hat{y} = 88.228$
$SSR = \sum(\hat{y}_i - \bar{y})^2$
$SSE = \sum(y_i - \hat{y}_i)^2$
r^2 = Explained Variation / Total Variation
Grades on Exam
10 20 30 40 50 60 70 80 90 100
Total Variation = SST
Unexplained Variation = SSE
Explained Variation = SSR
$\bar{y}=778/10 = 77.8$
$\hat{y} = 55.048 + 4.74x$
1 2 3 4 5 6 7 8 9 10 11
Hours Studied
r^2 = 95.05% of the variability in grades can be explained by the number of hours studied


Testing for Significance using the slope, β_1
$y = \beta_0 + \beta_1 x + \varepsilon$
If β_1=0
Then Y = β_0 no matter what value x is
y
Regression line
x
No Relationship
Slope β_1 is 0

Hypothesis Test of Significance, t test:

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

Test Statistic: $\boxed{t = \frac{b_1}{s_{b_1}}}$

Where: $s_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$

SSE =79.1215

$s = \sqrt{\frac{79.1215}{10-2}} = 3.1449$

And:

$s = \sqrt{\frac{\text{SSE}}{n-2}}$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
2	69	-2.8	-8.8	24.64	7.84
9	98	4.2	20.2	84.84	17.64
5	82	.2	4.2	.84	.04
5	77	.2	-.8	-.16	.04
3	71	-1.8	-6.8	12.24	3.24
7	84	2.2	6.2	13.64	4.84
1	55	-3.8	-22.8	86.64	14.44
8	94	3.2	16.2	51.84	10.24
6	84	1.2	6.2	7.44	1.44
2	64	-2.8	-13.8	38.64	7.84
$\Sigma x_i=48$	$\Sigma y_i=778$			320.6	67.6
$\bar{x}=48/10 = 4.8$	$\bar{y}=778/10 = 77.8$			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

Hypothesis Test of Significance, t test:

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

Test Statistic: $t = \dfrac{4.74}{.3825} = 12.3921$

Critical Value:

$\alpha = .01 \qquad \alpha/2 = .005$

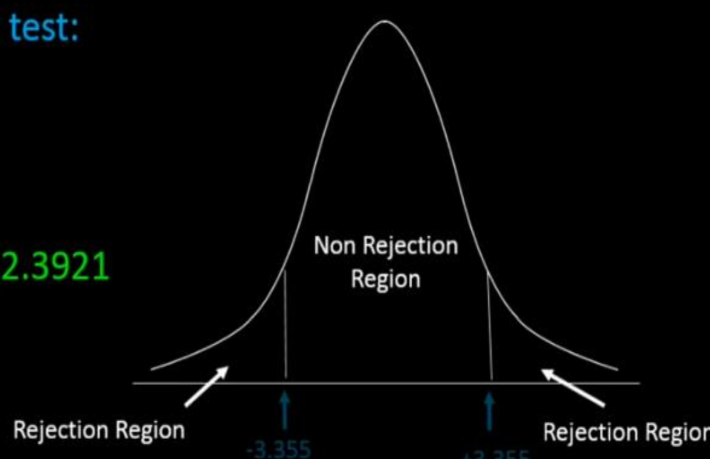
Hypothesis Test of Significance, t test:

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

Test Statistic: $t = \dfrac{4.74}{.3825} = 12.3921$

Critical Value: $= 3.355$



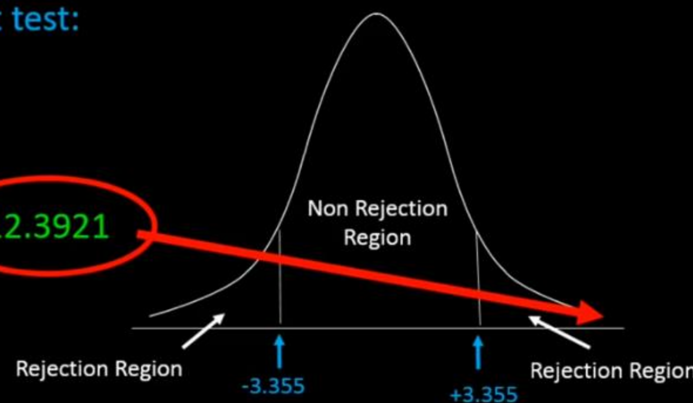
Hypothesis Test of Significance, t test:

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

Test Statistic: $t = \dfrac{4.74}{.3825}$ = 12.3921

Critical Value: $= 3.355$



Statistical Conclusion:

Reject H_0, there is evidence that β_1 is not equal to zero and that a significant relationship exists between grades and number of hours studied.

P value approach:

Test Statistic: = 12.3921

$\text{df} = \text{n-2} = 8$

For a Two-tailed Test: Double the area and compare to α

p-value =.0005 x 2 = .001

Rejection Rule:

Reject H_0 if p-value $\leq \alpha$

$\alpha = .01$

.001 < .01

Reject H_0, there is evidence that β_1 is not equal to zero and that a significant relationship exists between grades and number of hours studied.