Linear Regression

- a technique to find $y = ax + b$ for the line of best fit
- "Least Square Method"

slope $\longrightarrow$ $$a = \frac{n\sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

↙ mean of all y-values.

y-intercept $\longrightarrow$ $$b = \bar{y} - a\bar{x} \longleftarrow \text{mean of all } x\text{-values}$$

linear regression equation : $\boxed{y = ax + b}$

Covariance

- measure of correlation which reveals whether the relationship between the variables is positive or negative

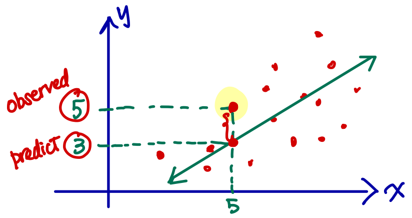
$$S_{xy} = \frac{1}{n-1}\left(\sum xy - \frac{\sum x \sum y}{n}\right)$$

Residual

- the difference between the values of y at the data point and at the point that lies on the line of best fit and has the same x-coordinate as the data point

$$Residual = \overset{\text{actual point}}{\underline{observed}} - \overset{\text{point of the linear regression}}{\underline{predict}}$$

positive = data above the line of best fit
negative = data below the line of best fit



when $\underline{\underline{x=5}}$, $\underline{\underline{\text{residual}}}$ is $\underline{\underline{5}}-\underline{\underline{3}}=2$ ✓

r-value = coefficient of correlation (for linear regression)
(line of best fit)

r^2 = Coefficient of Determination for (non-linear regression)
(curve of best fit)

- r^2 (the square of the correlation coefficient)
- it measures the proportion of the variation in y that is explained by the variation in x
- it tells what percentage of points that are on the line/curve of best fit
- $0 \leq r^2 \leq 1$, $r^2 = 1$ the curve is a perfect fit
- this applies to any type of regression

Linear Correlation and Regression Practice

Determine the linear regression and coefficient of correlation for the following examples.

Example #1

ıod"

x	y	x^2	y^2	xy
2	13	4	169	26
6	20	36	400	120
7	27	49	729	189
$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum xy$
15	60	89	1298	335

slope ⟶ $a = \dfrac{n\sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$

mean of all y-values.

y-intercept ⟶ $b = \bar{y} - a\bar{x}$ ⟵ mean of all x-values

linear regression equation : $\boxed{y = ax + b}$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

$$a = \frac{3(335) - (15)(60)}{3(89) - (15)^2}$$

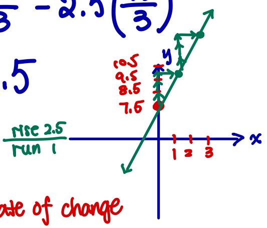
$= 2.5$

$$b = \frac{60}{3} - 2.5\left(\frac{15}{3}\right)$$

$= 7.5$

2.5y/x

∴ linear regression is $y = 2.5x + 7.5$



$$r = \frac{3(335) - (15)(60)}{\sqrt{(3(89) - (15)^2)(3(1298) - (60)^2)}}$$

$\doteq .945$ ∴ It is strong positive linear correlation

$5/3pumpkins

$1.67/pumkin

rate of change=slope

$m = \dfrac{y_2 - y_1}{x_2 - x_1}$ = rate of change

$\Delta x \rightarrow \Delta y$

For every 1 value of x increases, y-value will increase by 2.5 units.

Example #2:

Are the marks one receives in a course related to the amount of time spent studying the subject? To analyze this possibility, a student took a random sample of 10 students who had enrolled in an accounting class last semester. She asked each to report his or her mark in the course and the total number of hours spent studying accounting. These data are listed here.

Hours spent	40	42	37	48	25	44	41	48	35	28
Marks	77	63	79	86	51	78	83	90	65	47
x^2										
y^2										
xy										

$\sum x = 388$ $\sum y = 719$ $\sum x^2 = 15592$ $\sum y^2 = 53643$ $\sum xy = 28798$

a) Calculate the covariance.

$$S_{xy} = \frac{1}{n-1}\left(\sum xy - \frac{\sum x \sum y}{n}\right)$$

$$= \frac{1}{10-1}\left(28798 - \frac{(388)(719)}{10}\right)$$

$\doteq 100.089$

∴ there's a positive linear correlation.

b) Calculate the coefficient of correlation.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{10(28798) - (388)(719)}{\sqrt{(10(15592) - (388)^2)(10(53643) - (719)^2)}}$$

$\doteq 0.880$

∴ It is a strong positive linear correlation.

c) What do the statics you have calculated tell you about the relationship between marks and study times?

The r-value indicates a strong positive linear correlation between the two variables. It has suggested that the time spent in studying has a significant impact on the result as the more time spent on studying, the higher the marks can be achieved.

Example #3:

Least Squares Method for Finding Equation of a Line of Best Fit

Age (years) - (X)	Annual Income ($) - (Y)	x^2	y^2	xy
33	33			
25	31			
19	18			
44	52			
50	56			
54	60			
38	44			
29	35			
$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum xy$
292	329	11712	14975	13221

1) Substitute these totals into the formula for a.

$$a = \frac{n\sum xy-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2} = \frac{8(13221)-(292)(329)}{8(11712)-(292)^2}$$

$$= 1.150$$

2) To determine b, you also need the means of x and y.

$$b = \frac{329}{8} - a\left(\frac{292}{8}\right)$$

$$= -0.85$$

3) Now substitute the values of a and b into the equation:

$$y = 1.15x - 0.85$$

~~4) You can use the equation of the line of best fit as a model.~~

5) Predict the income for an employee who is 21 and an employee retiring at age 65.

a) For a 21 year old employee,

$$y = 1.15(21) - 0.85$$
$$= 23\,300$$

$\therefore$ a 21-year-old employee will earn approx. $23300.

b) For a 65 year old employee,

$$y = 1.15(65) - 0.85$$
$$= \$73900$$

$\therefore$ a 65-year-old employee will earn approx. $73900.

6) Determine which one is interpolation and which one is extrapolation from question #5.

Employee of 21-year-old is interpolation because the prediction is with the given interval of independent variable. Employee of 65-year-old is extrapolation because the prediction is out of the given interval of independent variable.

7) Comment on the accuracy of both estimates.

The model needs a restriction on the domain because children and retirees should be excluded.