Correlation and Regression

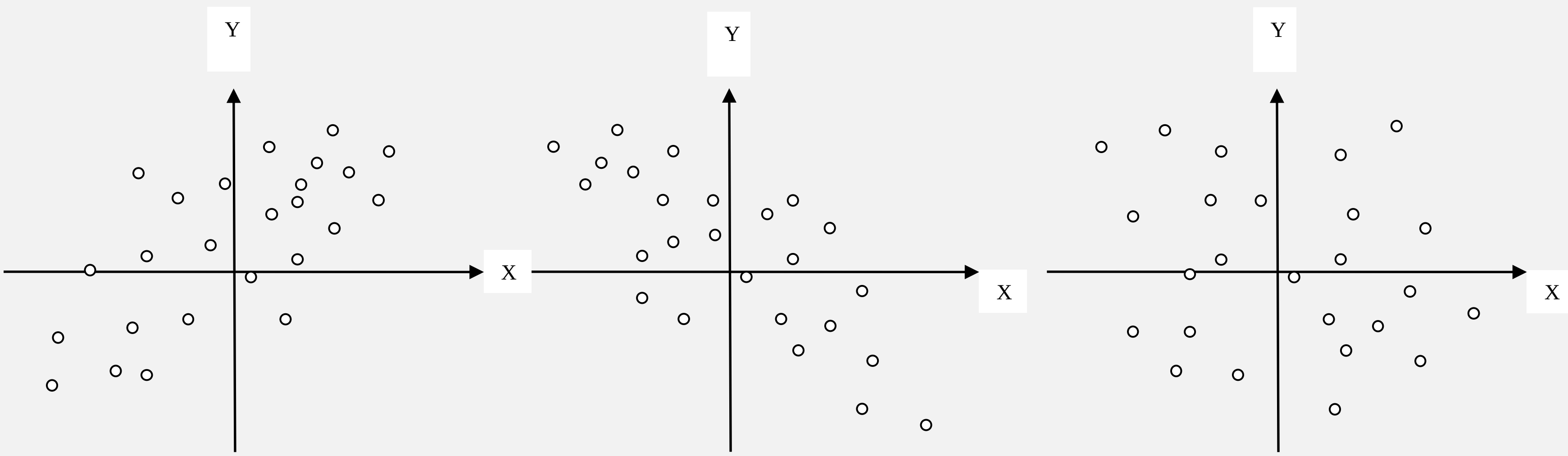
Topics Covered:

- Is there a relationship bivariate data(between *x* and *y)*?
- What is the strength of this relationship
  - Pearson's r
- Can we describe this relationship and use this to predict *y* from *x*?
  - Regression
- Is the relationship we have described statistically significant?
  - t-test

The relationship bivariate data between *x* and *y*

- Correlation: is there a relationship between 2 variables?
- Regression: how well a certain independent variable predict dependent variable?
- CORRELATION $\neq$ CAUSATION
  - In order to infer causality: manipulate independent variable and observe effect on dependent variable

Scattergrams



Positive correlation Negative correlation No correlation

Variance vs Covariance

- *First, a note on your sample:*
  - *If you're wishing to assume that your sample is representative of the general population (RANDOM EFFECTS MODEL), use the degrees of freedom (n – 1) in your calculations of variance or covariance.*
  - *But if you're simply wanting to assess your current sample (FIXED EFFECTS MODEL), substitute n for the degrees of freedom.*

Variance vs Covariance

- Do two variables change together?

Variance:

• Gives information on variability of a single variable.

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

Covariance:

• Gives information on the degree to which two variables vary together.

• Note how similar the covariance is to variance: the equation simply multiplies x's error scores by y's error scores as opposed to squaring x's error scores.

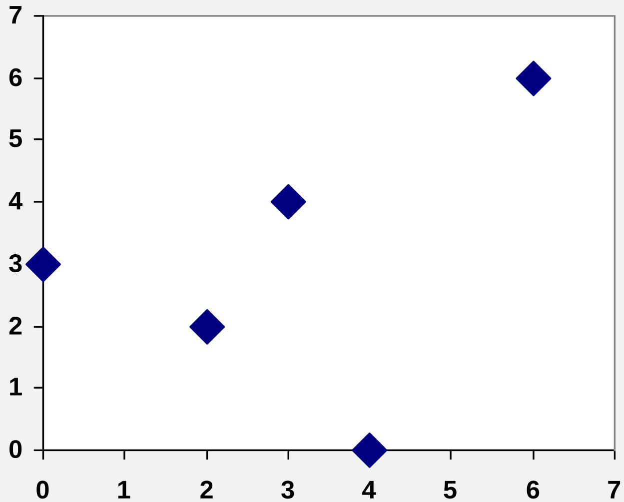
$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Covariance

$$\mathrm{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- When X $\uparrow$ and Y $\uparrow$: cov (x,y) = pos.
- When X $\downarrow$ and Y $\uparrow$: cov (x,y) = neg.
- When no constant relationship: cov (x,y) = 0

Example Covariance

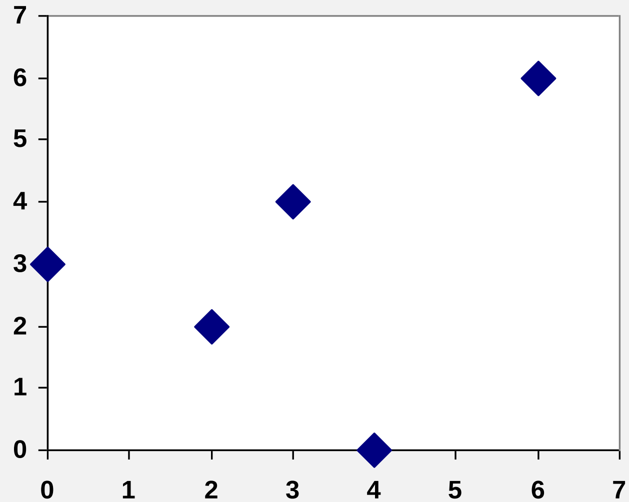


x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
0	3	-3	0	0
2	2	-1	-1	1
3	4	0	1	0
4	0	1	-3	-3
6	6	3	3	9
$\bar{x} = 3$	$\bar{y} = 3$			$\sum = 7$

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}))}{n-1} = \frac{7}{4} = 1.75$$

What does this number tell us?

Example Covariance



x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
0	3	-3	0	0
2	2	-1	-1	1
3	4	0	1	0
4	0	1	-3	-3
6	6	3	3	9
$\bar{x} = 3$	$\bar{y} = 3$			$\sum = 7$

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))}{n-1} = \frac{7}{4} = 1.75$$

What does this number tell us?

A positive covariance indicates that the two variables tend to move in the same direction, while a negative covariance indicates that they tend to move in opposite directions.

Problem with Covariance:

- The value obtained by covariance is dependent on the size of the data’s standard deviations: if large, the value will be greater than if small… *even if the relationship between x and y is exactly the same in the large versus small standard deviation datasets.*

Example of how covariance value relies on variance

	High variance data				Low variance data		
Subject	**x**	**y**	**x error * y error**		**x**	**y**	**X error * y error**
1	101	100	2500		54	53	9
2	81	80	900		53	52	4
3	61	60	100		52	51	1
4	51	50	0		51	50	0
5	41	40	100		50	49	1
6	21	20	900		49	48	4
7	1	0	2500		48	47	9
Mean	51	50			51	50	
Sum of x error * y error :			7000		**Sum of x error * y error :**		28
Covariance:			**1166.67**		**Covariance:**		**4.67**

Solution: Pearson’s r

- **Covariance does not really tell us anything**
    - ***Solution: standardise this measure***

- **Pearson’s R: standardises the covariance value.**
- **Divides the covariance by the multiplied standard deviations of X and Y:**

$$r_{xy} = \frac{cov(x,y)}{s_x s_y}$$

Pearson’s R continued

Formula for Correlation Coefficient

Population Correlation Coefficient

$$P_{xy} = \frac{\sigma_{xy}}{\sigma_x \; \sigma_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right)}}$$

Where,

$\sigma_x , \sigma_y \rightarrow$ Population Standard Deviation

$\sigma_{xy} \rightarrow$ Population Covariance

$\bar{X}, \bar{Y} \rightarrow$ Population Mean

Sample Correction, coefficient between x and y

$$r_{xy} = \frac{S_{xy}}{S_x \; S_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right.}}$$

Where,

$S_x , S_y \rightarrow$ Sample Standard Deviation

$S_{xy} \rightarrow$ Sample Covariance

$\bar{X}, \bar{Y} \rightarrow$ SampleMean

Sample Correlation Coefficient

The formula for pearson correlation coefficient for population of size **N** (written as $\boldsymbol{\rho_{X,Y}}$) is given as:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where **cov** is the covariance and (cov(X,Y)= $\frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N}$, $\boldsymbol{\sigma_X}$ is standard deviation of X and $\boldsymbol{\sigma_Y}$ is standard deviation of Y.

Given X and Y are two random variables.

Population Correlation Coefficient

The formula for pearson correlation coefficient for sample of size **n** (written as $\mathbf{r_{xy}}$) is given as:

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

where **n** is the sample size, $\mathbf{x_i}$ & $\mathbf{y_i}$ are the i^{th} sample points and $\bar{\mathbf{x}}$ & $\bar{\mathbf{y}}$ are the sample means for the random variables X and Y respectively.

Given X and Y are two random variables.

Linear Correlation Coefficient

It uses pearson's correlation coefficient to determine the linear relationship between two variables. Its value lies between -1 and 1. It is given as:

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

where **n** is the sample size, $\mathbf{x_i}$ & $\mathbf{y_i}$ are the i^{th} sample points and $\bar{\mathbf{x}}$ & $\bar{\mathbf{y}}$ are the sample means for the random variables x and y respectively.

Linear Correlation Coefficient

The sign of r indicates the strength of the linear relationship between the variables.

•If r is near 1, then the two variables have a strong linear relationship.

•If r is near 0, then the two variables have no linear relation.

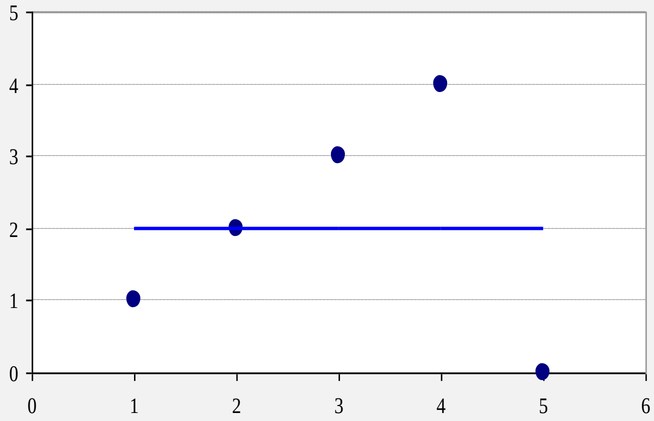
•If r is near -1, then the two variables have a weak (negative) linear relationship.

Limitations of r

- When r = 1 or r = -1:
  - We can predict y from x with certainty
  - all data points are on a straight line: y = ax + b
- r is actually $\hat{r}$

  r = true r of whole population

  $\hat{r}$ = estimate of r based on data
- r is very sensitive to extreme values:



Pearson's R Example

Calculate the Correlation coefficient of given data.

x	41	42	43	44	45
y	3.2	3.3	3.4	3.5	3.6

Pearson's R Example

Solution:

Here n = 5

Let us find $\sum x$, $\sum y$, $\sum xy$, $\sum x^2$, $\sum y^2$

x	y	xy	x^2	y^2
41	3.2	131.2	1681	10.24
42	3.3	138.6	1764	10.89
43	3.4	146.2	1849	11.56
44	3.5	154	1936	12.25
45	3.6	162	2025	12.96
$\sum x = 215$	**$\sum y = 17$**	**$\sum xy = 732$**	**$\sum x^2 = 9255$**	**$\sum y^2 = 57.9$**

R calculation:

$r = \sum((x - \bar{x})(y - \bar{y}))/\sqrt{((\sigma\sigma_x)(\sigma\sigma_y))}$

$r = 1/\sqrt{((10)(0.1))} = 1$

Since r = 1, this indicates significant relation between x and y.

X values

$\sum x = 215$

$\sum x^2 = 9255$

$\bar{x} = 43$

$\sum(x - \bar{x})^2 = \sigma\sigma_x = 10$

Y values:

$\sum y = 17$

$\sum y^2 = 57.9$

$\sum(y - \bar{y})^2 = \sigma\sigma_y = 0.1$

X and Y combined

N = 5

$\sum((x - \bar{x})(y - \bar{y})) = 1$

$\sum xy = 732$

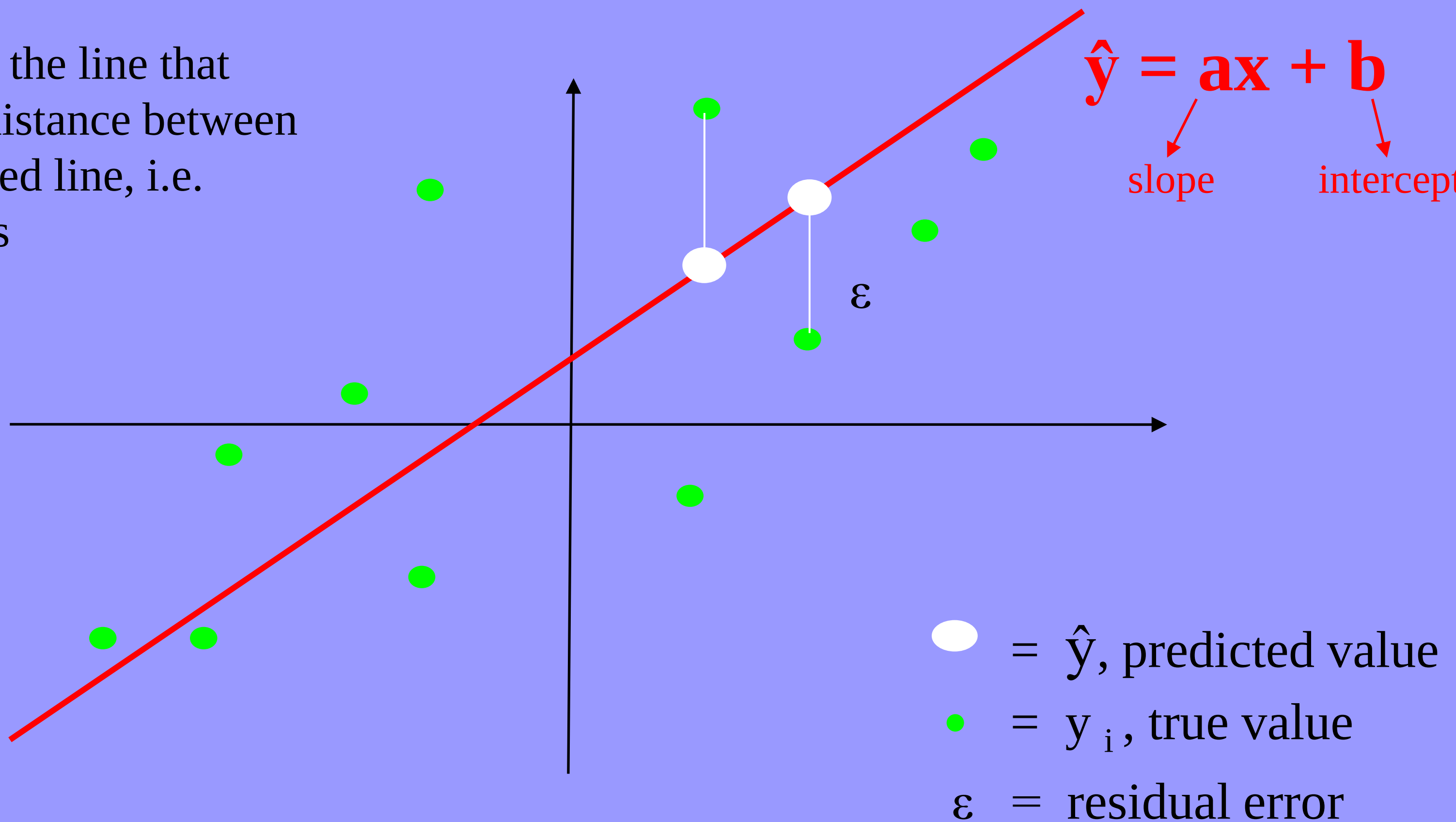
Regression

- Correlation tells you if there is an association between x and y but it doesn't describe the relationship or allow you to predict one variable from the other.

- To do this we need REGRESSION!

Best-fit Line

- Aim of linear regression is to fit a straight line, $\hat{y} = ax + b$, to data that gives best prediction of y for any value of x
- This will be the line that minimises distance between data and fitted line, i.e. the residuals



Least Squares Regression

- To find the best line we must minimise the sum of the squares of the residuals (the vertical distances from the data points to our line)

Model line: $\hat{y} = ax + b$ a = slope, b = intercept

Residual (ε) = $y - \hat{y}$

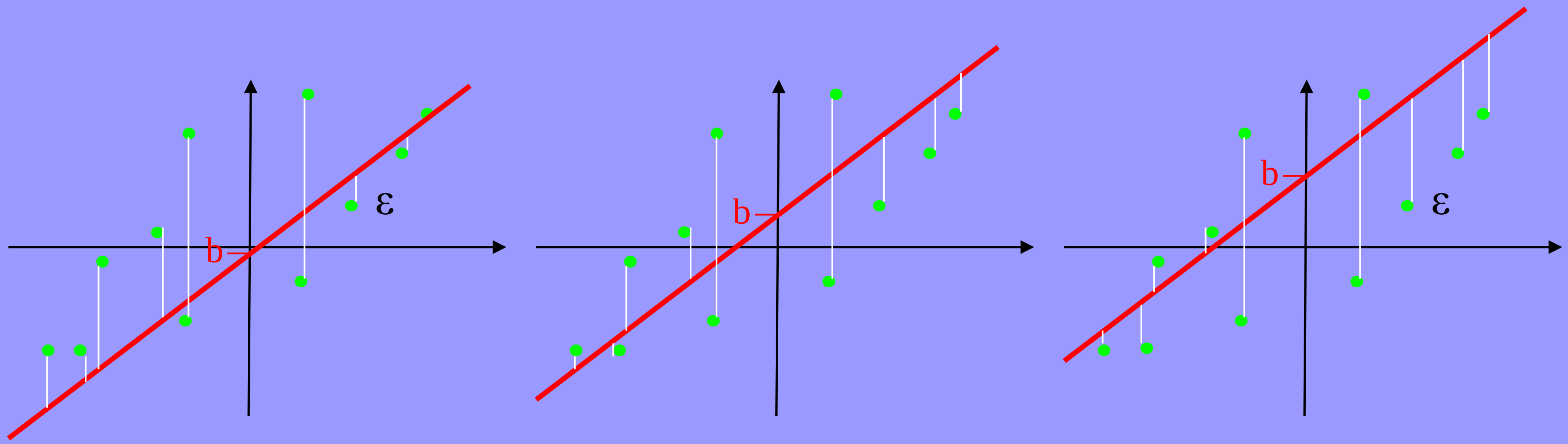
Sum of squares of residuals = $\Sigma (y - \hat{y})^2$

- we must find values of a and b that minimise

$$\Sigma (y - \hat{y})^2$$

Finding b

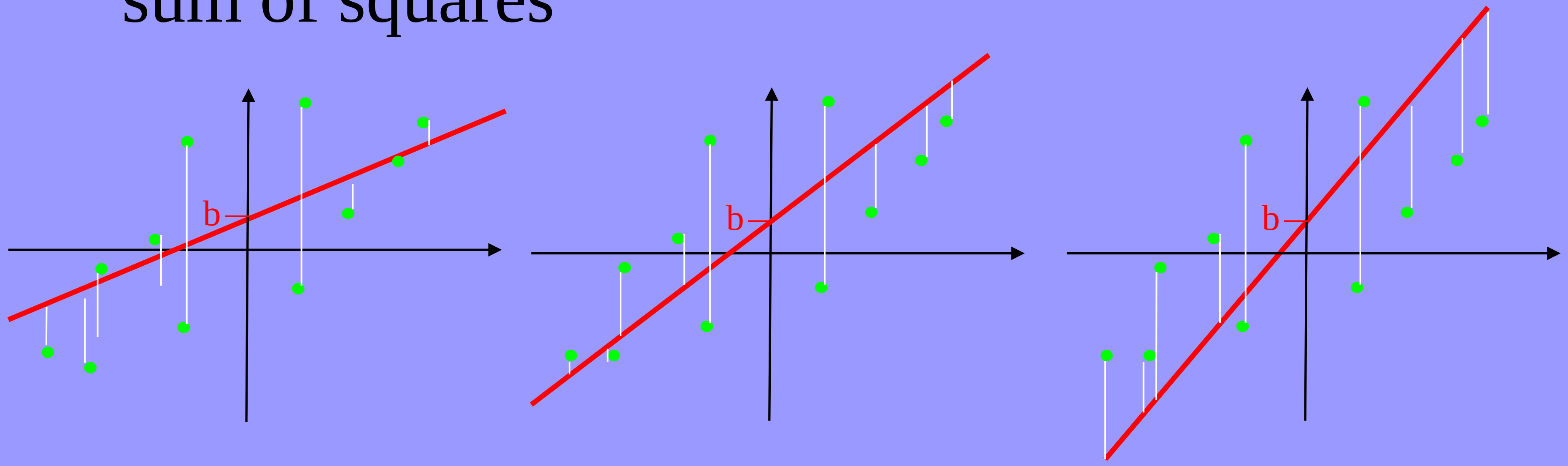
- First we find the value of b that gives the min sum of squares



- Trying different values of b is equivalent to shifting the line up and down the scatter plot

Finding a

- Now we find the value of a that gives the min sum of squares



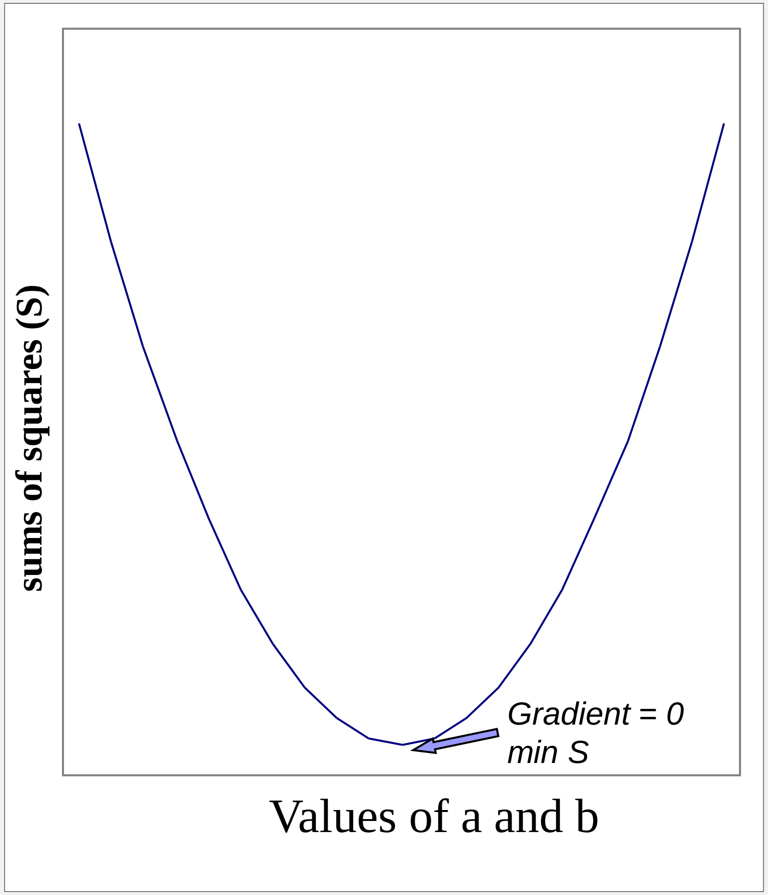
- Trying out different values of a is equivalent to changing the slope of the line, while b stays constant

Minimising sums of squares

- Need to minimise $\Sigma(y - \hat{y})^2$
- $\hat{y} = ax + b$
- so need to minimise:

$$\Sigma(y - ax - b)^2$$

- If we plot the sums of squares for all different values of a and b we get a parabola, because it is a squared term
- So the min sum of squares is at the bottom of the curve, where the gradient is zero.



The maths bit

- The min sum of squares is at the bottom of the curve where the gradient = 0

- So we can find a and b that give min sum of squares by taking partial derivatives of $\Sigma(y - ax - b)^2$ with respect to a and b separately

- Then we solve these for 0 to give us the values of a and b that give the min sum of squares

The solution

- Doing this gives the following equations for a and b:

$$a = \frac{r\, s_y}{s_x}$$

r = correlation coefficient of x and y
s_y = standard deviation of y
s_x = standard deviation of x

- From you can see that:
  - A low correlation coefficient gives a flatter slope (small value of a)
  - Large spread of y, i.e. high standard deviation, results in a steeper slope (high value of a)
  - Large spread of x, i.e. high standard deviation, results in a flatter slope (high value of a)

The solution cont.

- Our model equation is $\hat{y} = ax + b$
- This line must pass through the mean so:

$$\bar{y} = a\bar{x} + b \quad \Rightarrow \quad b = \bar{y} - a\bar{x}$$

- We can put our equation for a into this giving:

$$\mathbf{b = \bar{y} - \frac{r\, s_y}{s_x}\,\bar{x}}$$

r = correlation coefficient of x and y
s_y = standard deviation of y
s_x = standard deviation of x

- The smaller the correlation, the closer the intercept is to the mean of y

Back to the model

$$\hat{y} = ax + b = \underbrace{\frac{r\,s_y}{s_x}}_{a}\,x + \underbrace{\bar{y} - \underbrace{\frac{r\,s_y}{s_x}}_{a}\,\bar{x}}_{b}$$

Rearranges to: $$\hat{y} = \underbrace{\frac{r\,s_y}{s_x}}_{a}\,(x - \bar{x}) + \bar{y}$$

- If the correlation is zero, we will simply predict the mean of y for every value of x, and our regression line is just a flat straight line crossing the x-axis at y
- But this isn't very useful.
- We can calculate the regression line for any data, but the important question is how well does this line fit the data, or how good is it at predicting y from x

How good is our model?

- Total variance of y: $s_y^2 = \dfrac{\sum(y - \bar{y})^2}{n - 1} = \dfrac{SS_y}{df_y}$

- Variance of predicted y values (ŷ):

$$s_{\hat{y}}^2 = \frac{\sum(\hat{y} - \bar{y})^2}{n - 1} = \frac{SS_{pred}}{df_{\hat{y}}}$$

This is the variance explained by our regression model

- Error variance:

$$s_{error}^2 = \frac{\sum(y - \hat{y})^2}{n - 2} = \frac{SS_{er}}{df_{er}}$$

This is the variance of the error between our predicted y values and the actual y values, and thus is the variance in y that is NOT explained by the regression model

How good is our model cont.

- Total variance = predicted variance + error variance

$$s_y^2 = s_{\hat{y}}^2 + s_{er}^2$$

- Conveniently, via some complicated rearranging

$$s_{\hat{y}}^2 = r^2 s_y^2$$

$$\Downarrow$$

$$r^2 = s_{\hat{y}}^2 / s_y^2$$

- so r^2 is the proportion of the variance in y that is explained by our regression model

How good is our model cont.

- Insert $r^2 s_y^2$ into $s_y^2 = s_{\hat{y}}^2 + s_{er}^2$ and rearrange to get:

$$
\begin{aligned}
s_{er}^2 &= s_y^2 - r^2 s_y^2 \\
&= s_y^2 (1 - r^2)
\end{aligned}
$$

- From this we can see that the greater the correlation the smaller the error variance, so the better our prediction

Is the model significant?

- i.e. do we get a significantly better prediction of y from our regression equation than by just predicting the mean?

- F-statistic:

complicated rearranging

$$F_{(df_{\hat{y}},df_{er})} = \frac{s_{\hat{y}}^{2}}{s_{er}^{2}} = \ldots\ldots = \frac{r^2 (n-2)^2}{1 - r^2}$$

- And it follows that:

(because $F = t^2$)

$$t_{(n-2)} = \frac{r\,(n-2)}{\sqrt{1 - r^2}}$$

So all we need to know are r and n

General Linear Model

- Linear regression is actually a form of the General Linear Model where the parameters are a, the slope of the line, and b, the intercept.

$$y = ax + b + \varepsilon$$

- A General Linear Model is just any model that describes the data in terms of a straight line

Multiple regression

- Multiple regression is used to determine the effect of a number of independent variables, x_1, x_2, x_3 etc, on a single dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = a_1x_1 + a_2x_2 + \ldots\ldots + a_nx_n + b + \varepsilon$$

- The a parameters reflect the independent contribution of each independent variable, x, to the value of the dependent variable, y.
- i.e. the amount of variance in y that is accounted for by each x variable after all the other x variables have been accounted for

Calculate the regression coefficient and obtain the lines of regression for the following data

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

Solution:

X	Y	X^2	Y^2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
$\sum X = 28$	$\sum Y = 77$	$\sum X^2 = 140$	$\sum Y^2 = 875$	$\sum XY = 334$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{28}{7} = 4,$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{77}{7} = 11$$

Regression coefficient of *X* on *Y*

$$b_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\,\Sigma Y^2 - (\Sigma Y)^2}$$

$$= \frac{7(334) - (28)(77)}{7(875) - (77)^2}$$

$$= \frac{2338 - 2156}{6125 - 5929}$$

$$= \frac{182}{196}$$

$$b_{xy} = 0.929$$

Regression equation of *X* on *Y*

$$X - \overline{X} = b_{xy}(Y - \overline{Y})$$

$$X - 4 = 0.929(Y - 11)$$

$$X - 4 = 0.929Y - 10.219$$

$\therefore$ The regression equation X on Y is $X = 0.929Y - 6.219$

Calculate the regression coefficient and obtain the lines of regression for the following data

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

Solution:

X	Y	X^2	Y^2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
$\sum X = 28$	$\sum Y = 77$	$\sum X^2 = 140$	$\sum Y^2 = 875$	$\sum XY = 334$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{28}{7} = 4,$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{77}{7} = 11$$

Regression coefficient *of Y* on *X*

$$b_{yx} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\,\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{7(334) - (28)(77)}{7(140) - (28)^2}$$

$$= \frac{2338 - 2156}{980 - 784}$$

$$= \frac{182}{196}$$

$$\therefore \quad b_{yx} = 0.929$$

Regression equation of *Y* on *X*

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

$$Y - 11 = 0.929\,(X - 4)$$

$Y = 0.929X - 3.716 + 11$

$= 0.929X + 7.284$

The regression equation of Y on X is $Y = 0.929X + 7.284$

Calculate the two regression equations of X on Y and Y on X from the data given below, taking deviations from a actual means of X and Y.

Price(Rs.)	10	12	13	12	16	15
Amount demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs.20.

Solution:

X	$x = (X - 13)$	x^2	Y	$y = (Y - 41)$	y^2	xy
10	–3	9	40	–1	1	3
12	–1	1	38	–3	9	3
13	0	0	43	2	4	0
12	–1	1	45	4	16	–4
16	3	9	37	–4	16	–12
15	2	4	43	2	4	4
$\sum X = 78$	$\sum x = 0$	$\sum x^2 = 24$	$\sum Y = 246$	$\sum y = 0$	$\sum y^2 = 50$	$\sum xy = -6$

Calculate the two regression equations of X on Y and Y on X from the data given below, taking deviations from a actual means of X and Y.

Price(Rs.)	10	12	13	12	16	15
Amount demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs.20.

Solution:

Regression equation of X on Y

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

$$\overline{X} = \frac{78}{6} = 13\ ,\ \overline{Y} = \frac{246}{6} = 41$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-6}{50} = -0.12$$

$$X - 13 = -0.12\,(Y - 41)$$

$$X - 13 = -0.12Y + 4.92$$

$$X = -0.12Y + 17.92$$

Regression Equation of Y on X

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = -\frac{6}{24} = -0.25$$

$$Y - 41 = -0.25\,(X - 13)$$

$$Y - 41 = -0.25X + 3.25$$

$$Y = -0.25X + 44.25$$

When X is 20, Y will be

= –0.25 (20)+44.25

= –5+44.25

= 39.25 (when the price is Rs. 20, the likely demand is 39.25)

The following table shows the sales and advertisement expenditure of a form

	Sales	Advertisement expenditure (Rs. Crores)
Mean	40	6
SD	10	1.5

Coefficient of correlation r= 0.9. Estimate the likely sales for a proposed advertisement expenditure of Rs. 10 crores.

Solution:

Given $\overline{X}=40, \overline{Y}=6, \sigma_x=10, \sigma_y=1.5$ and $r=0.9$

Equation of line of regression x on y is

$$X-\overline{X} = r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

$$X-40 = (0.9)\frac{10}{1.5}(Y-6)$$

$$X-40 = 6Y-36$$

$$X = 6Y+4$$

When advertisement expenditure is 10 crores i.e., Y=10 then sales X=6(10)+4=64 which implies sales is 64.

The two regression lines are $3X+2Y=26$ and $6X+3Y=31$. Find the correlation coefficient.

Solution:

Let the regression equation of Y on X be $3X+2Y = 26$

$$3X+2Y = 26$$

$$2Y = -3X+26$$

$$Y = \frac{1}{2}(-3X+26)$$

$$Y = -1.5X+13$$

$$r\frac{\sigma_y}{\sigma_x} = -1.5$$

Implies $\quad b_{yx} = r\frac{\sigma_y}{\sigma_x} = -1.5$

Let the regression equation of X on Y be

$$6X+3Y = 31$$

$$X = \frac{1}{6}(-3Y+31) = -0.5Y+5.17$$

$$r\frac{\sigma_x}{\sigma_y} = -0.5$$

Implies $\quad b_{xy} = r\frac{\sigma_x}{\sigma_y} = -0.5$

$$r = \pm\sqrt{b_{xy}.b_{yx}}$$

$= -\sqrt{(-1.5).(-0.5)}$ (Since both the regression coefficient are negative r is negative)

$\therefore \quad r = -0.866$

In a laboratory experiment on correlation research study the equation of the two regression lines were found to be $2X-Y+1=0$ and $3X-2Y+7=0$. Find the means of X and Y. Also work out the values of the regression coefficient and correlation between the two variables X and Y.

Solution:

Solving the two regression equations we get mean values of X and Y

$$2X-Y = -1 \qquad \ldots (1)$$

$$3X-2Y = -7 \qquad \ldots (2)$$

Solving equation (1) and equation (2) We get $X=5$ and $Y=11$

Therefore the regression line passing through the means $\overline{X}=5$ and $\overline{Y}=11$

The regression equation of Y on X is $3X-2Y=-7$

$$2Y = 3X+7$$

$$Y = \frac{1}{2}(3X+7)$$

$$Y = \frac{3}{2}X+\frac{7}{2}$$

$$\therefore \qquad b_{yx} = \frac{3}{2}\ (>1)$$

In a laboratory experiment on correlation research study the equation of the two regression lines were found to be $2X–Y+1=0$ and $3X–2Y+7=0$. Find the means of X and Y. Also work out the values of the regression coefficient and correlation between the two variables X and Y.

Solution:

The regression equation of X on Y is

$$2X-Y = -1$$

$$2X = Y-1$$

$$X = \frac{1}{2}(Y-1)$$

$$X = \frac{1}{2}Y-\frac{1}{2}$$

$$\therefore \quad b_{xy} = \frac{1}{2}$$

The regression coefficients are positive

$$r = \pm\sqrt{b_{xy}.b_{yx}} = \pm\sqrt{\frac{3}{2}\times\frac{1}{2}}$$

$$= \sqrt{\frac{3}{2}\times\frac{1}{2}}$$

$$= \sqrt{\frac{3}{4}}$$

$$= 0.866$$

$$r = 0.866$$