Application of Statistics in Educational Research I

MPU1034

REGRESSION*

Prof. Dr. Mohd Salleh Abu
Dr. Hamidreza Kashefi

main source: Vernoy & Vernoy (1997)

Some Commonly Used Jargons…

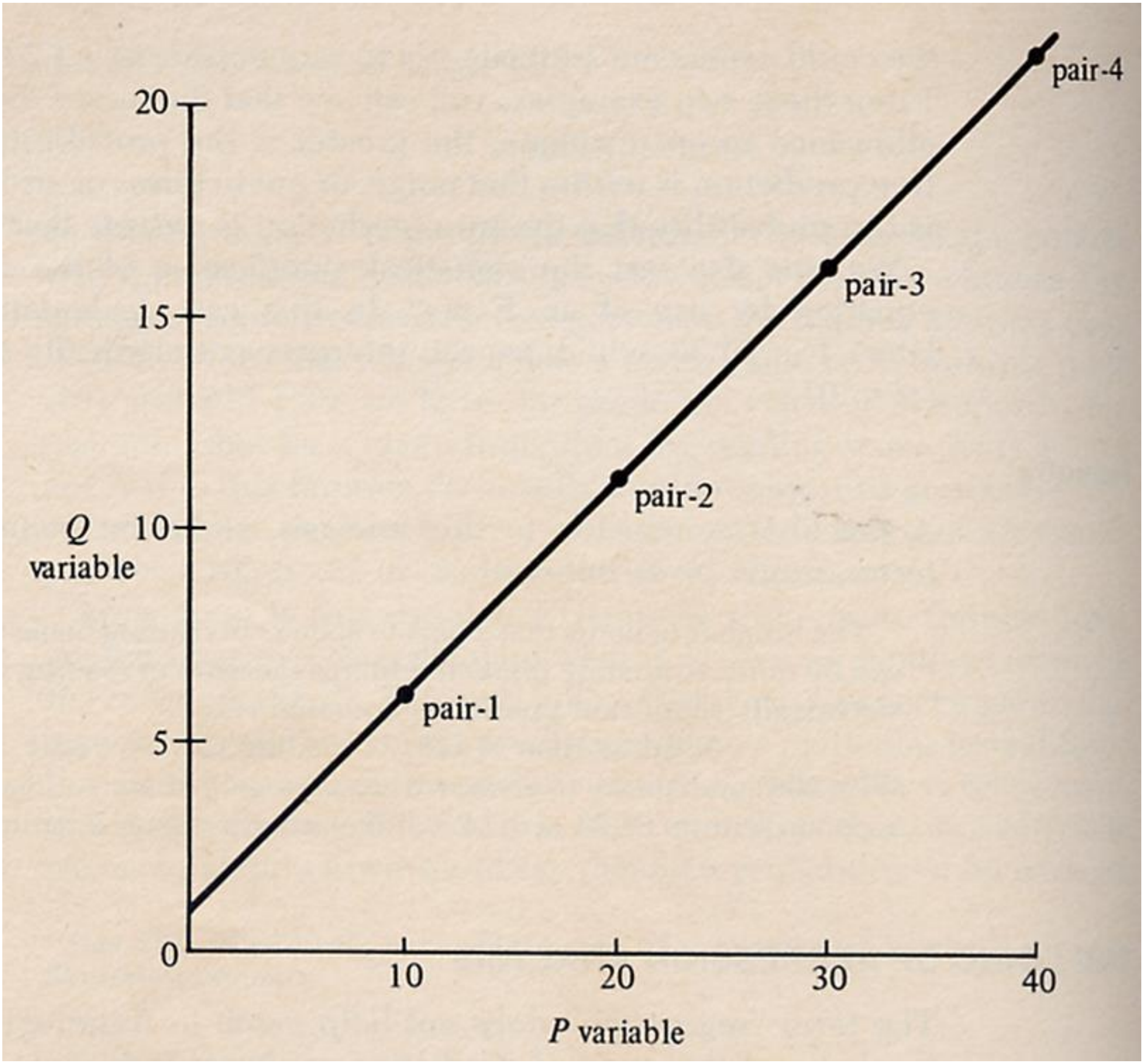
- Linear Regression
- Line of Best Fit
- Regression Equation
- Standard Error of Estimates

The General Idea About Regression

Suppose we were asked to investigate the relationship between two variables namely Variable P (being the independent) and Variable Q (being the dependent):

Pair	Variable P	Variable Q
Pair 1	10	7
Pair 2	20	12
Pair 3	30	17
Pair 4	40	22

What would be the predicted value of Q if P = 15?   If P = 25? How do you predict these?



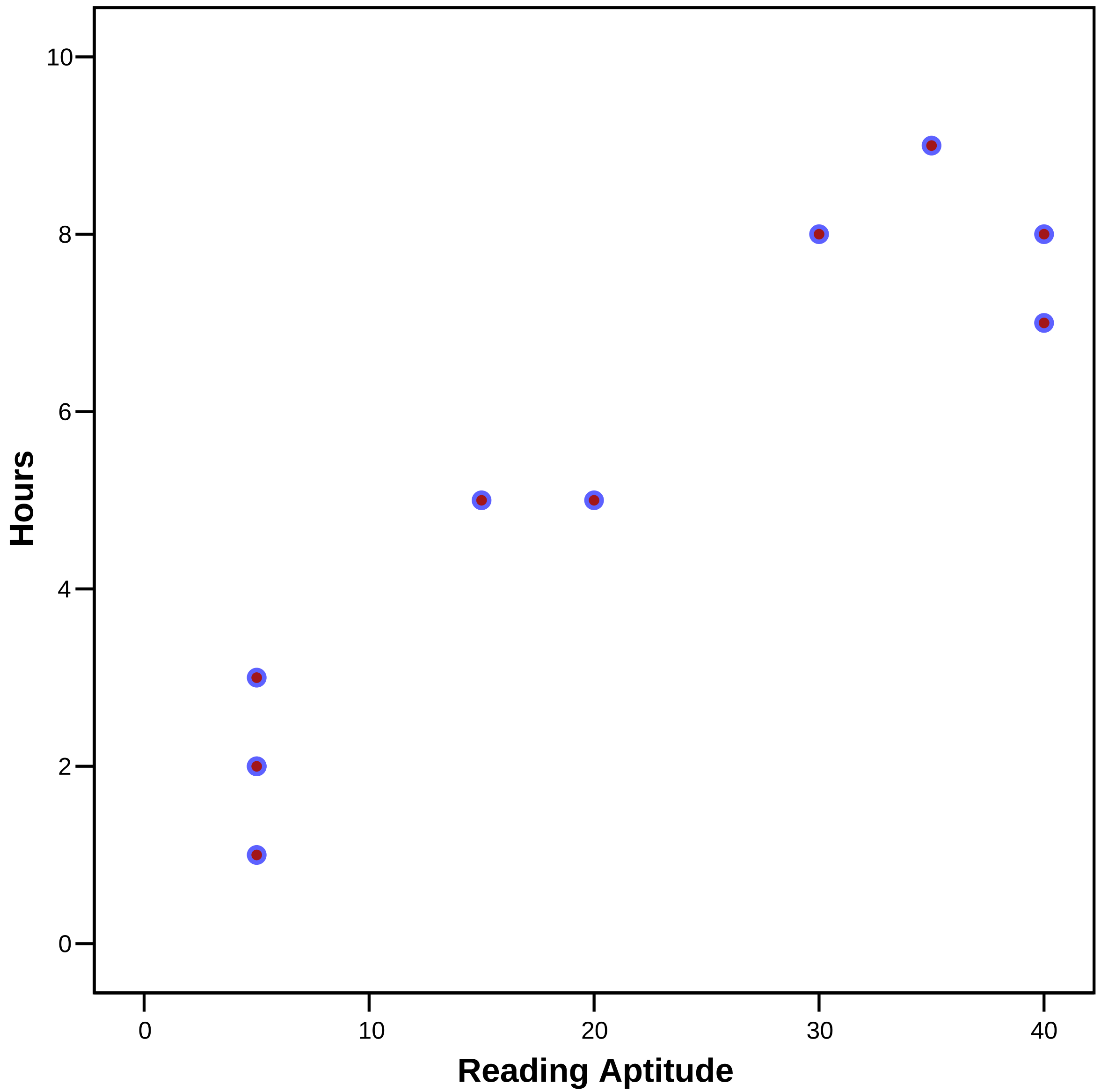
20
15
10
5
0
Q
variable
10
20
30
40
P variable
pair-1
pair-2
pair-3
pair-4

- Notice that if we connect these points, we would get a straight line. This line fits ALL the ‘observed’ points.
- This straight line is called the line of best fit or regression line
- The line of best fit defines a basis for <u>predicting</u> values of Q, given values of P (and vice versa)
- The concept of the line of best fit can be extended to form a basis for linear regression as well as non-linear regression (see the following examples)

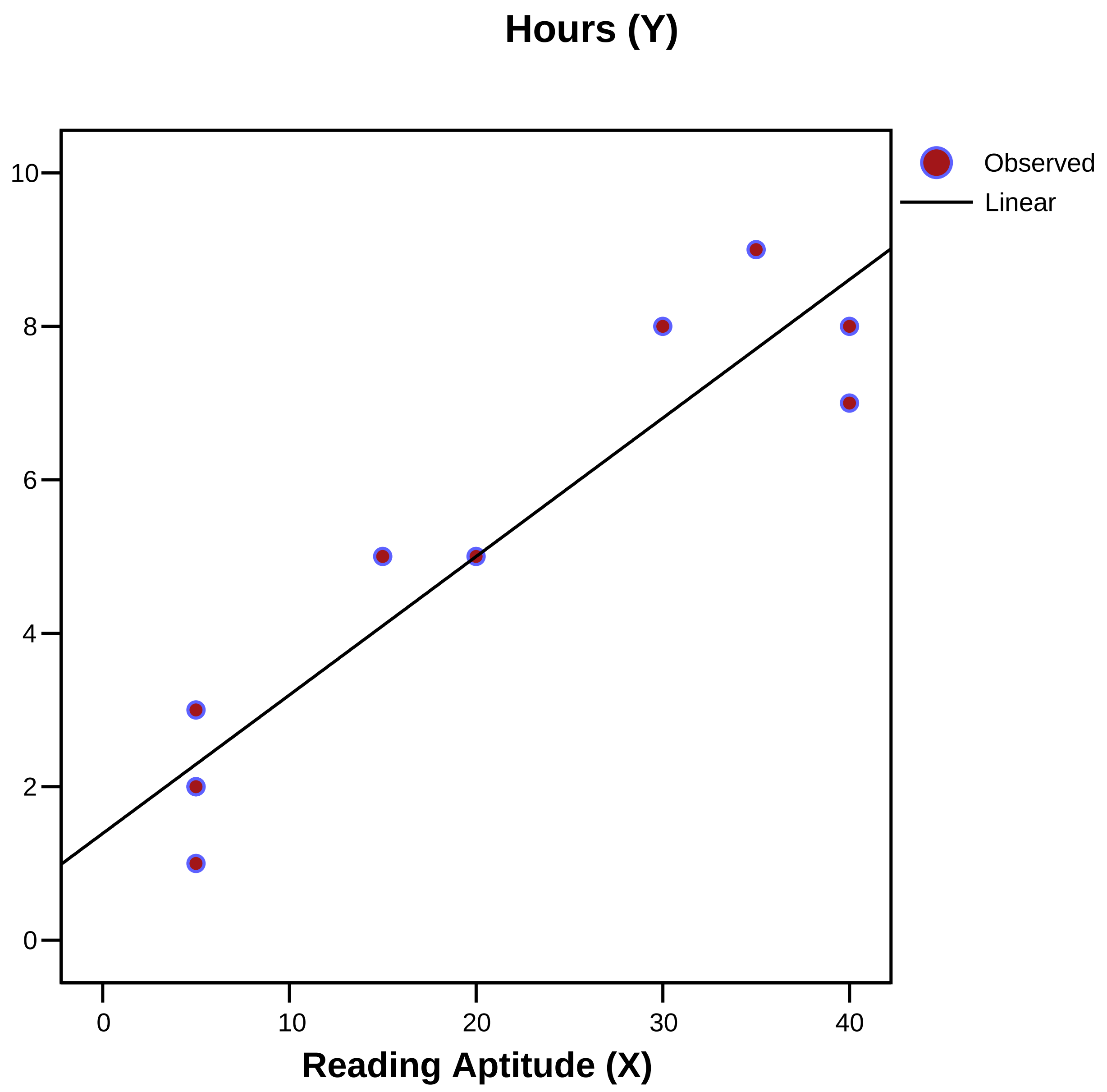
A researcher investigates the relationship between individual's score on a Reading Aptitude Test and the average amount of hours he/she spends for reading (simply called Hours). The data gathered from 10 students are as follows:

Student	Score on Reading Aptitude Test (X)	Hours (Y)
S1	20	5
S2	5	1
S3	5	2
S4	40	7
S5	30	8
S6	35	9
S7	5	3
S8	5	2
S9	15	5
S10	40	8

What would be the predicted number of hours spent by a student with a Reading Aptitude score of 27? How do you predict this?

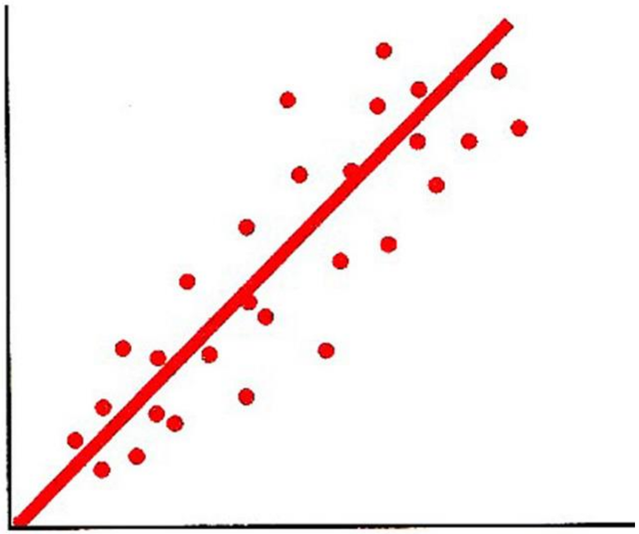


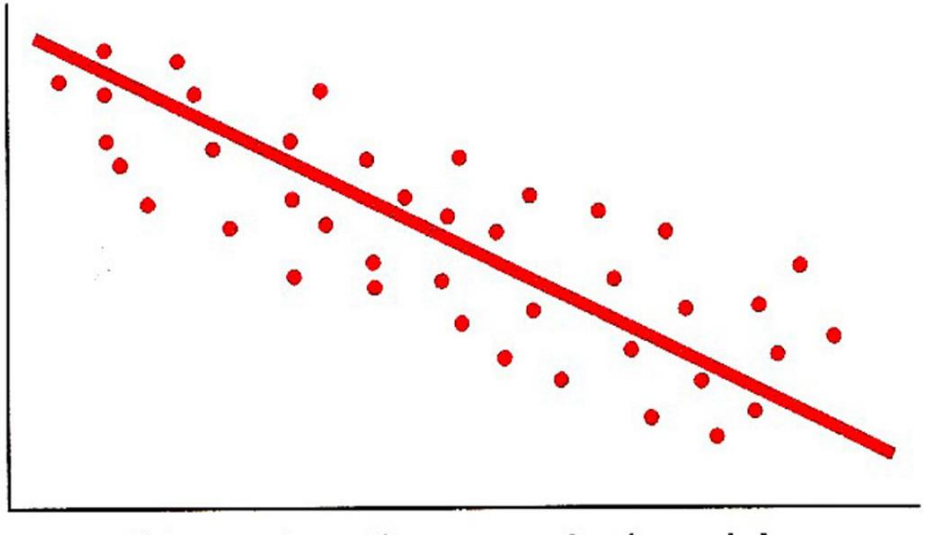
Hours
Reading Aptitude
0
2
4
6
8
10
0
10
20
30
40



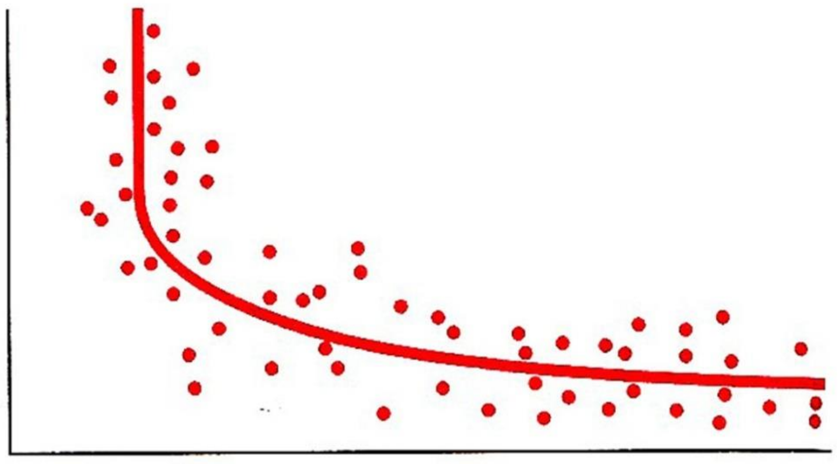
Hours (Y)
10
8
6
4
2
0
0
10
20
30
40
Reading Aptitude (X)
Observed
Linear

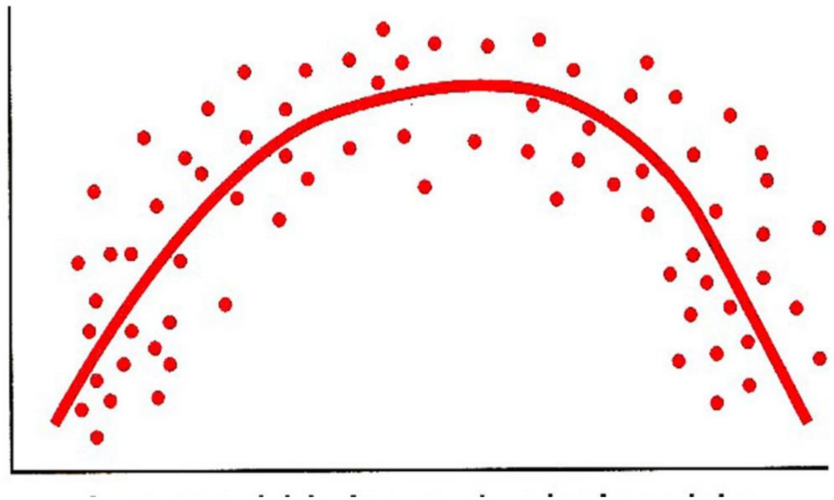
- Notice that in this case, we cannot connect these points with a straight line. This is because the relation between X and Y values is not a perfect positive relationship (in fact, *r* = 0.94).
- However, the concept of the line of best fit can be employed with some adjustments; draw a regression line such that, given the observed relation between X and Y, we can attempt to make predictions of Y which, although not perfect, would involved the least degree of prediction error (i.e. this regression line is a compromise in getting the line of best fit. It may not pass through ANY of the 'observed' points!)
- The procedure which produces a linear regression line (which provides us with a basis for the best prediction) is called linear regression.
- In using this type of prediction, we make a vital assumption: that the variables used to make the prediction are linearly related





Linear Regression





Non-Linear Regression

LINEAR REGRESSION

Eg.1

A researcher investigates the relationship between individual's score on a Reading Aptitude Test and the average amount of hours he/she spends spend for reading (simply called Hours). The data gathered from 10 students are as follows:

Student	**Score on Reading Aptitude Test (X)**	**Hours (Y)**
S1	20	5
S2	5	1
S3	5	2
S4	40	7
S5	30	8
S6	35	9
S7	5	3
S8	5	2
S9	15	5
S10	40	8

Find the linear regression equation for Y and sketch the curve.

The Regression Equation

The linear regression equation is given by

$$\hat{Y} = a_{YX} + b_{YX} X$$

where

X is the score of independent variable

$\hat{Y}$ is the predicted value of dependent variable with respect to X value

$$b_{YX} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

$$a_{YX} = \bar{Y} - b_{YX} \bar{X}$$

$\bar{X}$ is the mean of X

$\bar{Y}$ is the mean of Y

n is the number of pairs

	X	Y	X^2	XY
	20	5	400	100
	5	1	25	5
	5	2	25	10
	40	7	1600	280
	30	8	900	240
	35	9	1225	315
	5	3	25	15
	5	2	25	10
	15	5	225	75
	40	8	1600	320
Σ	200	50	6050	1370

$$b_{YX} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

$$= \frac{1370 - \frac{(200)(50)}{10}}{6050 - \frac{(200)^2}{10}}$$

$$= \frac{1370 - 1000}{6050 - 4000}$$

$$= 0.180$$

With $\bar{X} = \frac{\sum X}{n} = 5.0$ and $\bar{Y} = \frac{\sum Y}{n} = 20$

$$a_{YX} = \bar{Y} - b_{YX}\bar{X}$$

$$= 5.0 - 0.180(20)$$

$$= 1.40$$

The linear regression equation is

$$\hat{Y} = 1.40 + 0.18X$$

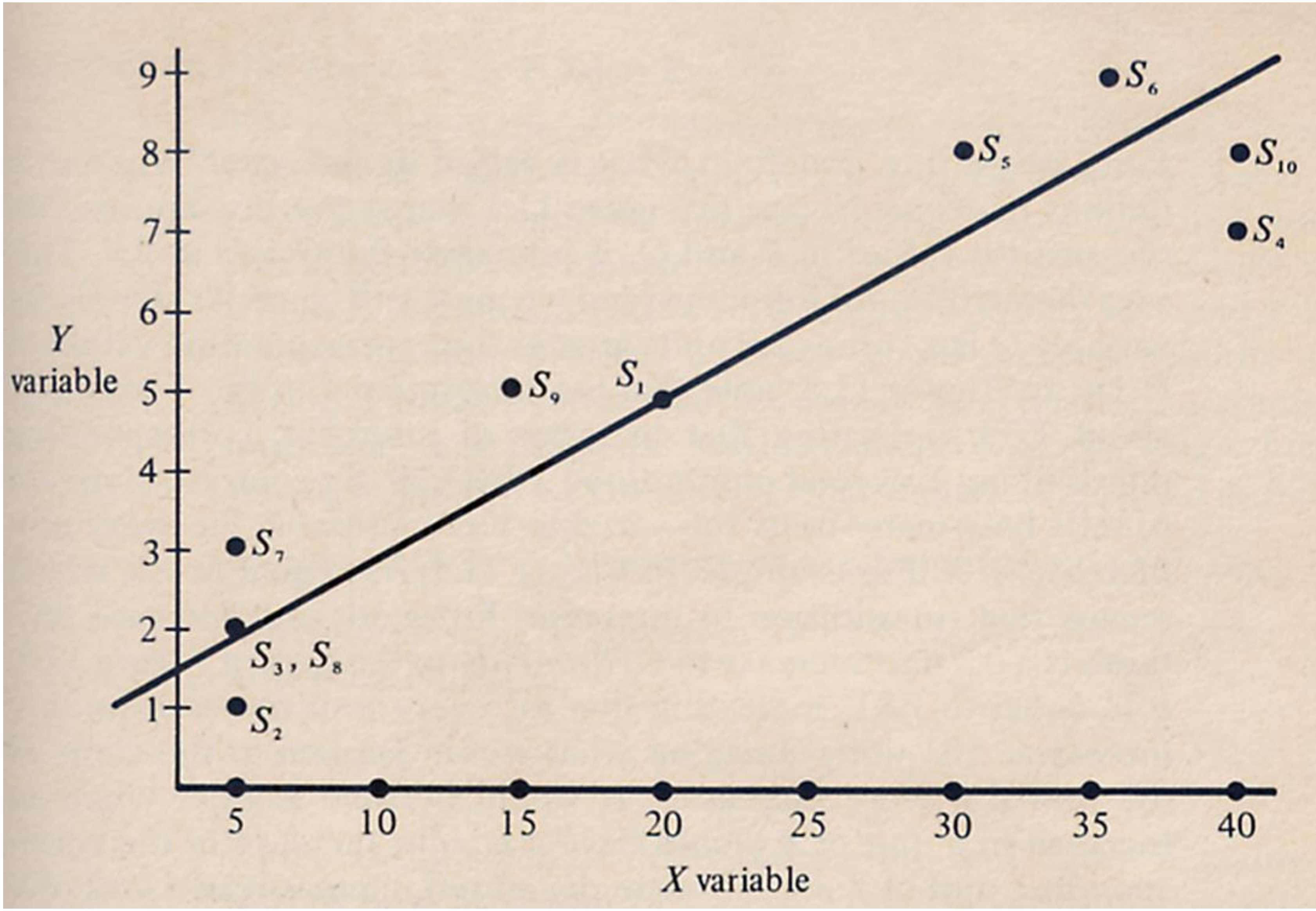
To plot the regression line, you must determine a few pairs (at least two pairs) that pass the line, e.g.

If $X = 0 \quad \hat{Y} = 1.40 + 0.18(0) = 1.40$

$X = 10 \quad \hat{Y} = 1.40 + 0.18(10) = 3.20$

$X = 20 \quad \hat{Y} = 1.40 + 0.18(20) = 5.00$

(You can now sketch the regression line)



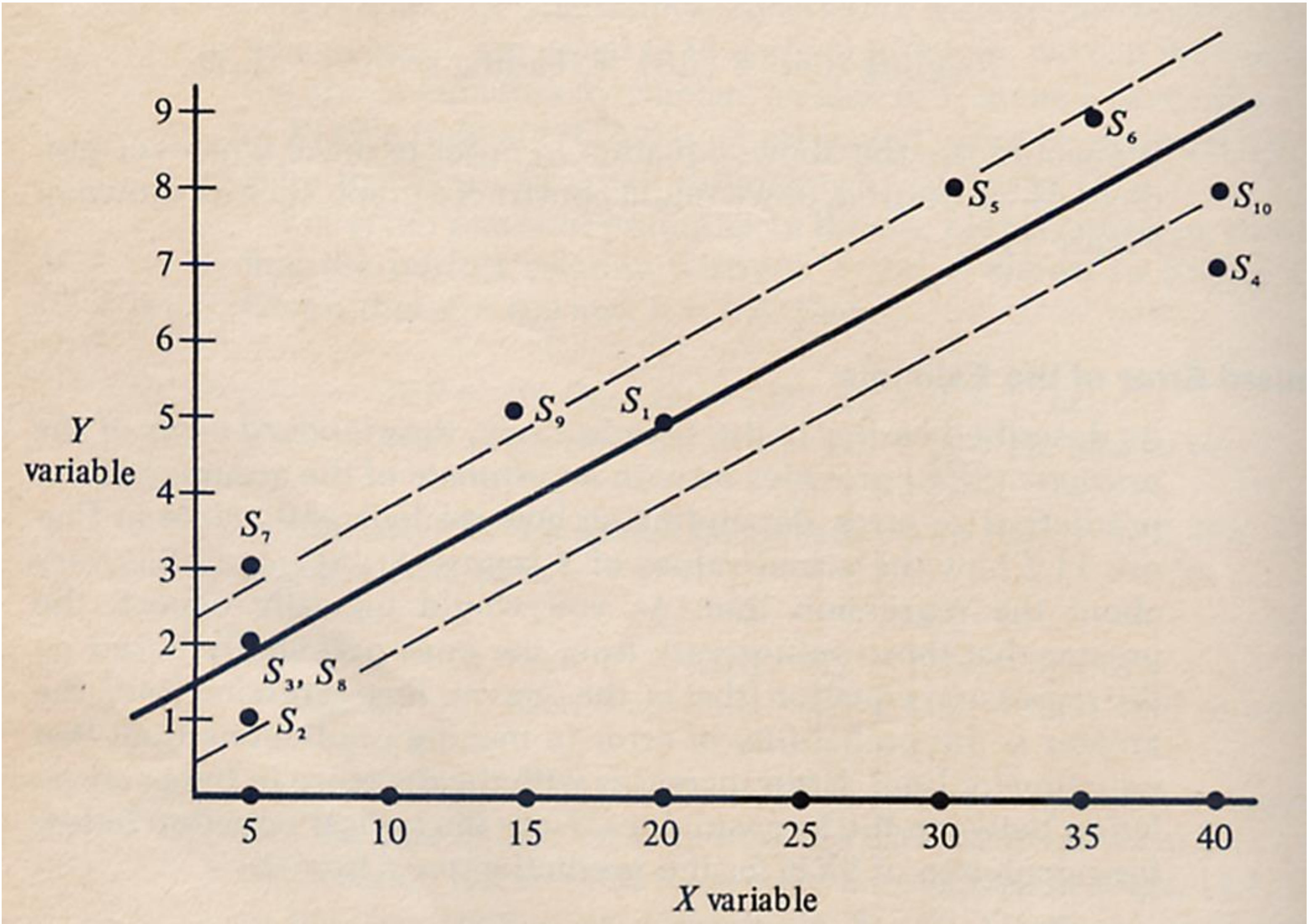
Y variable
X variable
9
8
7
6
5
4
3
2
1
5
10
15
20
25
30
35
40
S_1
S_2
S_3, S_8
S_4
S_5
S_6
S_7
S_9
S_{10}

The Standard Error of Estimates

- ☑ Regression equations are more accurate than just guessing. But even if you are using a regression equation, you should not fall into the trap of thinking that these equations will give you predictions that are 100% accurate. As long as the correlation coefficients you're using are not perfect, your predictions will not be perfect. In all psychological and biological populations some unexplained variance will cause inaccuracies in predictions made using regression equations.

- ☑ As you may recall, the coefficient of determination expresses the amount of variance in one variable that is explained by the other variable. Whenever the correlation coefficient is not +1.00 or -1.00, an unexplained variance will always cause some error in prediction. This error of prediction is called the *standard error of estimate.*

- ☑ The standard error of estimate (SEE) is the standard deviation of actual values from the value estimated from the regression equation. The smaller the standard error of estimate, the closer to the actual real-world values the prediction is likely to be. Since there are two regression equations, there are also two standard errors of estimate, one for the prediction of Y from a known X and another for the prediction of X from a known Y.



Range of deviation of Y scores one SEE above and one SEE below the line of best fit for sample study where SEE = 0.963