Regression Analyses

- Regression: technique concerned with predicting some variables by knowing others
- The process of predicting variable Y using variable X

Regression

- Uses a variable (x) to predict some outcome variable (y)
- Tells you how values in y change as a function of changes in values of x

Correlation and Regression

- Correlation describes the strength of a **linear** relationship between two variables
- **Linear** means "**straight line**"

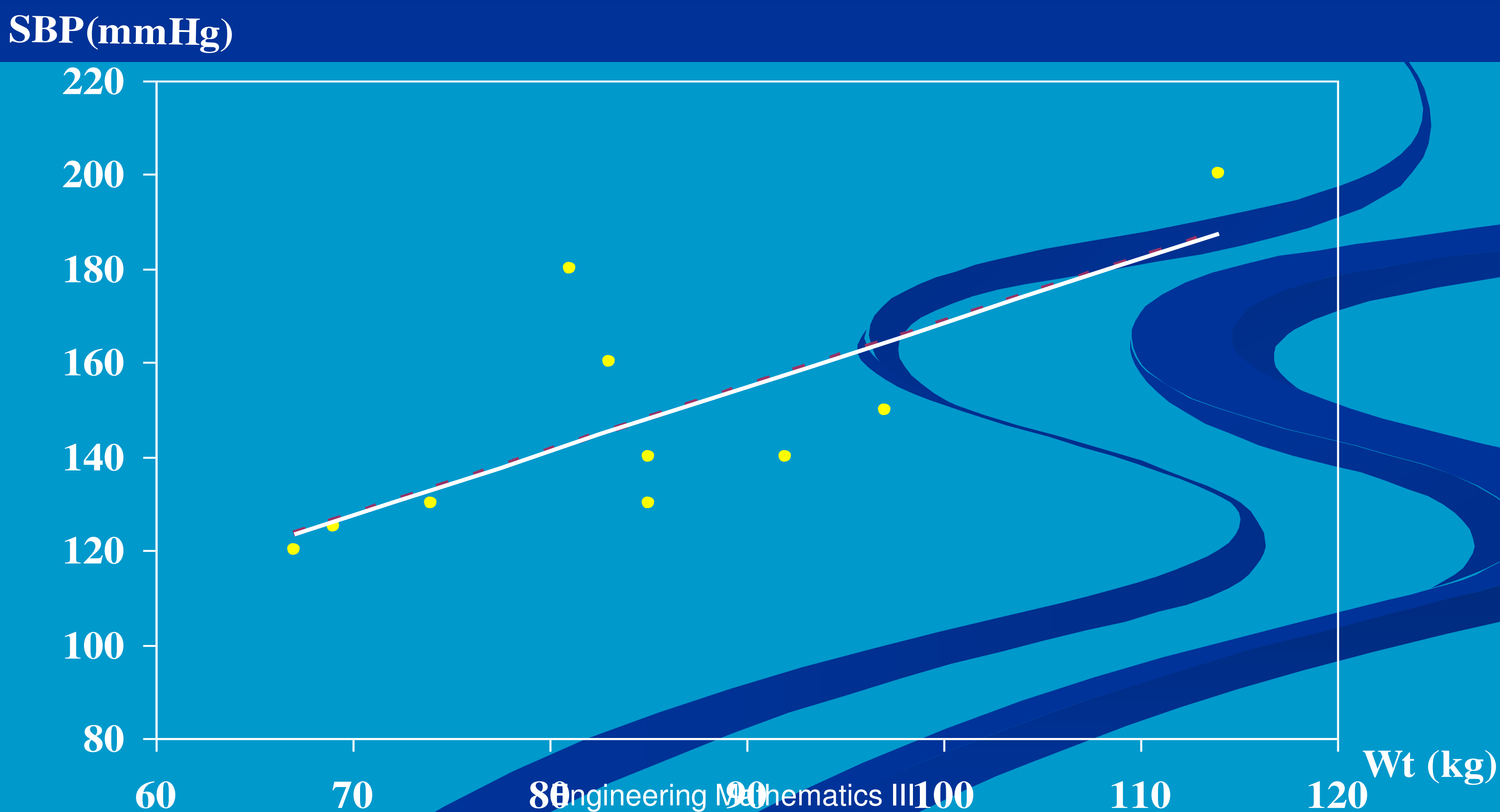
- **Regression** tells us how to draw the straight line described by the correlation

Regression

- Calculates the "best-fit" line for a certain set of data

The regression line makes the sum of the squares of the residuals smaller than for any other line

Regression minimizes residuals



By using the least squares method (a procedure that minimizes the vertical deviations of plotted points surrounding a straight line) we are able to construct a best fitting straight line to the scatter diagram points and then formulate a regression equation in the form of:

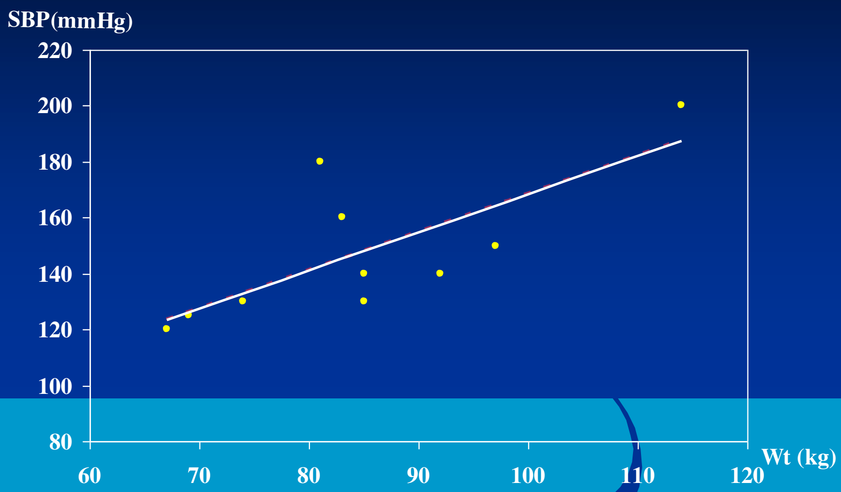
$$\hat{y} = a + bX$$

$$\hat{y} = \bar{y} + b(x - \bar{x})$$

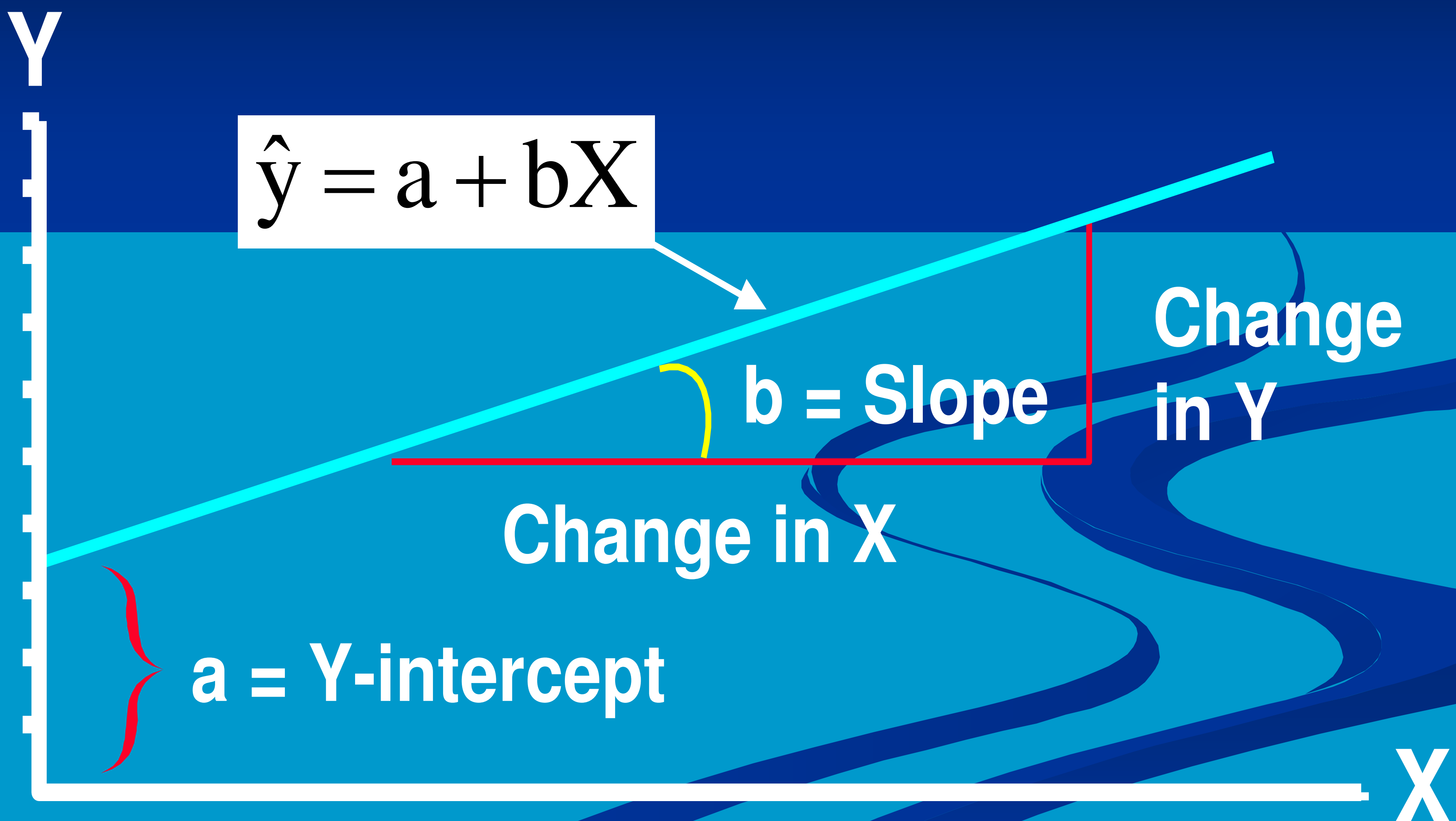
$$\mathbf{b} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Regression Equation

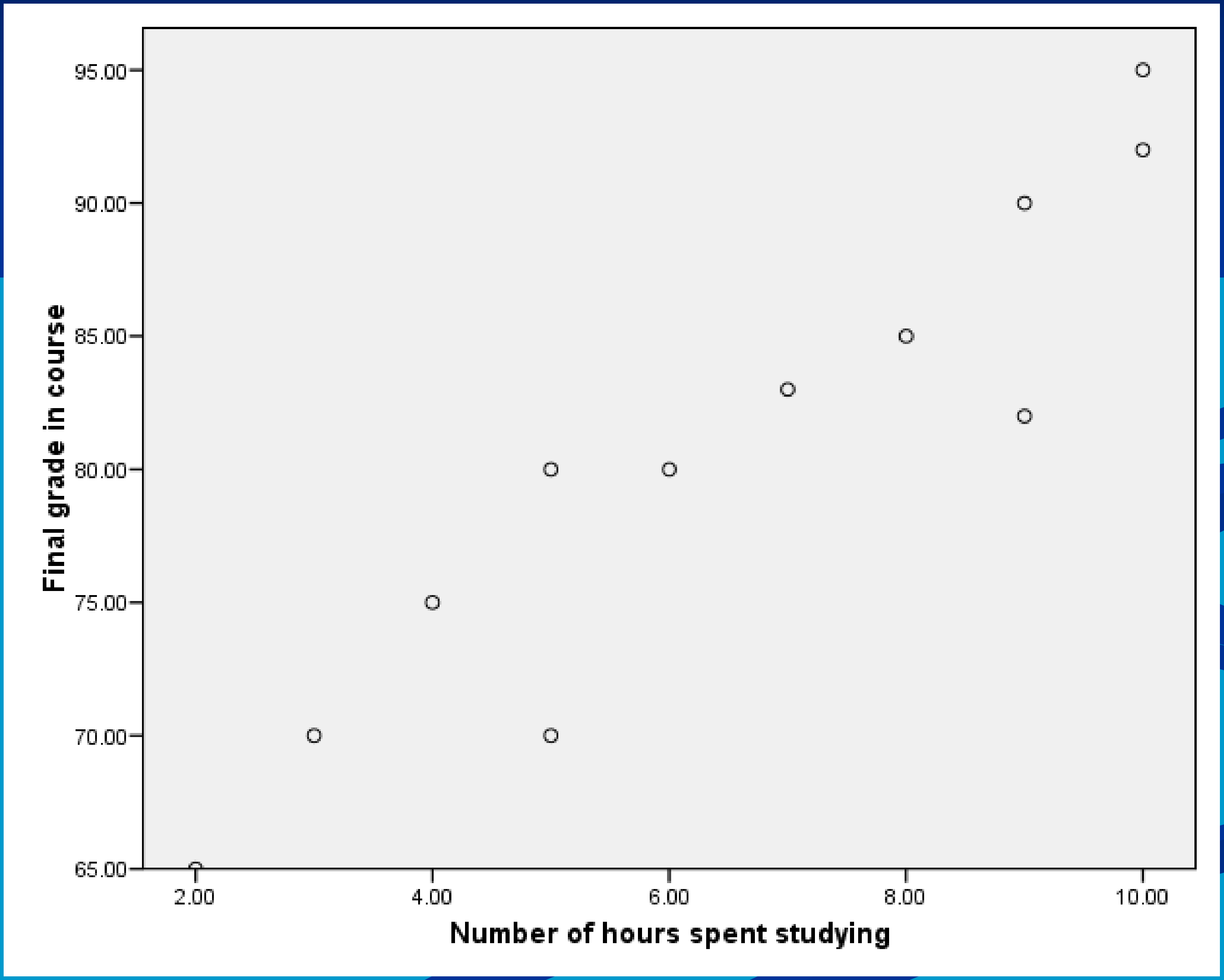
- Regression equation describes the regression line mathematically
  - Intercept
  - Slope



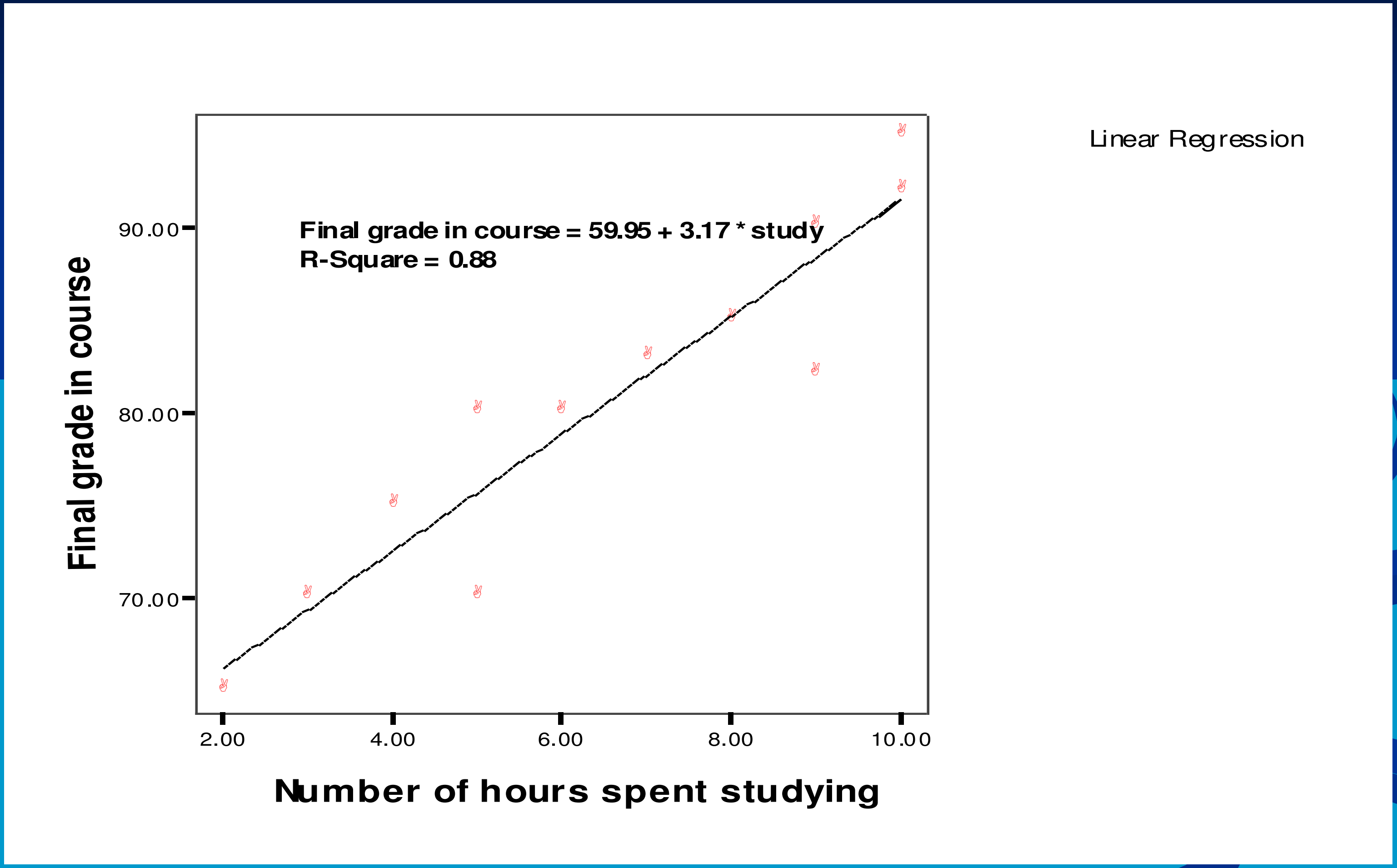
Linear Equations



Hours studying and grades



Regressing grades on hours



Predicted final grade in class =

59.95 + 3.17*(number of hours you study per week)

Predicted final grade in class = 59.95 + 3.17*(hours of study)

Predict the final grade of…

- Someone who studies for 12 hours
- Final grade = 59.95 + (3.17*12)
- Final grade = 97.99

- Someone who studies for 1 hour:
- Final grade = 59.95 + (3.17*1)
- Final grade = 63.12

Things to remember

Regressions are still focuses on association, not causation.

Association is a necessary prerequisite for inferring causation, but also:

- The independent variable must preceded the dependent variable in time.
- The two variables must be plausibly lined by a theory,
- Competing independent variables must be eliminated.

Exercise

A sample of 6 persons was selected the value of their age (x variable) and their weight is demonstrated in the following table. Find the regression equation and what is the predicted weight when age is 8.5 years.

Serial no.	Age (x)	Weight (y)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

Answer

Serial no.	Age (x)	Weight (y)	xy	X^2	Y^2
1 2 3 4 5 6	7 6 8 5 6 9	12 8 12 10 11 13	84 48 96 50 66 117	49 36 64 25 36 81	144 64 144 100 121 169
Total	41	66	461	291	742

$$\bar{x} = \frac{41}{6} = 6.83$$

$$\bar{y} = \frac{66}{6} = 11$$

$$b = \frac{461 - \frac{41 \times 66}{6}}{291 - \frac{(41)^2}{6}} = 0.92$$

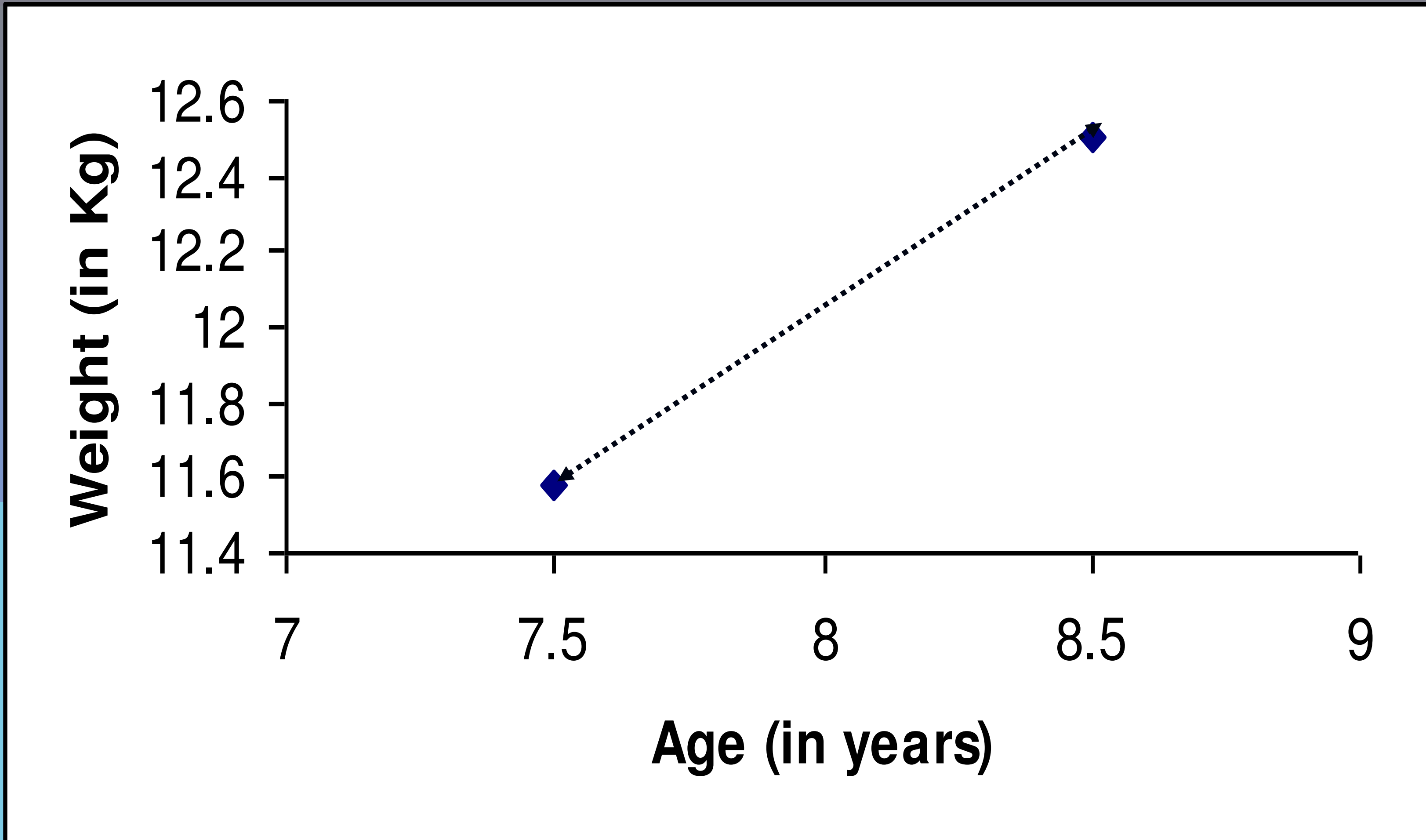
Regression equation

$$\hat{y}_{(x)} = 11 + 0.9(x - 6.83)$$

$$\hat{y}_{(x)} = 4.675 + 0.92x$$

$$\hat{y}_{(8.5)} = 4.675 + 0.92 * 8.5 = 12.50\text{Kg}$$

$$\hat{y}_{(7.5)} = 4.675 + 0.92 * 7.5 = 11.58\text{Kg}$$



we create a regression line by plotting two estimated values for y against their X component, then extending the line right and left.

Exercise 2

The following are the age (in years) and systolic blood pressure of 20 apparently healthy adults.

Age (x)	B.P (y)	Age (x)	B.P (y)
20	120	46	128
43	128	53	136
63	141	60	146
26	126	20	124
53	134	63	143
31	128	43	130
58	136	26	124
46	132	19	121
58	140	31	126
70	144	23	123

- Find the correlation between age and blood pressure using simple and Spearman's correlation coefficients, and comment.
- Find the regression equation?
- What is the predicted blood pressure for a man aging 25 years?

Serial	x	y	xy	x2
1	20	120	2400	400
2	43	128	5504	1849
3	63	141	8883	3969
4	26	126	3276	676
5	53	134	7102	2809
6	31	128	3968	961
7	58	136	7888	3364
8	46	132	6072	2116
9	58	140	8120	3364
10	70	144	10080	4900

Serial	x	y	xy	x2
11	46	128	5888	2116
12	53	136	7208	2809
13	60	146	8760	3600
14	20	124	2480	400
15	63	143	9009	3969
16	43	130	5590	1849
17	26	124	3224	676
18	19	121	2299	361
19	31	126	3906	961
20	23	123	2829	529
Total	852	2630	114486	41678

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

=

$$\frac{114486 - \frac{852 \times 2630}{20}}{41678 - \frac{852^2}{20}} = 0.4547$$

$\hat{y}$ =112.13 + 0.4547 x

for age 25

B.P = 112.13 + 0.4547 * 25=123.49 = 123.5 mm hg

Multiple Regression

Multiple regression analysis is a straightforward extension of simple regression analysis which allows more than one independent variable.

Multiple Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.736[a]	.542	.532	2760.003

a. Predictors: (Constant), Percent of Population 25 years and Over with Bachelor's Degree or More, March 2000 estimates

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.32E+08	1	432493775.8	56.775	.000[a]
	Residual	3.66E+08	48	7617618.586		
	Total	7.98E+08	49			

a. Predictors: (Constant), Percent of Population 25 years and Over with Bachelor's Degree or More, March 2000 estimates

b. Dependent Variable: Personal Income Per Capita, current dollars, 1999

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10078.565	2312.771		4.358	.000
	Percent of Population 25 years and Over with Bachelor's Degree or More, March 2000 estimates	688.939	91.433	.736	7.535	.000

a. Dependent Variable: Personal Income Per Capita, current dollars, 1999

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.849[a]	.721	.709	2177.791

a. Predictors: (Constant), Population Per Square Mile, Percent of Population 25 years and Over with Bachelor's Degree or More, March 2000 estimates

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.75E+08	2	287614518.2	60.643	.000[a]
	Residual	2.23E+08	47	4742775.141		
	Total	7.98E+08	49			

a. Predictors: (Constant), Population Per Square Mile, Percent of Population 25 years and Over with Bachelor's Degree or More, March 2000 estimates

b. Dependent Variable: Personal Income Per Capita, current dollars, 1999

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13032.847	1902.700		6.850	.000
	Percent of Population 25 years and Over with Bachelor's Degree or More, March 2000 estimates	517.628	78.613	.553	6.584	.000
	Population Per Square Mile	7.953	1.450	.461	5.486	.000

a. Dependent Variable: Personal Income Per Capita, current dollars, 1999