Statistical Methods

Simple Linear Regression and Correlation

Linear Regression Analysis…

- Regression analysis is used to predict the value of one variable (the ***dependent variable***) on the basis of other variables (the ***independent variables***).

- Dependent variable: denoted **Y**
- Independent variables: denoted $X_1, X_2, \ldots, X_k$
- **If we only have ONE independent variable, the model is**

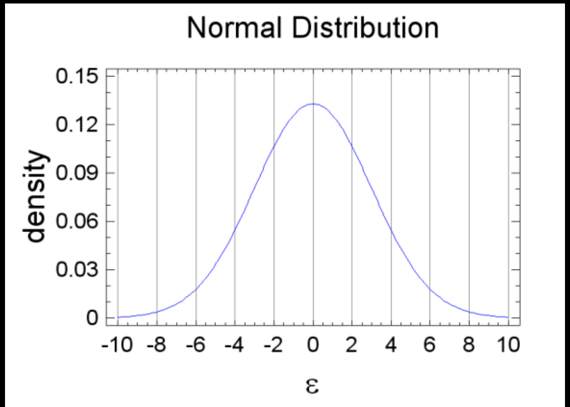
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- which is referred to as simple linear regression. We would be interested in estimating β_0 and β_1 from the data we collect.

Linear Regression Analysis

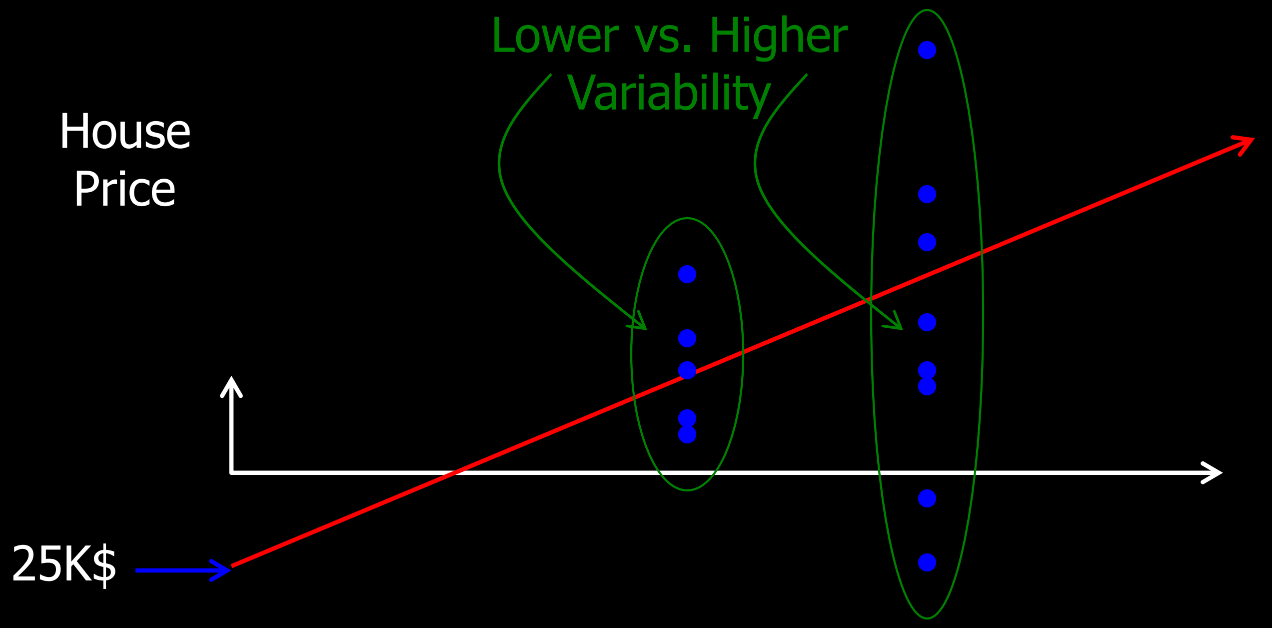
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Variables:
- X = Independent Variable (we provide this)
- Y = Dependent Variable (we observe this)
- Parameters:
- β_0 = Y-Intercept
- β_1 = Slope
- ε ~ Normal Random Variable ($\mu_\varepsilon = 0$, σ_ε = ???) [Noise]



Effect of Larger Values of σ_ε

•

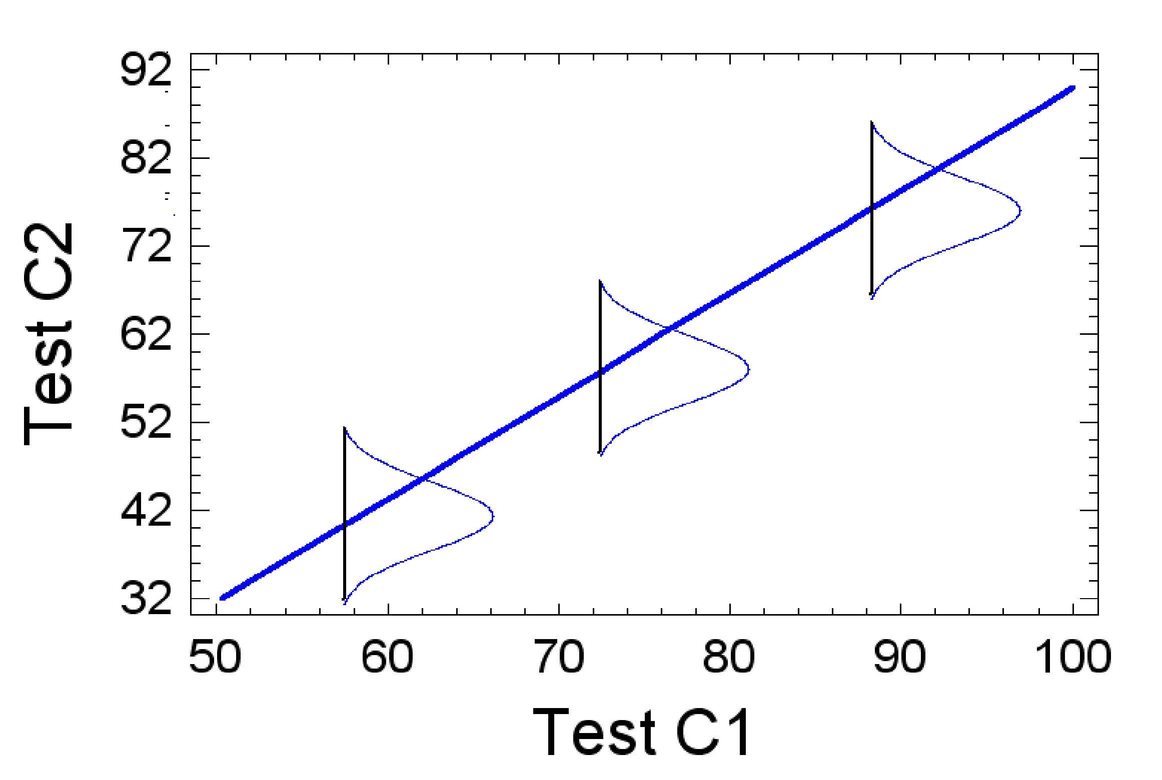


$$\text{House Price} = 25{,}000 + 75(\text{Size}) + \boxed{\varepsilon}$$

Same square footage, but different price points
(e.g. décor options, cabinet upgrades, lot location…)

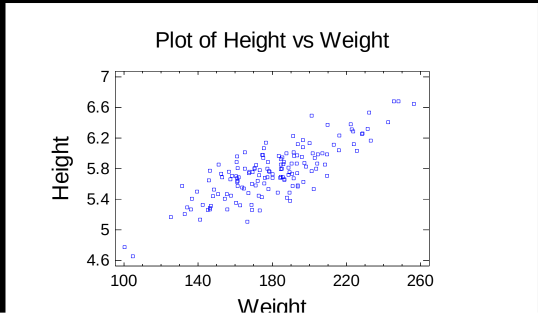
Theoretical Linear Model

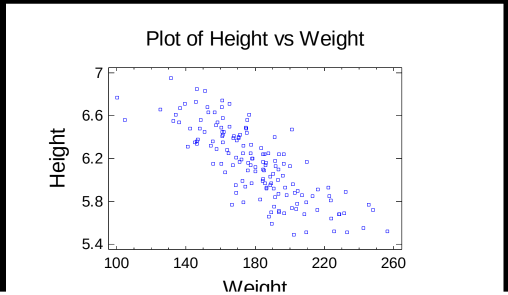
$$y = \beta_0 + \beta_1 x + \varepsilon$$

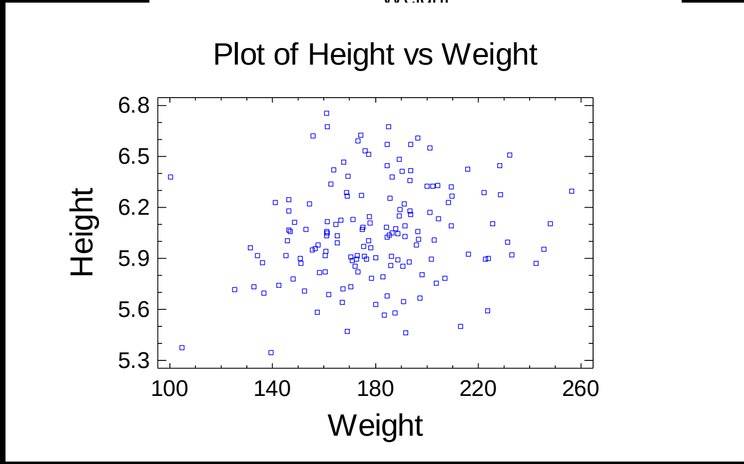


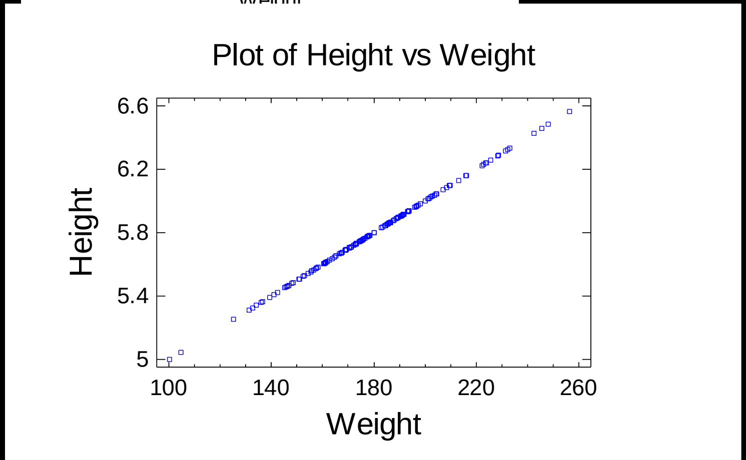
Correlation Analysis… “-1 $\leq$ ρ < 1”

- If we are interested ***only*** in determining whether a relationship ***exists***, we employ ***correlation analysis***. Example: Student’s height and weight.









- The sign of r denotes the nature of association
- The value of r denotes the strength of association
- It’s not just about the relationship strength but the direction too!

- If r = Zero this means NO LINEAR association or correlation between the two variables.
- If $0 < r < 0.25$ = weak correlation.
- If $0.25 \le r < 0.75$ = intermediate correlation.
- If $0.75 \le r < 1$ = strong correlation.
- If r = l = perfect correlation.

Correlation Analysis… “$-1 \leq \rho < 1$”

- If the correlation coefficient is close to +1 that means you have a strong positive relationship.
- If the correlation coefficient is close to -1 that means you have a strong negative relationship.
- If the correlation coefficient is close to 0 that means you have no correlation.
- WE HAVE THE ABILITY TO TEST THE HYPOTHESIS

$$\mathbf{H_0: \rho = 0 \text{ vs } H_0: \rho \neq 0}$$

Assessing the Model…

- The least squares method will always produce a straight line, even if there is no relationship between the variables, or if the relationship is something other than linear.

- Hence, in addition to determining the coefficients of the least squares line, we need to assess it to see how well it "fits" the data. We'll see these evaluation methods now. They're based on the what is called sum of squares for errors (SSE).

Coefficient of Determination

- r^2
- say $r = 0.7836$, $r^2 = 61.40\%$ [**always in percentage**]. Thus 61.40% of the variation in DV can be explained by your regression model. The remaining 38.60% is ***unexplained***, i.e. due to error.

- Unlike the value of a test statistic, the ***coefficient of determination*** does **not** have a ***critical value*** that enables us to draw conclusions.
- In general the higher the value of R^2, the ***better*** the model fits the data.
- $r^2 = 1$: Perfect match between the line and the data points.
- $r^2 = 0$: There are NO LINEAR (only) relationship between x and y.

How to compute the simple correlation coefficient (r)

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \cdot \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Example:

A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table . It is required to find the correlation between age and weight.

serial No	Age (years)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

These 2 variables are of the quantitative type, one variable (Age) is called the independent and denoted as (X) variable and the other (weight) is called the dependent and denoted as (Y) variables to find the relation between age and weight compute the simple correlation coefficient using the following formula:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \cdot \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Serial n.	Age (years) (x)	Weight (Kg) (y)	xy	X^2	Y^2
1	7	12			
2	6	8			
3	8	12			
4	5	10			
5	6	11			
6	9	13			
Total	$\sum x=$	$\sum y=$	$\sum xy=$	$\sum x2=$	$\sum y2=$

$$r = \frac{461 - \dfrac{41 \times 66}{6}}{\sqrt{\left[291 - \dfrac{(41)^2}{6}\right] \cdot \left[742 - \dfrac{(66)^2}{6}\right]}}$$

r = 0.759

direct/positive strong correlation

Testing for the significance of the correlation coefficient, *r*

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

* r is the correlation coefficient

* n is the sample size (small)

* t is the computed t-statistic

* degree of freedom is n-2

EXAMPLE: Relationship between Anxiety and Test Scores

Anxiety (X)	Test score (Y)	X^2	Y^2	XY
10	2			
8	3			
2	9			
1	7			
5	6			
6	5			
$\sum X = 32$	$\sum Y = 32$	$\sum X^2 =$	$\sum Y^2 =$	$\sum XY=$

Calculating Correlation Coefficient

$$r = \frac{(6)(129) - (32)(32)}{\sqrt{\left(6(230) - 32^2\right)\left(6(204) - 32^2\right)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -.94$$

r = - 0.94

Indirect/negative strong correlation

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

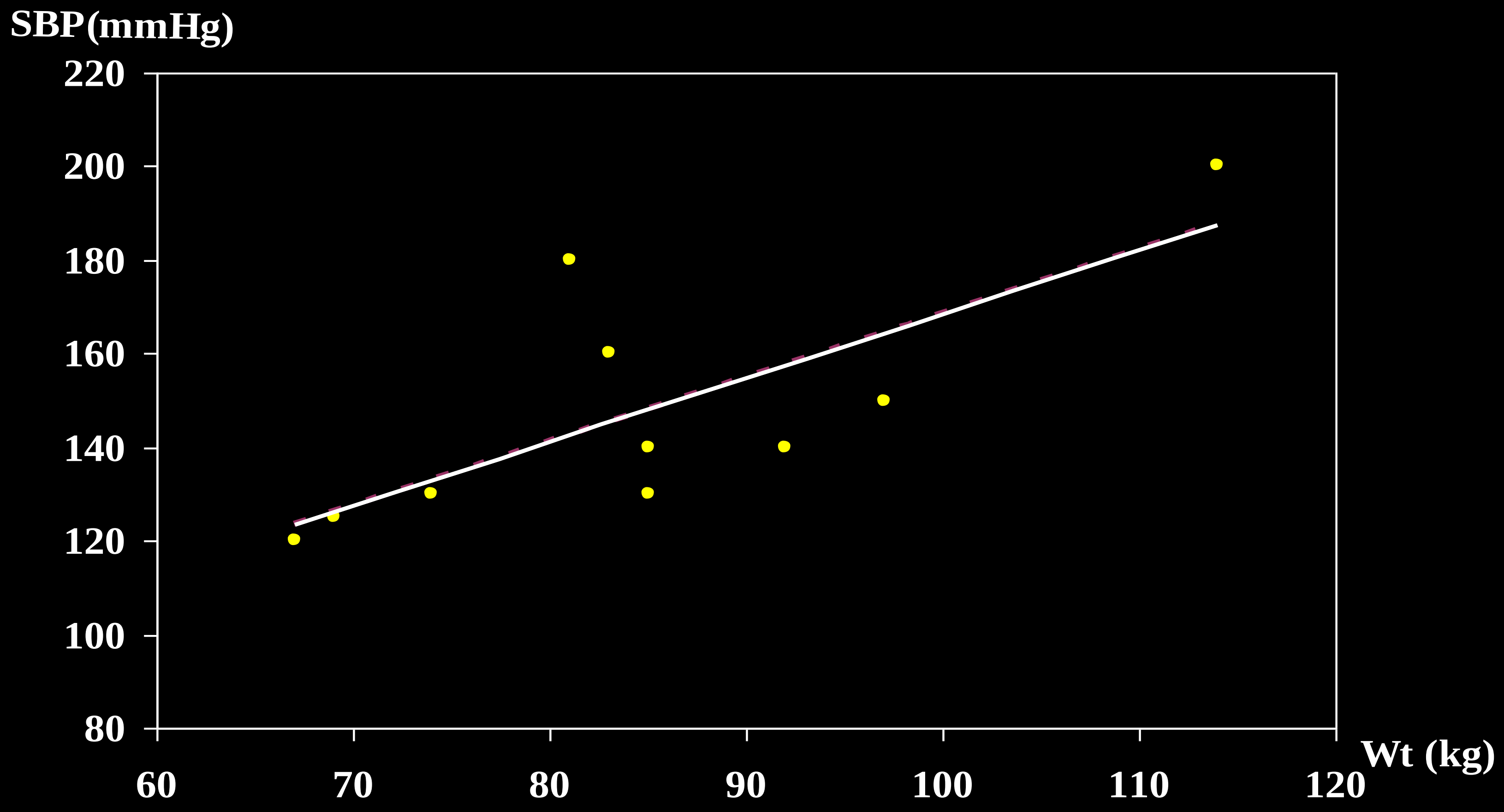
- Correlation describes the strength of a **linear** relationship between two variables
- **Linear** means "**straight line**"
- **Regression** tells us how to draw the straight line described by the correlation

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:

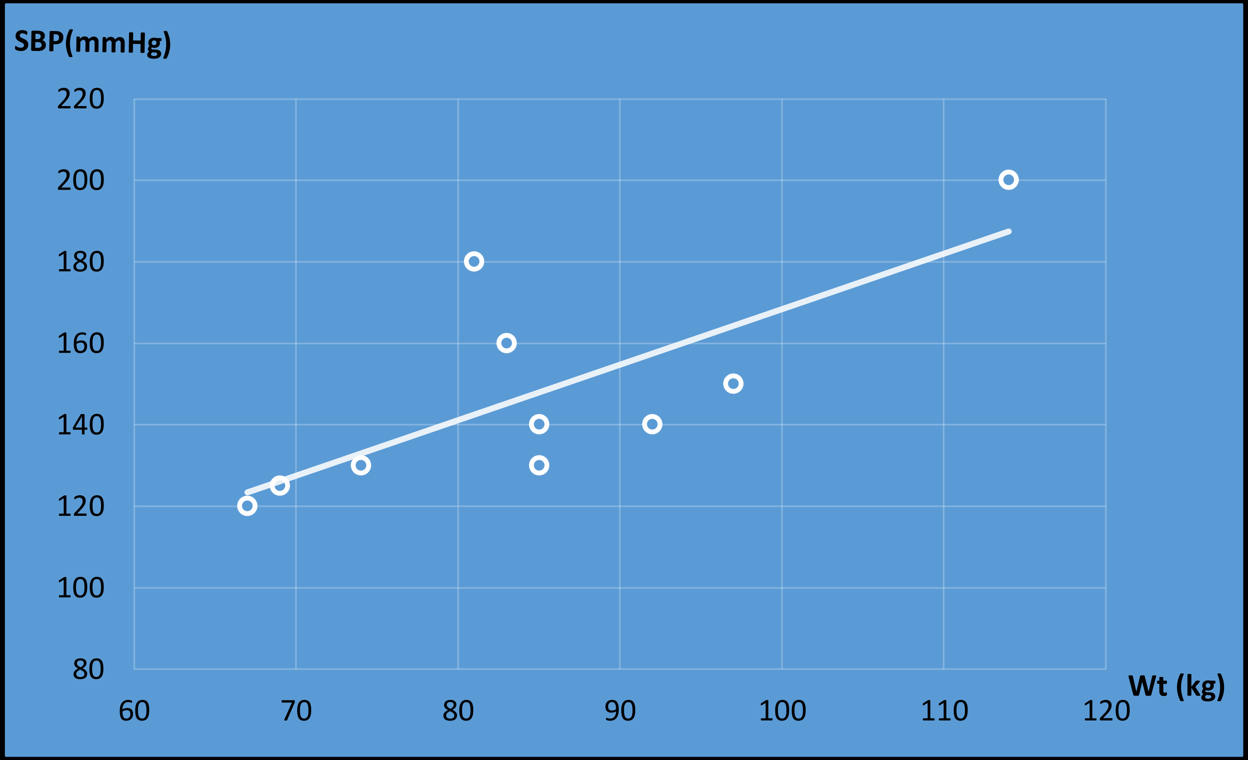
$$\hat{y} = a + bX$$

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

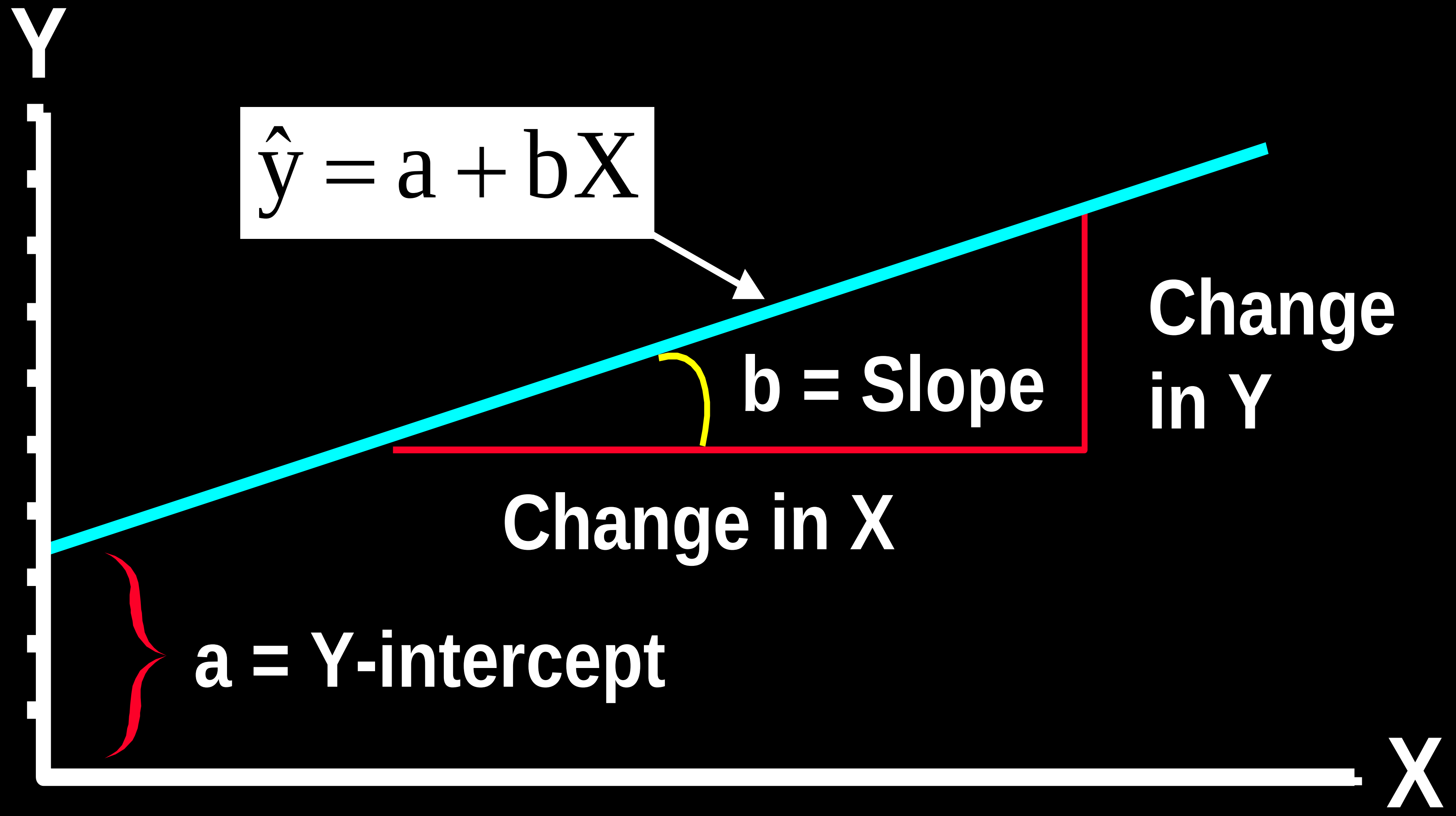
$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Regression Equation

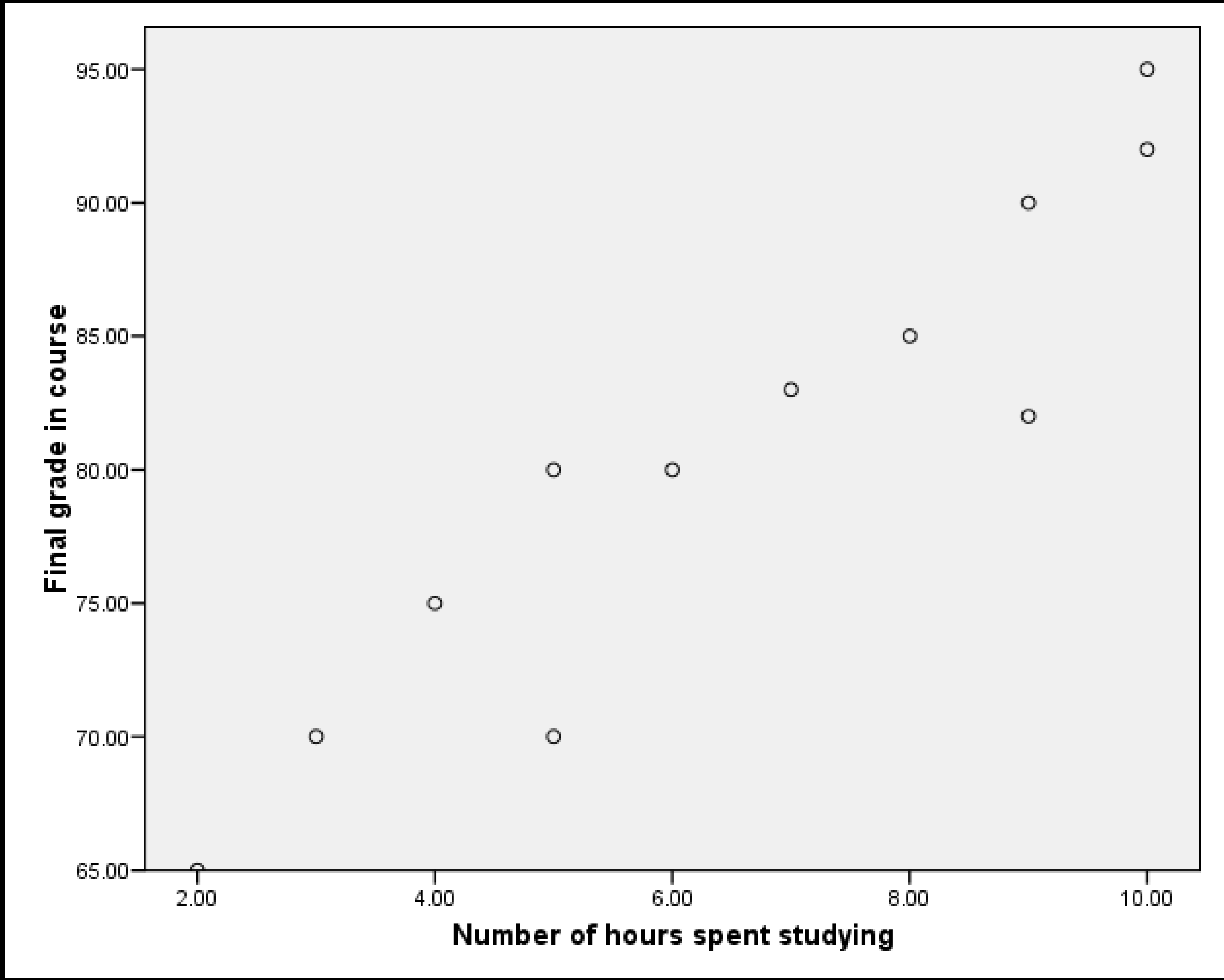
- Regression equation describes the regression line mathematically
  - Intercept
  - Slope



Linear Equations



Hours studying and grades



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	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	41	66	461	291	742

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

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

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

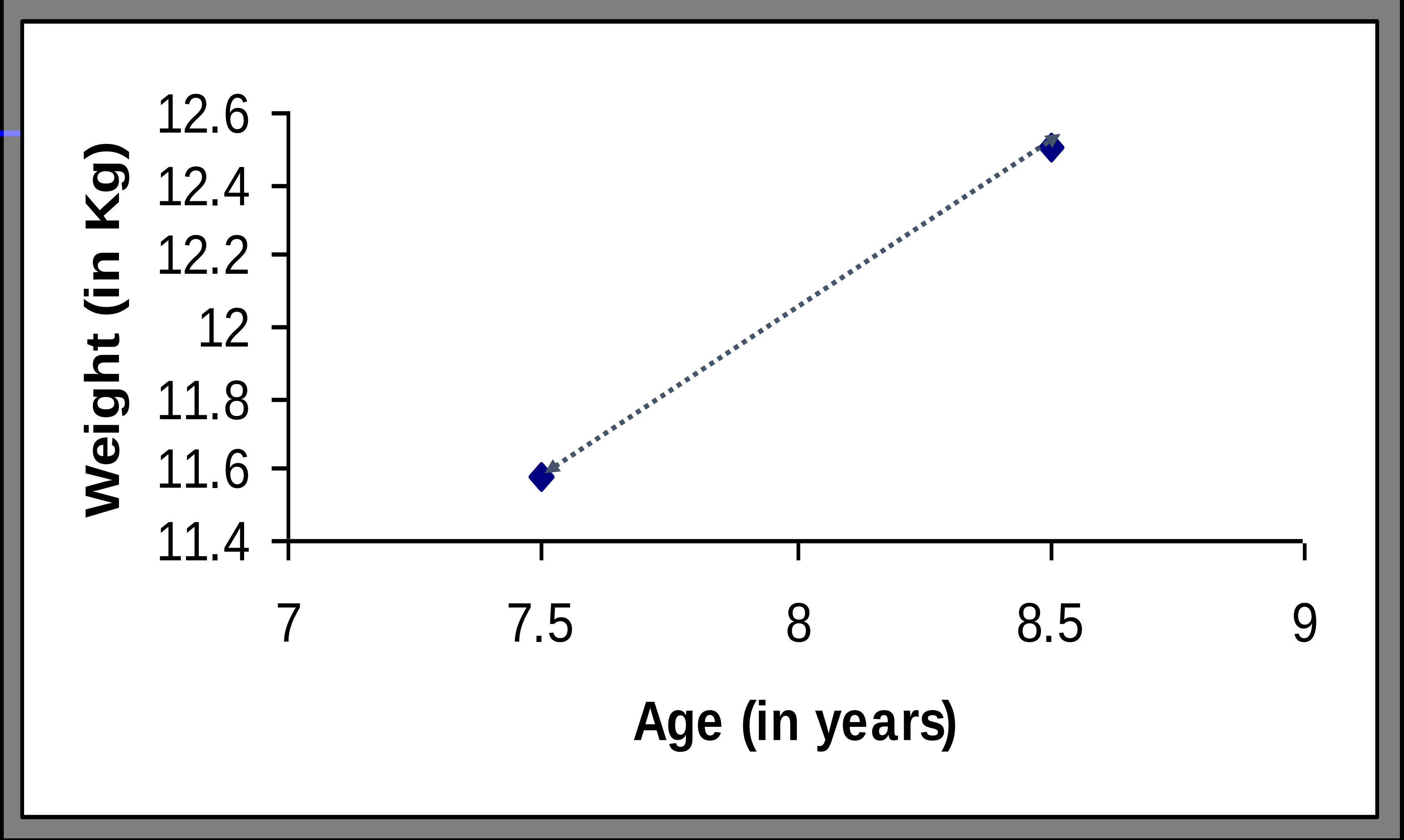
Regression equation:

$$\hat{y}_{(x)} = 11 + 0.92(x - 6.83)$$

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

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

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



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)
20	120
43	128
63	141
26	126
53	134
31	128
58	136
46	132
58	140
70	144

Determine and interpret the following:

1. The correlation between age and blood pressure and if it is significant.
2. The coefficient of determination.
3. the regression equation.
4. The predicted blood pressure for a man aging 25 years.
5. The predicted blood pressure for a man aging 18 years.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	**0.966627344**					
R Square	**0.934368422**					
Adjusted R Square	**0.926164475**					
Standard Error	**2.051293377**					
Observations	**10**					
ANOVA						
	df	**SS**	**MS**	**F**	**Significance F**	
Regression	**1**	**479.2376**	**479.2376**	**113.8925**	**5.212447E-06**	
Residual	**8**	**33.66244**	**4.207805**			
Total	**9**	**512.9**				
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**
Intercept	**112.4324713**	**2.024594**	**55.53333**	**1.23E-11**	**107.7637485**	**117.1012**
Age (x)	**0.437340358**	**0.04098**	**10.67205**	**5.21E-06**	**0.342840318**	**0.53184**