Biostatistics
Correlation and Regression

Deepa Anwar

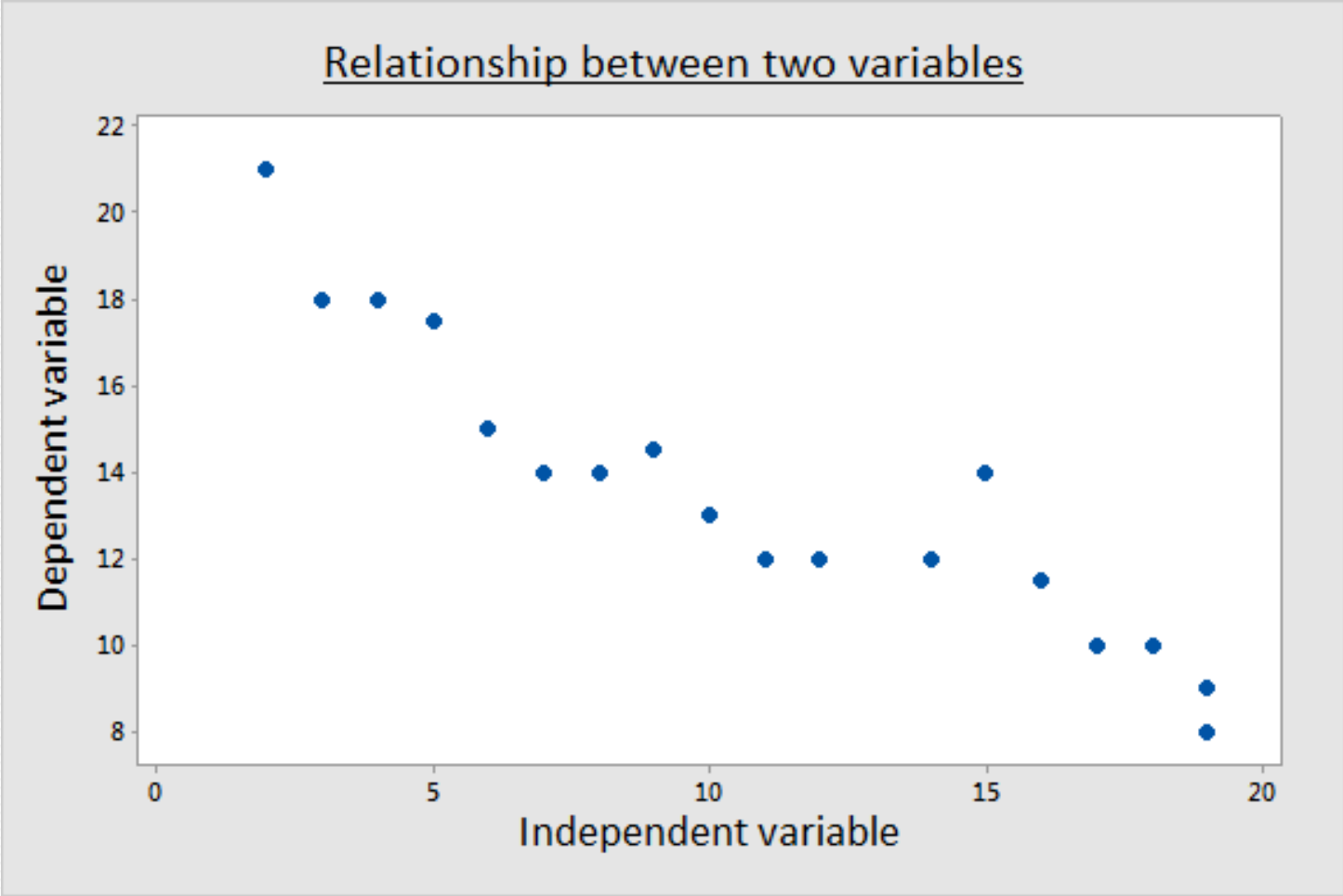
Relationships between Variables

- Common aim of research is to try to relate a variable of interest to one or more other variables.
- Demonstrate an association between variables (not necessarily causal) eg dose of drug and resulting systolic blood pressure, advertising expenditure of a company and end of year sales figures, etc.
- Establish a theoretical model to predict the value of one variable from a number of known factors.

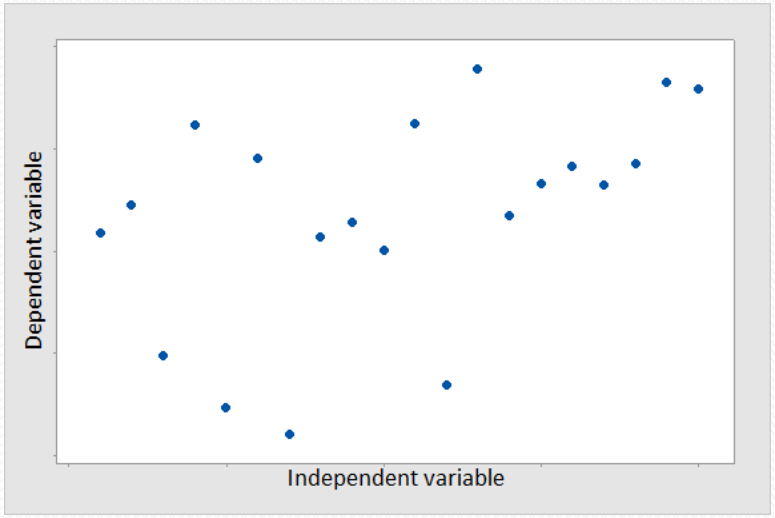
Dependent and Independent Variables

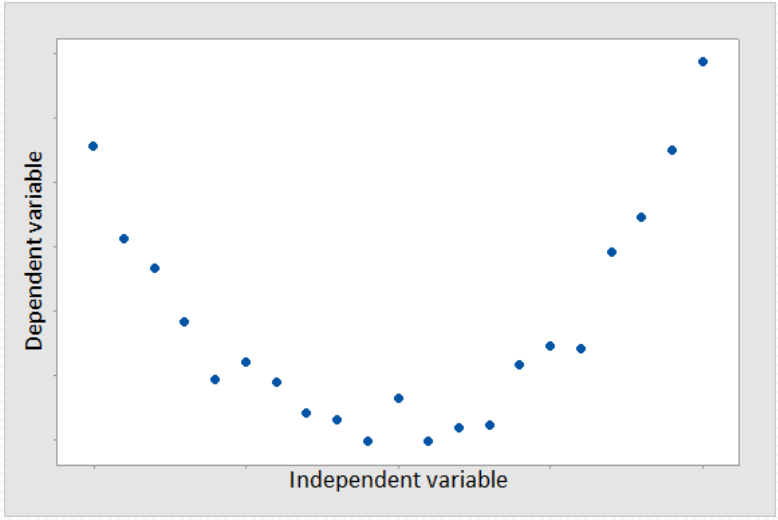
- The *independent* variable is the variable which is under the investigator's control (denoted *x*).
- The *dependent* variable is the one which the investigator is trying to estimate or predict (denoted *y*).
- Can a relationship be used to predict what happens to *y* as *x* changes (ie what happens to the dependent variable as the independent variable changes)?

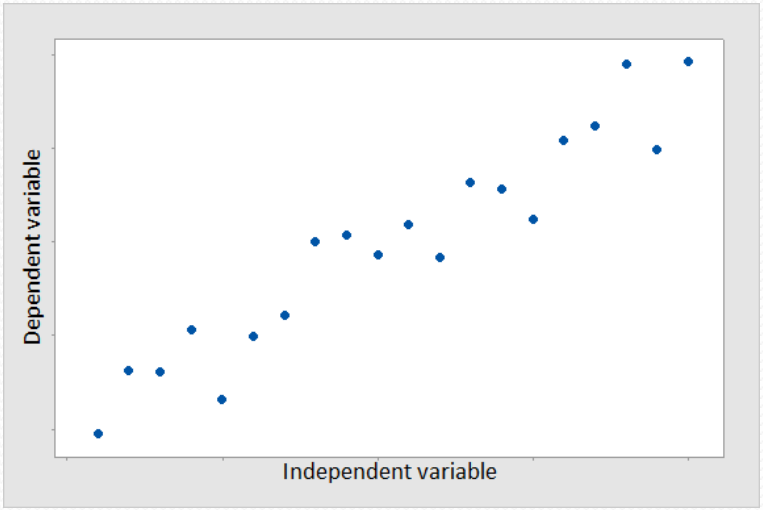
Exploring Relationships

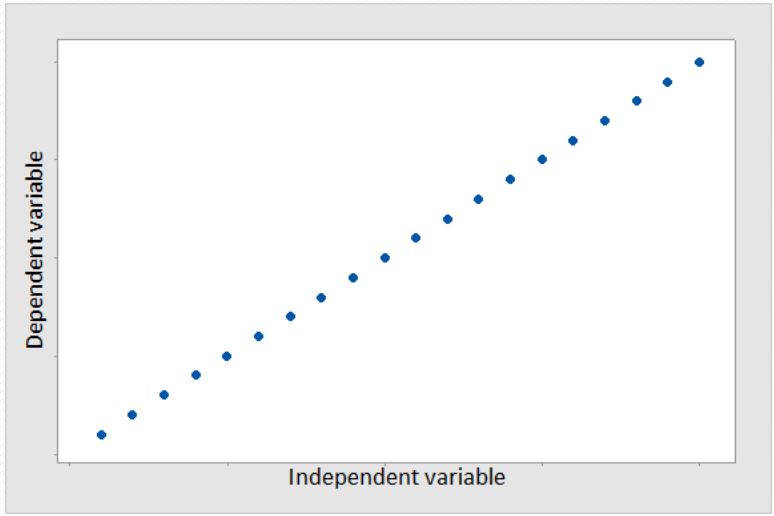


Scatter Plots





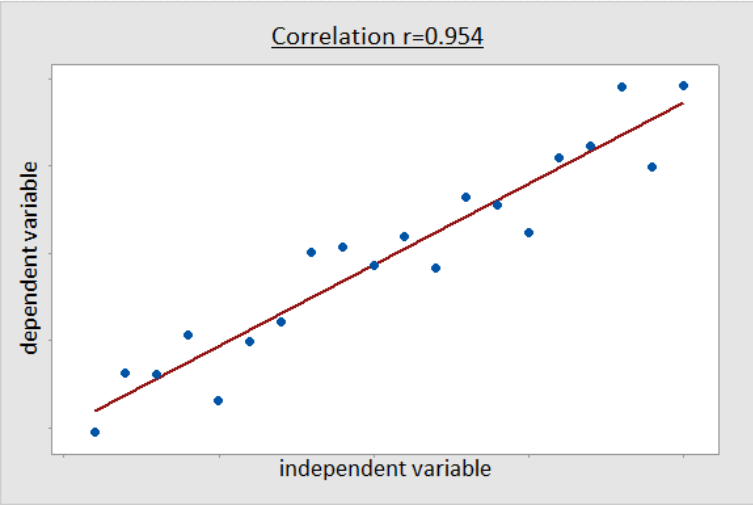


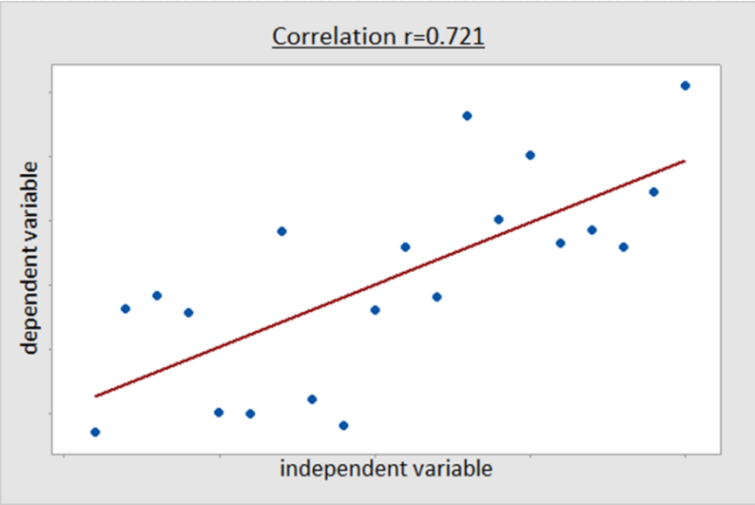


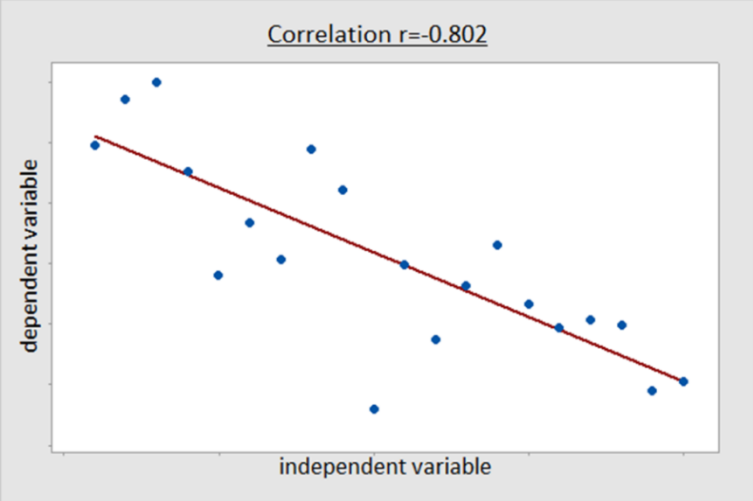
Correlation

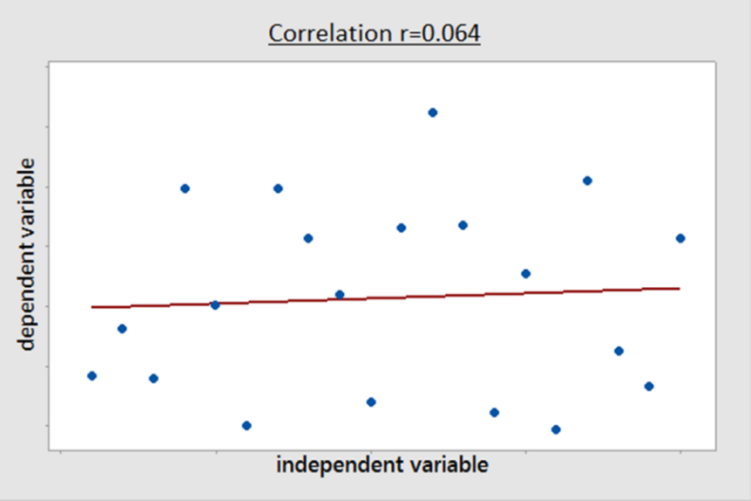
- Correlation is a measure of the degree of **linear association** between *two numerical* variables.
- The correlation is said to be *positive* if ‘large’ values of both variables occur together and *negative* if ‘large’ values of one variable tend to occur with ‘small’ values of the other.
- The range of possible values is from -1 to +1
- The correlation is high if observations lie close to a straight line (ie values close to +1 or -1) and low if observations are widely scattered (correlation value close to 0).
- It does not indicate a causal effect between the variables.

Correlation









Correlation

The correlation coefficient is computed as –

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum(x_i - \bar{x})^2)(\sum(y_i - \bar{y})^2)}}$$

Additional points

- Low correlation does not necessarily mean a low degree of association (relationship may be non-linear).
- Method comparison studies.
- Causal effects.

Problem

If the correlation between body weight and annual income were high and positive, we would conclude that…

(a) High incomes cause people to eat more food

(b) Low incomes cause people to eat less food

(c) High income people tend to spend a greater proportion of their income on food than low income people, on average

(d) High income people tend to be heavier than low income people, on average

(e) High incomes cause people to gain weight

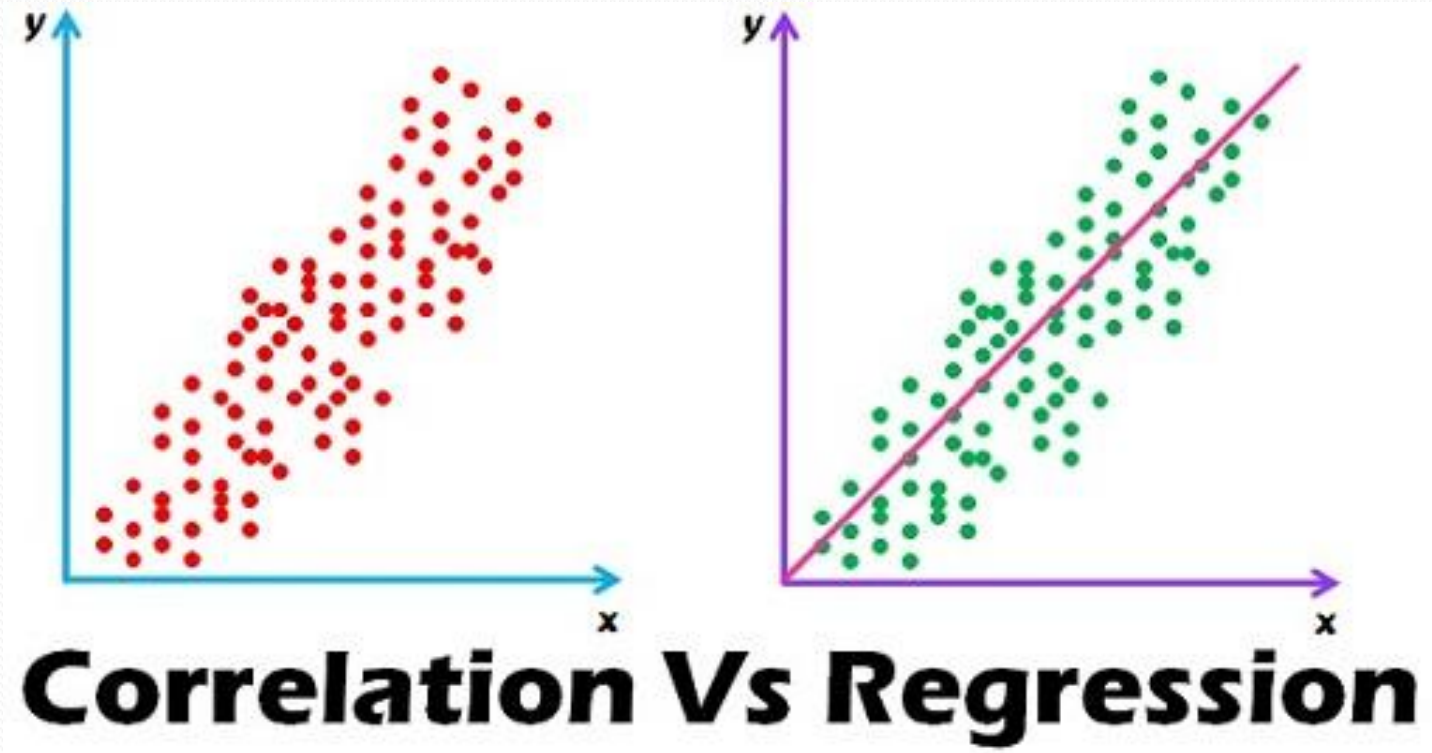
Hypothesis Testing

- The linear relationship between two variables is significant if there is evidence to suggest that r is significantly different from zero.
- Null hypothesis: $r=0$
- Alternative hypothesis: $r\neq0$
- Observed value of r has an associated p-value.
- Conclusion: if $p<0.05$, reject the null hypothesis and conclude that there is evidence to suggest that there is a linear relationship.

Summary

- Correlation (ρ) is a measure of the linear relationship between two variables.
- It gives an actual measure of the strength of the linear relationship seen in a scatter plot.
- Values range from –1 to +1
- The closer *r* is to ±1, the stronger the linear relationship.
- Note: high correlation does not indicate a causal effect!
- Correlation is not an appropriate measure for testing the equivalence of two methods.

Correlation vs. Regression



Regression

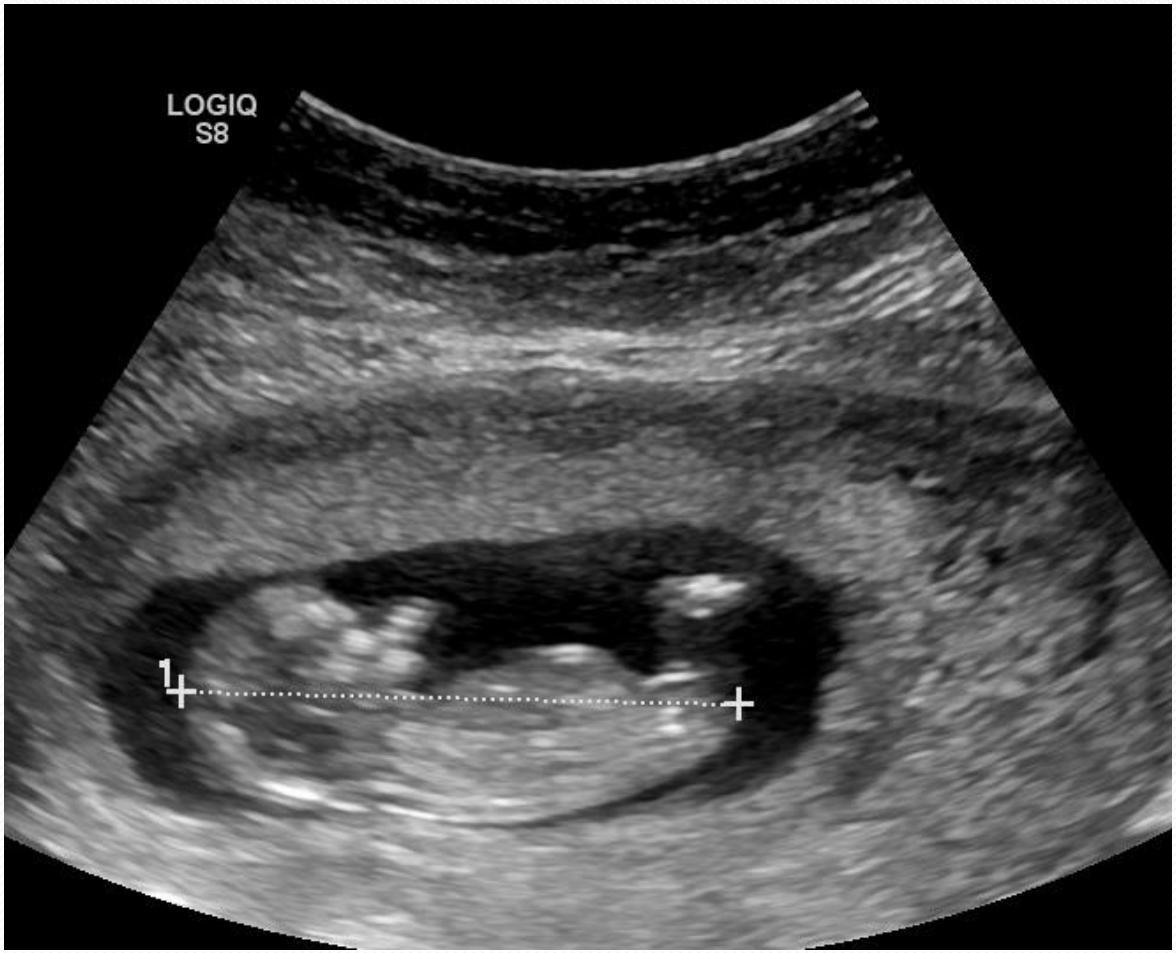
- Develop an equation to predict the dependent variable from knowledge of predictor variable(s).
- Linear regression fits a straight line to the data.
- General equation of a straight line …

$$y = a + bx$$

- Where a is the intercept and b is the slope, or gradient, of the line.
- Fit this line by eye – subjective.
- Method of least squares.

Growth of a Foetus

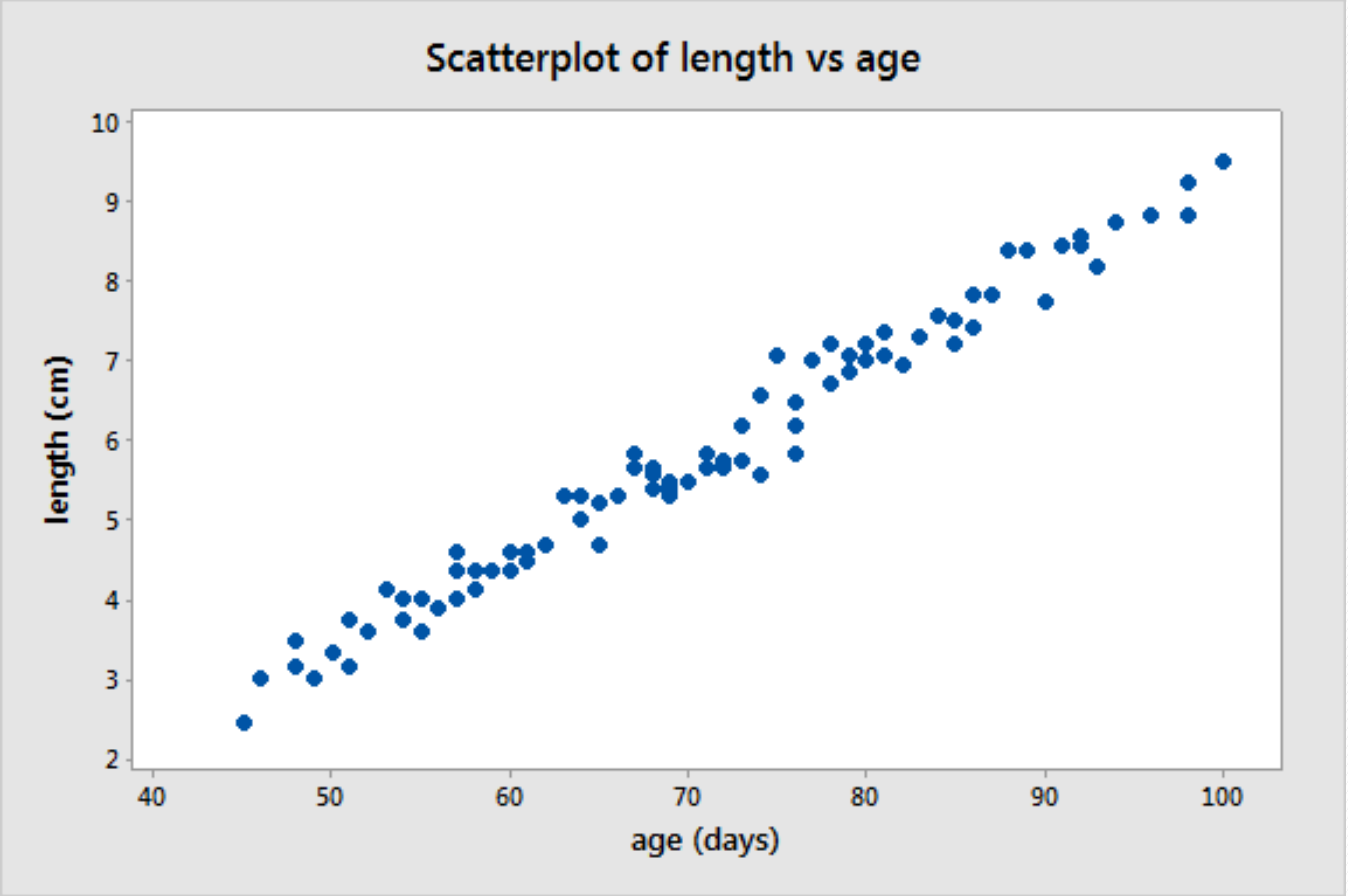
Practical Example



Practical Example

- Data from a study of foetal development.
- Date of conception (and hence age) of the foetus is known accurately.
- Height of the foetus (excluding the legs) is known from ultrasound scan.
- Age and length of the foetus are clearly related.
- Aim is to model the length and age data and use this to assess whether a foetus of known age is growing at an appropriate rate.

Graphical Assessment of Data



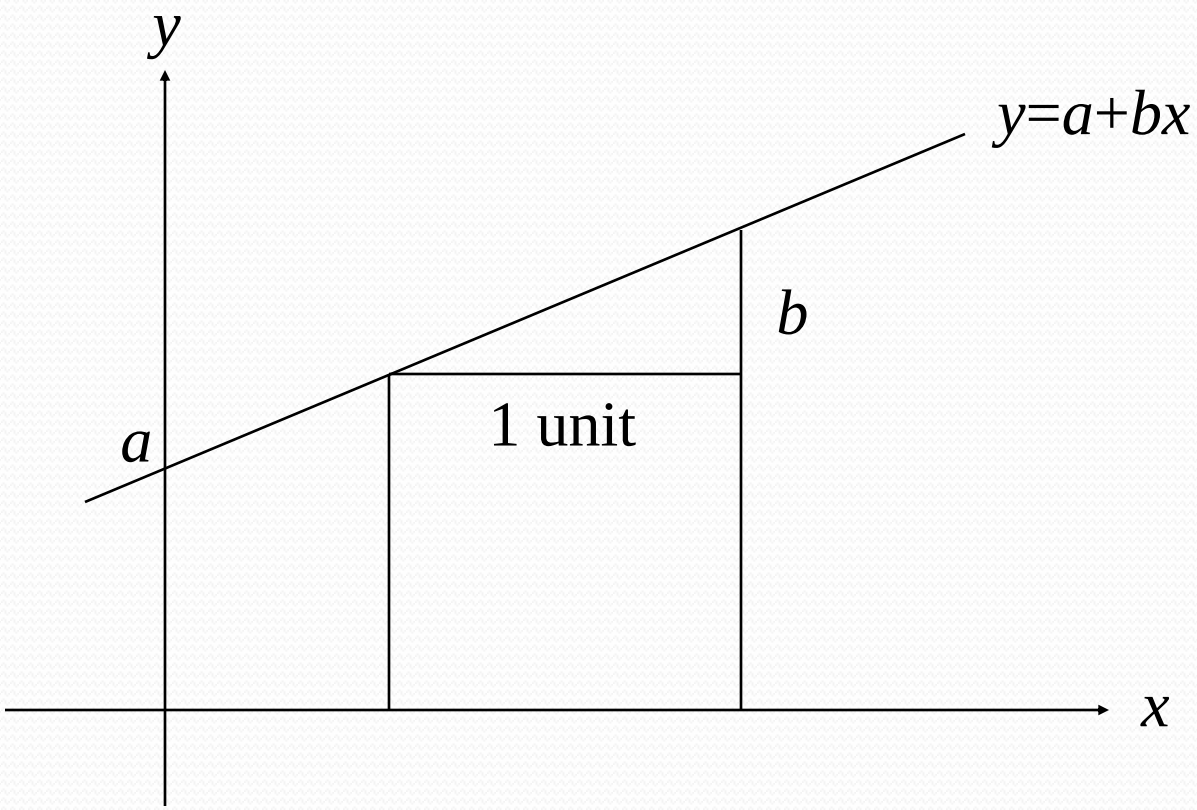
Linear Regression Model

- From the plot it would appear that age and length are strongly related, possibly in a linear way.
- A straight line can be expressed mathematically in the form

$$y = a + bx$$

- Where b is the slope, or gradient of the line, and a is the intercept of the line with the y-axis.

Modelling a Straight Line



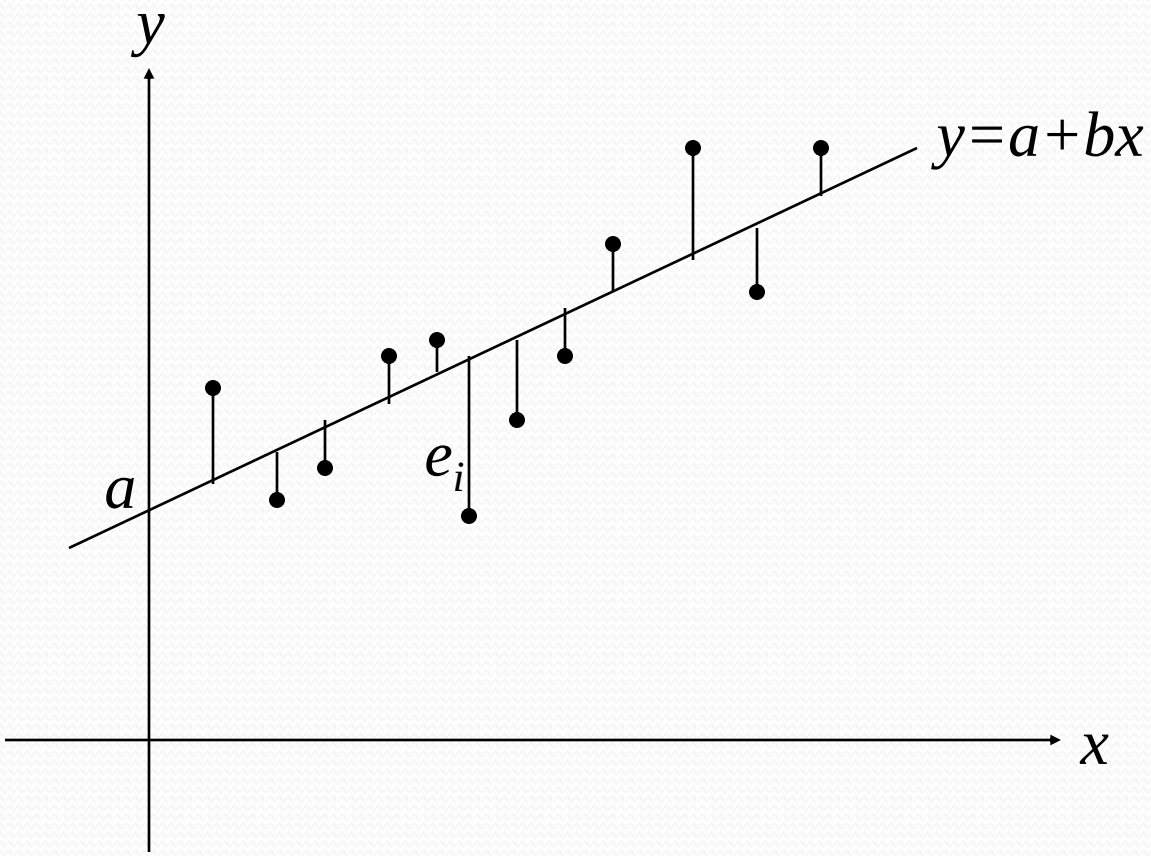
Fitting a Regression Line

- If the data lay on a straight line and there was no random variation about that line, it would be simple to draw an approximate straight line on the scatter-plot.
- This is not the case with real data.
- For a given value of the explanatory variable there will be a range of observed values for the response variable.
- Different assessors would estimate different regression lines.
- In order to have objective results, it is necessary to define some criteria in order to produce the ‘best fitting straight line’ for a given set of data.

Method of Least Squares

- Define the position of the line which, on average, has all the points as close to it as possible.
- The method of *least squares* finds the straight line which minimises the sum of the squared vertical deviations from the fitted line.
- The best fitting line is called the *least squares linear regression line.*
- The vertical distances between each point and the fitted line are called the *residuals* and are used in estimating the variability about the fitted line.

Least Squares Line



Parameter Estimates

- The least squares estimates of a and b are obtained by choosing the values which minimise the sum of squared deviations e_i
- The sum of squared deviations is given by …

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (a + bx_i) \right]^2$$

- Which is a function of the unknown parameters a and b.

Interpretation of Results

- The regression equation is …

$$length = -2.66 + 0.12 \times age$$

- This implies that as the age of the foetus increases by one day, the length increases by 0.12cm.
- For a foetus of age 85 days, the estimated length would be

$$length = -2.66 + (0.12 \times 85) = 7.51$$

- A prediction interval gives the range of values between which the value for an individual is likely to lie: (7.01 to 8.08cm).

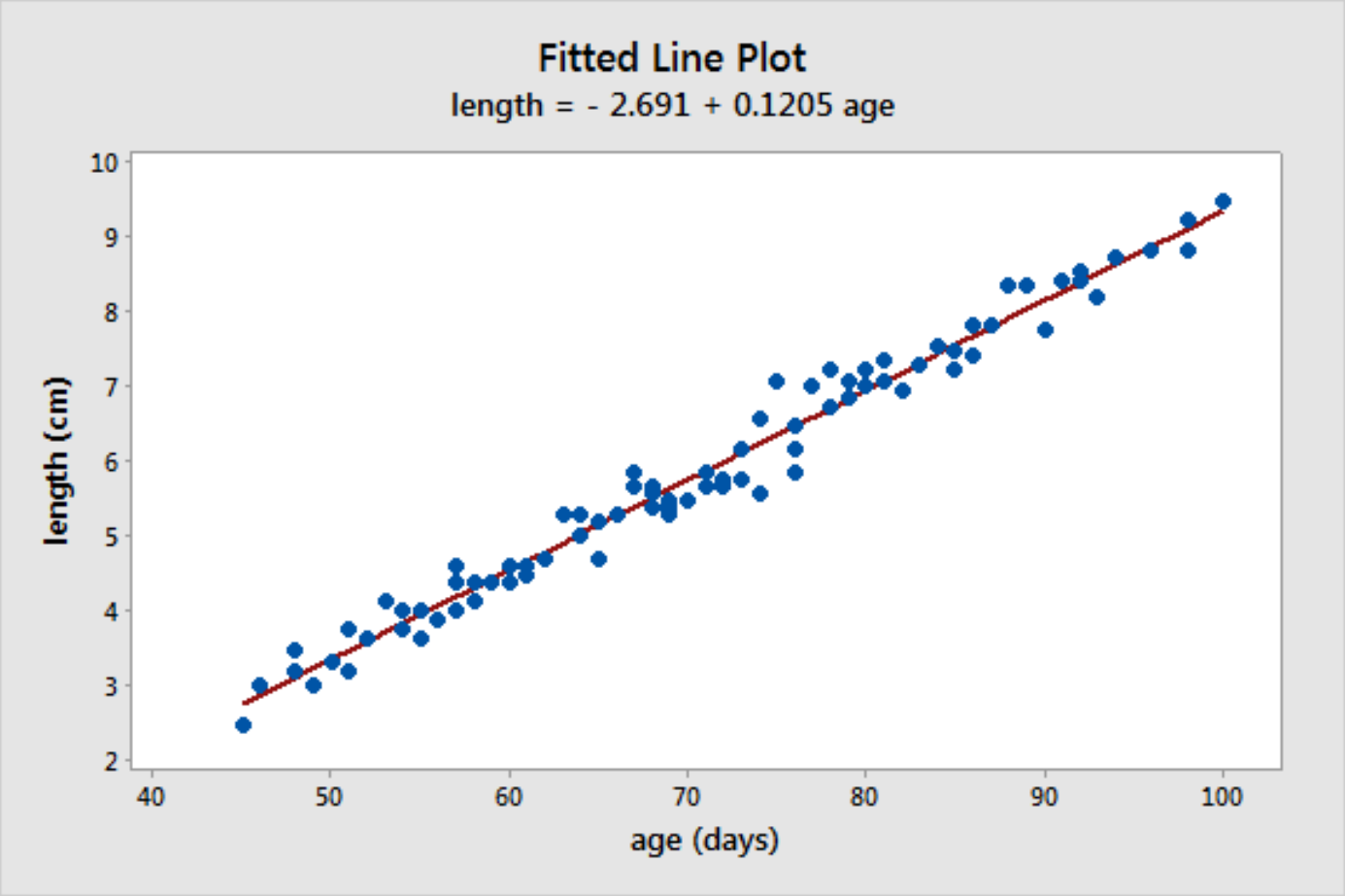
Model Predictions

- Use the regression model to assess whether a foetus of known age is growing at an appropriate rate
- For example, consider a foetus of age 85 days
- Does the measured length lie within the normal range ie between 7.01 and 8.08cm?
- If measured length is <7.01cm, there is evidence that the foetus is not growing as it should
- If measured length is >8.08cm, is the foetus larger than expected? Is the actual age (and due date) wrong?

Uses of Regression Lines

- The least squares regression line may be used to estimate a value of the dependent variable given a value of the independent variable.
- The value of the independent variable (x) should be within the range of the given data.
- The predicted value of the dependent variable (y) is only an estimate.
- Even though the fit of the regression line is good, it does not prove there is a relationship between the variables outside of the values from the given experiment (use care in predicting).

Fitted Line


Fitted Line Plot
length = - 2.691 + 0.1205 age
length (cm)
age (days)
10
9
8
7
6
5
4
3
2
40
50
60
70
80
90
100

Exercise

- Use the calculated least squares linear regression line to estimate the size of a foetus at the following gestation times:

  (a) 2 days

  (b) 60 days

  (c) 100 days

  (d) 300 days

- For each of your estimated lengths, state whether or not you believe the estimate to be accurate.
- Comparison data is available at: https://www.babycenter.com/average-fetal-length-weight-chart