 Year 12 Mathematics Standard 2

| MS-S4 Bivariate data analysis | Unit duration |
| --- | --- |
| Statistical Analysis involves the collection, display, analysis and interpretation of data to identify and communicate key information. Knowledge of statistical analysis enables the careful interpretation of situations and raises awareness of contributing factors when presented with information by third parties, including the possible misrepresentation of information. Study of statistical analysis is important in developing students’ appreciation of how conclusions drawn from data can be used to inform decisions made by groups, such as scientific investigators, business people and policy-makers. | 2 weeks |

| Subtopic focus | Outcomes |
| --- | --- |
| The principal focus of this subtopic is to introduce students to a variety of methods for identifying, analysing and describing associations between pairs of numerical variables. Students develop the ability to display, interpret and analyse statistical relationships related to bivariate numerical data analysis and use this ability to make informed decisions. | A student:* analyses representations of data in order to make inferences, predictions and draw conclusions MS2-12-2
* solves problems requiring statistical processes, including the use of the normal distribution, and the correlation of bivariate data MS2-12-7
* chooses and uses appropriate technology effectively in a range of contexts, and applies critical thinking to recognise appropriate times and methods for such use MS2-12-9
* uses mathematical argument and reasoning to evaluate conclusions, communicating a position clearly to others and justifying a response MS2-12-10

Related Life Skills outcomes: MALS6-2, MALS6-9, MALS6-13, MALS6-14 |

| Prerequisite knowledge | Assessment strategies |
| --- | --- |
| This unit links to the Stage 5 units MA5.2-16SP Bivariate Data Analysis and MA5.2-9NA Linear relationships. In addition, students should have studied the Stage 6 topic MS-A2 Linear relationships. | **Is mathematical modelling better than guessing?** is an investigative task in which students interact with the statistical analysis process to solve to problem of their choice. |

All outcomes referred to in this unit come from [Mathematics Standard 2019](https://syllabus.nesa.nsw.edu.au/mathematics-standard-stage6/) Syllabus © NSW Education Standards Authority (NESA) for and on behalf of the Crown in right of the State of New South Wales, 2017

Glossary of terms

| Term | Description |
| --- | --- |
| Association | An association is a relationship/interconnection between two variables, their values change according to a pattern |
| Bias | Bias generally refers to a systematic favouring of certain outcomes more than others, due to unfair influence (knowingly or otherwise). |
| Bivariate data | Bivariate data is data relating to two variables that have both been measured on the same set of items or individuals. For example, the arm spans and heights of 16-year-olds, the sex of primary school students and their attitude to playing sport. |
| Correlation | measures the strength of the linear relationship between a pair of variables or datasets |
| Dependant | A dependent variable within a statistical model is one whose value depends upon that of another. It is represented on the vertical axis of a scatterplot. The dependent variable is also known as the outcome variable or the output of a function. |
| Extrapolation | Extrapolation occurs when the fitted model is used to make predictions using values that are outside the range of the original data upon which the fitted model was based. Extrapolation far beyond the range of the original data is a dangerous process as it can sometimes lead to quite erroneous predictions. |
| Independent | An independent variable within a statistical model is one whose outcomes are not due to those of another variable and is represented on the horizontal axis of a scatterplot. The independent variable is also referred to as the input of a function. |
| Interpolation | Interpolation occurs when a fitted model is used to make predictions using values that lie within the range of the original data. |
| Least squares regression line | Least-squares regression is a method for finding a straight line that summarises the relationship between two variables, within the range of the dataset.The least-squares regression line is the line that minimises the sum of the squares of the residuals. Also known as the least-squares line of best fit. |
| Line of best fit | A line of best fit is a line drawn through a scatterplot of data points that most closely represents the relationship between two variables. |
| Pearson’s correlation coefficient | Pearson’s correlation coefficient is a statistic that measures the strength of the linear relationship between a pair of variables or datasets. Its value lies between -1 and 1 (inclusive). Also known as simply the correlation coefficient. For a sample, it is denoted by rr. |

| Lesson sequence | Content | Suggested teaching strategies and resources  | Date and initial | Comments, feedback, additional resources used |
| --- | --- | --- | --- | --- |
|  |  | Sources of data* Staff may like to use these sources of data throughout the unit to contextualise the concepts being taught:
* [Australian Bureau of Statistics](http://www.abs.gov.au/) (ABS)
* [Australian Bureau of Meteorology](http://www.bom.gov.au/) (BOM)
* [Australian Sports Commission](http://www.ausport.gov.au/)
* [Australian Institute of Health and Welfare](http://www.aihw.gov.au/) (AIHW)
* [World bank](https://data.worldbank.org/)
* [Google trends](https://trends.google.com/trends/?geo=US)
* [Statista](https://www.statista.com/)
 |  |  |
| Describing bivariate associations(3 lessons) | * construct a bivariate scatterplot to identify patterns in the data that suggest the presence of an association (ACMGM052) **AAM** **Paperclip icon**  Information and communication technology capability icon
* use bivariate scatterplots (constructing them when needed) to describe the patterns, features and associations of bivariate datasets, justifying any conclusions **AAM** Paperclip icon
* describe bivariate datasets in terms of form (linear/non-linear) and, in the case of linear, the direction (positive/negative) and strength of any association (strong/moderate/weak)
* identify the dependent and independent variables within bivariate datasets where appropriate
* describe and interpret a variety of bivariate datasets involving two numerical variables using real-world examples from the media or freely available from government or business datasets  Information and communication technology capability icon Civics and citizenship icon
* calculate and interpret Pearson’s correlation coefficient $\left(r\right)$ using technology to quantify the strength of a linear association of a sample (ACMGM054)  Information and communication technology capability icon
 | Introducing associations of bivariate data sets* Define bivariate data sets (refer to glossary) and introduce students to the concept of an association in the context of statistics.
* Students to examine scatterplots to:
* identify and describe the type of relationships between the two variables. (linear or non-linear)
* identify dependent and independent variables and that the independent variable is shown on the horizontal axis, the dependant on the vertical.
* identify that bivariate data sets can contain time as a variable and that time is always an independent variable, shown on the horizontal axis. Refer to the sources of data.

Resource: associations-of-bivariate-data-sets.DOCXExamining the direction and strength of linear associations* Teacher to lead discussions on the direction (positive or negative) of the trend and the strength of the association (strong/moderate/weak). Approximating a line of best fit may assist students to identify the direction of the association.
* Students to examine scatterplots to determine the direction of a trend and strength of the association. Resource: direction-and-strength-of-a-linear-association.docx
* Students complete [scatter plot capture](https://teacher.desmos.com/activitybuilder/custom/58cc26d4c722f106146a8310) and use observations to make predictions about future points in the plot. Students focus on linear vs nonlinear association, strong vs weak association, and positive vs negative plots
* Students use the [GapMinder](https://www.gapminder.org/tools/#$state$marker$color$data=data_fasttrack&which=income_groups&spaceRef=entities;;;&chart-type=bubbles) website to examine real life trends and correlations between variables such as life expectancy, education and income of countries.
* Teacher to lead discussions on causality:
* Students to examine [spurious correlation](http://www.tylervigen.com/) website and discuss whether there is causality.
* Students to conclude that correlation does not equal causation.
* Note: Teachers need to be mindful of any data that may cause distress to students in their class.

Quantifying the strength of associations* Teacher to explain that the strength of an association can be quantified, Pearson’s correlation coefficient ($r$). Teachers should discuss:
* Values of $r$: 1, -1, 0 positive and negative values.
* Why is this important? It enables us to determine if the association is significant. If there is an association, then we can examine the reasons why this association exists.
* Students to match scatterplots to the appropriate correlation co-efficient by examining the strength and direction of the association. Resource: match-the-correlation-coefficient.DOCX
* Teacher to demonstrate the calculation of Pearson’s correlation coefficient using Excel, Geogebra and the calculator.
* Students to calculate Pearson’s correlation coefficient
* Students source data using google trends and calculate Pearson’s correlation coefficient.
* Students relate the value of Pearson’s correlation coefficient to scatterplots.

Resources: how-to-guide-sourcing-bivariate-data-using-google-trends.DOCX, how-to-guide-ms-excel-regression-analysis.DOCX, how-to-guide-calculator-regression-analysis.DOCXNESA exemplar question1. The height and length of the right foot of $10$ high school students were measured. The results were tabulated as follows:

A table comparing the heights and length of the right foot of 10 high school students* 1. Using technology, calculate the Pearson correlation coefficient for the data.
	2. Describe the strength of the association between height and length of the right foot

Resource: ms-s4-nesa-exemplar-question-solutions.DOCX |  |  |
| Developing linear models(3 lessons) | * model a linear relationship by fitting an appropriate line of best fit to a scatterplot and using it to describe and quantify associations AAM Paperclip icon
* fit a line of best fit both by eye and by using technology to the data (ACMEM141, ACMEM142)  Information and communication technology capability icon
* fit a least-squares regression line to the data using technology  Information and communication technology capability icon
* interpret the intercept and gradient of the fitted line (ACMGM059)
 | Introducing lines of best fit* Teacher to lead discussions on what a line of best fit is and why it is important:
* It enables predictions to be made using the model.
* Example: If we have sales history we can then predict how much we will sell in the future.
* Students to fit a line of best fit by eye and consider the concept of residuals using the [Desmos - line of best fit](https://teacher.desmos.com/activitybuilder/custom/56fab6bc1ab86b1f0600369d) activity.

Developing a line of best fit by eye* Teacher to model the fitting of a line of best fit by eye using a digital scatterplot and finding the equation of the line of best fit. Resource: Part 2, how-to-guide-desmos-regression-analysis.DOCX.
* Students to fit a line of best fit by eye:

Part 1: Students use technology to produce lines of best by eye. After completion of Part 1, teacher to lead a discussion on * Which graphs they found more difficult to fit by eye? (compare this to the correlation coefficient/or description of the strength of association for each data set)
* Why do students obtain different lines?

Part 2: Students develop lines of best fit by eye for practical data and interpret the intercept and gradient of the line with respect to the variables graphed.Resources: fitting-a-line-of-best-fit-by-eye-activity.DOCX, data-file-1.XLSX, data-file-2.XLSX, how-to-guide-desmos-regression-analysis.DOCXDeveloping the least squares regression line* The teacher rediscusses the methods used by students to fit a line of best fit by eye and introduces the idea of residuals and the least-squares regression line. Please note that residuals are not mentioned in the syllabus and only serve the purpose of developing the concept for least squares regressions.
* The teacher demonstrates finding the equation of the least squares regression line using a calculator then fit it to a scatterplot. The scatterplot could be constructed and the line graphed using graphing software.

Resource: how-to-guide-calculator-regression-analysis.DOCX* The teacher demonstrated fitting a least-squares regression line using digital technology such as Desmos, Geogebra or Excel and record its equation.
* Students practice the methods using basic examples. Students will reinforce these methods when interpolating and extrapolating in the following lesson.

**Resources:** how-to-guide-desmos-regression-analysis.DOCX, how-to-guide-ms-excel-regression-analysis.DOCX |  |  |
| Applying and interpreting the linear model to make predictions(1 lesson) | * use the appropriate line of best fit, both found by eye and by applying the equation, to make predictions by either interpolation or extrapolation Paperclip icon
* recognise the limitations of interpolation and extrapolation, and interpolate from plotted data to make predictions where appropriate (ACMGM062)  Information and communication technology capability icon
 | Introducing interpolation and extrapolation* Interpolation: A fitted model is used to make predictions using values that lie within the range of the original data.
* Extrapolation: A fitted model is used to make predictions using values that are outside the range of the original data.
* Relate interpolation and extrapolation to the application of Hawk-eye to tennis and cricket. Refer to The Guardian article [hawk-eye at Wimbledon](https://www.theguardian.com/science/sifting-the-evidence/2013/jul/08/hawk-eye-wimbledon).

Predicting by Interpolating and Extrapolating* The teacher models methods of making predictions by interpolating and extrapolating. Teachers may like to use this [Geogebra app](https://ggbm.at/XwC4p4Qc) linking Fuel Use and Engine Size to illustrate the concept:
* The graphical method: use the graph of a line of best fit to read predicted values.
* Students may need to extend their line of best fit to allow extrapolation.
* The algebraic method: substitute into the equation of a line of best fit and then evaluate the resulting expression or solve the resulting equation to make a prediction.
* Students compare the predictions obtained using the two methods and consider the benefits of each.
* Student activity: Students interpolate and extrapolate to make predictions, including in context, and examines variations in the predictions obtained using the graphical and algebraic method as well as using a line of best by eye and the least squares regression line.

Resources: interpolating-and-extrapolating-activity.DOCX, data-file-1.XLSX, data-file-2.XLSX, how-to-guide-calculator-regression-analysis.DOCX, how-to-guide-ms-excel-regression-analysis.DOCX, how-to-guide-desmos-regression-analysis.DOCXExamining the limitations of interpolation and extrapolation* Student activity: Students use the [Alligator Investigation](https://teacher.desmos.com/activitybuilder/custom/5670aca10255cd610798be54) to look at the limits of linear models. “An enormous alligator lurks in the swamp. Can scatterplots and least-squares regression tell you if you have enough animal tranquilizer to stay safe?”
* Teacher to lead students to identifying the limitations of making predictions using a model. For example:
* If a model’s correlation is weak, the accuracy of predictions will diminish.
* How confident are we that the model will behave in the same pattern? When extrapolating, we need to determine if it is reasonable to assume the model is valid outside of our data range.

NESA exemplar question1. Ahmed collected data on the age ($a$) and the height ($h$) of males aged $11$ to $16$ years. He created a scatterplot of the data and constructed a line of best fit to model the relationship between the ages and height of males.

A graph showing the age and height of males and a line of best fit for the data.* 1. Determine the gradient of the line of best fit shown on the graph.
	2. Explain the meaning of the gradient in the context of the data.
	3. Determine the equation of the line of best fit shown on the graph.
	4. Use the line of best fit to predict the height of a typical $17$-year-old male.
	5. Why would this model not be useful for predicting the height of a typical $45$-year-old male?

Resource: ms-s4-nesa-exemplar-question-solutions.DOCX |  |  |
| Applying linear models to investigate real world issues(3 lessons) | * solve problems that involve identifying, analysing and describing associations between two numerical variables **AAM** **Paperclip icon**
* construct, interpret and analyse scatterplots for bivariate numerical data in practical contexts AAM **Paperclip icon** Aboriginal and Torres Strait Islander histories and cultures icon Asia and Australia’s engagement with Asia icon Ethical understanding icon  Information and communication technology capability icon
* demonstrate an awareness of issues of privacy and bias, ethics, and responsiveness to diverse groups and cultures when collecting and using data
* investigate using biometric data obtained by measuring the body or by accessing published data from sources including government organisations, and determine if any associations exist between identified variables Critical and creative thinking icon  Information and communication technology capability icon
 | The statistical investigation process* The teacher is to model the complete statistical investigation process:
* Identifying a problem
* Posing a statistical question
* Collecting or obtaining data: Demonstrating awareness of issues of privacy and bias, ethics, and responsiveness to diverse groups and cultures.
* Representing and analysing that data (Completing regression analysis – represent bivariate data on a scatterplot, examine the strength of an association informally and using Pearson’s correlation coefficient, fit a line of best fit, find its equation, use the line or equation to interpolate or extrapolate)
* Communicating and interpreting findings
* Suggested investigations:
* Sales Forecasting Activity

Resources: sales-forecasting-activity.DOCX, sales-forecasting-activity-teachers-guide.DOCX, sales-forecasting-activity-teachers-resource.DOCX* The Ebola crisis

Resource: ebola-crisis-activity.docx* Mobile phone battery life

Resource: mobile-phone-battery-life-activity.docx* Sales forecasting using demographics

Resource: popmobile-activity.docx |  |  |

Reflection and evaluation

Please include feedback about the engagement of the students and the difficulty of the content included in this section. You may also refer to the sequencing of the lessons and the placement of the topic within the scope and sequence. All information and communication technologies (ICT), literacy, numeracy and group activities should be recorded in the ‘Comments, feedback, additional resources used’ section.