Linear Relationships

(…and how Maths can help you choose a great car)

An AMSI Schools CHOOSE**MATHS** Rich Task

for Students in Years 9 & 10

Image credit: License CC0 Public Domain. URL https://www.publicdomainpictures.net/en/view-image.php?image=268518&picture=buying-new-car. Downloaded 18 November 2018.

This unit was developed in collaboration with
Ms Natalie Hammond from Bayside P-12 College, Victoria.

PART 1 – WHOLE CLASS ACTIVITY

Introduction

Linear relationships – that is, situations in which one value is directly influenced by *only one* other value – rarely occur in nature. Usually, any *naturally* occurring variable such as the daily temperature, the rate of growth of a plant or animal or the spread of bacterium are influenced by a *range* of other factors.

Linear relationships are more common in areas such as finance, for example, when the revenue raised by sales of an item is influenced only by the number of sales.

Nevertheless, there are some situations in which natural phenomena come *close to* exhibiting linear relationships. This is because, in some situations, a variable may change mainly due to the influence of a strongly dominant other variable. An example is the number of cricket chirps per minute amongst some species of cricket, which is *strongly influenced by* the temperature of the air where the crickets are active (crickets tend to chirp at a faster rate when the air is warmer).

In such situations, a plot of the data points (e.g. 'chirps per minute' against 'air temperature') will not necessarily form a perfect straight line. However, we can form a 'line of best fit' which most closely approximates the expected relationship between the two variables. This line will thus give us a kind of average or expected relationship between two variables. We can then use this estimated linear relationship to predict how an *independent variable* (e.g. a given air temperature on a Summer's night) is likely to affect a *dependent variable* (e.g. the rate of chirping of the crickets on that same night).

In this unit of work, we are going to examine how we can use a fairly simple statistical technique of plotting two related variables together on a scatterplot and finding a 'line of best fit' which approximates a linear relationship between the two variables. We can then find an equation that expresses this graphical relationship. We'll then use both the graph and the algebraic linear equation to help us analyse and predict expected values.

Background Reading

The following online module from VisionLearning will provide you with further information and background on Linear Relationships in nature. Read through this before you go on to examine the linear equation example below, which explores the relationship between humans' femur (upper leg bone) length and their overall height.

https://www.visionlearning.com/en/library/Math-in-Science/62/Linear-Equations-in-Science/194/reading

Example Bivariate Analysis and Linear Equation:

Femur Length and Overall Height in Humans

You have hopefully read the case study in the above VisionLearning regarding the relationship between the length of a person's femur bone and their overall height. You are now going to collect data from your class together, to demonstrate this relationship. We'll use a Microsoft Excel spreadsheet to record the data and generate the graph.

The sample spreadsheet supplied by your teacher with this unit of work will generate a scatterplot for you and will also find a 'line of best fit' that uses an averaging technique called 'regression analysis' (more about that in your advanced maths course when you choose that in your senior years at school!).

From this, we can use the features of this graphed line to find an algebraic equation that expresses an approximated linear relationship between between the femur lengths and overall heights of you and your classmates.

This equation will express the y value (the overall height of a person) of the linear trendline in terms of its x value (the length of that person's femur bone). This equation will be in the form of:

$y = mx + c$,

where y is the *dependent* value (that is, the value that is *affected* by something else);

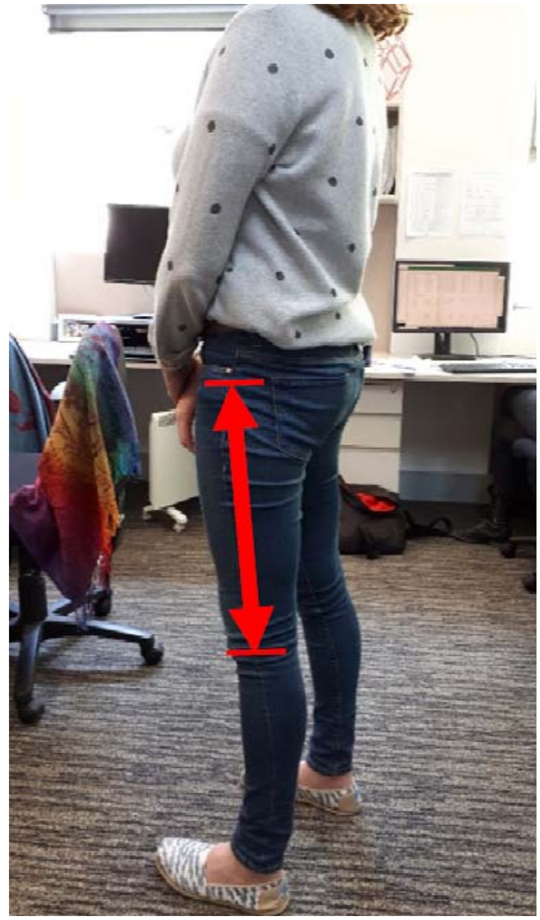
x is the *independent* value (the value that changes, but is not affected by anything else of interest in the analysis);

m is the *gradient* of the trendline (that is, the value we *multiply* x by); and

c is a fixed or *constant* value that is the same for all the data points, that we add to (or subtract from) the multiplied x value.

Activity:

Using a tape measure, working together as a class, measure the distance between just below the hip to the middle of the knee for each of your classmates, and record this in centimetres (to the closest centimetre).



Record your results in the table below.

Name (optional)	**Length of Femur** (cm)	**Overall Height** (cm)

Your teacher will now transcribe the class's results into the spreadsheet.

In the spreadsheet, you will see something like this as your class's results:

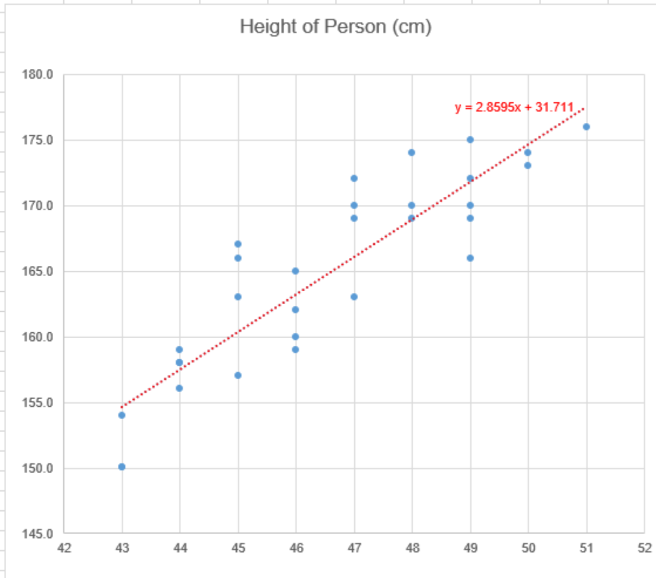
*(Note: The data in this example is **fictional**!)*

Relationship between Femur Length and Height in Homo Sapiens

Note that we have chosen the x-axis for the femur length and the y-axis for the person's height. Why is this?

Person ID	x = Femur Length (cm)	y = Height of Person (cm)
1	46	165.0
2	44	159.0
3	47	169.0
4	48	169.0
5	49	170.0
6	49	169.0
7	46	159.0
8	43	154.0
9	43	150.0
10	49	172.0
11	45	163.0
12	47	163.0
13	46	162.0
14	48	170.0
15	51	176.0
16	49	166.0
17	44	158.0
18	47	172.0
19	50	173.0
20	48	174.0
21	44	158.0
22	45	166.0
23	45	167.0
24	49	175.0
25	44	156.0
26	46	160.0
27	45	157.0
28	47	170.0
29	43	150.0
30	50	174.0

Insert a scatterplot here to analyse this bivariate data set:



The data set in the picture above shows that in this **fictional** sample of 30 humans, the relationship between a person's height (y) and their femur length (x) (as estimated by the distance between the bottom of their hip and the middle of their knee) is given by:

$$y = 2.8595x + 31.711$$

Let's round the decimals to one decimal place, for the sake of simplicity. So,

$$y = 2.9x + 31.7$$

This would mean that, according to this fictional sample, we could estimate that a person's height in centimetres could be determined as '2.9 times their femur length (in cm), PLUS 31.7 cm'.

Thus, if a person had a femur length of 44 cm, we could estimate their expected height as being:

$$\text{Height } (y) = 2.9 \text{ x } 44\text{cm} + 31.7\text{cm}$$

$$= 127.6 \text{ cm} + 31.7 \text{ cm}$$

$$= 159.3 \text{ cm.}$$

Look at your class's data on the spreadsheet graph you have generated with your teacher.

What is the linear equation that expresses the trendline for your class?

(round your numbers to one decimal place)

____

Discuss this with the class.

What does your answer above tell you about the relationship between femur length and overall height for your class? Try to express this in words.

____

____

Imagine a new student joins your class. Their femur length is measured at 41.5 cm.

From this information and your own data, calculate the new class member's expected height in centimetres. Show your working.

____

____

____

The accepted linear relationship for the human population overall, according to the VisionLearning module you read online, is given by the equation:

Height (cm) = 1.88 x femur length (cm) + 32.01 cm

OR

$y = 1.88\,x + 32.01\text{cm}$

How close is this to your class's linear equation predicting height from femur length?

Have a class discussion about possible reasons why your equation may differ from the one for the general overall human population.

Enrichment Explanation: What is 'Regression Analysis' and a 'Line of Best Fit'?

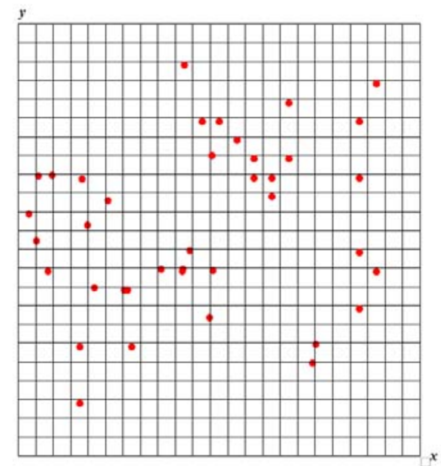
Regression Analysis is a statistical technique designed to find a line of best fit through a set of data points on a scatter plot. It allows us to find an approximated or 'averaged' linear relationship through a set of data points that are scattered across a Cartesian plane (i.e. a graph with x and y axes).

For many data samples, graphed onto a scatter plot, there is not a *readily apparent* linear relationship through the data.

Regression analysis allows us to find a line through the data which is the optimal or 'best possible' approximation of any relationship that might be applied to a data set.

As we will see in this unit, fitting a trend line 'by the eye' is not usually a very accurate way of finding the actual 'line of best fit' through data.

Regression analysis uses a technique called 'minimisation of the sum of squared errors' to ensure that the actual line of best fit is one *from which the differences away from this line are minimised overall*. The 'errors' here are the distances from the actual data point to the line of best fit.



Where would the 'line of best fit' go through this scatterplot?

Regression analysis minimises the sum of the squares of all the x and y values on a scatter plot and then finds an *averaged* linear relationship between x and y from which *the differences between this linear relationship and the total of the squared values of all the actual data points is minimised*.

This linear relationship will have the form $y = mx + b$. The regression analysis performed on the data set pushes and pulls the constant value 'b' and the gradient 'm' up and down until the averaged-out distances of all the points above and below line $y = mx + b$ are at their very smallest.

When we ask a spreadsheet such as Microsoft Excel to find the line of best fit on a scatterplot, it is in fact performing this little piece of mathematics on the data to generate the line.

For some further explanation of how this works mathematically, have a look at the article 'Least Squared Regression' on the 'Math Is Fun' website, and look at the video from Khan Academy, 'Introduction to residuals and least-squares regression', both linked below:

https://www.mathsisfun.com/data/least-squares-regression.html

https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/least-squares-regression/v/regression-residual-intro

Regression analysis is an interesting and highly useful piece of statistical mathematics and used frequently in all sorts of areas of endeavour, including science, engineering, health, economics, finance and information technology.

PART 2 - INDIVIDUAL TASK

Background

In this task, we are going to be looking at an estimated (or predicted) linear relationship between the 'asking price' (that is, how much a seller of a car wants) for a used car, and the distance that car has travelled in kilometres.

To do this, we are going to collect some data, plot it on a scatter plot and use this to find a linear algebraic relationship, just as we did in our whole class activity.

The distance a car has travelled overall during its life is measured by an instrument on the dashboard called an 'odometer'.

A new car will usually have an odometer reading between 0 and 500 km (it may have been driven from the port or factory to the car yard), and much older cars can have odometer readings of over 300,000 km! (Usually by this stage, however, the car is pretty worn out and unlikely to travel much further reliably.)

Getting Started

Grab your SmartPhone or a laptop with a browser, and go to the 'Car Sales.com' website by clicking on the following URL:

https://www.carsales.com.au/

1. Choose a make and model of car from the following: Suzuki Swift, Mazda 2, Toyota Yaris, Hyundai i20.
2. Select the 'Used' car tab;
3. Set your search for 'Melbourne' region only;
4. Select vehicles ranging in age between the years 2000 and 2016 *only* (if that option is not available on the front page menu, you can set the date range later, after you have hit 'Show Me Cars'
5. *Don't* add any other 'key words'.

Choose 5 or 6 cars from the results and note down (a) the price of each car, and (b) the kilometres it has travelled (from the 'Odometer' field).

Write your sets of these two data variables into the table below.

Now… select a different make of car from the four makes above, repeat the process, select another 5 or 6 car prices and odometer readings and write these into the table.

Repeat this until you have 20 sets of data filled in the table on the next page.

Each of these 20 sets of data (car price and car odometer reading) we will refer to as ‘data points’.

Data Point	Km Travelled (‘Odometer’) (x)	Car Price (y)
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		

Complete the following activities relating to the data set you have collected from CarSales.com and the relationship between used car price and kilometres travelled.

(A) *Understanding* Linear Relationships

(i) Considering these two variables (y = Price of Used Vehicle and x = Odometer Reading or 'km travelled'), which of these do you think is the '*independent* variable' and which is the '*dependent* variable', and why?

___

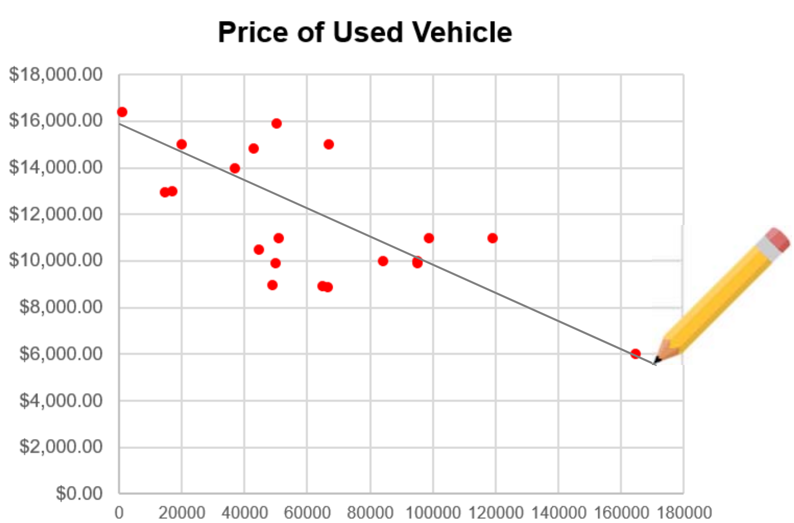
___

___

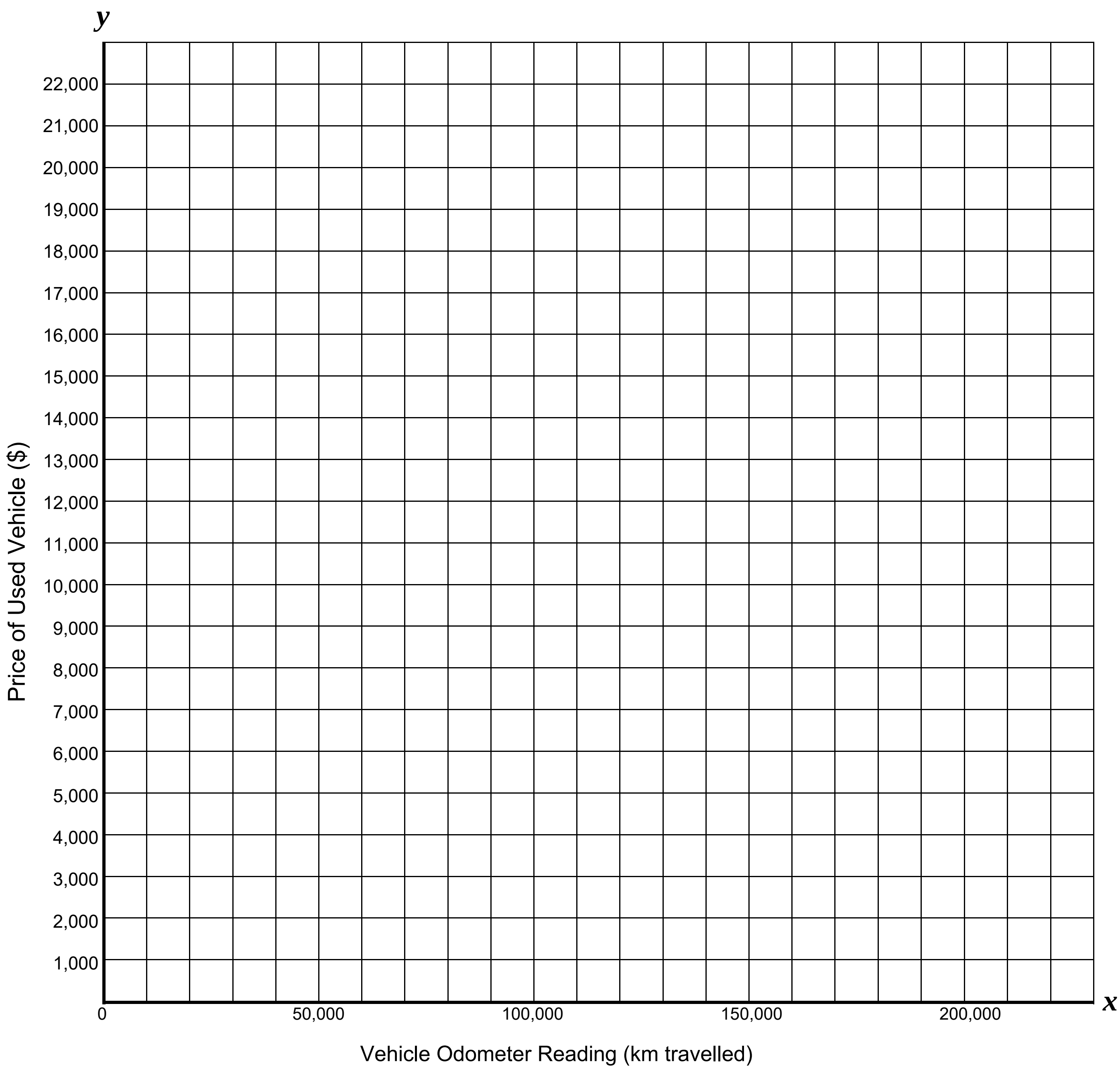
(ii) Use the graph paper on the next page to create a scatterplot with the 20 data points you have collected from CarSales.com. Look carefully at the overall pattern of the distribution of points on your scatterplot.

(iii) Using your own visual judgement, **draw a straight line of 'best fit'** on your scatterplot that you think most closely approximates the trendline of your data.

For example, if your scatterplot was as follows, you might use a ruler and a pencil to draw a line a bit like this:



Scatterplot – Price of Used Vehicle and Odometer Reading of Used Vehicle



(iv) Derive (work out) an algebraic equation for your 'line of best fit', in the form $y = mx + c$,

where m is the *gradient* of the trendline (that is, the value we *multiply* 'x' by); and

a fixed or constant value 'c' that is the same for all the data points (that we add (or subtract) to the multiplied x value).

(v) In Part 1, your teacher has shown you how to create a scatterplot with a trendline, showing the algebraic linear equation, using Microsoft Excel.
Use Excel yourself to compose an electronic version of the graph for your data (collected and recorded above in Part 2 'Getting Started'.

Your Excel scatterplot should include each of:

- Labels and title for the scatterplot graph;
- trend line for the 'line of best fit' through the data; and
- the algebraic linear equation that expresses the trendline.

(vi) Look at your scatterplot graph and the trendline you have inserted.
Using your Excel data as a guide, how much would you expect to pay (all other factors being equal), for a car that has travelled 100,000 km?

(vii) What do each of the 'x' and the 'y' values represent in your linear equation and what does the linear relationship between y and x *mean* in the context of buying a used car?

(viii) What is the 'constant' ('c') in your electronically generated linear equation, and what does this tell us about buying this type of used car in the local used car market?

(B) Attaining *Fluency* with Linear Relationships

(i) Look carefully at both your hand drawn scatterplot and trendline (line of best fit), and then at your electronic version, and compare them both.
What are the similarities between your two linear trendlines? What are the differences?
For example, were there any differences between:

- The constants (‘c’)?
- The gradients (‘m’)?

Why might this be?

(ii) What does the value of the ‘y’ in the y ***intercept*** represent?

(iii) What does the value of the ‘x’ in the x ***intercept*** represent?

(iv) In the same used car market for a *different* vehicle class, a buyer can expect to pay $16,500, minus $3 for every 100 km travelled (that is, minus 3 cents, or $0.03, for every km travelled).

Write the linear equation that describes or models what a buyer can expect to pay for this type of vehicle in the used car market and *draw it onto your graph* on page 9.
*(Important: Use a **different colour** to draw this linear function, and label it carefully).*

(v) If the vehicle mentioned above in question (iv) had been driven 250,000 km, what would you expect to pay for this car in the used car market? *Show or explain your working.*

(C) Solving Problems using Linear Relationships

Select another make and model of vehicle (eg. Holden Commodore or Hyundai i40…) of your own choice within the same year of manufacture range (eg. 2010 – 2017).

Collect 20 data points, enter it into a new Excel spreadsheet and then run the same type of linear analysis for this new selected vehicle.

You have a budget of $15,000 with which to purchase a used car, excluding the stamp duty and other transfer of ownership fees.

Is it likely that you could afford to purchase this new type of car within your budget?

Think carefully about whether the car would really be 'a good buy', given the amount of distance it has travelled. You may need to do some research on what makes for 'high mileage' in a car, and when a car's distance travelled makes it really quite risky to purchase.

Give reasons and an explanation for your response by referring to the results of your linear analysis.

(D) Applying Reasoning using Linear Relationships

(i) The *selling price* of one of the vehicles within the set of parameters is given on CarSales as $12,990.

Using your Excel graph of the linear equation from Part 2 (A) above, *estimate the number of kilometres you would expect this car to have travelled.*
Now, *rearrange the linear equation* to isolate the 'x' *(km travelled)* variable, to provide a *numerical proof* for the value of x when y (price) = $12,990.

Estimated odometer reading for car priced at $12,990: ________________________________

Mathematical proof for exact value of x when y = $12,990:

____

____

____

____

____

____

____

(ii) What other factors might be affecting the **variability** of this data, that is, the fact that there are some data points above this line and some below it?
In other words, what other factors might be affecting the relationship between specific used vehicles and their price, *besides* the distance it has travelled?

____

____

____

____

(iii) What general *effect* would you expect each of these other factors to have on the dependent variable y in this analysis (i.e., would the value of y increase or decrease)?

(iv) How could using a simple graphical analysis like this help you make a clear, well informed decision about buying a used car?

Give examples by talking about:

- Why you would or would not consider a vehicle with a data point that sits *above* the linear equation line;
- Why you would or would not consider a vehicle with a data point that sits *below* the linear equation line; and
- How you could use this information to help you *negotiate* the price of a vehicle with a seller of a similar vehicle.

References

CarSales.com.au (2018). URL: https://www.carsales.com.au/#{searchmakemodel:{Service:Carsales,SiloType:siloTypeUsed,State:New

Gloag, A. and Gloag, A. (2015). Modified from *'Introductory Algebra'*, by CC-BY 2012, CK-12 Foundation, www.ck12.org. Licensed under a Creative Commons Attribution 3.0 Unported License; Modified from '*Precalculus: An Investigation of Functions'*, by David Lippman and Melonie Rasmussen, CC-BY 2015, under a Creative Common Attribution-Share Alike 3.0 United States License (http://creativecommons.org/licenses/by-nc-sa/3.0).

Hoekenga, C., Carpi, A. (Ph.D) and Egger, A.E. (Ph.D) (2013), *'Linear Equations in Science: Relationships with Two Variables'.* Article on VisionLearning.com, URL: https://www.visionlearning.com/en/library/Math-in-Science/62/Linear-Equations-in-Science/194. Accessed 18 November 2018.