



Using a least squares line to model a relationship between two numerical variables

Year 12 General Maths
Units 3 and 4

www.maffsguru.com

Learning Objectives

By the end of the lesson, I would hope that you have an understanding and be able to apply to questions the following concepts:

- To be able to interpret the intercept and slope of the least squares line.
- To be able to use the equation of the least squares line to make predictions.
- To be able to use the coefficient of determination in a regression analysis.
- To be able to use a residual plot to investigate the linearity assumption.
- To be able to report a regression analysis.



Recap

In the previous lesson we looked at how we can use “by hand” and the CAS to find the least squared regression line for two numerical data items which have been plotted on a scatter plot. We learned why we do this and how to interpret the slope and the intercept.

As this is a really important topic this lesson will look at how we can tie it all together and package it in report.

Note: This is the main fodder for SACs. I would suggest you take the time to completely understand how to do this.

$$= a + bx$$

$$b = \frac{r \cdot S_y}{S_x} \quad a = \bar{y} - b \cdot \bar{x}$$



The data

The age (in years) and price (in dollars) of a selection of second-hand cars of the same brand and model have been collected and are recorded in a table (shown).

Age (years)	Price (dollars)	Age (years)	Price (dollars)	Age (years)	Price (dollars)
1	32 500	3	22 000	5	18 400
1	30 500	4	22 000	6	6 500
2	25 600	4	23 000	7	6 400
3	20 000	4	19 200	7	8 500
3	24 300	5	16 000	8	4 200

During the course of this lesson we are going to:

- constructing a scatterplot to investigate the nature of an association
- calculating the correlation coefficient to indicate the strength of the relationship
- determining the equation of the regression line
- interpreting the coefficients of the y -intercept (a) and the slope (b) of the least squares regression line $y=a+bx$
- calculating and interpreting the coefficient of determination
- using the regression line to make predictions
- calculating residuals and using a residual plot to test the assumption of linearity
- **writing a report on your findings.**



The scatter plot

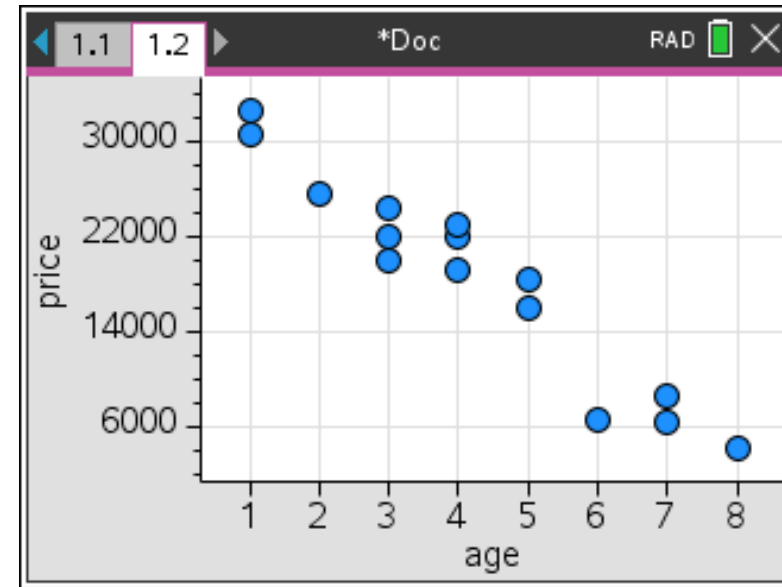
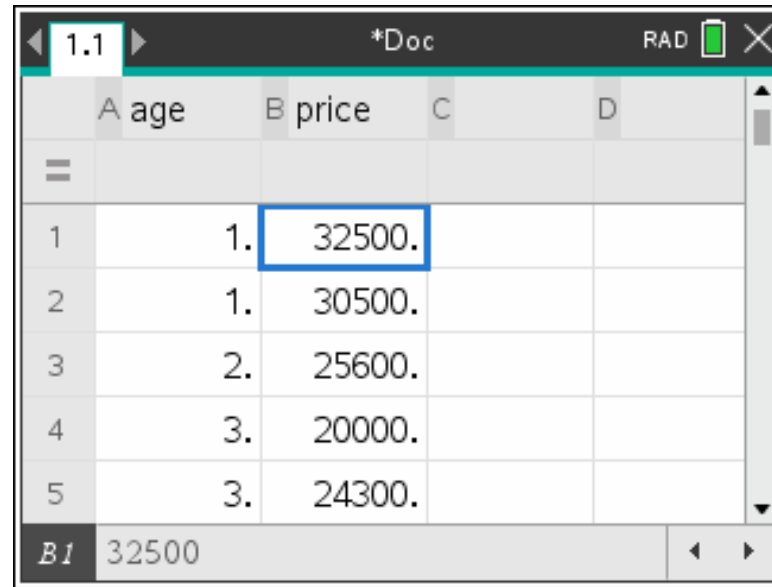
Make sure that you always put the table headings first.

Check you have the correct number of data items and scroll up and down to make sure they "look" correct.

It should be obvious if there is a wrong number.

Make sure you have the EV and RV on the correct axes.

Age (years)	Price (dollars)	Age (years)	Price (dollars)	Age (years)	Price (dollars)
1	32 500	3	22 000	5	18 400
1	30 500	4	22 000	6	6 500
2	25 600	4	23 000	7	6 400
3	20 000	4	19 200	7	8 500
3	24 300	5	16 000	8	4 200



Calculate the correlation coefficient

This can be found from the CAS.

You can either do a regression analysis or use two variable statistics.

Use the value to write down the strength of the relationship.

$$r = -0.964$$

Age (years)	Price (dollars)	Age (years)	Price (dollars)	Age (years)	Price (dollars)
1	32 500	3	22 000	5	18 400
1	30 500	4	22 000	6	6 500
2	25 600	4	23 000	7	6 400
3	20 000	4	19 200	7	8 500
3	24 300	5	16 000	8	4 200

TI-84 Plus calculator screen showing data entry for age and price. The screen displays a table with columns A (age) and B (price). The data points are: (1, 32500), (2, 30500), (3, 25600), (4, 20000), (5, 24300). The value 32500 in cell B1 is highlighted with a blue box.

A	age	B	price	C	D
1	1.	32500.			
2	1.	30500.			
3	2.	25600.			
4	3.	20000.			
5	3.	24300.			

TI-84 Plus calculator screen showing linear regression results. The screen displays a table with columns price, C, D, and E. The data points are: (3, 25600., a, 35134.0...), (4, 20000., b, -3935.0...), (5, 24300., r², 0.92979...), (6, 22000., r, -0.9642...), (7, 22000., Resid, {1301.03...}). The value -0.9642... in cell E6 is highlighted with a red box.

price	C	D	E
3	25600.	a	35134.0...
4	20000.	b	-3935.0...
5	24300.	r ²	0.92979...
6	22000.	r	-0.9642...
7	22000.	Resid	{1301.03...



Find the equation of the regression line

This can be done in lots of ways.

The screen below shows the values of 'a' and 'b' which allow you to write directly into the formula

Age (years)	Price (dollars)	Age (years)	Price (dollars)	Age (years)	Price (dollars)
1	32 500	3	22 000	5	18 400
1	30 500	4	22 000	6	6 500
2	25 600	4	23 000	7	6 400
3	20 000	4	19 200	7	8 500
3	24 300	5	16 000	8	4 200

price	C	D	E
=			=LinRegB
3 25600.	a	35134.0...	
4 20000.	b	-3935.0...	
5 24300.	r ²	0.92979...	
6 22000.	r	-0.9642...	
7 22000.	Resid	{1301.03...	
E1	="Linear Regression (a+bx)"		

$$\text{price} = 35134 - 3935 \times \text{age}$$



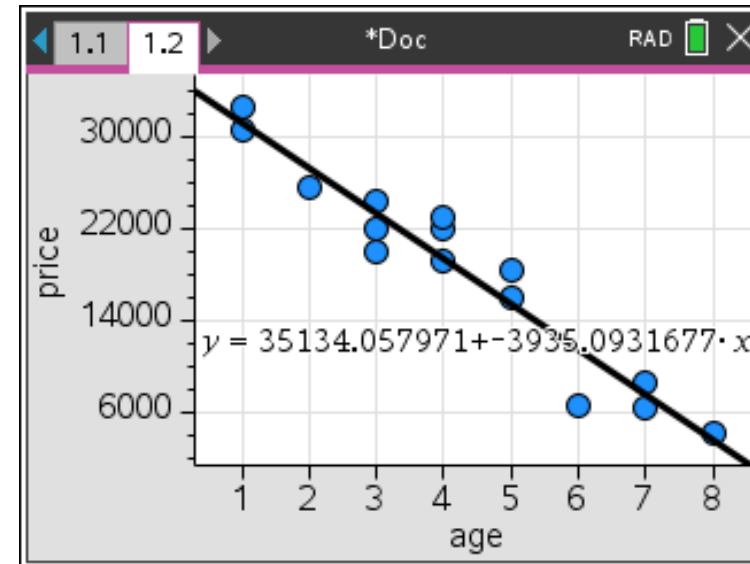
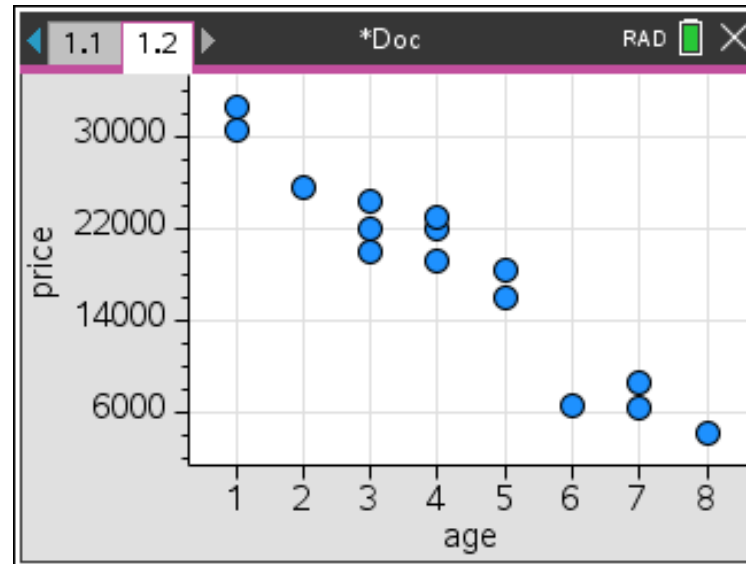
Find the equation of the regression line

This can be done in lots of ways.

Read the values from the regression line on the graph.

BEWARE: Your CAS does not know what are on the 'x' and 'y' axes. You must write the names of the EV and RV and not 'x' and 'y'.

Age (years)	Price (dollars)	Age (years)	Price (dollars)	Age (years)	Price (dollars)
1	32 500	3	22 000	5	18 400
1	30 500	4	22 000	6	6 500
2	25 600	4	23 000	7	6 400
3	20 000	4	19 200	7	8 500
3	24 300	5	16 000	8	4 200



Interpret the coefficients

It is vitally important that you use the correct wording when explaining the values of 'a' and 'b'.

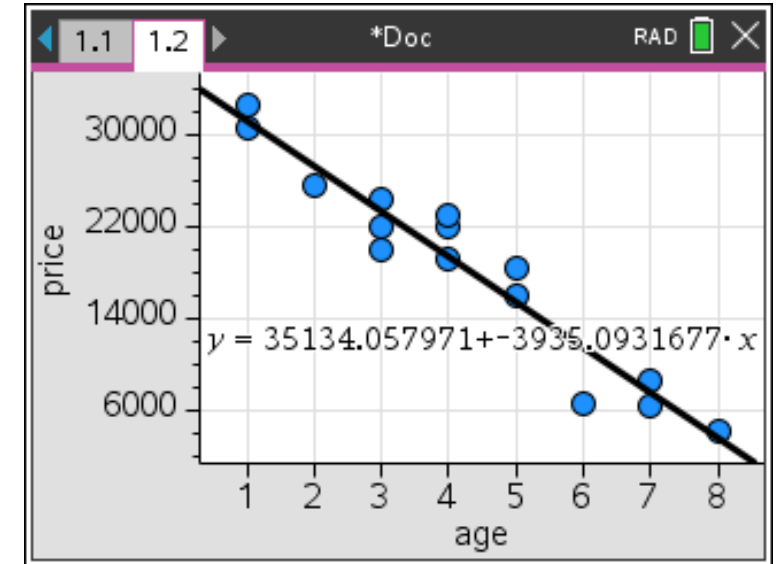
The slope of the regression line predicts that, on average, the **price** of these **second-hand cars decreased** by **\$3940 each year**.

The intercept predicts that, on average, the **price** of these cars when **new** was **\$35100**

The above sentences are scaffolded to ensure you just change the wording in red.

a

b



Calculating and interpreting the coefficient of determination

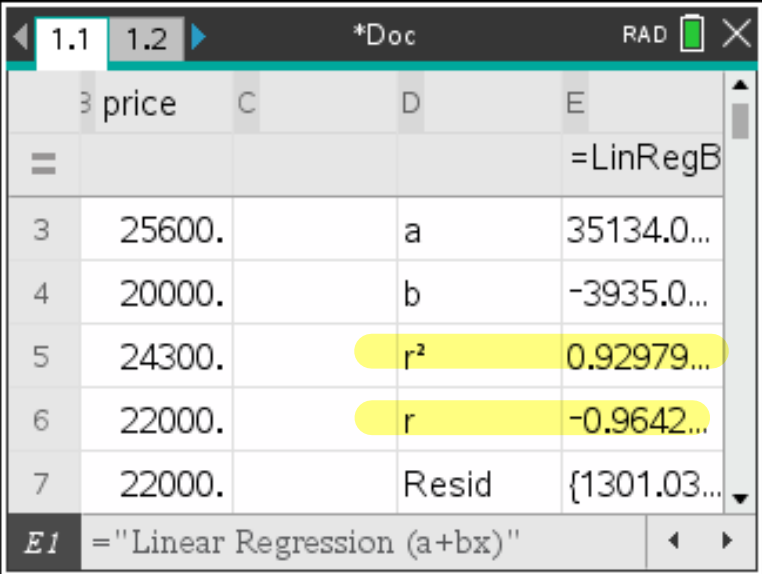
This can, once again, be done using the CAS.

If you already have the value of 'r' then the coefficient of determination is simply this value squared (r^2)

Alternatively, you can read it from the screen if you used the linear regression menu option.

Again, you must use the correct wording when describing the coefficient of determination.

The coefficient of determination indicates that 93% of the variation in the price of these second-hand cars is explained by the variation in their age.



	price	C	D	E
=				=LinRegB
3	25600.		a	35134.0...
4	20000.		b	-3935.0...
5	24300.		r^2	0.92979...
6	22000.		r	-0.9642...
7	22000.		Resid	{1301.03...
E1	="Linear Regression (a+bx)"			

$$r^2 = 0.92979\dots$$

$$r^2 = 92.98\%$$



Making predictions

The whole point of doing this is to allow us to make predictions.

So, when we have the equation of our regression we can use it to predict values.

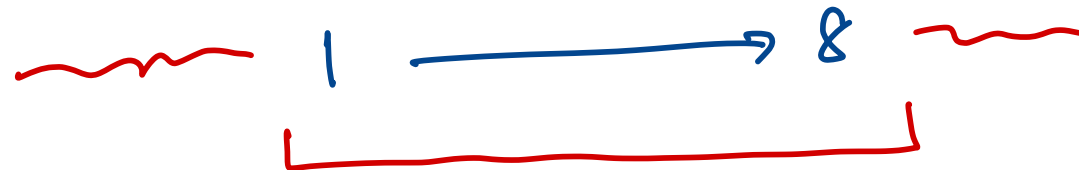
Note: When we are predicting values within the range of the x-values we are said to be interpolating. When we are using x-values outside the range of our data, we are said to be extrapolating.

Extrapolating isn't great as we can't say with certainty that our regression line holds true for data we don't have.

$$price = 35100 - 3940 \times age.$$

$$age = 3.5$$

$$price = 35100 - 3940 \times \underline{\underline{3.5}}$$



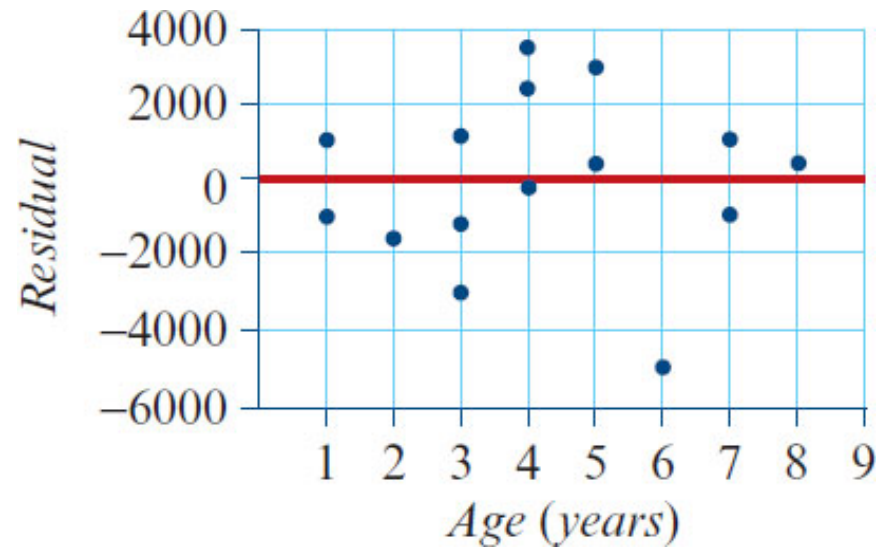
Calculate residuals

A residual is the difference between the **actual data item** and the **point shown on the regression line**.

$$price = 35100 - 3940 \times age.$$

A point above the line will have a positive residual. This means the regression line is **under predicting the value**.

A point below the line will have a negative residual. This means the regression line is over-predicting the value.



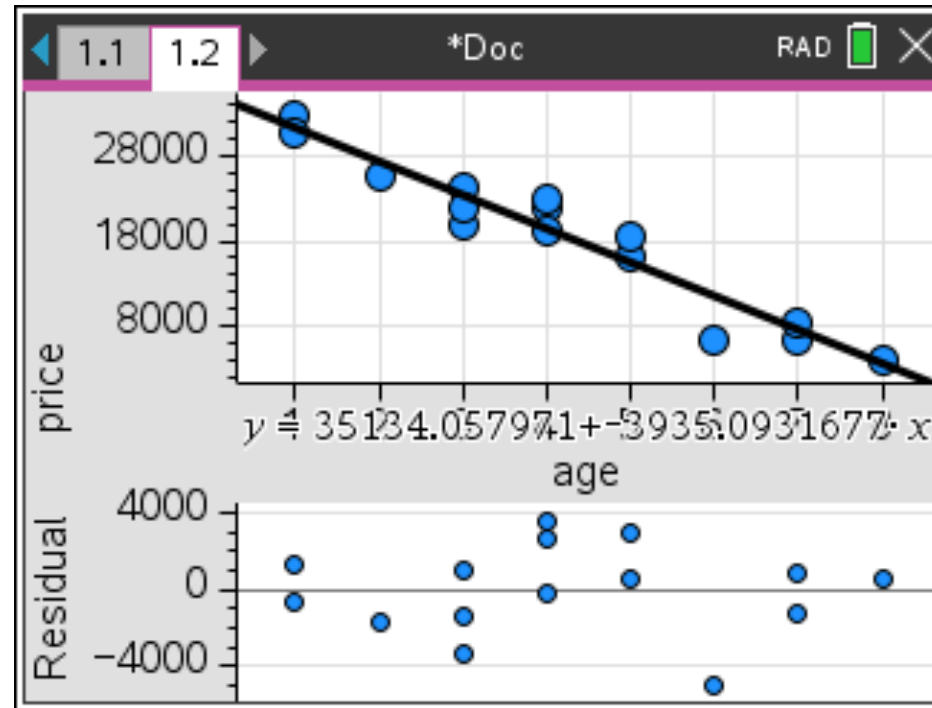
The CAS can show residuals

You can get the CAS to show the residual plot for you.

Where there is a pattern it suggests that the data isn't linear.

When there is no pattern we can suggest that the data is linear. This is called the **assumption of linearity**.

$$price = 35100 - 3940 \times age.$$



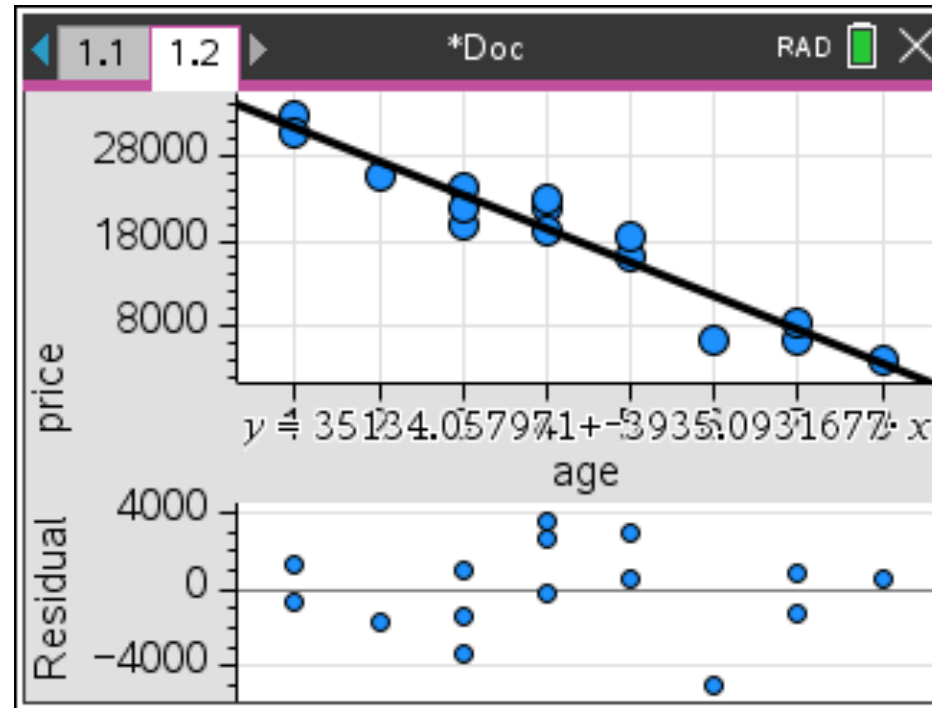
The CAS can show residuals

You can get the CAS to show the residual plot for you.

Where there is a pattern it suggests that the data isn't linear.

When there is no pattern we can suggest that the data is linear. This is called the **assumption of linearity**.

$$price = 35100 - 3940 \times age.$$



We can use the residual plot to see how many data items have been over predicted and underpredicted



Writing a report

The whole point of doing this is to write a report of your findings. Hence, an example report might be:

Construct a report to describe the association between the price and age of secondhand cars.

From the scatterplot we see that there is a strong negative, linear association between the price of a second hand car and its age, $r = -0.964$. There are no obvious outliers.

The equation of the least squares regression line is: $price = 35100 - 3940 \times age$.

The slope of the regression line predicts that, on average, the price of these second-hand cars decreased by \$3940 each year.

The intercept predicts that, on average, the price of these cars when new was \$35100.

The coefficient of determination indicates that 93% of the variation in the price of these second-hand cars is explained by the variation in their age.

The lack of a clear pattern in the residual plot confirms the assumption of a linear association between the price and the age of these second-hand cars.



Making Maths Easy, Engaging Educational, Entertaining



Navigation: [Home](#)

- Latest uploads
- Years 6 to 10
- VCE Courses
- Exam Solutions
- Buy Merchandise

Why choose MaffsGuru?

I hate talking about myself. So, here are some of the amazing comments I receive about the videos and content I produce followed by reasons to use the resource:

“ I wish I watched your videos before naplan
— Overjoyed Cherry (youtube)



VCAA exam questions

VCE lessons, where possible, include the use of past VCAA exam questions to



Professional Development

This resource isn't just meant for students. I hope it will be useful for teachers both new



Downloadable notes

Every lesson has downloadable notes. Whatever I write on the screen, you can download for



Respected Presenter

I currently present for Cambridge University Press and Nelson - as well as produce my own content for