Performing a regression analysis

Tuesday, 26 February 2019 6:00 PM

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

- Know how to perform a regression analysis
- Know how to complete the required number of steps for a good regression analysis
- Know how to write a detailed report

RECAP:

In the last lessons we looked at how we can create a least squares line to some bivariate data.
We learned how to create the equation of the least squares.
Later in the course we are going to use this equation to help us predict values.
But, for now, we need to know how to take some data and perform a regression analysis.

A Recipe for Success

We all love a good recipe when baking. We trust that someone has thought of the very best way to make the cake or meal we are preparing. So, when Maths gives us a recipe, then lets make sure we follow it.

Here is the recipe for a perfect regression analysis:

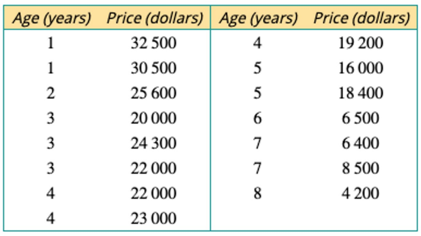
1. Construct a scatterplot to investigate the nature of an association
2. Calculate the correlation coefficient to indicate the strength of the relationship
3. Determine the equation of the regression line
4. Interpret the coefficients the y-intercept (a) and the slope (b) of the least squares line $y=a+bx$
5. Use the coefficient of determination to indicate the predictive power of the association
6. Use the regression line to make predictions
7. Calculate residuals and use a residual plot to test the assumption of linearity
8. Write a report on your findings.

The good news is … much of the above you already know how to do.
I will fill in the gaps for the things you don't know as we proceed through the lesson.

Some data

Always a good idea to have some data to use to analyse.

As this course is linked to the Cambridge Further Mathematics Units 3 and 4 course, I'm going to use the data they provide. It relates age and price (dollars) of a second hand car.



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

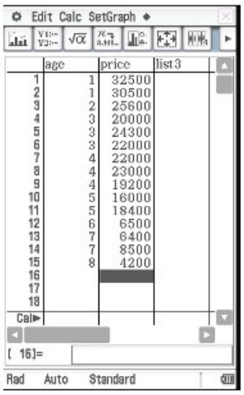
Remember: It's important to get the **explanatory and response** variables the correct way.

Explanatory: Age
Response: Price (dollars)

STEP 1: Construct a scatterplot to investigate the nature of an association

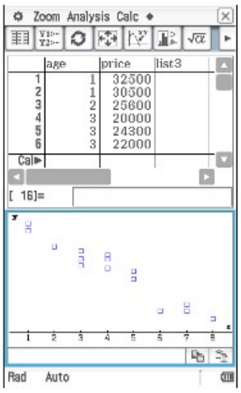
We can use the CAS for this …

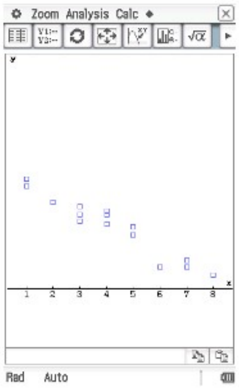
This is what we end up with:



	age	price	list3
1	1	32500	
2	1	30500	
3	2	25600	
4	3	20000	
5	3	24300	
6	3	22000	
7	4	22000	
8	4	23000	
9	4	19200	
10	5	16000	
11	5	18400	
12	6	6500	
13	7	6400	
14	7	8500	
15	8	4200	
16			
17			
18			

[16]=

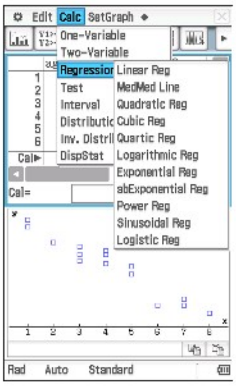


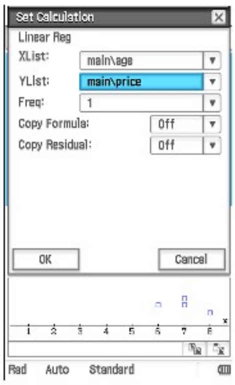


From the scatter plot we can see there is **a negative linear association** but is it weak, moderate or strong?

STEP 2: Calculate the correlation coefficient to indicate the strength of the relationship

Ask the calculator to calculate the value of the Pearson's correlation coefficient!




Set Calculation

Linear Reg

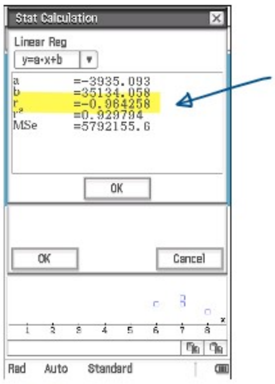
XList: main\age

YList: main\price

Freq: 1

Copy Formula: Off

Copy Residual: Off


Stat Calculation

Linear Reg

y=a·x+b

```
a    =-3935.093
b    =35134.058
r    =-0.964258
r²   =0.929794
MSe  =5792155.6
```

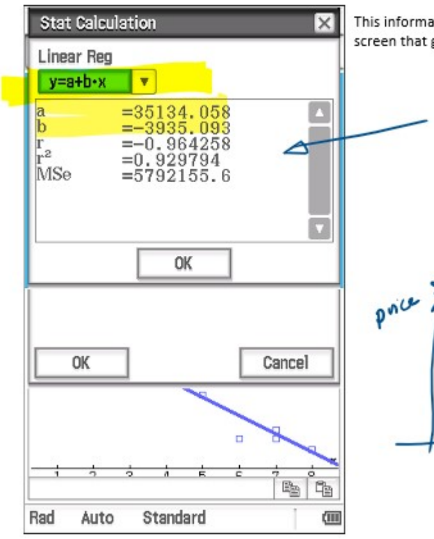
Hence we can now describe the association as **a strong negative linear association** between the price of the car and its age with no clear outliers.

We now have the first part of the report!

There is a strong negative linear association between the price of these second-hand cars and their age (r=−0.964).

STEP 3: Determine the equation of the regression line

Now we use the calculator to find the values for the equation of the least squares line.


Stat Calculation

Linear Reg

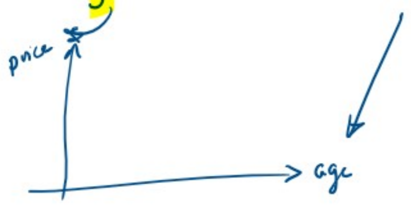
y=a+b·x

```
a    =35134.058
b    =-3935.093
r    =-0.964258
r²   =0.929794
MSe  =5792155.6
```

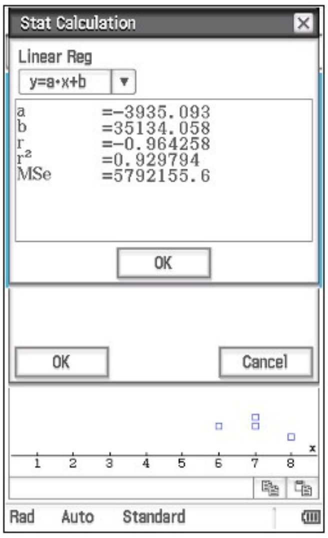
This information has already been given to us with the screen that gave me the value of r.

$$y = a + bx$$

$$y = 35134 - 3935 \times x$$



Notice the difference it makes when we have the formula the "wrong way around"



```
Stat Calculation
Linear Reg
y=a·x+b
a       =-3935.093
b       =35134.058
r       =-0.964258
r²      =0.929794
MSe     =5792155.6
```

Hence, the equation of the least squares regression line using the formula $y = a + bx$:

$$price = 35134 - 3935 \times age$$

I note, with interest, that the Cambridge book has this as:

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

Response variable

Explanatory variable

STEP 4: Interpret the coefficients the y-intercept (a) and the slope (b) of the least squares line y=a+bx

This is one of the most importants steps in writing the report.
This shows you **understand** the work and are not just **regurgitating** it.

To be able to fully understand you need to understand what **gradient** really is.
We know that it's rise over run but it's more important than that.
It is really telling us by how much the **response variable** changes for one unit increase in the **explanatory variable**.

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

1
→ inc age − 3940 price

inc
− dec

Gradient relates how one unit **increase** in the explanatory variable changes the **response** variable.

In the above example we know the gradient is -3940.

This means for every increase of one year in age, the price of the car will reduce by $3940.

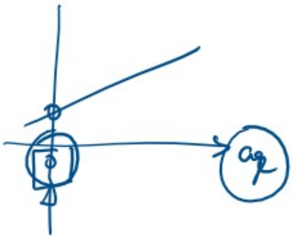
Now we need to know how to interpret the intercept.

The intercept is used to describe the size of the **response variable** for a starting value of the **explanatory variable**.

Hence, in the example:

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

When the age of the car was 0 years, the price was $35, 100



STEP 6: Use the regression line to make predictions

The whole point of doing this is to **predict** values of the response or explanatory variable outside of the data items we have been given.

We can now do this using the formula for the least squares regression line.

Using the least squares regression line from our example, what would the price be of a car which is 5.5 years old?

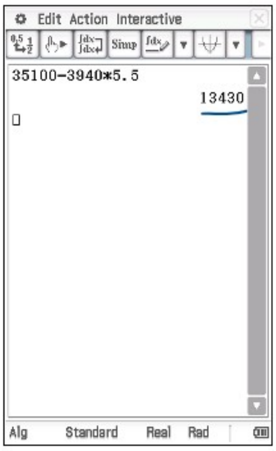
data items we have been given.

We can now do this using the formula for the least squares regression line.

Using the least squares regression line from our example, what would the price be of a car which is 5.5 years old?

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

They have given us the age in the question.
Use the CAS to find the answer.



```
35100-3940*5.5
                13430
```

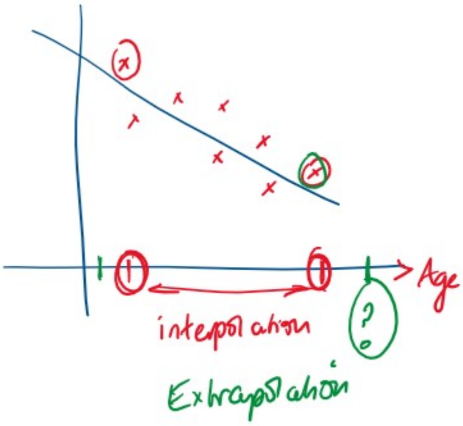
Hence, the price of the car at 5.5 years old would be $13, 430

Warning: Will Robinson

The Further Maths exams love asking questions about whether the data you predict is good or not!
What they are really asking is whether you understand what it means by **interpolation** and **extrapolation**.

When we predict using data inside of the range of values we were given, then we are **interpolating** it.
This is generally seen as more accurate than is we predict from data outside of the range we were given.
When we do this, we are **extrapolating** the data.

So, using ages between 1 and 8 years old, we are interpolating the data.
If we are asked to use ages older than 8 years old we are extrapolating the data.



STEP 5: Use the coefficient of determination to indicate the predictive power of the association

We are nearing the end of collecting the data for the analysis.
If you remember, we have also learned about the coefficient of determination which is the percentage value of how accurate it is to predict the value of the response variable from the explanatory variable.

We know that:

r

$$Coefficient\ of\ determination = (correlation\ coefficient)^2$$

Using our example, we know that:

0.930
× 100

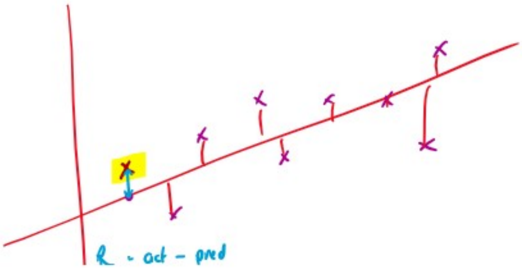
$$Coefficient\ of\ determination = r^2 = 0.964^2 \approx 0.930 \text{ or } 93\%$$

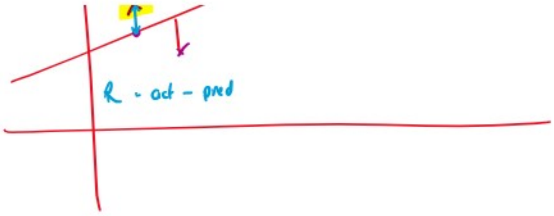
Interpreting this value we know that we can now write the following in our report.

The value of the coefficient of determination suggests that 93% of the variation observed in the price in dollars can be explained by the variation in the age of the car.

STEP 7: Calculate residuals and use a residual plot to test the assumption of linearity

We looked, in the last video, at how the least squares line comes from the residual values.
These values were the difference between the actual data point and the value the least squares line would "predict".





To be able to find the residual value of any point, we use a simple formula.

$$Residual\ value = actual\ data\ value - predicted\ value$$

Continuing to use the example, if I wants to find the residual value of a car which is 6 years old:

We know that the actual data item from the table shows the car to be $6, 500

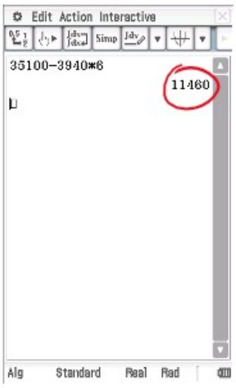
Age (years)	Price (dollars)	Age (years)	Price (dollars)
1	32 500	4	19 200
1	30 500	5	16 000
2	25 600	5	18 400
3	20 000	6	6 500
3	24 300	7	6 400
3	22 000	7	8 500
4	22 000	8	4 200
4	23 000		

6 6,500

We can use the least squares regression line to find the predicted value:

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

6



```
35100−3940*6
                11460
```

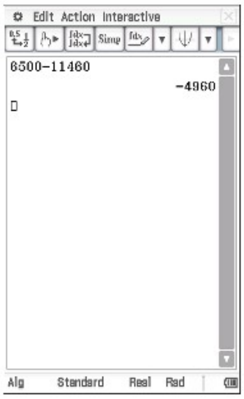
$ 11 460

The predicted value is $11, 460

Hence the residual is -$4960
Note: These values can be negative!

$$Residual\ value = actual\ data\ value - predicted\ value$$

$$Residual\ value = 6500 - 11460 = -4960$$



```
6500−11460
                -4960
```

This means the value of the car is actually $5000 **less than we would have predicted!**

Testing the assumption of linearity

We always need to test to see if the data is linear

This means the value of the car is actually $5000 **less than we would have predicted!**

Testing the assumption of linearity

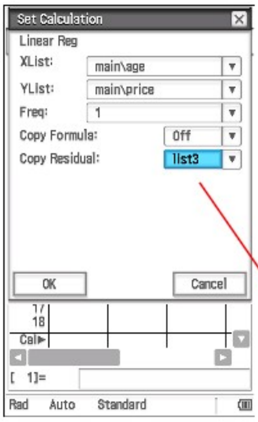
We always need to test to see if the data is linear.
Remember, one of the key assumptions made comes from the following list:

- The data is numerical
- The association is linear
- There are no clear outliers

We can use the CAS to do a simple test for the assumption of linearity using a residual plot.

The steps are:

1. Add a residuals column to your data.
2. Graphs the residuals
3. Look for any pattern (or lack of!)



Set Calculation

Linear Reg

XList: main\age

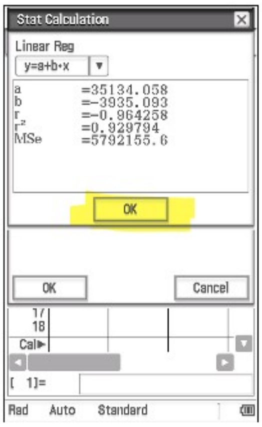
YList: main\price

Freq: 1

Copy Formula: Off

Copy Residual: list3

Make sure this is empty and has a title



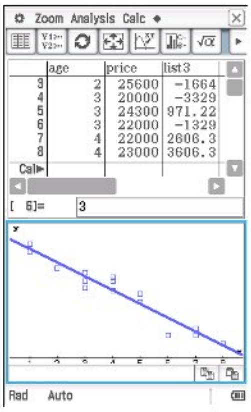
Stat Calculation

Linear Reg

y=a+b·x

```
a      =35134.058
b      =-3935.093
r      =-0.964258
r²     =0.929794
MSe    =5792155.6
```

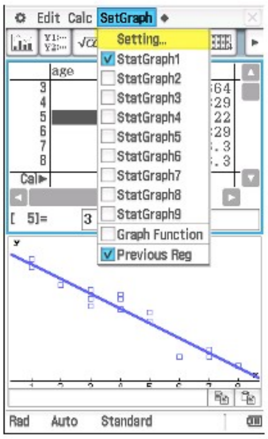
This will then plot the data for you of age against price.

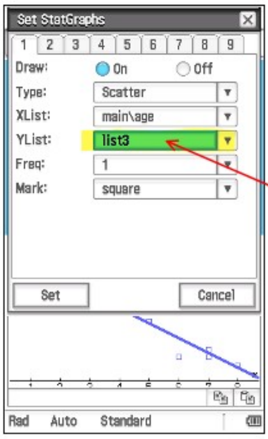


	age	price	list3
3	2	25600	−1664
4	3	20000	−3329
5	3	24300	971.22
6	3	22000	−1329
7	4	22000	2606.3
8	4	23000	3606.3

[6]= 3

We need the CAS to change this to Age vs Residual (list3) so we need to SetGraph and change the Setting ...





Set StatGraphs

Draw: ● On ○ Off

Type: Scatter

XList: main\age

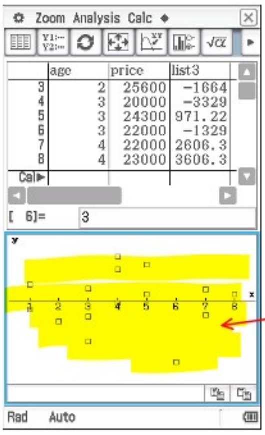
YList: list3

Freq: 1

Mark: square

Change this to list3
Then click Set.

Finally click the graph button



	age	price	list3
3	2	25600	−1664
4	3	20000	−3329
5	3	24300	971.22
6	3	22000	−1329
7	4	22000	2606.3
8	4	23000	3606.3

[6]= 3

Here is your residual plot!

Linearity is assumed when the residual plot has no clear pattern (essentially they are randomly scattered around the zero line).

As the above has no clear pattern, we can **assume linearity**.

STEP 8: Write a report on your findings.

Writing a clear and detailed report covering all the steps above with appropriate statistics is very important!

Below is the suggested report from Cambridge:

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× 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.

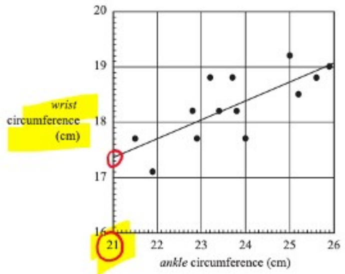
The lack of a clear pattern in the residual plot confirms the assumption of a linear association between the price and the age of these second-hand cars.

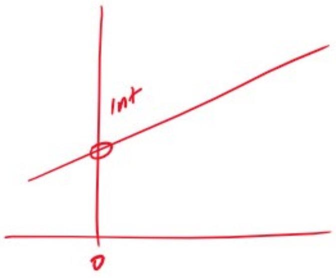
Repeat the above for any question, and I don't see how you can't score full marks!
Remember: The values placed in the report are what makes this a good report.

VCAA Exam Question on this concept
2017 Paper 1

Use the following information to answer Questions 8–10.

The scatterplot below shows the *wrist* circumference and *ankle* circumference, both in centimetres, of 13 people. A least squares line has been fitted to the scatterplot with *ankle* circumference as the explanatory variable.





Question 8

The equation of the least squares line is closest to

~~**A.** *ankle* = 10.2 + 0.342 × *wrist*~~
[✓ **B.**] *wrist* = 10.2 + 0.342 × *ankle*
~~**C.** *ankle* = 17.4 + 0.342 × *wrist*~~
~~**D.** *wrist* = 17.4 + 0.342 × *ankle*~~
~~**E.** *wrist* = 17.4 + 0.731 × *ankle*~~

wrist = 10·2 + 0·342 x ankle
= 10·2 + 0·342 x 24
= 18·408

Question 9

When the least squares line on the scatterplot is used to predict the wrist circumference of the person with an ankle circumference of 24 cm, the residual will be closest to

[✓ **A.**] −0.7
B. −0.4
C. −0.1
D. 0.4
E. 0.7

Res = Act − Pre
= 17·7 − 18·408

VCAA Exam Question on this concept

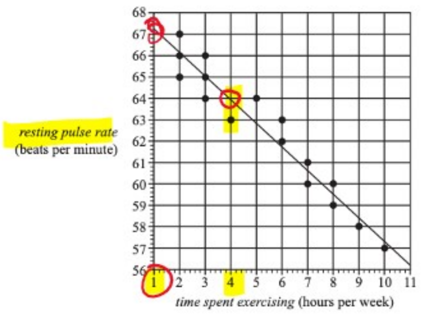
E. 0.7

Res = Act − Pr

= 17.7 − 18.408

VCAA Exam Question on this concept
2018 Paper 1

The scatterplot below displays the *resting pulse rate*, in beats per minute, and the *time spent exercising*, in hours per week, of 16 students. A least squares line has been fitted to the data.



Res = Act − pred
$= 63 - 64$
$= -1$

$(1, 67.2)$ $(4, 64)$

$$m = \frac{64 - 67.2}{4 - 1}$$

=

Question 7

Using this least squares line to model the association between *resting pulse rate* and *time spent exercising*, the residual for the student who spent four hours per week exercising is closest to

A. −2.0 beats per minute.

[✓ B.] −1.0 beats per minute.

C. −0.3 beats per minute.

D. 1.0 beats per minute.

E. 2.0 beats per minute.

Question 8

The equation of this least squares line is closest to

~~A. *resting pulse rate* = 67.2 − 0.91 × *time spent exercising*~~

~~B. *resting pulse rate* = 67.2 − 1.10 × *time spent exercising*~~

C. *resting pulse rate* = 68.3 − 0.91 × *time spent exercising*

[✓ D.] *resting pulse rate* = 68.3 − 1.10 × *time spent exercising*

~~E. *resting pulse rate* = 67.2 + 1.10 × *time spent exercising*~~