Conducting a regression analysis using raw data

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 conduct a regression analysis using raw data
- Use the skills from the previous lesson to complete the regression analysis

RECAP:

In the last lesson we looked at all the steps which are needed to perform a regression analysis to enable us to write a report.

The steps given were:

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.

We are going to use these (one last time) to conduct a regression analysis on some data.

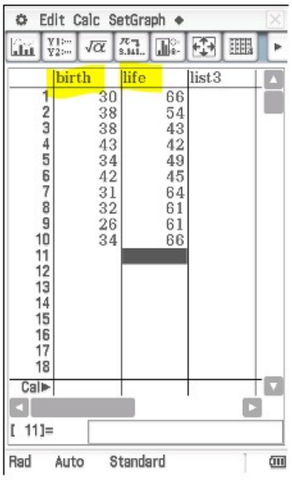
The question

The following data is provided from the Cambridge Further Mathematics Units 3 and 4 textbook and is used with permission

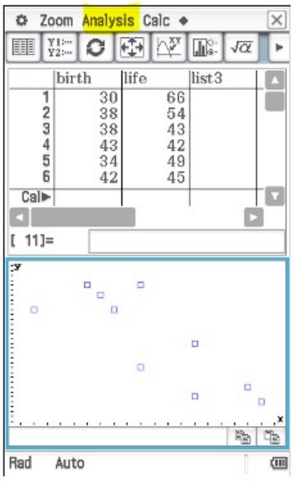
Using the following data, conduct a regression analysis for birth rate (per thousand) compared with Life Expectancy (years).

Birth rate (per thousand)	30	38	38	43	34	42	31	32	26	34
Life expectancy (years)	66	54	43	42	49	45	64	61	61	66

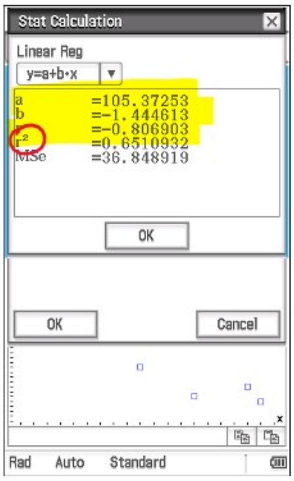
Expected: Birth rate
Response: Life expectancy



```
   birth  life  list3
1     30    66
2     38    54
3     38    43
4     43    42
5     34    49
6     42    45
7     31    64
8     32    61
9     26    61
10    34    66
11
12
13
14
15
16
17
18
Cal▸
[ 11]=
```



```
   birth  life  list3
1     30    66
2     38    54
3     38    43
4     43    42
5     34    49
6     42    45
Cal▸
[ 11]=
```



```
Linear Reg
y=a+b·x
a       =105.37253
b       =-1.444613
r       =-0.806903
r²      =0.6510932
MSe     =36.848919
```

=-0.807

The following section of the report can be written from the above three screens

There is a strong negative linear association between life expectancy and birth, $r = -0.8069$. There are no obvious outliers.

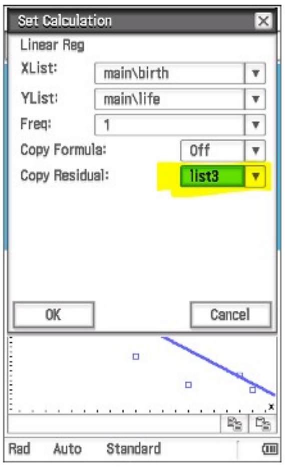
The equation of the least squares regression line is: $life\ expectancy = 105 - 1.4446 \times birth\ rate$

1

The slope of the regression line predicts that, on average, the life expectancy (years) will decreased by 1.4446 per one thousand births.

The intercept predicts that, on average, the life expectancy at birth was 105 years..

The coefficient of determination indicates that 65.1% of the variation in the life expectancy (years) is explained by the variation in their birth rate (per thousand).



Set Calculation

Linear Reg

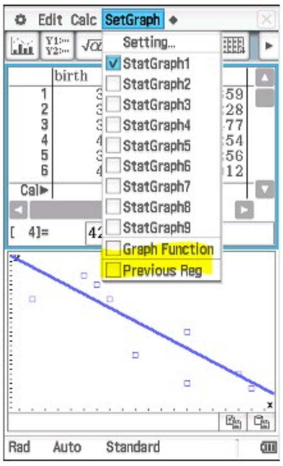
XList: main\birth

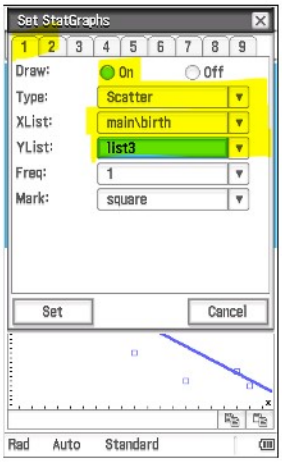
YList: main\life

Freq: 1

Copy Formula: Off

Copy Residual: list3





Set StatGraphs

1 2 3 4 5 6 7 8 9

Draw: ● On ○ Off

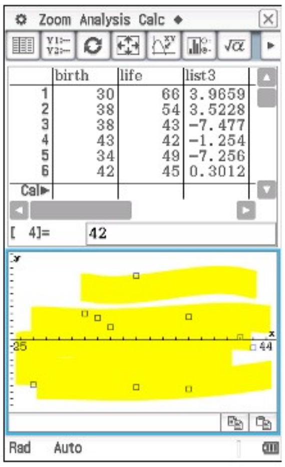
Type: Scatter

XList: main\birth

YList: list3

Freq: 1

Mark: square



	birth	life	list3
1	30	66	3.9659
2	38	54	3.5228
3	38	43	-7.477
4	43	42	-1.254
5	34	49	-7.256
6	42	45	0.3012

[4]= 42

Using the above screen I can now complete the report:

The lack of a clear pattern in the residual plot confirms the assumption of a linear association between the life expectancy (years) and the birth rate (per thousand).

VCAA Exam Question on this concept
2018 Paper 2

Question 2 (7 marks)

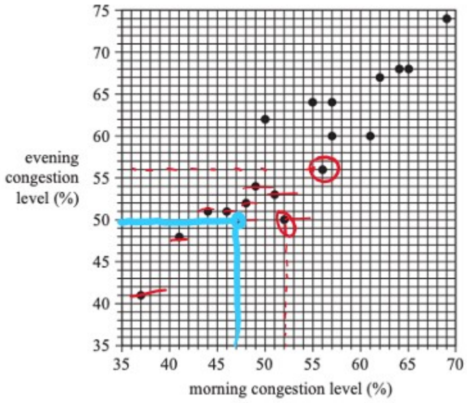
The congestion level in a city can also be recorded as the percentage increase in travel time due to traffic congestion in peak periods (compared to non-peak periods).

This is called the percentage congestion level.

The percentage congestion levels for the morning and evening peak periods for 19 large cities are plotted on the scatterplot below.

$\frac{19+1}{2} = \frac{20}{2}$

$= \boxed{10}$



a. Determine the median percentage congestion level for the morning peak period and the evening peak period.

Write your answers in the appropriate boxes provided below. 2 marks

Median percentage congestion level for morning peak period 52 %

Median percentage congestion level for evening peak period 56 %

A least squares line is to be fitted to the data with the aim of predicting evening congestion level from morning congestion level.

The equation of this line is

evening congestion level = 8.48 + 0.922 × *morning congestion level*

b. Name the response variable in this equation. 1 mark

Evening congestion level

c. Use the equation of the least squares line to predict the evening congestion level when the morning congestion level is 60%. 1 mark

$E = 8.48 + 0.922 \times 60$

$= 63.8$

d. Determine the residual value when the equation of the least squares line is used to predict the evening congestion level when the morning congestion level is 47%.

Round your answer to one decimal place. 2 marks

Pre : $E = 8.48 + 0.922 \times 47$

$= 51.814$

Res = Act - Pred = $50 - 51.814 = -1.8$

e. The value of the correlation coefficient *r* is 0.92

What percentage of the variation in the evening congestion level can be explained by the variation in the morning congestion level?

Round your answer to the nearest whole number. 1 mark

85%

$r = 0.92$

$r^2 = 0.8464$

$= 84.64\%$

85%