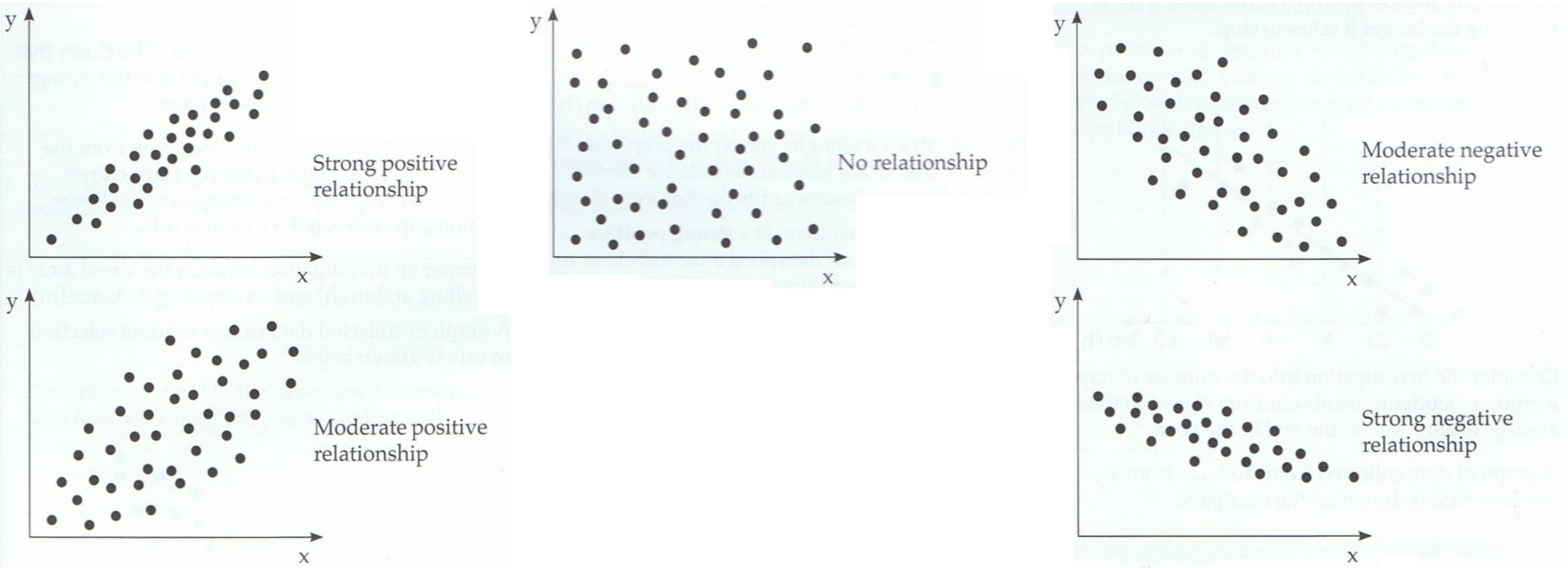
Analysing the Data

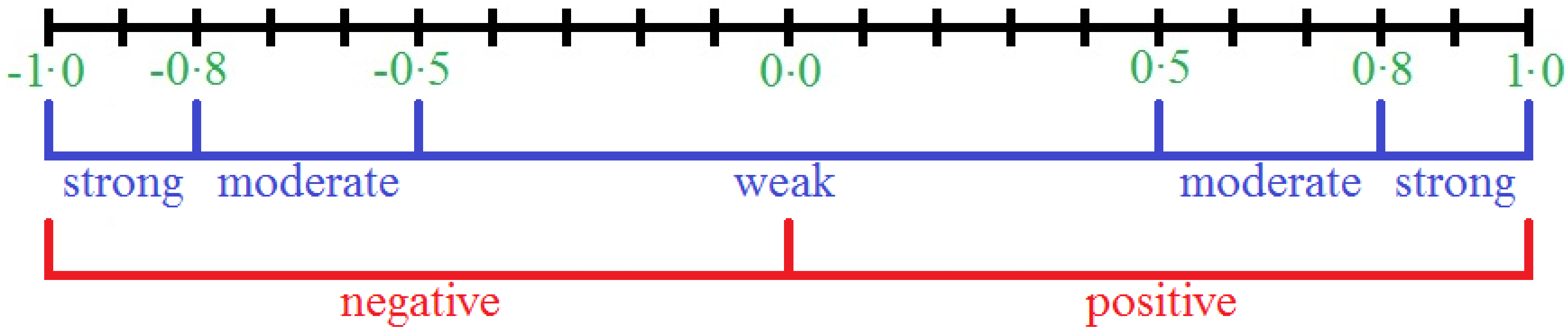
Start by commenting of the general trend.

- Is there or is there not a relationship between the two variables?
- If there is a relationship, is it linear or non-linear?
- If there is a linear relationship, is it positive or negative?
- If there is a linear relationship, is it strong or weak?



Analysing the Data

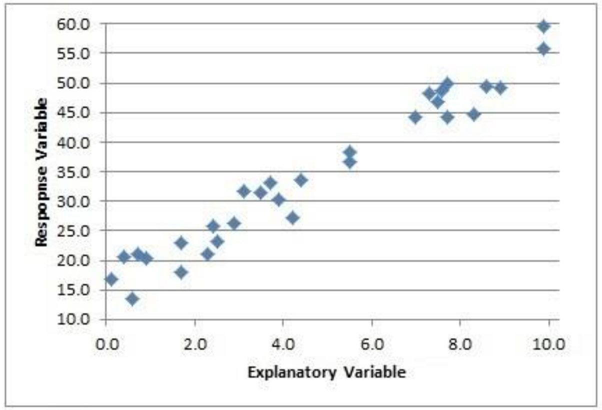
Finding the correlation coefficient (r) from the coefficient of determination (R^2) can help describe the strength of the relationship. They only measure the strength of linear relationships.



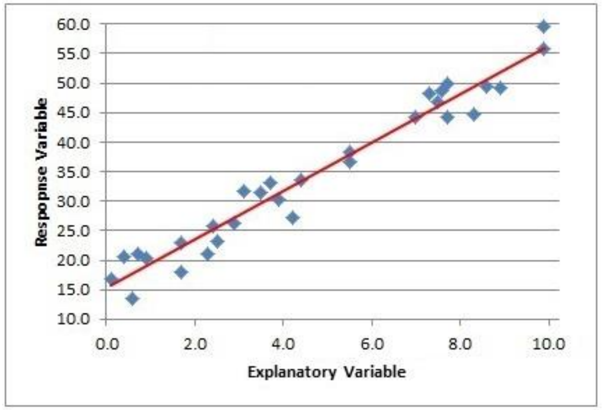
As well as describing the nature of the relationship you need to discuss any features you can see in your scatter plot.

- Are there changes in the variation?
- Are there any unusual points?
- Are there any groups in the data?

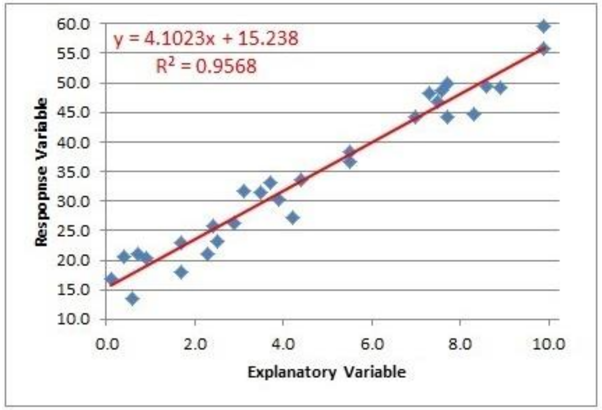
Analysing the Data



In a positive linear relationship as one variable increases, the other tends to increase.



If the data seems to be packed close to the line then the relationship is a strong one.

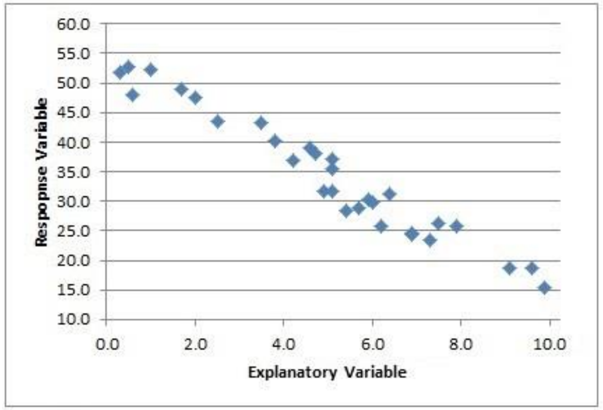


$R^2 = 0.9568 \Rightarrow r = 0.9782$

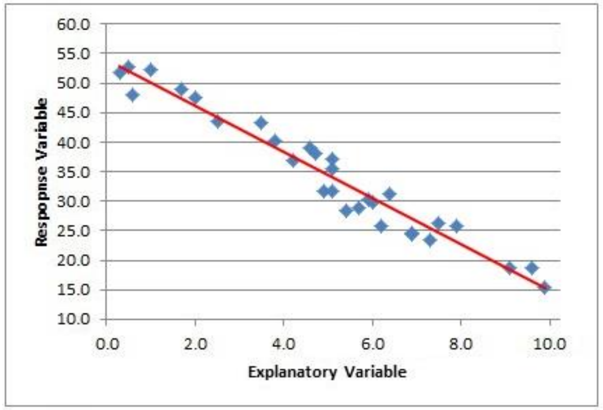
There is a very strong positive relationship between the explanatory variable and the response variable.

The model suggests that for each increase of 1 unit in the explanatory variable there is an increase of 4.1023 units in the response variable.

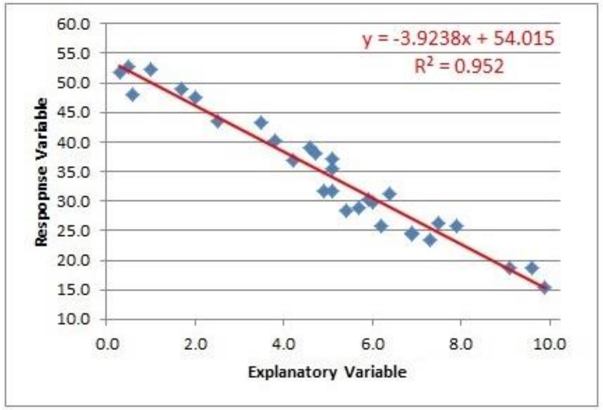
Analysing the Data



In a negative linear relationship as one variable increases, the other tends to decrease.



If the data seems to be packed close to the line then the relationship is a strong one.

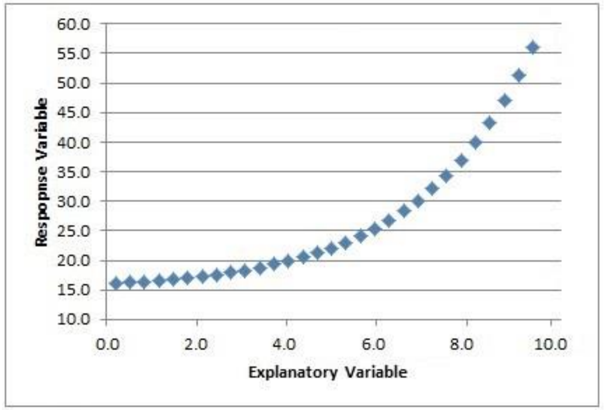


$R^2 = 0.952 \Rightarrow r = -0.9757$

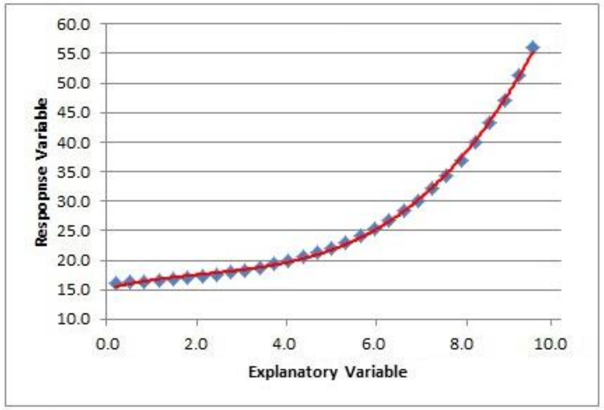
There is a very strong negative relationship between the explanatory variable and the response variable.

The model suggests that for each increase of 1 unit in the explanatory variable there is a decrease of 3.9238 units in the response variable.

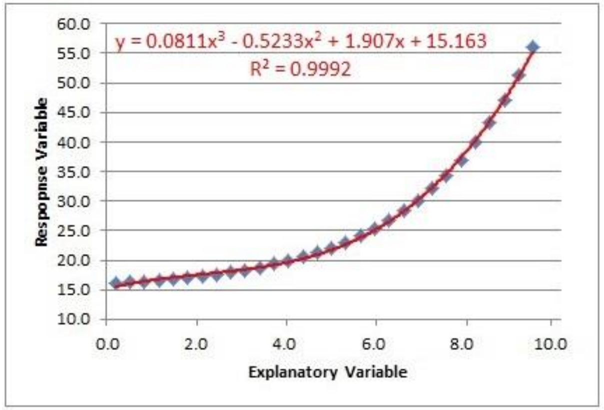
Analysing the Data



If a line does not seem to fit the data as well as a curve would then the data is non-linear.



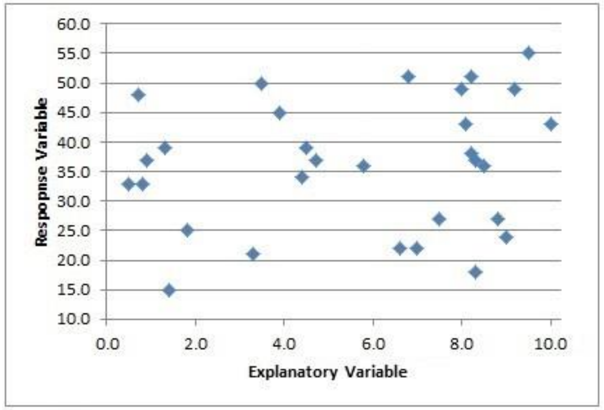
This data seems to be packed close to the curve but if we had data outside this domain it may not be close to this particular curve.



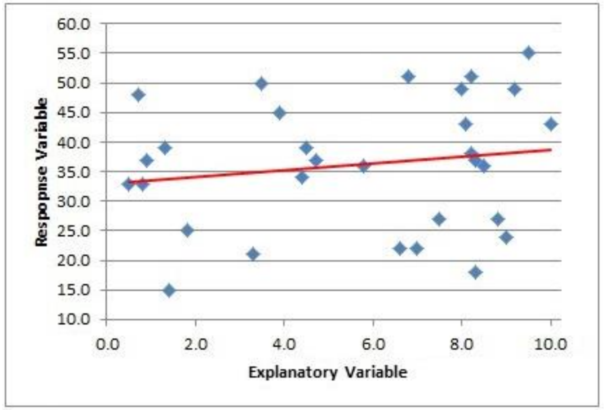
$R^2 = 0.9992 \Rightarrow r = 0.9996$

There would suggest that the data is closely related the cubic curve that has been fitted. However the coefficient of determination (R^2) and the correlation coefficient (r) only measure the strength of linear relationships. In this case the computer has fitted a cubic model but the data is in fact exponential.

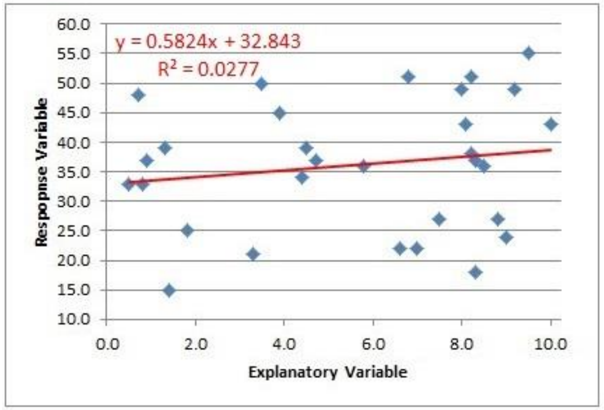
Analysing the Data



If there is no discernible trend line and the data does not seem to fit a curve then there is no relationship or it is a weak relationship.



The computer will place a trend line if you ask it to. However if the data does not seem to be packed at all close to the line then there is no relationship or it is a weak one.

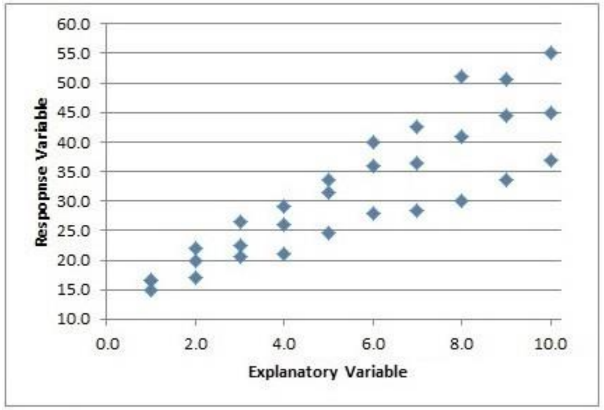


$R^2 = 0.0277 \Rightarrow r = 0.1664$

There is a very weak positive relationship between the explanatory variable and the response variable.

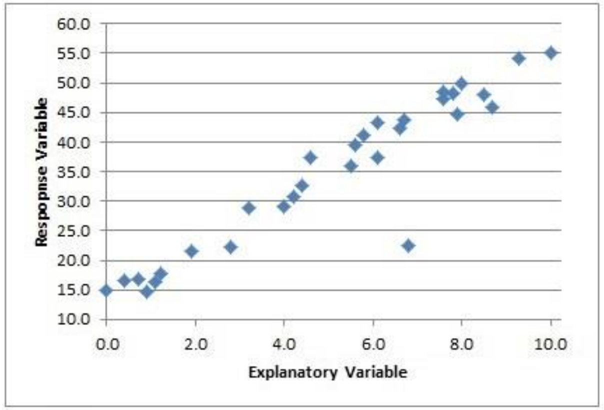
The model suggests that for each increase of 1 unit in the explanatory variable there is an increase of 0.5824 units in the response variable.

Analysing the Data



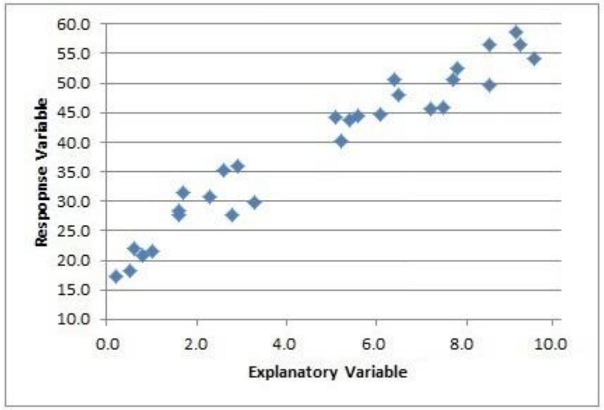
Changes in the variation: Does the variation in the response variable change as the predictor variable changes? If so, then there is variation in the scatter.

It appears that as the explanatory variable increases, the variation in the response variable tends to be greater.



Unusual points: These are points that appear to be very different from the general pattern, they are likely to lie outside the general mass of data.

The point (6.8, 22.5) appears to be unusual.

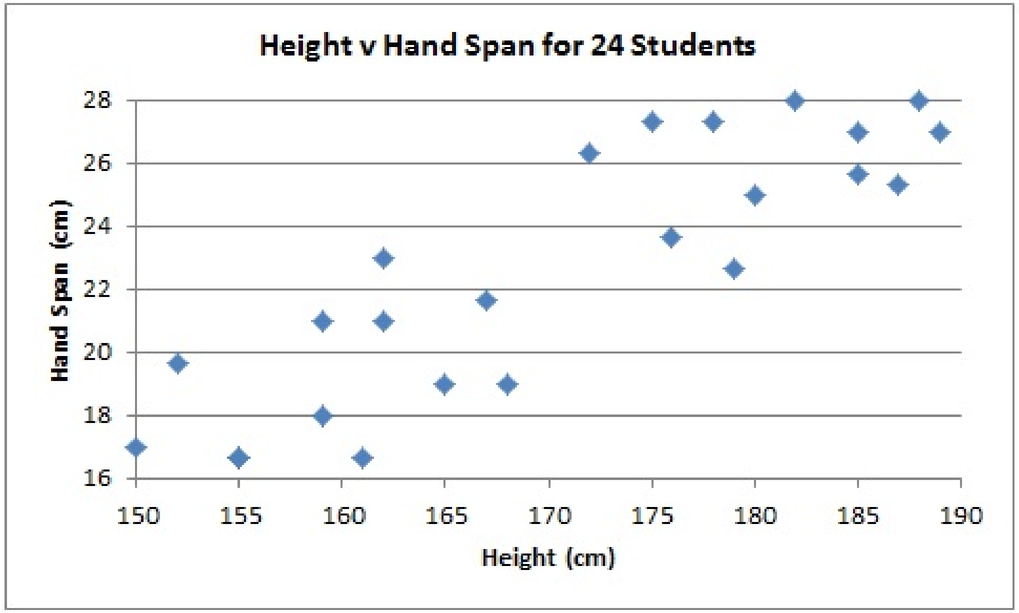


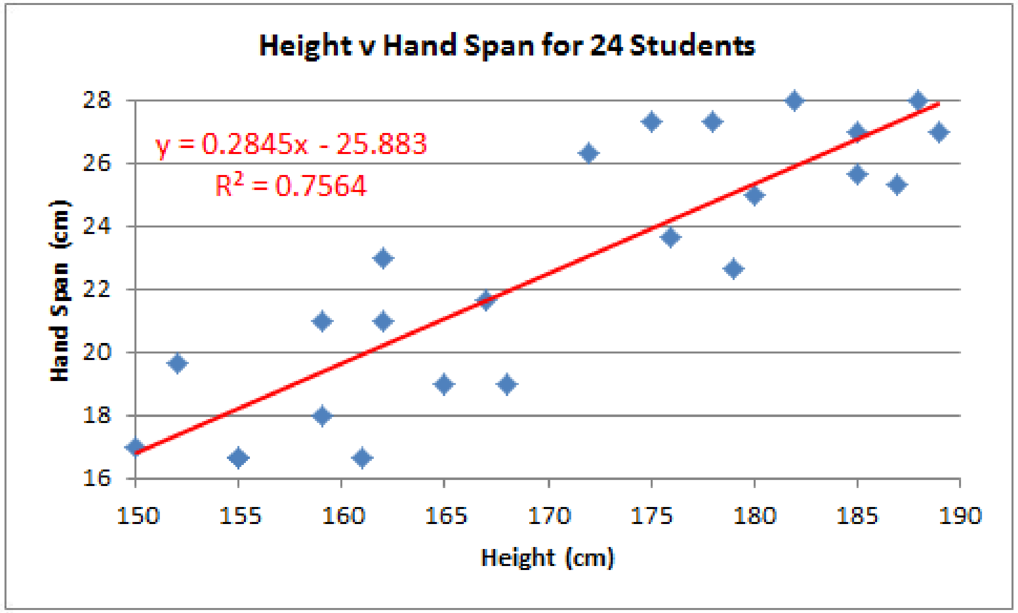
Groups: This would be seen if there are different, separated bulges in the mass of data.

There appears to be two groups, one for where the explanatory variable is less than 4 and where the explanatory variable is greater than 4.

Analysing the Data

Example 1: I wonder if there is a relationship between height and hand span for high school students.





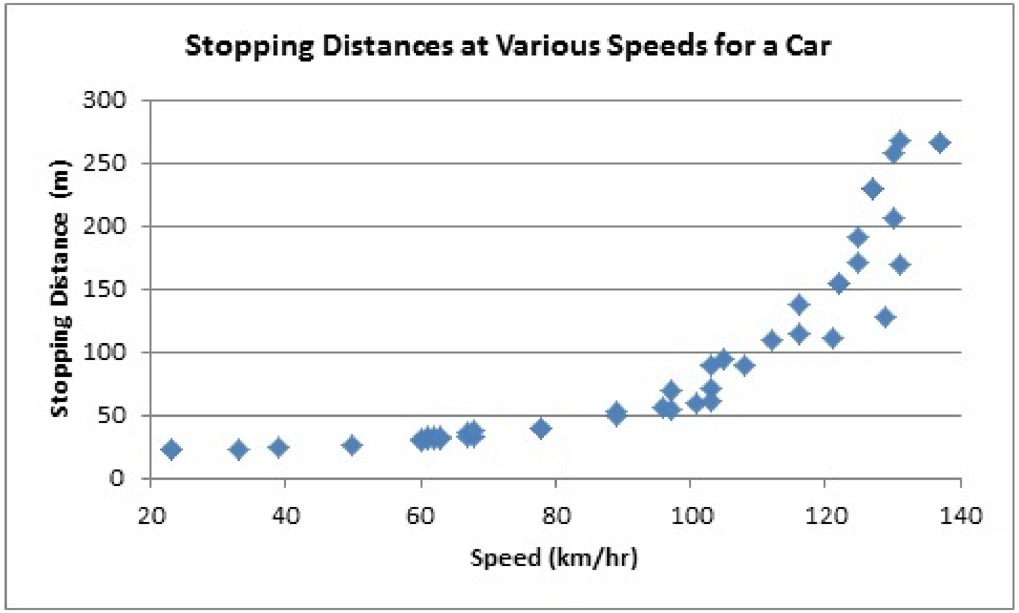
There is a positive linear relationship between a student's height and span. Hand spans of students tend to be going up as height increases. There appear to be two groups, one with heights less then 170cm and the other with heights greater than 170cm. It could be that the former is mainly girls and the latter mainly boys or the former mainly juniors and the latter mainly seniors.

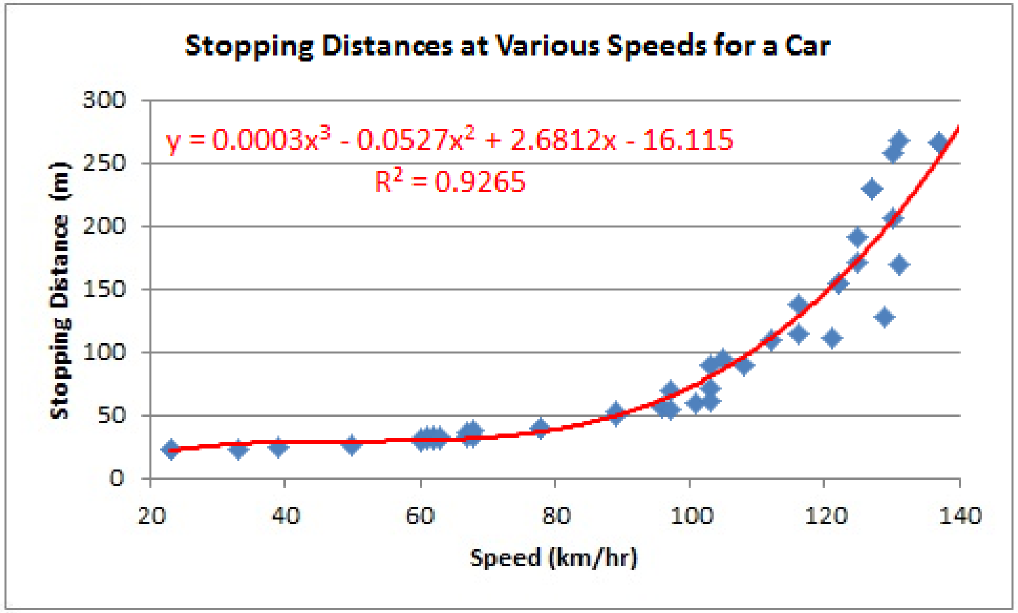
The trend line shows that as height increases, hand span also tends to increase. The points are reasonably close to the trend line but there is some scatter, suggesting that the relationship is of moderate strength.

$R^2 = 0.7564 \Rightarrow r = 0.8697$. This suggests that there is a moderate to strong positive relationship between height and hand span for these 24 students. The model suggests that for each increase of 1cm in height there is an increase of 0.2845cm in hand span.

Analysing the Data

Example 2: I wonder if there is a relationship between driving speed and stopping distances for a particular car.





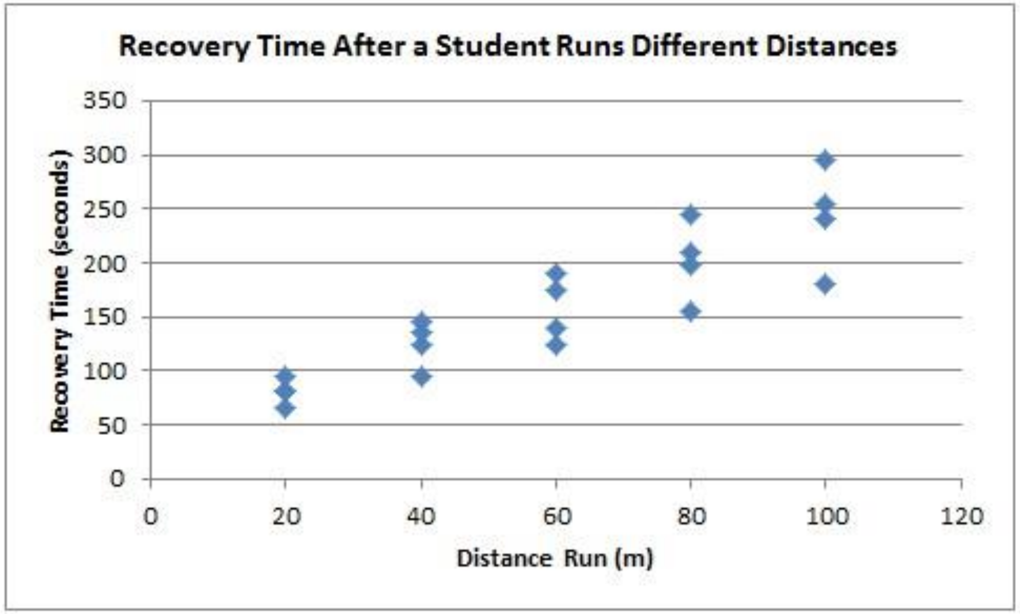
There is a non-linear relationship between the speed a car is going and its stopping distance. As speed increases stopping distances also tend to increase, but the amount by which they tend to increase goes up as speeds increases.

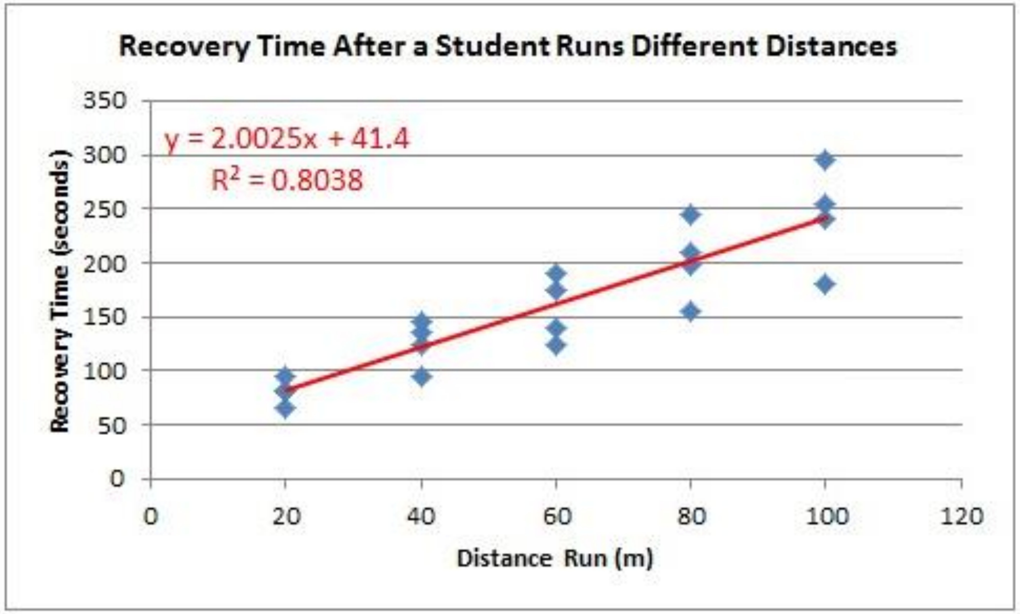
The points are close to the curve, particularly for lower speeds. This suggests there is strong relationship between speed and stopping distance. For higher speeds the stopping distance is more varied. This could be because of differences in the reaction time of the driver at higher speeds or that the effect of braking is more variable at higher speeds.

$R^2 = 0.9186 \Rightarrow r = 0.9584$. No comment can be made on the strength of the relationship from this as R^2 and r measure the strength of linear relationships.

1. What would you write in your "Analysing the Data" section? Discuss the features, including the strength of the relationship.

I wonder if there is a relationship between the distance a student runs and the amount of time it takes him/her to recover after the run.





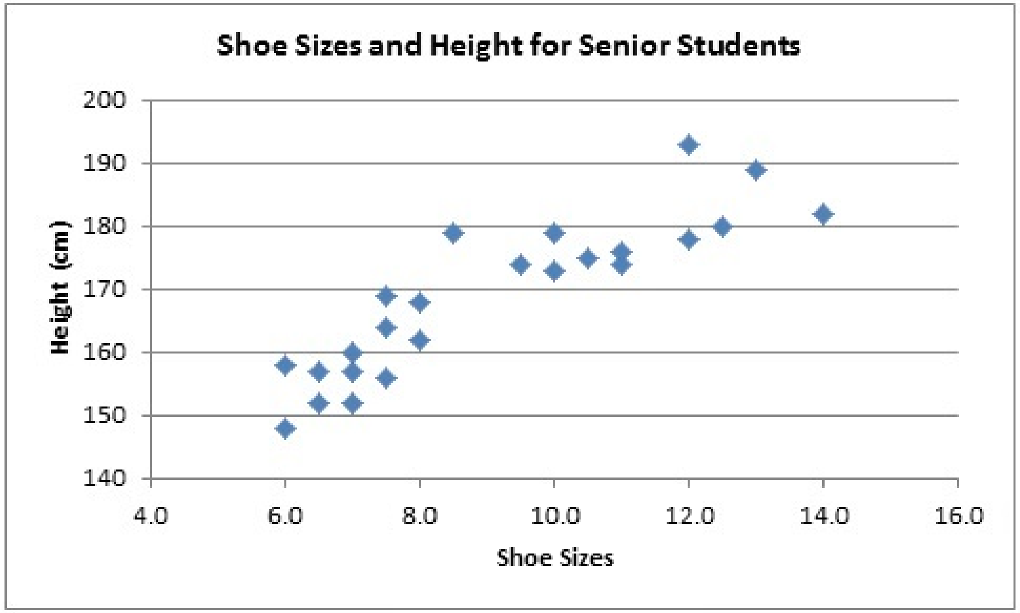
There is a positive linear relationship between the distance the student runs and his recovery time. Recovery times tend to be higher for longer distances. There is more variation in recovery times for longer distances. This could be because the student varies his running speed more over longer runs.

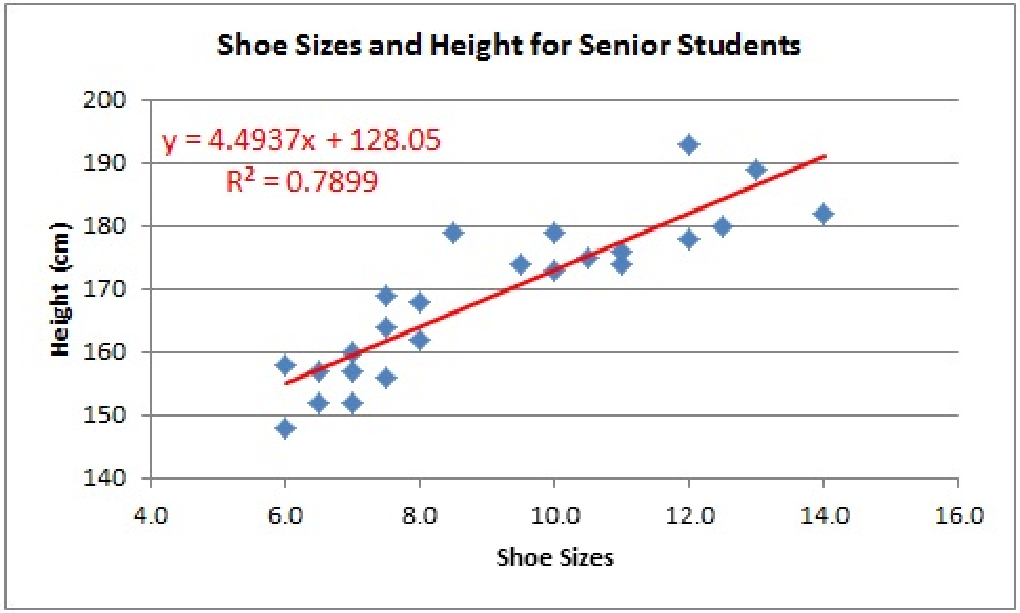
The trend line shows that as distance run increases, recovery time also tends to increase. Most of the dots seem to be reasonably close to the line which suggests there is a strong relationship between the distance the student runs and his recovery time.

$R^2 = 0.8038 \Rightarrow r = 0.8965$. This suggests that there is a strong positive relationship between distance run and recovery time for this student. The model suggests that for each increase of 1m run there is an increase of 2.0025 seconds required for recovery time.

2. What would you write in your “Analysing the Data” section? Discuss the features, including the strength of the relationship.

I wonder if there is a relationship between the shoe sizes and the heights of senior students.





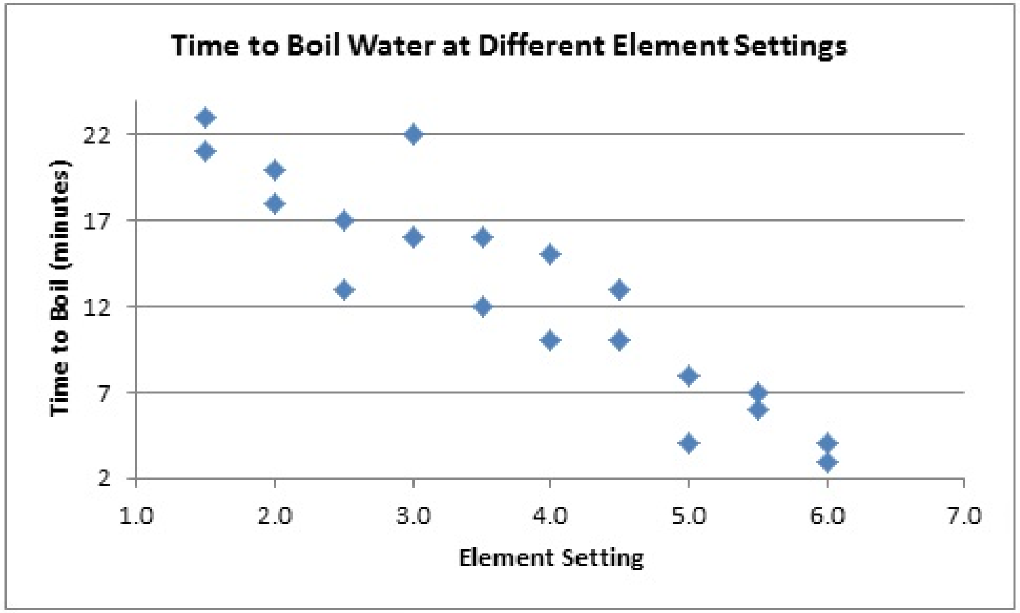
There is a positive linear relationship between the shoe sizes and heights of senior students. By eye I wold says it is possible that the relationship is nonlinear, although all of the dots for the given data are reasonably close to the given trend line.

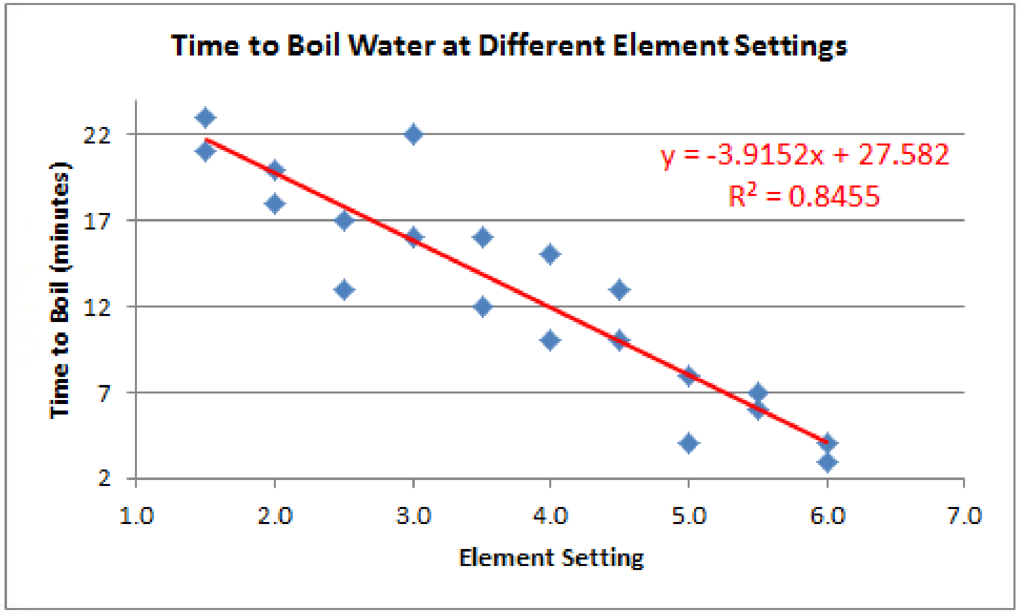
The trend line shows that as shoe size increases, height also tends to increase. Shoe sizes tend to be bigger for taller students. There is some scatter about the trend line, indicating a moderate relationship. It is possible there are two groups, those students with shoe sizes less than 9 (probably mostly girls) and those with shoe sizes greater than 9 (probably mostly boys).

$R^2 = 0.7899 \Rightarrow r = 0.8888$. This suggests that there is a strong positive relationship between shoe size and height for these 24 senior students. The model suggests that for each increase of 1 in shoe size there is an increase of 4.4937 cm in height.

3. What would you write in your “Analysing the Data” section? Discuss the features, including the strength of the relationship.

I wonder if there is a relationship between the element setting on the stove used to boil water and the time it takes for the water to boil.





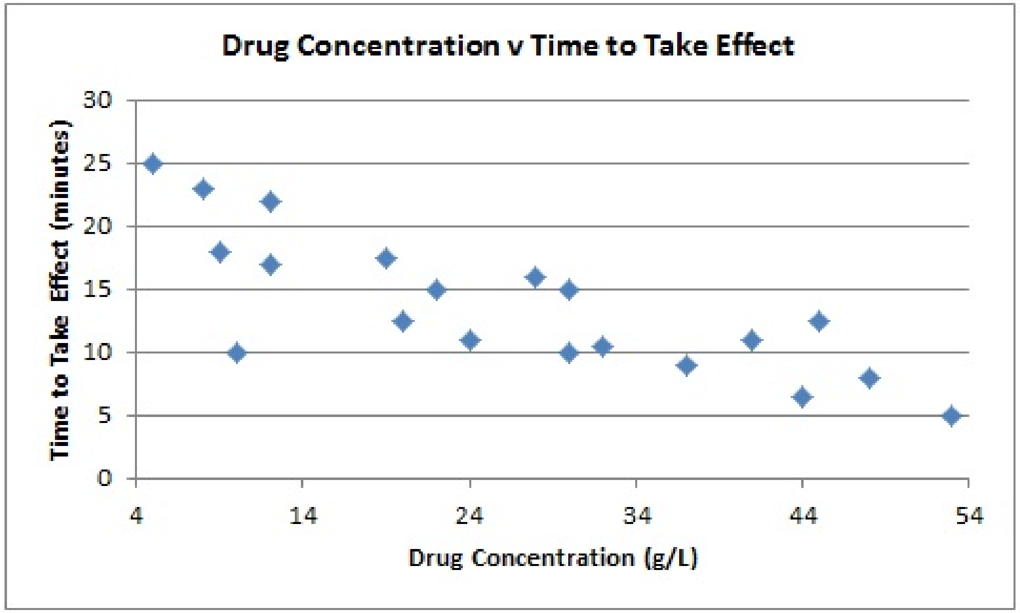
There is a negative linear relationship between the element setting and the time it takes to boil the water.

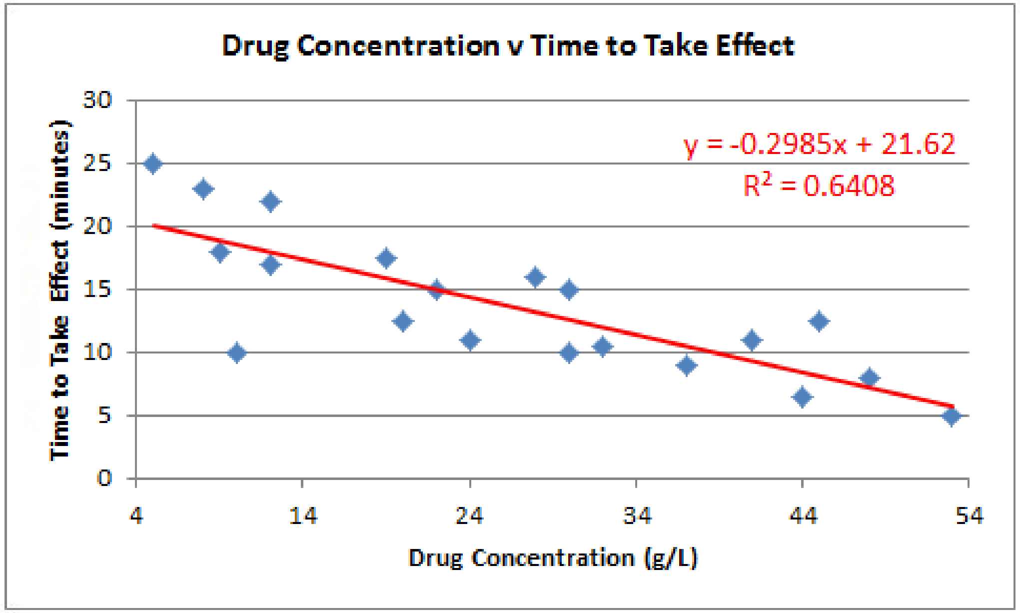
The trend line shows that as element setting increases, boiling time tends to decrease. Higher settings tend to result in shorter boiling times. There is some scatter about the trend line, indicating a moderate relationship, The point (3, 22) appears to be unusual, for a setting of 3 it has a longer than expected time or 22 minutes to boil, We need to go back to the data to see if it is a data entry error, if there has been a mistake (for example the time the pot started to boil was missed), or if it is a valid data point.

$R^2 = 0.8455 \Rightarrow r = -0.9195$. This suggests that there is a strong negative relationship between element setting and time to boil for the stove and quantity of water used in this study. The model suggests that for each increase of 1 in the element setting there is a 3.9152 minute decrease in the time required for the water to boil.

4. What would you write in your "Analysing the Data" section? Discuss the features, including the strength of the relationship.

I wonder if there is a relationship between the concentration of a drug and the time it takes for the drug to take effect.





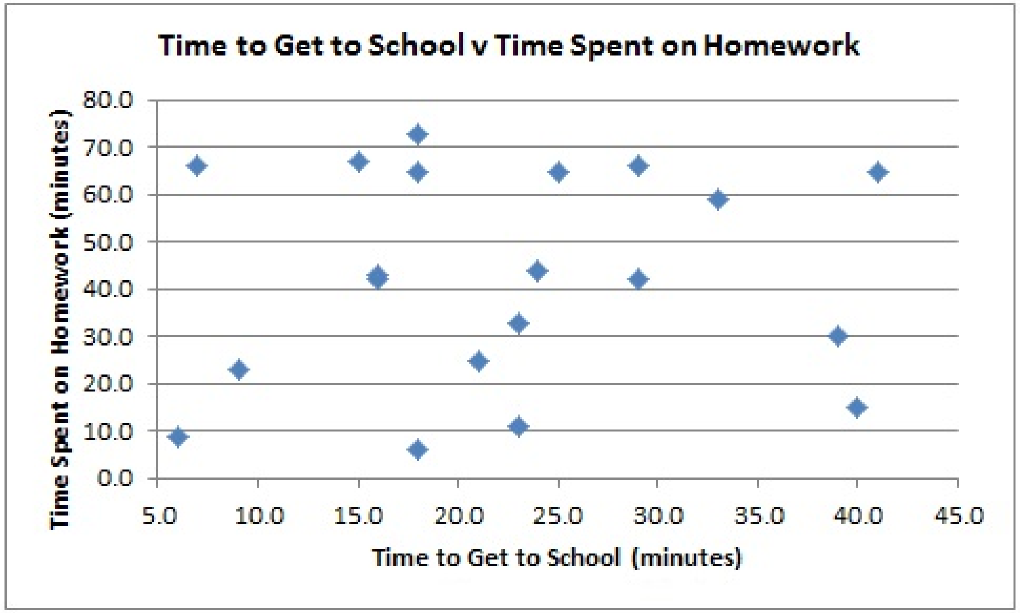
There is a negative linear relationship between the concentration of a drug and the time it takes for it to take effect.

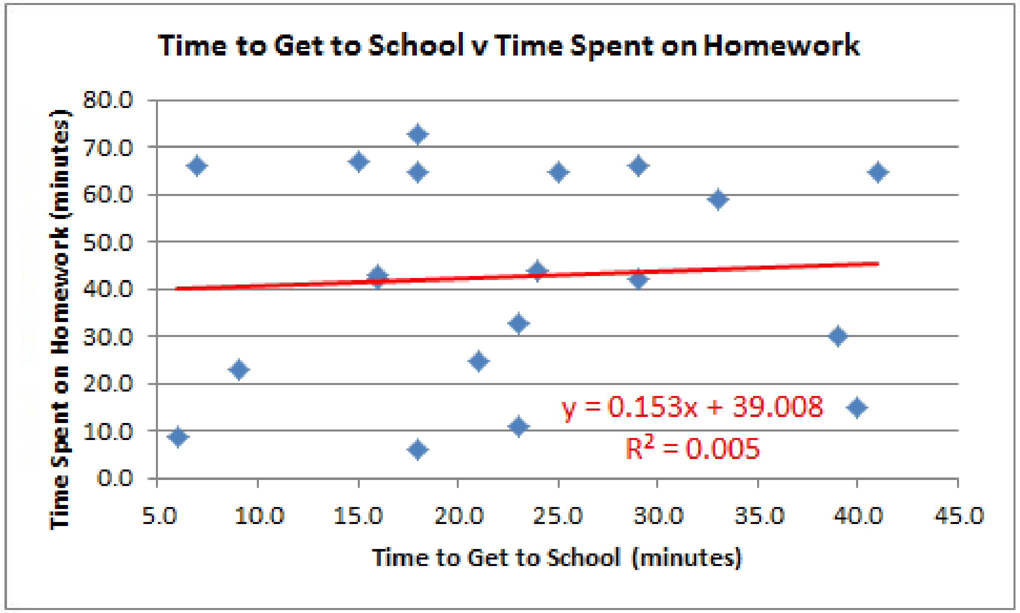
The trend line shows that as drug concentration increases, time to take effect tends to decrease. Stronger doses tend to take shorter times to take effect. There is some scatter about the trend line, indicating a moderate relationship. The point (10, 10) appears to be unusual, for a concentration of 10 g/L the time to take effect of 10 minutes is less than expected. We need to go back to the data to see if it is a data entry error, if there has been a mistake (for example the wrong amount of drug might have been mixed), or if it is a valid data point.

R^2 = 0.6408 => r = 0.8005. This suggests that there is a moderate to strong negative relationship between drug concentration and time to take effect for the 20 people and the drug used in this study. The model suggests that for each increase of 1 gram per litre in concentration of the drug there is a 0.2985 minute decrease in the time required for the drug to take effect.

5. What would you write in your “Analysing the Data” section? Discuss the features, including the strength of the relationship.

I wonder if there is a relationship between the time students take to get to school each day and they amount of time they spend on homework each day.



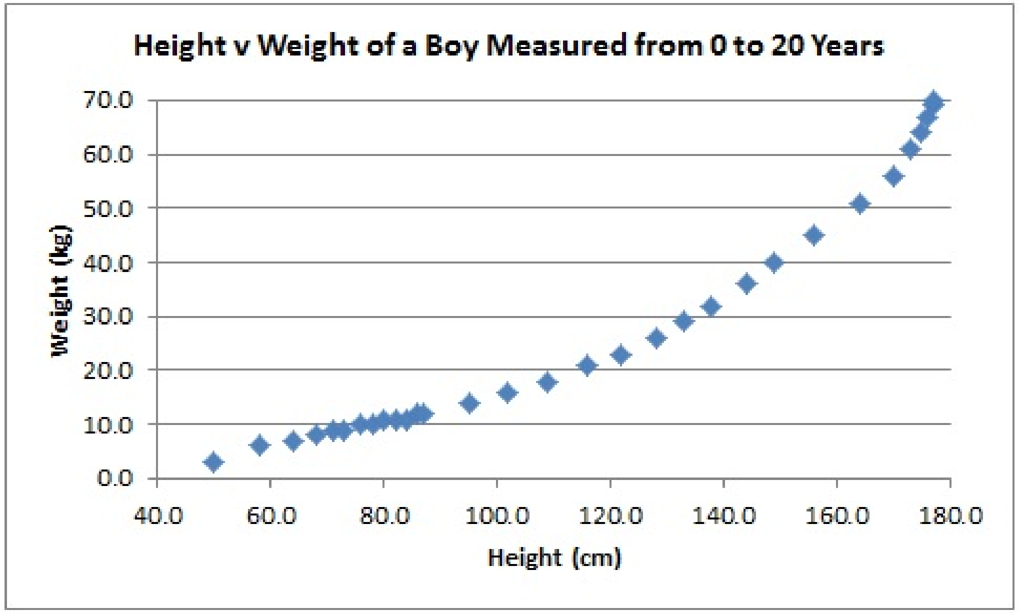


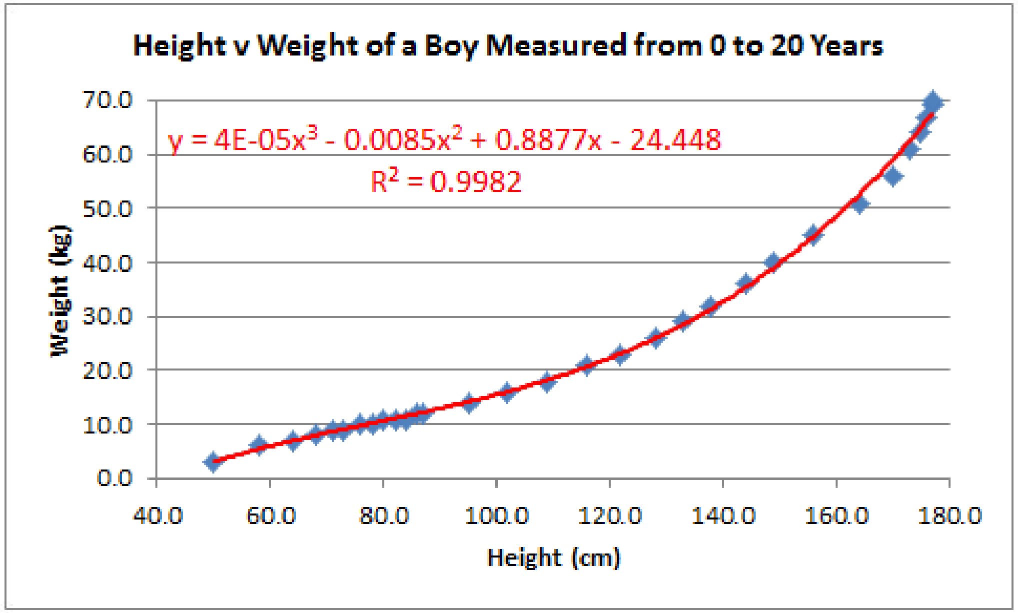
There is no relationship between the time students took to get to school and the amount of time students spent on homework. The points appear to be randomly scattered throughout the plot. Although the computer places a trend line it cannot be meaningfully interpreted.

$R^2 = 0.005 \Rightarrow r = 0.0707$. This suggests that there is a no relationship between time to get to school and time spent on homework.

6. What would you write in your “Analysing the Data” section? Discuss the features, including the strength of the relationship.

I wonder if there is a relationship between a boy’s height and his weight measured at significant times from 0 to 20 years of age.



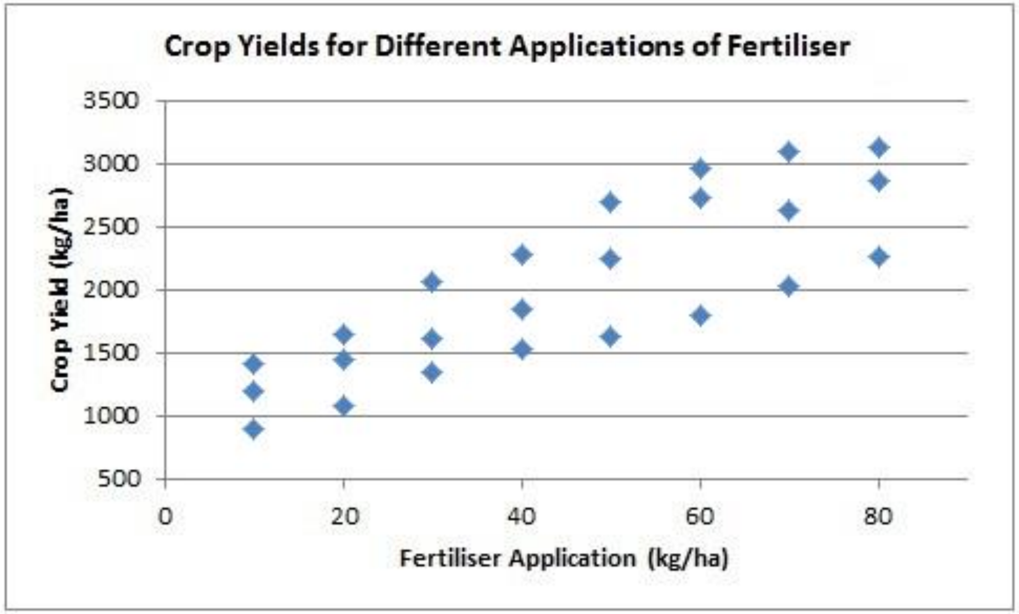


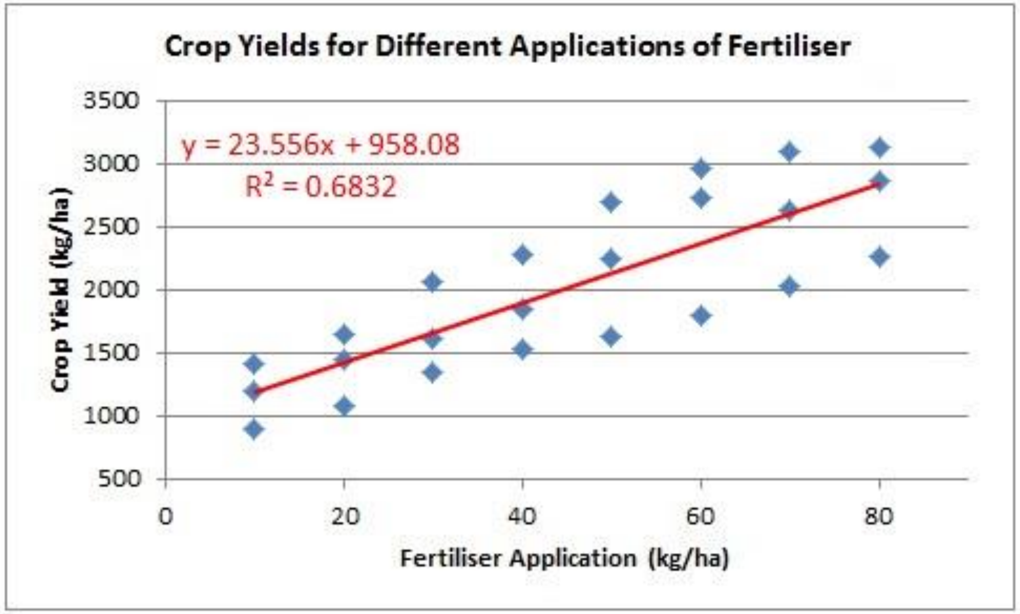
There is a positive non-linear relationship between the height and weight of this boy at significant points of his life from 0 to 20 years of age. Weights go up at an increasing rate as heights increase. The relationship is a strong one as all data values are very close to the curve.

$R^2 = 0.9982 \Rightarrow r = 0.9991$. No comment can be made on the strength of the relationship from this as R^2 and r measure the strength of linear relationships.

7. What would you write in your "Analysing the Data" section? Discuss the features, including the strength of the relationship?

I wonder if there is a relationship between the amount of fertiliser applied per hectare and the crop yield per hectare.





There is a positive linear relationship between the amount of fertiliser that is applied and the crop yield.

The trend line shows that as fertiliser application increases, crop yield also tends to increase. Yields tend to increase with higher applications of fertiliser. There is some scatter about the trend line, indicating a moderate relationship. It appears that there is more variation in the yields for higher application rates. It is possible the yield for the highest two application rates is levelling off.

$R^2 = 0.6832 \Rightarrow r = 0.8266$. This suggests that there is a moderate to strong positive relationship between fertiliser application and crop yield for the particular crop and field used in this study. The model suggests that for each increase of 1 kilogram per hectare in fertiliser application there is an increase of 23.556 kilogram per hectare in crop yield.