Level 3 – AS91581 – 4 Credits – Internal

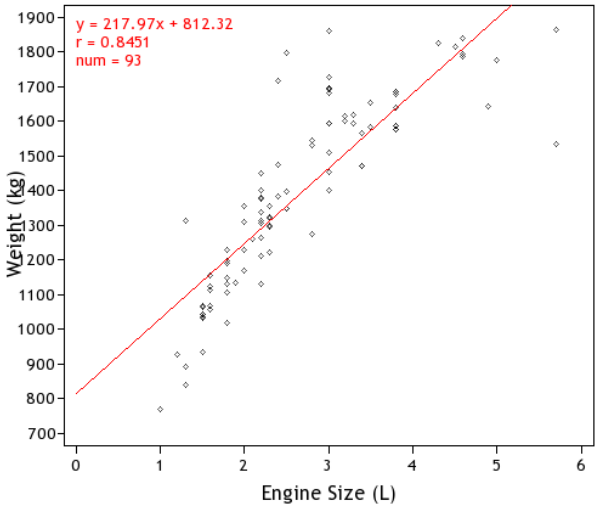
Investigate Bivariate Measurement Data

Written by J Wills – MathsNZ – jwills@mathsnz.com

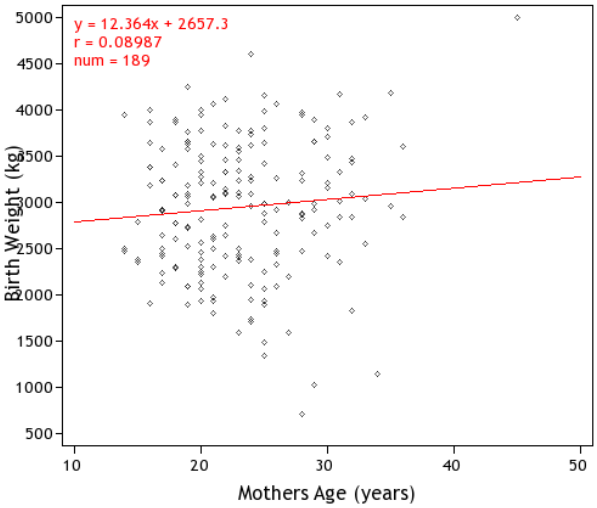
Edited by D Starshaw

Achievement	Achievement with Merit	Achievement with Excellence
Investigate bivariate measurement data.	Investigate bivariate measurement data, with justification.	Investigate bivariate measurement data, with statistical insight.

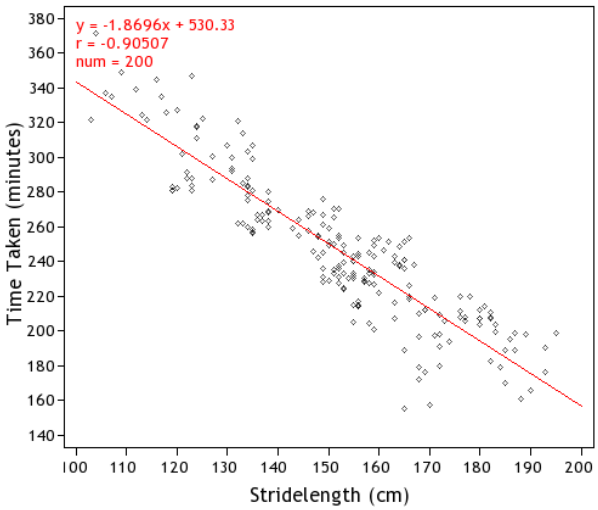
1. Cars Weight by Engine Size



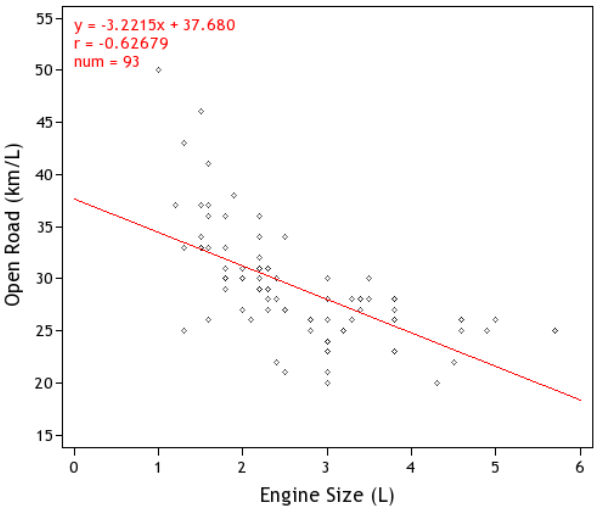
4. Babies Birth Weight by Mothers Age



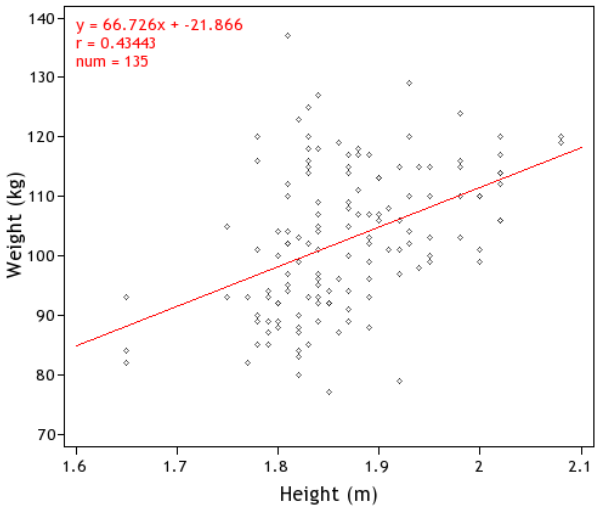
2. Marathon Time by Stride Length



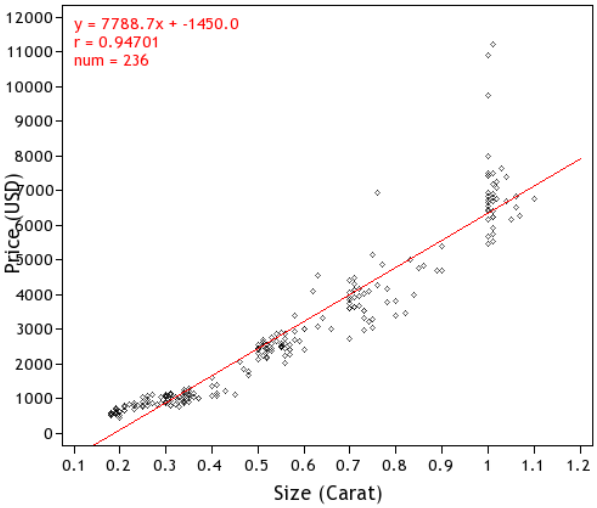
5. Cars Open Road by Engine Size



3. Rugby Players Weight by Height



6. Diamond Price by Size



Problem and Plan

Question and Purpose

For each of the graphs on the previous page write a good relationship question. A question should have:

- What you are trying to find a relationship between and
- What you are planning on predicting (what is on the y-axis).

The first one has been done for you.

1. I wonder if there is a relationship between the engine size of a car and the car's weight for the purpose of predicting the weight of a car.
2. _____
3. _____
4. _____
5. _____
6. _____

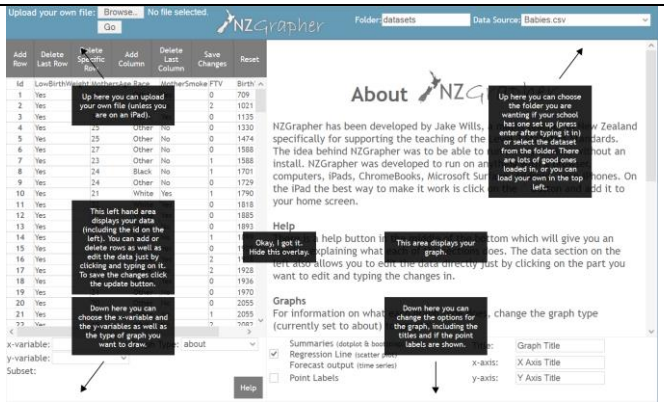
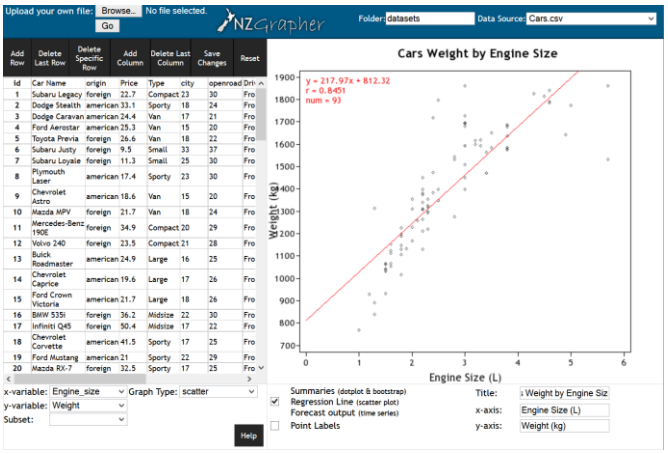
Identifying the Variables

The next thing that we need to do is identify our variables and say what units are being used. The independent variable is the variable on the x-axis, and the dependent variable is the variable on the y-axis that we are wanting to predict.
Identify the variables for each of the graphs on page 2. The first one has been done as an example for you.

1. The independent variable is engine size, which is the displacement of the engine in litres. The dependent variable is the weight of the car, measured in kilograms.
2. _____
3. _____
4. _____
5. _____
6. _____

Data – Using NZGrapher

The next section that we need to do is the data section. This is reproducing the graphs on Page 2 using NZGrapher. The example below will go through using the cars dataset for weight by engine size. NZGrapher runs on anything with a browser… Macs, PCs, iPad, Android, ChromeBooks and more.

First up we need to start NZGrapher by going to the link in the box to the right.	www.jake4maths.com/grapher
The first time you load NZGrapher it will display an overlay with descriptions as to what all the different areas do as shown to the right. To load your data in either select it from the dropdown in the top right, or upload it in the top left corner and press go.	
To draw a scatter plot there are just three things you need to do. 1. Select the x-variable… this is your independent variable that will be on the x-axis, in this case it's engine size. 2. Select the y-variable… this is your response or dependent variable, in this case it's weight. 3. Select the graph type… for this we want the scatter. You then just need to check the graph title and axis labels to make sure they are appropriate (include units where necessary) and press update graph to save the title changes. To add in the regression line press the 'Regression Line' check box. To save the graph just right click on it and press 'Save Image As' or whatever your device says that is similar. *Note 1: The summary statistics are automatically overlaid in red, if you want to remove them just un-tick the summary statistics box.* *Note 2: sometimes you may want to only use some of the dataset… you can either delete each row you don't want in the data viewer, or open it in excel and delete the parts you do not want.* *Note 3: If you want to identify the outliers, if you click the 'Point Labels' checkbox this will add little numbers next to the points that correspond with the point id.*	

Now it is your turn. For each dataset you need to produce the scatter plot for each dataset on page 2. Don't forget to add appropriate titles and units to your graph and axis.

Analysis

We now start on the Analysis section of our report. The acronym we use for this section is TARSOG. The most important thing that you need to remember in this section is that what you can see with your eyes in the most important, not just the numbers.

Trend

The first comment we need to make is about the trend. There are three statements we need to make about the trend.

- How strong the trend is: weak, moderate, strong (or somewhere in between),
- If the trend is positive or negative (does it go up or down) and
- Is the trend linear (most circumstances we look at – forming a straight line) or non-linear.

Write a trend statement for each of the datasets. Again the first one has been done for you.

1. From the graph I can see a strong positive linear relationship.
2. _____
   _____
   _____
   _____
   _____
3. _____
   _____
   _____
   _____
   _____
4. _____
   _____
   _____
   _____
5. _____
   _____
   _____
   _____
6. _____
   _____
   _____
   _____

Association

Association is about explaining why the relationship is either positive or negative, and it is important to link back to the context.

Discuss the association for each of the sets of data, the first one has been done for you.

1. I can see that the association is positive because as the engine size of the car increases, the weight of the car also increases.
2. _____
3. _____
4. _____
5. _____
6. _____

Relationship

The statement about the relationship is about justifying the strength of the trend that you can see on your graph. It is important that you are commenting on what you can see. You can use the correlation coefficient (r-value) to back up your strength statement, but it should only be used as a backup… what you can see is the most important.

There is an activity matching correlation coefficients with graphs at:
http://www.wilderdom.com/301/int/cor-guess.html

The table below is just a guide, and a positive r-value indicates that the trend is positive, a negative r-value indicated the trend is negative, so when you use the table below ignore the negative sign if there is one.

r-value	**Strength**
0.4 – 0.6	Weak
0.6 – 0.8	Moderate
0.8 – 1.0	Strong

Discuss the relationship for each of the sets of data, the first one has been done for you.

1. The relationship is strong and linear as I can see most the points are fairly close to the trend line. This is confirmed by the correlation coefficient of 0.8451, indicating that the linear relationship is quite strong.
2. _____
   _____
   _____
3. _____
   _____
   _____
4. _____
   _____
   _____
5. _____
   _____
   _____
6. _____
   _____
   _____

Scatter

In the scatter section you need to look and see how consistent the scatter is. Are there any areas that are denser or sparser than others?

Discuss the scatter for each of the sets of data, the first one has been done for you.

1. The scatter appears to be reasonably consistent throughout the data range with no obvious areas with more points.
2. _____
3. _____
4. _____
5. _____
6. _____

Outliers

In a large number of graphs there will be points that do not follow the trend. These are called outliers. When you identify an outlier you need to find it on the data list and find out as much information about it as you can in order to explain why it might be an outlier.

Discuss the outliers for each of the sets of data, the first one has been done for you.

1. There is a possible outlier that has a lower weight than expected. This is the Chevrolet Corvette which has a 5.7 litre engine but only weighs 1533kg. The Chevrolet Corvette is a sports car, so hence doesn't weigh as much as the other 'Large' type vehicles that weigh that much.
2. _____
3. _____
4. _____
5. _____
6. _____

Grouping

Sometimes in graphs you can end up with two groups (or clusters) of data. If this happens you need to comment on it and what might be causing it, otherwise you can comment that there is not any obvious grouping. Again link it to what you can see.

Discuss the grouping for each of the sets of data, the first one has been done for you.

1. Looking at the graph I cannot see any obvious groupings.
2. _____
   _____
   _____
   _____
   _____
3. _____
   _____
   _____
   _____
   _____
4. _____
   _____
   _____
   _____
   _____
5. _____
   _____
   _____
   _____
   _____
6. _____
   _____
   _____
   _____
   _____

Note: to be going for a Merit Grade in this TARSOG section you need to be justifying these features in context.

Interpretation of Regression Line – Merit Level Skill

One of the key bits of information that we get given from NZGrapher or iNZight is the equation of the regression line. Interpreting the gradient of this regression line is an important comment to make. It is vital that you realise that this is only giving the **average** increase over the whole graph, and not a fixed amount for every unit.

Interpret the regression line for each of the sets of data, the first one has been done for you.

1. The regression line of Weight = 217.97 x Engine_size + 812.32 means that for every one litre increase in engine size, the weight of the car increases by 217.97 kg on average.
2. _____
3. _____
4. _____
5. _____
6. _____

Predictions

As well as interpreting the regression line we need to use this line to make **at least two** predictions and comment on how reliable you think the predictions are.

In order to do this you need to substitute two different values into the equation. It is normally good to have one that is inside your data range (interpolation), and one that is outside of your data range (extrapolation). It is also vital that you **round the prediction sensibly** (usually the same as the original data for that variable was rounded to)

Make a prediction inside the data range for each of the sets of data, the first one has been done for you.

1. *217.97 * 3 + 812.32 = 1466.23*

   Based on my regression line I would predict that a car with a 3 litre engine would weigh approximately 1470kg. I am confident in this prediciton as it is right in the middle of my data range and I have similar values to this result.

2. _____

   _____

   _____

   _____

   _____

3. _____

   _____

   _____

   _____

   _____

4. _____

   _____

   _____

   _____

5. _____

   _____

   _____

6. _____

   _____

   _____

   _____

Make a prediction outside the data range for each of the sets of data, the first one has been done for you.

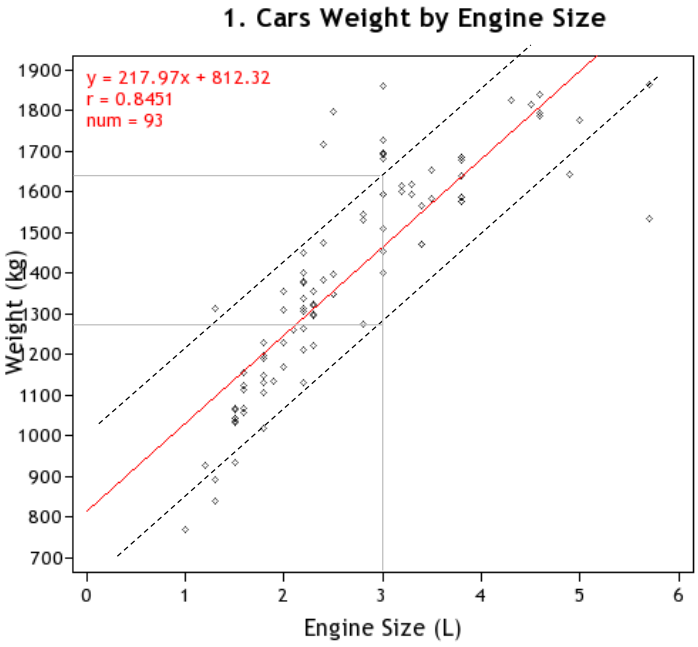
1. Based on my regression line I would predict that a car with a 10 litre engine would weigh approximately 2990kg (217.97*10 + 812.32). I cannot be confident in this predicion as I do not have any cars to compare it to.
2. _____
3. _____
4. _____
5. _____
6. _____

Note: to be going for a Merit level grade you need to be discussing the confidence in the predictions (see page 15) and linking this to the residuals (see page 16) and to be getting Excellence you need to be reflecting on how widely these predictions can be used.

Using the Graph for Confidence in Predictions – Merit Level Skill

Looking at the graph and how big the scatter is can give us a good indication as to how accurate our predictions are likely to be, and therefore create an interval that we think our predictions might be in.

If we look at this graph here we can see that most of the points are between the dotted lines (you put these in manually by eye.)
From this we can see that the prediction that I made earlier for a car with a 3L engine is likely not to be exactly 1470kg, but is likely to between 1280kg and 1650kg (as shown by the light grey lines)



Causation and Correlation – Merit Level Skill

While there is a relationship between two variables, there are two reasons why we cannot make causal statements:

- We don't know the direction of the cause – Does X cause Y or does Y cause X?
- A third variable may be involved that is responsive for the covariance between X and Y, we call this a lurking variable.

Causal relationships can only be determined by controlled experiments, which we will look at later in the year in a different standard.

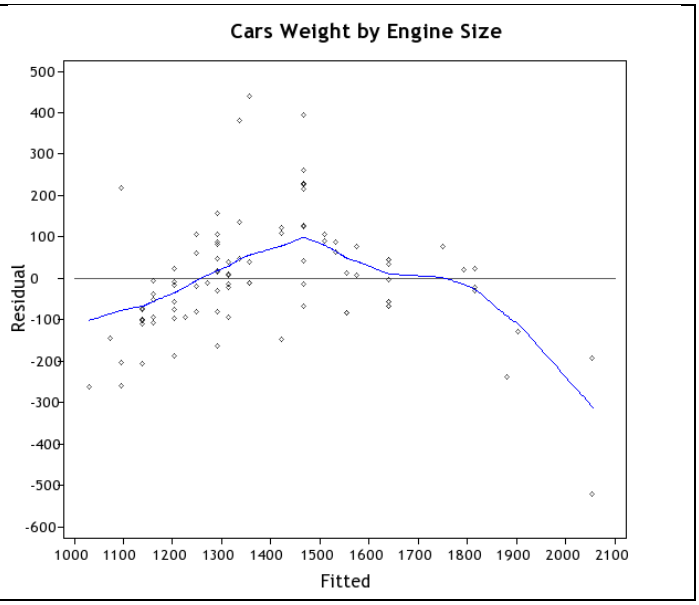
Residuals – Merit Level Skill

One of the ways that we can analyse how well our model fits the data and therefore how accurate our predictions are is by looking at the residuals. We can create a plot of the residuals using NZGrapher.

This part is really easy… all you need to do is change the graph type from the graph that we did earlier to 'residuals'.

This gives the output shown to the right, which shows the expected (or fitted) values on the x-axis and the difference between the fitted and the actual (the residual) on the y-axis.

The line that is put in is a weighted average curve that shows the overall trend of the data.



Now it's your turn. For each of the sets of data produce a residuals plot, and discuss it on the following page.

For each of the sets of data look at the residual plots and use this to justify how accurate you think your predictions are, the first one has been done for you.

1. Looking at my residuals plot I think that the first car I predicted to weigh 1470kg might actually weigh slightly more than this due to most of the values being above the predicted line in the middle of the range. Based on looking at the residuals the car that I predicted would weigh 2990kg probably will not weigh this much as the weighted average line shows a clear curve downwards, but I do not have any other cars to compare it to.
2. _____
3. _____
4. _____
5. _____
6. _____

Conclusion

We now need to make a concluding statement to summarise our report. You need to include a statement around the relationship, and it needs to be linked back to what you are predicting.

Make a conclusion for each of the sets of data, the first one has been done for you.

1. In conclusion I think there is a strong positive relationship between the engine size and weight of cars – the larger the car's engine is the more that they will weigh, therefore if we know the size of a car's engine we should be able to predict how heavy the car will be.
2. _____
3. _____
4. _____
5. _____
6. _____

Congratulations, you now have written up a report for 5 different sets of data.

Writing Your Own Internal 1

Using the framework below write a report on the kiwi data. You can use the sample internal on pages 23 and 24 to help you if you need it.

Kiwi Birds

Problem

_____

_____

_____

Question and Purpose.

Plan

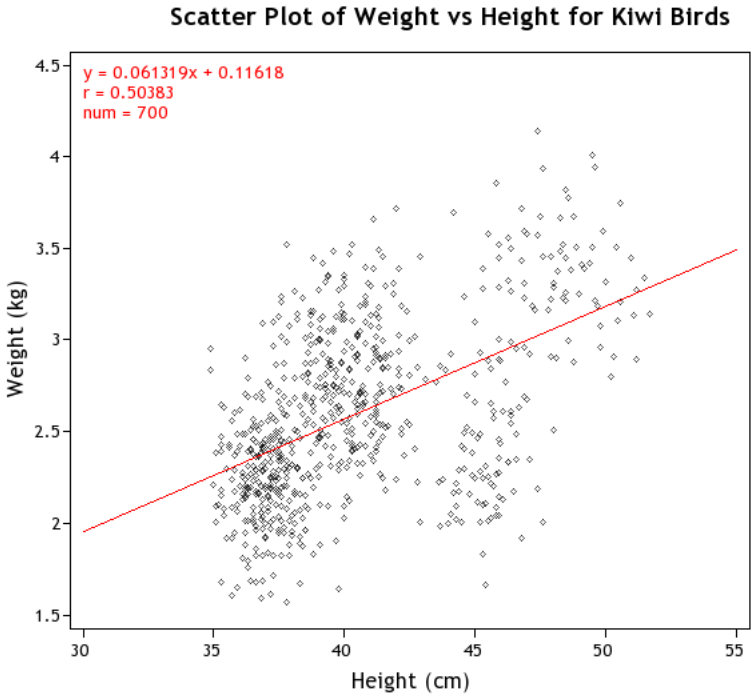
_____

_____

_____

Variables Identified

Data



Graph with Regression Line and Equation Given

Analysis

_____

_____

Trend

_____

_____

_____

Association

_____

_____

_____

_____

Relationship

Scatter

Outliers

Grouping

Interpretation of Regression Line

Predictions

Conclusion

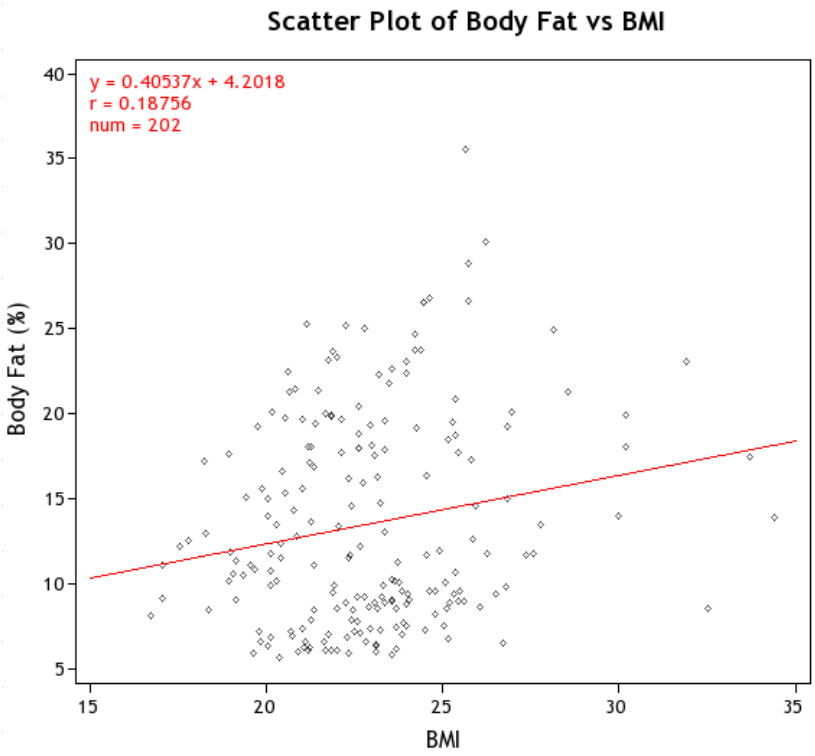
Summarise and link back to the purpose

Writing your own internal.

Writing Your Own Internal 2

This time you have just been provided with a title and graphs. Using these write your own internal.

BMI and Body Fat



Weight and Height of Athletes

Problem

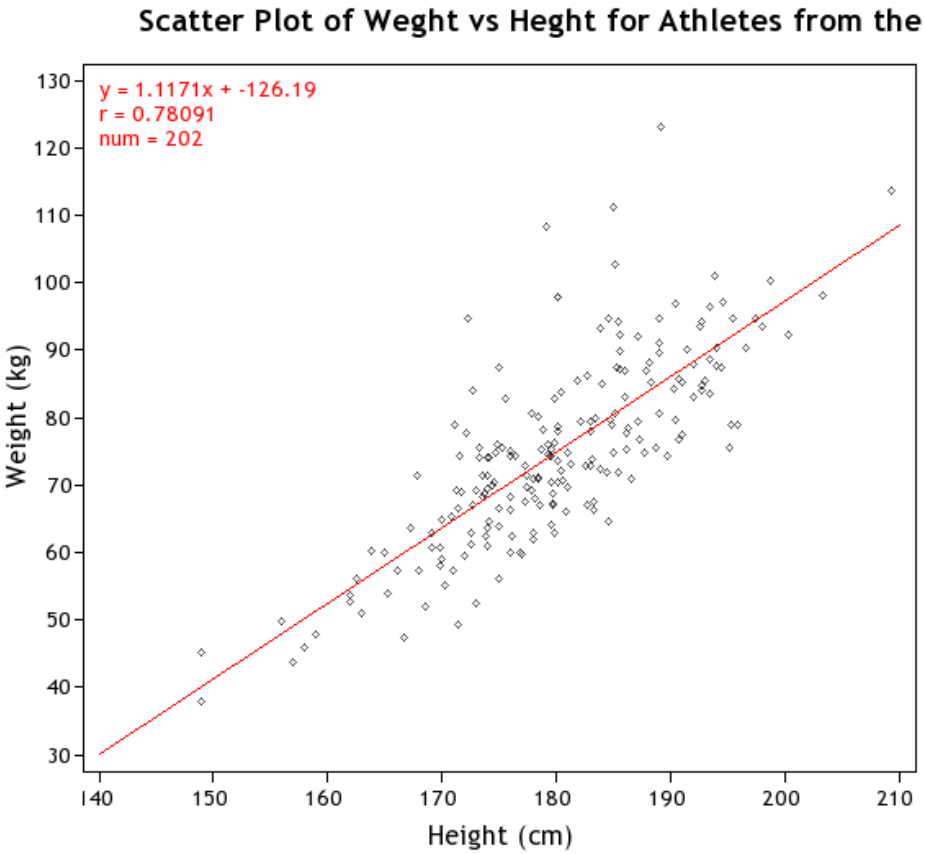
I wonder if there is a relationship between the weight and height of athletes for the purpose of predicting the weight of an athlete. | Question and Purpose.

Plan

The independent variable is the height of the athlete in centimetres. The dependent variable is the weight of the athlete in kilograms. | Variables Identified

The data used in this investigation is from the Australian Institute of Sport. | Source Named

Data



Graph with Regression Line and Equation Given

Analysis

From the graph I can see a moderate positive linear relationship. | Trend

I can see that the association is positive because as the height of the athlete increases, the weight of the athlete increases. This is also backed up by the positive gradient of the trend line. | Association

The relationship is moderate and linear because I can see most of the points are fairly close to the trend line. This is confirmed by the correlation coefficient of 0.78, | Relationship

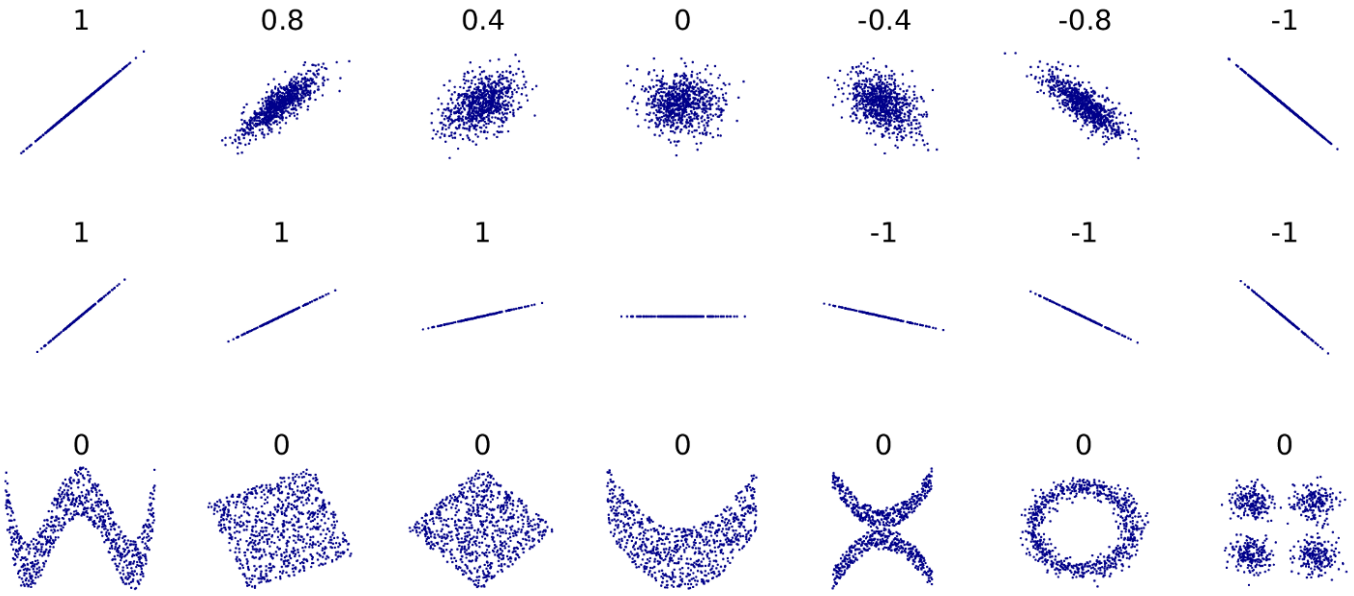
indicating a moderate strength to the linear relationship.	
The scatter seems reasonably consistent, but with more points in the middle than on the edges.	Scatter
There is one possible outlier that has a higher weight than expected. This is a male athlete who is 189.2cm tall and 123.2kg. He is a field athlete, so this could be due to him being a shotputter.	Outliers
Looking at the graph I cannot see any obvious groupings.	Grouping
The regression line of Weight = 1.1171 x Height + -126.19 means that for every one centimetre increase in height, the weight of the athlete increases by 1.1171 kg on average.	Interpretation of Regression Line
*1.1171 * 180 - 126.19 = 74.888* Based on my regression line I would predict that an athlete that was 180cm tall would weigh approximately 74.9kg. I am confident in this prediciton as it is right in the middle of my data range and I have similar values to this result. *1.1171 * 130 - 126.19 = 19.033* Based on my regression line I would predict that an athlete that was 130cm tall would weigh approximately 19.0kg. I cannot be confident in this predicion as I do not have any athletes to compare it to.	Predictions

Conclusion

In conclusion I think there is a moderate positive relationship between the height and weight of athletes – the taller an athlete is the more that they will weigh, therefore if we know how tall an athlete is we should be able to predict how tall they are.	Summarise and link back to the purpose

Examples of Correlation Coefficient (r)

Below are graphs showing the different correlation coefficients (r values) above each one.

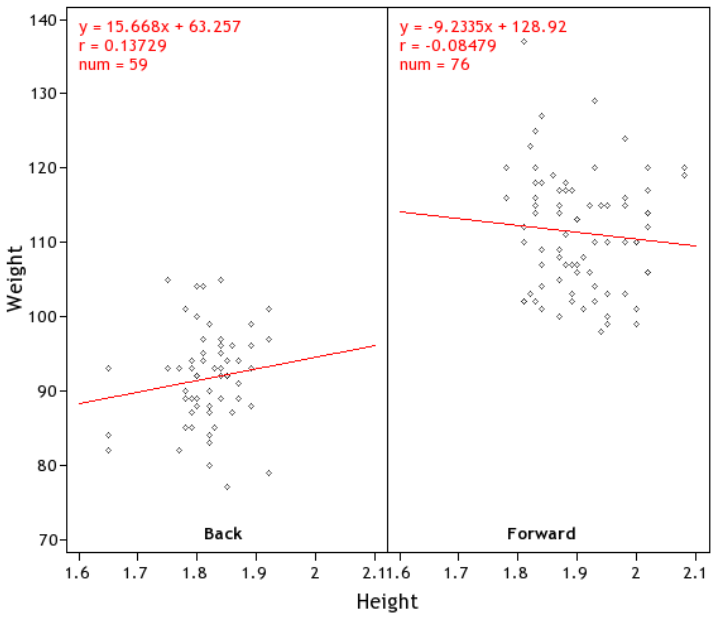


Source: http://upload.wikimedia.org/wikipedia/commons/d/d4/Correlation_examples2.svg

The important thing to note here is that correlation coefficients, or r-values, only tell you about **linear** relationships. If the relationship is non-linear then it is not useful (see the graphs at the bottom).

Separating Variables

For Merit and Excellence one of the things you can do is separate out the data by a categorical variable to see if this is affecting the trend. You do this by choosing a variable in the 'Subset' section on NZGrapher. This gives the output on the right.



Data Set Information

Babies

Babies

The data on 189 births were collected at Baystate Medical Center, Springfield, Mass. during 1986.

The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data was collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies.

Variable	Description
LowBirthWeight	No = Birth Weight >= 2500g Yes = Birth Weight < 2500g
MothersAge	Age of the Mother in Years
Race	Race of the mother
MotherSmoke	Smoking Status During Pregnancy
FTV	Number of Physician Visits During the First Trimester
BirthWeight	Birth Weight in Grams

Cars

With rising costs of owning and running a car, and environmental awareness, buyers are becoming more conscious of the features when purchasing new cars. The data supplied is for new vehicles sold in America in 1993.

Variable	Description
Vehicle Name	
Origin	Country of manufacture • America • Foreign
Price	US $1000
Type	Small, midsize, large, compact, sporty, van
City	Fuel efficiency in kilometres per litre in cities and on motorways
OpenRoad	Fuel efficiency in kilometres per litre on country and open roads
Drive Train	Front Wheel Drive Rear Wheel Drive
Engine Size	Size in litres
Manual Transmission	Yes No
Weight	Weight of car in Kg

Diamonds

Every diamond is unique, and there are a variety of factors which affect the price of a diamond. Insurance companies in particular are concerned that stones are valued correctly.

Data on 308 round diamond stones was collected from a Singapore based retailer of diamond jewellery, who had the stones valued.

Variable	Description
Carat	Weight of diamond stones in carat units 1 carat = 0.2 grams
Colour	Numerical value given for quality of colour ranging from 1=colourless to 6=near colourless
Clarity	Average = score 1, 2 or 3 Above average = score 4, 5 or 6
Lab	Laboratory that tested & valued the diamond 1 = laboratory 1 2 = laboratory 2
Price	Price in US dollars

Kiwi

A sample of kiwi birds around New Zealand was collected in order to help with conservation efforts. The original data is from: http://www.kiwisforkiwi.org/ and was sourced from the secondary school guides (*http://seniorsecondary.tki.org.nz/Mathematics-and-statistics/Achievement-objectives/AOs-by-level/AO-S7-1*)

Variable	Description	
Species	GS-Great Spotted NIBr-NorthIsland Brown Tok-Southern Tokoeka	
Gender	M-Male F-Female	
Weight(kg)	The weight of the kiwi bird in kg	
Height(cm)	The height of the kiwi bird in cm	
Location	NWN-North West Nelson CW-Central Westland EC-Eastern Canterbury StI-Stewart Island NF-North Fiordland	SF-South Fiordland N-Northland E-East North Island W-West North Island

Teachers note: this is a synthesised dataset based on real data. At the time of creating the data set there were around 25,000 brown, 17,000 great spotted and 34,500 southern tokoeka. These numbers formed the basis of the data set, but instead of being out of around 76,000 the data set contains around 700 birds.

The data was generated using the population parameters, including gender, location, height, weight and species in Fathom. The size of the population was so that it was too big to use all the data (when doing by hand) but not too big that it couldn't be created for students to use as a "population" to sample from.

Marathon

The data is a sample taken from marathons in NZ.

It is a simple random sample of 200 athletes.

Variable	Description
Minutes	How many minutes they completed the marathon in
Gender	Male (M) or Female (F)
AgeGroup	Younger (under 40) or older (over 40)
StridelengthCM	The persons average stride length over the marathon in cm.

Rugby

The data is real data and comes from http://www.rugby-sidestep-central.com/

Variable	Description
Country	New Zealand or South Africa
Position	Forward or Back
Weight	The weight of the player in kilograms (kg)
Height	The height of the player in metres (m)

Assessment Guidelines – 91581 – Investigate Bivariate Measurement Data

	Achieved (all compulsory)	Merit… Achieved PLUS	Excellence… Merit PLUS
Overview	Evidence is shown of using each component of the statistical enquiry cycle.	Evidence is shown of linking components of the statistical enquiry cycle to the context and supported by referring to evidence such as statistics, data values, trends, or features of the displays.	Evidence is show of integrating statistical and contextual knowledge throughout the process. This may include reflecting on the process, considering other relevant variables, or evaluating the model(s).
Problem	An appropriate relationship question is posed.	The question is justified in context.	The choice of variables is reflected on and linked to the context and research.
Plan	Variables and data source are identified. The explanatory and response variables are clear.		
Data	Scatter plot(s) is produced with tile and labelled axis and regression line fitted.	Residuals plot may be produced.	
Analysis	Features in the data are identified and described. This should include Trend, Association, Relationship, Scatter, Grouping and Outliers. Other features and unusual points have been identified.	Features are described in context and justified with visual references to the displays and possible contextual reasons for the features are given. Causation may be discussed in context. The appropriateness of the model is justified by discussion of fit throughout the range of x-values in the data or the number of data points. An analysis of the residuals may be used.	Contextual evidence and research is used to support discussion about the features of the data. A discussion may be given as to how other (named) factors will impact the variable. They have reflected on features by discussing their relevance to a wider population. The strength of the model is discussed with consideration for aspects such as the number of data points. Improvements to the model may be considered by looking at alternative models or separating the data into relevant subsets.
Predictions	A prediction is made that is sensible with respect to the context and uses units and sensible rounding.	Predictions are interpreted in context and are justified with discussion on how precise they might be, supported with references to statistical evidence from the analysis.	The choice of variable used for predictions is justified by giving reasons for using the selected one rather than others. Reflection is made on predictions by discussing their relevance to a wider population. The prediction may be made using alternative models and the accuracy of these has been discussed in the context of the investigation.
Conclusion	A conclusion is given that is consistent with the question.	The conclusion is linked to the question with contextual support. Consideration may be given as to ways to extend the investigation (for example sub-setting).	The conclusion shows an extension of the original investigation and research and contextual reasons are made to support findings and justify the extension and how this is linked to the original question.

Final grades will be decided using professional judgement based on a holistic examination of the evidence provided against the criteria in the Achievement Standard.