Level 3 – AS91581 – 4 Credits – Internal

Investigate Bivariate Measurement Data

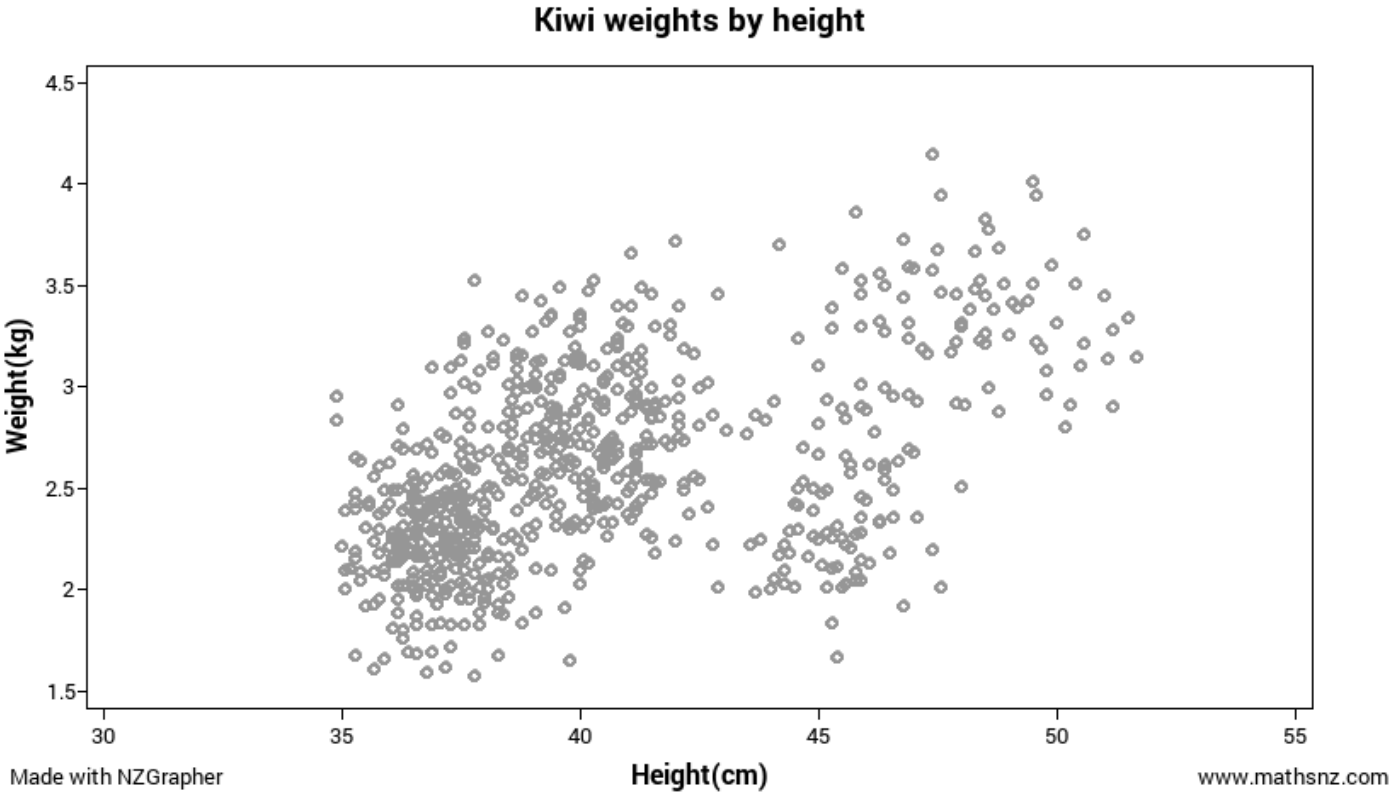
Context

Level 1 **Level 2** **Level 3 statistics**

3.9 Bivariate data
4 credits (Int)

Sample question: Kiwi birds

Does a Kiwi's weight depend on its height? Use the scatter graph below to help you answer this question.

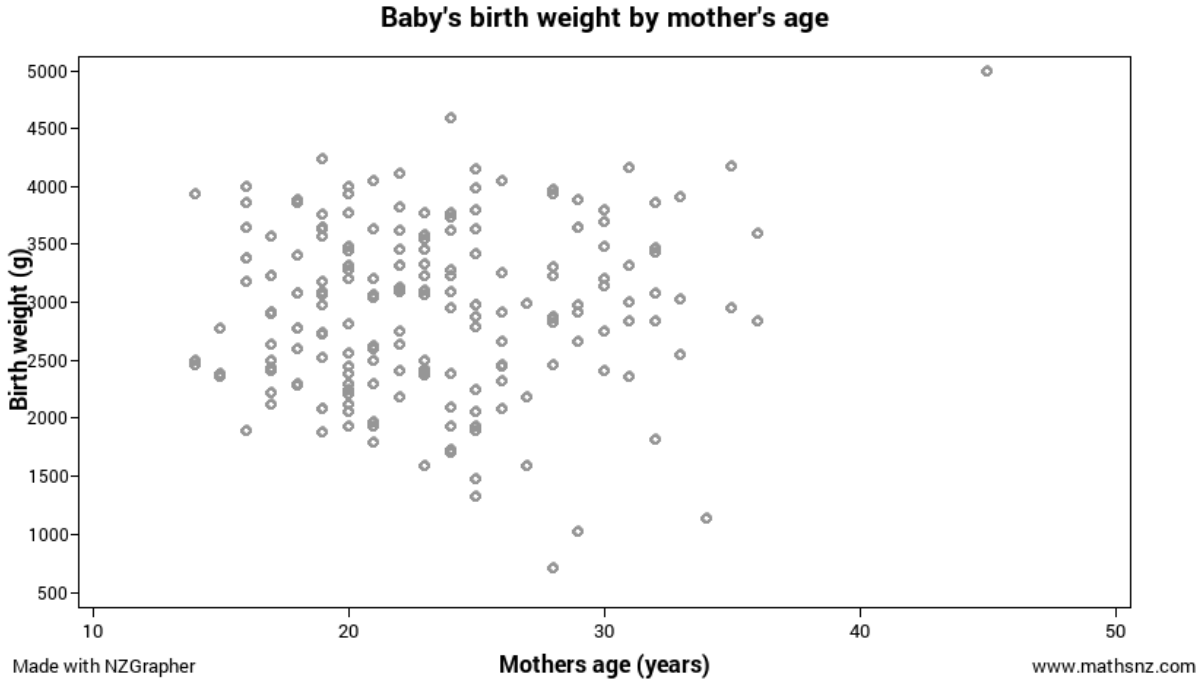


Practice internal

As you progress through this workbook, you'll do two practice internals.

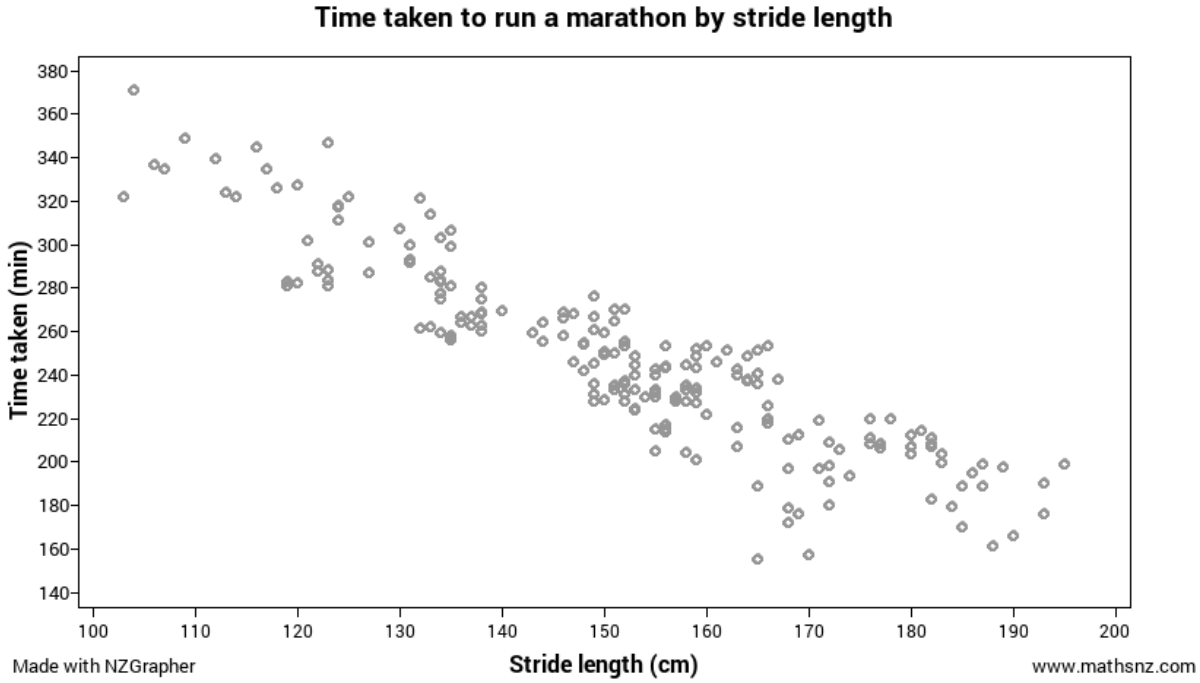
Babies

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.



Marathon

The data is a sample taken from marathons in NZ.
It is a simple random sample of 200 athletes.



Grade requirements

	Achieved	Merit	Excellence
Problem	An appropriate relationship question is posed and linked to research. The explanatory and response variables are clear.	An appropriate relationship question is posed, **justified in context,** and linked to research. The explanatory and response variables are clear.	An appropriate relationship question is posed and justified in context. **The choice of variables is reflected on and linked to the context and research.** The explanatory and response variables are clear.
Plan	Variables have been elaborated upon and the data source is described.		
Data	Scatter plot is produced with an appropriate title, labelled axes, and a regression line fitted. The explanatory and response variables are on the correct axis.		
Analysis	Features in the data are identified from a visual inspection and described. This should include: • Trend • Strength of the relationship • Scatter • Clusters • Outliers • Other features and unusual points have been identified. Found a linear model using linear regression. Made a prediction using the regression line.	Features in the data are identified from a visual inspection and described. This should include: • Trend • Strength of the relationship • Scatter • Clusters • Outliers • Other features and unusual points have been identified. **Findings are justified with reference to evidence from the displays and statistics, then links findings to their research and purpose.** Found **an appropriate model using appropriate regression. The appropriateness of the model is justified.** Made **and justified** a prediction using the regression line, strength of the relationship, and residuals plot.	Features in the data are identified from a visual inspection and described. This should include: • Trend • Strength of the relationship • Scatter • Clusters • Outliers • Other features and unusual points have been identified. Findings are justified with reference to evidence from the displays and statistics **and research**, then links findings to their research and purpose. **Features are reflected on by discussing their relevance.** Found an appropriate model using appropriate regression **and by comparing different linear and non-linear models**. The appropriateness of the model is justified. Made and justified a prediction using the regression line, strength of the relationship, and residuals plot. **Improvements to the model have been considered by considering other variables (eg: separating the data into relevant subsets or looking at another related variable).**
Conclusion	A conclusion is given that is consistent with the question and linked to the purpose.	The conclusion is linked to the question with contextual support.	The conclusion shows a deeper understanding of the data and research and contextual reasons are made to support findings.

Lesson One: Introduction

Bivariate [noun]: Two variables.
Bivariate data [noun]: Data involving two quantitative variables.
Quantitative variable [noun]:

If we want to know how a variable[1] affects another variable[2], we're asking a bivariate question. For example:

- Does a person's weight depend on their height?
- Does the fuel efficiency of a car depend on its engine size?
- Does your BMI depend on the number of hours you spend playing sports?
- Does the fat content of food depend on the salt content?

Ask your own bivariate questions:

-
-
-
-

In each bivariate question, one of the variables depends on the other. The variable that depends on the other is called the dependent variable. The other variable is the independent variable.

Independent variable [noun]: The variable that is causing the change. AKA the explanatory variable.
Dependent variable [noun]: The variable that is being changed. AKA the response variable.

For each of the eight questions above, identify the dependent/response variable and the independent/explanatory variable.

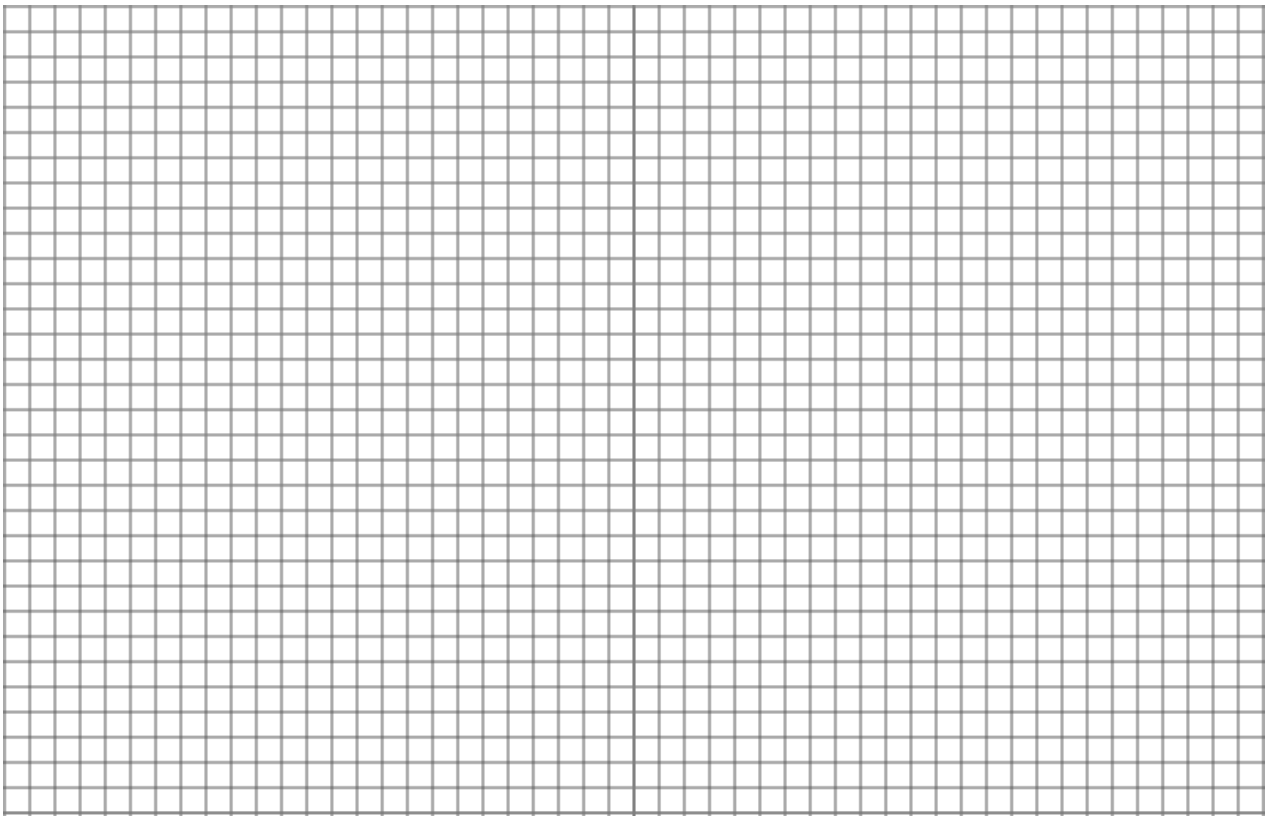
1.
2.
3.
4.
5.
6.
7.
8.

[1] A quantitative variable
[2] Another quantitative variable

Scatter plots

To investigate a bivariate question, you'll create a scatter plot, AKA a scatter graph. Create a scatter graph for the data below. You'll be answering the question "Does the time taken to run a marathon depend on your stride length?"

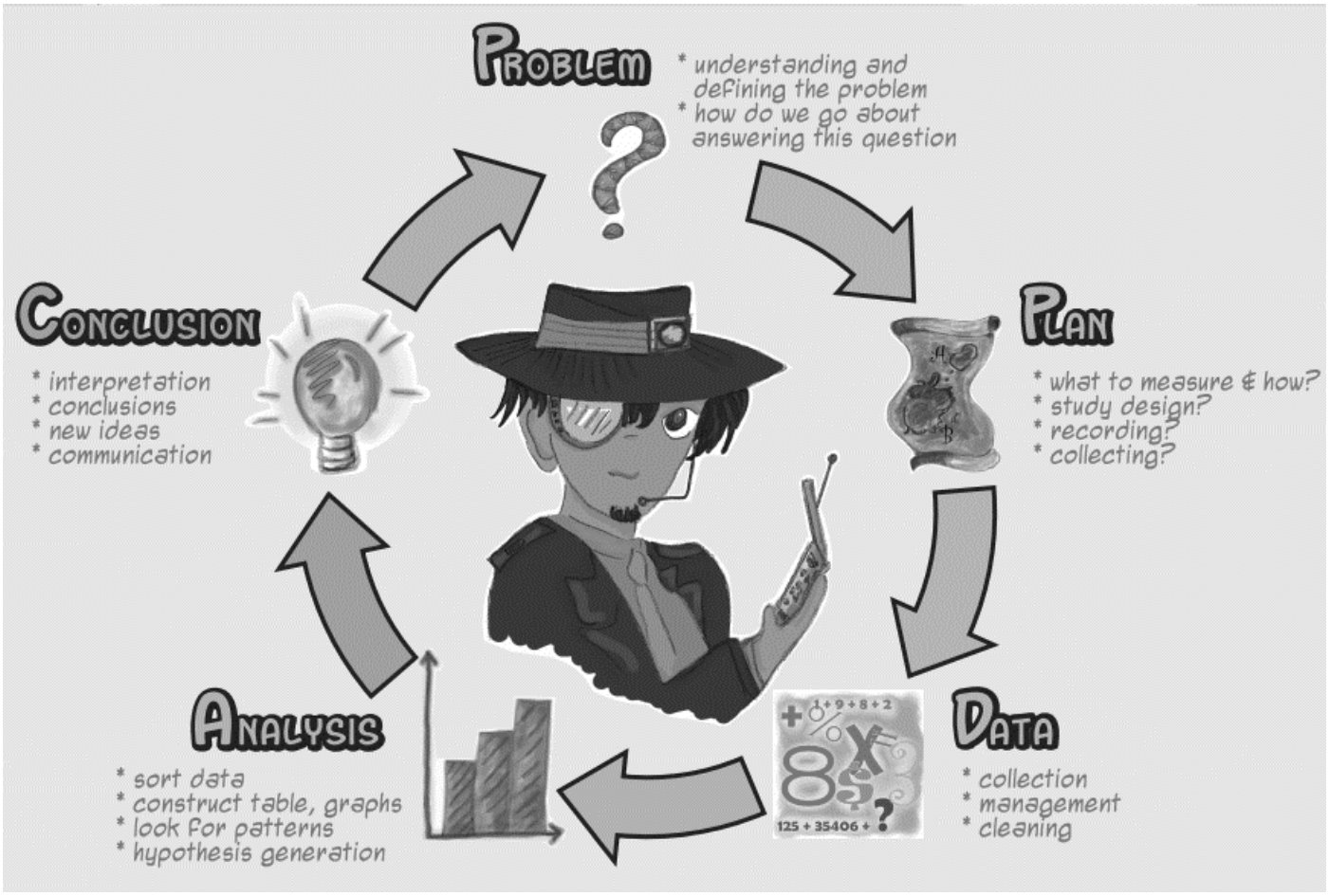
id	Time taken (minutes)	Stride length (cm)
1	321.7	114
2	302.9	134
3	283	134
4	233	153
5	171.8	168
6	195	186
7	259.6	134
8	279.9	138
9	269.5	140
10	263.8	144
11	257.7	146
12	269	146
13	245.7	147
14	253.8	148
15	267	149
16	253	156
17	248.7	164
18	155	165
19	240.7	165
20	211	176



(6)

Lesson Two: Problem

Recall the PPDAC cycle.



Any statistical report starts with a problem. You need to state:

- What's the benefit of your research? Who does it benefit?
- The two variables you're comparing
- Which variable affects which variable
- Findings from looking at others' research

Benefit

At level 3, we're not doing statistical investigation for the fun of it. We're doing them because it has a believable use in the real world. For example, by comparing fuel efficiency and engine size of cars, we can help car buyers [especially inexperienced buyers] make an informed choice.

Look back at the bivariate questions from lesson one. For each question, write a believable benefit of researching that question on the next page.

1.

2. It can be confusing buying a car because you're not sure who to believe. It's especially confusing if it's your first time buying a car. My research will help car buyers make an informed choice.

3.

4.

5.

6.

7.

8.

Variables
You know that you'll be comparing two variables. What kind of variables are you comparing?[3] You need to state what both of these variables are and which one affects the other, i.e. which is the independent variable and which is the dependent variable. For example, "I will investigate if the fuel efficiency of a car depends on its engine size."

Others' research
You're not the first person to be investigating this problem. Sorry. There are going to be a lot of sources that have really good knowledge about your topic[4] [e.g. cars]. Use them! You are required to search for existing information about your context. For example:

"A large engine often has more cylinders and a greater capacity to burn fuel than a small one, so it uses more fuel as a result."[5]

I suggest you put the source of your information as a footnote rather than in the body of your report as I've done here.

Practice internals
Write an appropriate question for both of the practice internals. Check page 5 for the requirements.

[3] Quantitative variables, AKA measurement variables
[4] AKA context
[5] http://www.carbuyer.co.uk/tips-and-advice/146778/what-do-engine-sizes-actually-mean

Lesson Three: Plan

In the plan, you would normally describe how you intend to collect the data. In this standard, the data is already given to you but because you're a quality statistician, you want to check a few things rather than blindly accepting data given to you.

You want to check:

- How the variables are measured
- Where the data came from

How the variables are measured

Usually, this only involves stating the units. For example, "the weight of the car is measured in kilograms and the fuel efficiency is measured in kilometres per litre". But sometimes you'll need to elaborate on what the variable actually is. For example, BMI is a confusing term. You may find it helpful to research what it means, e.g.:

"BMI stands for Body Mass Index. It measures your weight and height and compares them to guess how much body fat you have."[6] It is measured in kilograms per metre squared[7], i.e. $\frac{kg}{m^2}$

Where the data comes from

This is known as the source of the data. Describe as much as you can about:

- Who collected the data?
- Why they collected it?
- When they collected it?
- Where the data applies to? [is the data all from one country?]
- How they collected it? [are there any obvious mistakes in their method?]

You will be given the information on the source of the data; you won't have to research it. It's almost certain that the source information you're given won't answer all of these questions, that's okay. Describe what you *have* been given.

Practice internals

Check how the variables are measured [and elaborate on them if needed] and check the source of the data for both practice internals. The dataset information can be found in appendix A.

[6] http://www.medicalnewstoday.com/info/obesity/what-is-bmi.php

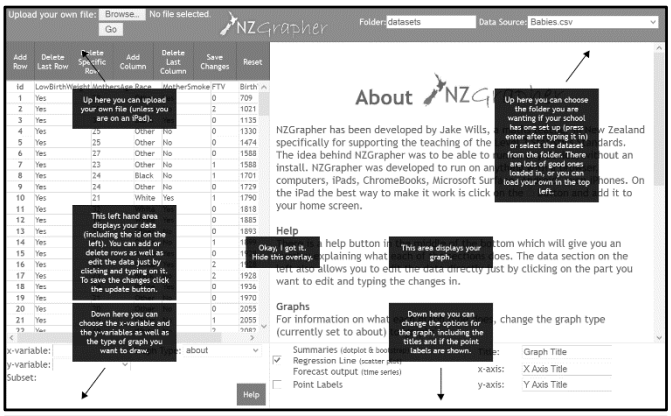
[7] http://www.physiologyweb.com/calculators/body_mass_index_calculator.html

Lesson Four: Data

You'll need to produce a scatter graph to answer your question [like in the sample question at the start of this workbook]. You could do it by hand as you did in lesson one. But to be honest, I cut that dataset down from 200 data points to 20. In other words, it will be much harder in reality so let a computer do it for you.

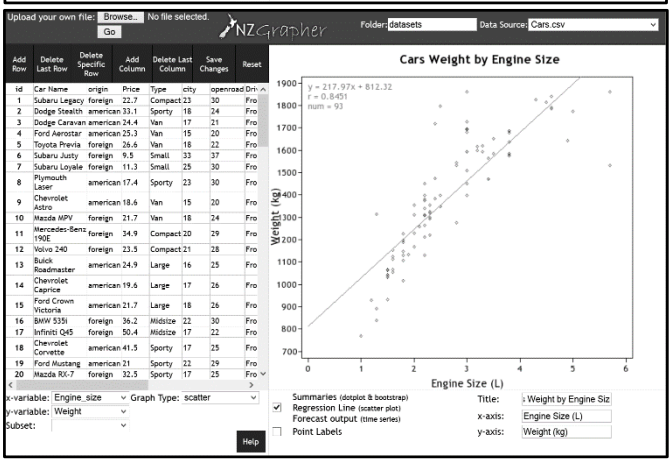
First up we need to start NZGrapher by going to www.jake4maths.com/grapher

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.



To draw a scatter plot there are three things you need to do.

1. Select the graph type… for this we want a scatter graph
2. Select variable 1… this is the x-variable, the explanatory or independent variable, in this case it's engine size.
3. Select variable 2… this is the y-variable, the response or dependent variable, in this case it's weight.



Check the graph title and axis labels to make sure they are appropriate (include units where necessary) and press update graph to save the changes. To add in the regression line press the 'Regression Line' check box.

To save the graph, right click on it and press 'Save Image As'.

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.

Practice internals

Recreate the scatter plot for both of the practice internals. You don't need to save the graphs, just show that you can do it. Check the grade requirements.

Lesson Five: Analysis: Features of the relationship

Here's where things get real. In the sample question about Kiwi birds' heights and weights, you were doing analysis. We'll do it more thoroughly now.

A large emphasis is placed on what you can see in the graph rather than relying on numbers like averages or the correlation coefficient[8]. Use your eyes to comment on:

- Trend
- Strength of the relationship
- Scatter
- Clusters
- Outliers

Trend

The trend describes the overall tendency, e.g. "Overall, the weight of a Kiwi bird seems to increase as the height increases". But we can be more specific than that. When you describe the trend, comment on:

- How well the data fits the trend (strongly, moderately, weakly, or something in between)
- If the trend goes up or down, i.e. a positive or negative relationship
- If the trend is linear or non-linear

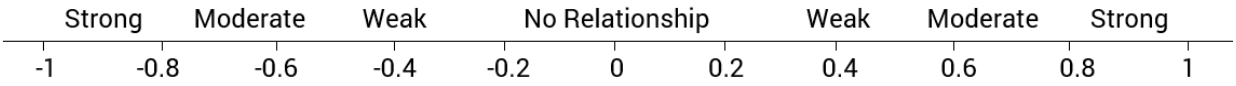
For example, "Overall, there is a moderate, positive, linear relationship between a Kiwi bird's weight and height."[9]

Or alternatively, "A Kiwi bird's height seems to be moderately related to its weight. As the weight increases, the height tends to increase as well. This trend is linear."

Strength of the relationship

Make a statement justifying why you think the trend is weak, moderate, or strong. Again, use your eyes. But you may also back up your judgement with the correlation coefficient.[10] For example "It is a strong relationship because the data is tightly packed around the regression line and has a correlation coefficient of 0.89; a high score."

As a guide to correlation coefficients, see the following diagram:



There is an activity that gives you practice at comparing r-values at http://guessthecorrelation.com/

For more information on correlation coefficients, see appendix B.

[8] You'll learn about this soon.

[9] If you make a statement like this, you must explain what you mean by a positive relationship, e.g. "The trend is positive because as the Kiwi bird's height increases, the weight tends to increase as well."

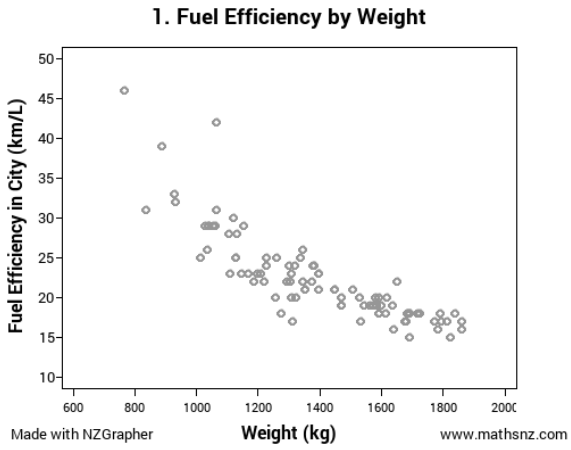
[10] Told you you'd learn about this soon.

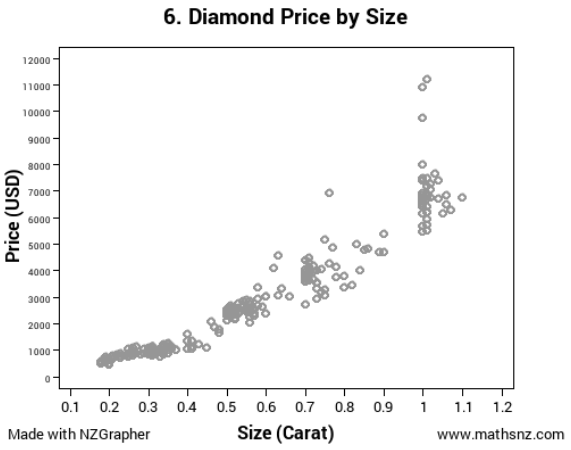
Scatter

The scatter refers to how spread the data is horizontally. In other words, it's referring to the spread of the independent/explanatory variable.[11]

Are there any ranges along the independent variable that are denser or sparser than others?

For example:





"There is a consistent spread of data from 1000kg to 1900kg. But there are only a few data points between 800kg and 1000kg. Presumably it's rare to have cars in this weight range."

"There is a reasonably consistent spread of data across the entire range of sizes from 0.2 carats to 1.1 carats. However, there are some clusters visible at certain sizes: 0.2-0.4, 0.5-0.6, 0.7, and 1.0 carats. There is an odd lack of data for sizes between 0.9 and 1.0 carats. Perhaps the manufacturers often round these sizes up to 1.0"

Clusters

AKA groups. These describe any tight bunches of data. It often makes sense to combine your description of any clusters with your description of the scatter as in the second example above. If there are no obvious clusters, say "there are no obvious clusters of data". Why do you think data is grouped around these values?

[11] The spread of the dependent/response variable is talked about as the strength of the relationship.

Outliers

Outliers are data that are significantly different from the rest of the data. They are not necessarily right or wrong but they require further scrutiny. If you suspect an outlier, examine the data in the table. Can you see any explanation for it being so different? It could be different for a number of reasons:

- This data point doesn't belong in the population, e.g. if you're researching the weight of Kiwi birds and weigh a Kākāpō accidentally[12] then you'd expect a significantly different weight.
- Magnitude error.[13] If the decimal point has been put in the wrong place. In this case, the data will be 10x, 100x, 1/10x, 1/100x the expected range of values, e.g. instead of 171cm for human height you might have 17.1cm or 1710cm.

You decide whether you can estimate what the value *should* have been and edit it or just to delete the outlier. Or perhaps it's not an outlier at all, it's just a really light Kiwi. Whatever you decide, justify your decision.

Deleting or editing a single outlier can make a great difference to your analysis, particularly your regression line and correlation coefficient. It would be sensible to compare these before and after changing your outlier to further justify your decision.

Practice internals

Analyse the two practice internals by referring to the features of the relationship: trend, strength of the relationship, scatter, clusters, and outliers.

[12] In which case you'd be a muppet.

[13] I just made this up.

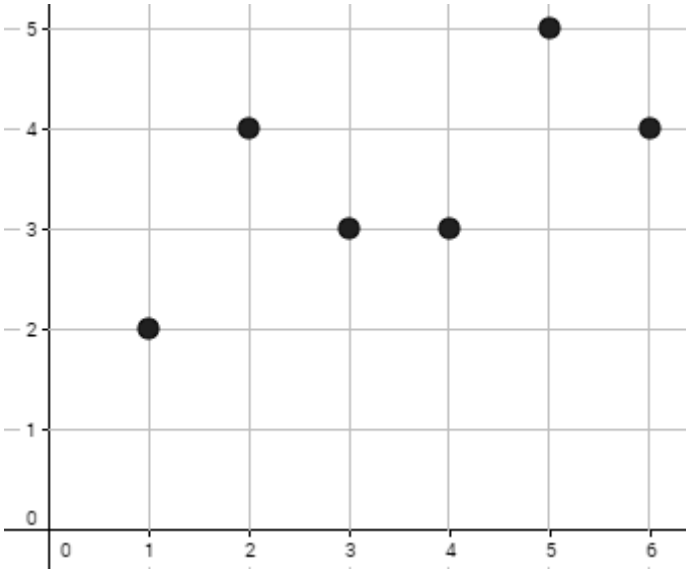
Lesson Six: Analysis: Regression

Regression [noun]: A return to a former or less developed state; going backwards. In statistics, it is a process that measures how well one variable affects another, i.e. is the height of the Kiwi causing it's weight or is it something else?[14] We'll look at:

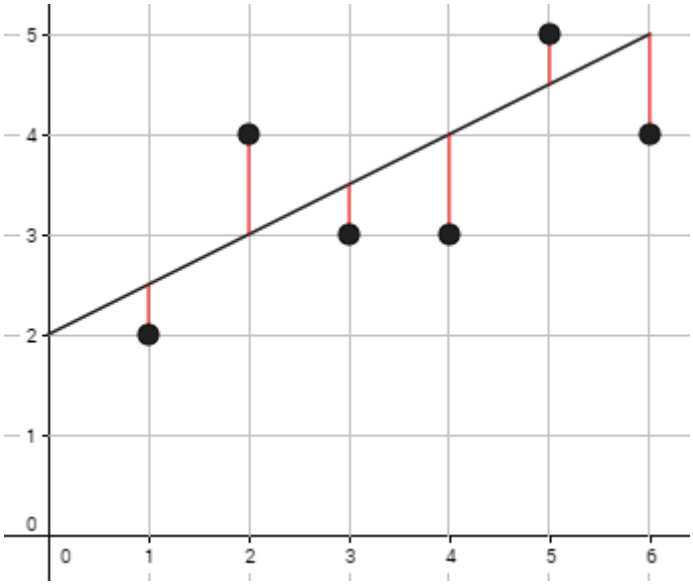
- The regression line
- Making a prediction
- Residuals

Linear regression

If we are convinced that our two variables are linked by a linear relationship, we can proceed to perform linear regression. Take an arbitrary dataset: {(1,2), (2,4), (3,3), (4,3), (5,5), (6,4)}. We can plot this on a scatter graph.



We are reasonably convinced that this relationship is a linear one so we proceed to perform a linear regression analysis. We begin by attempting to fit a line of best fit and calculating the residuals.



[14] Spoiler alert, it's both.

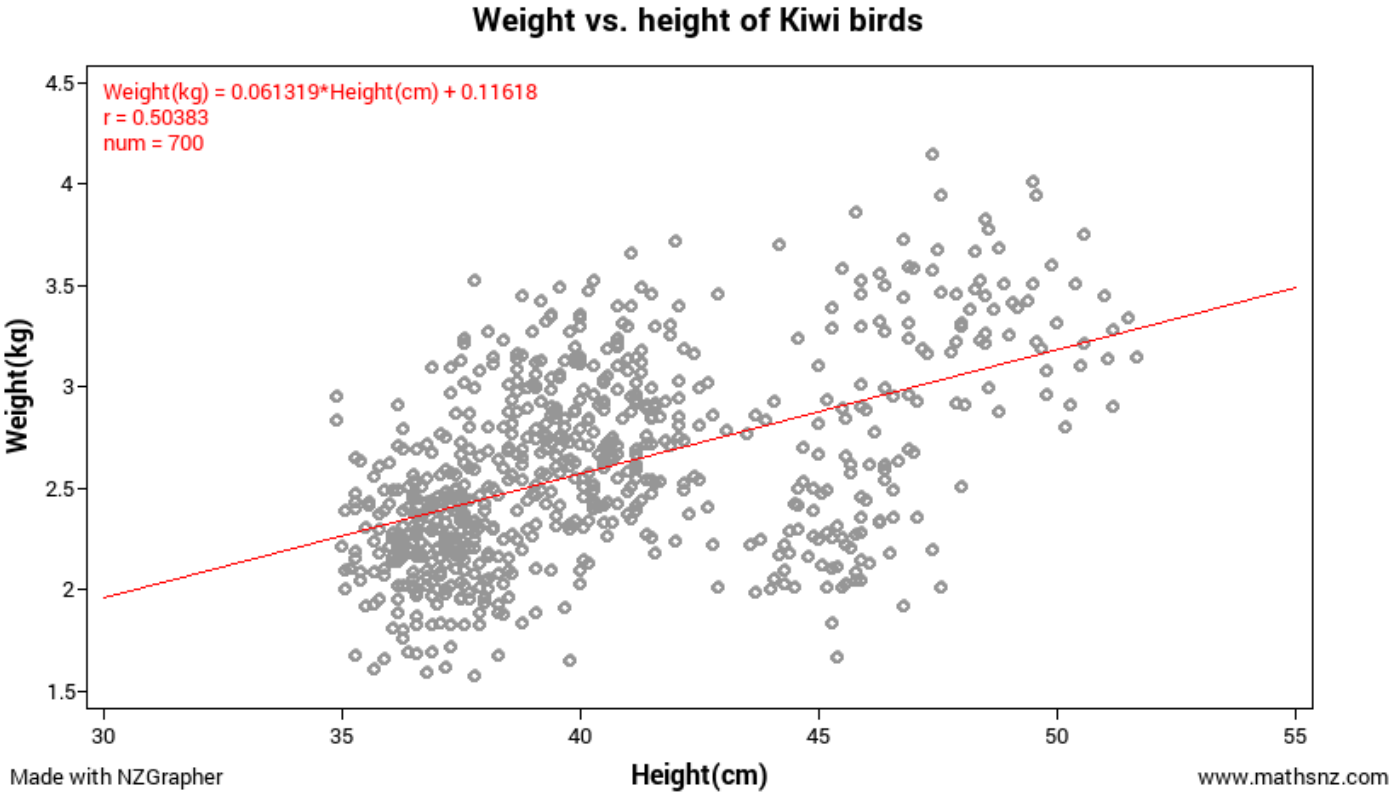
We want to minimise the sum of the squared residuals to get a *best line of best fit* which we call the "regression line". We can get a computer to do this for us. If you're curious as to how the regression line is calculated, see appendix C.

You need to interpret the regression line. In particular, the gradient of the line.

What do you think the gradient describes?

Bear in mind that the gradient of the regression line only gives the *average* increase, not a fixed amount of change in the response variable for every unit increase in the explanatory variable.

For example, "my regression line has the equation $weight(kg) = 0.061319 \times height(cm) + 0.11618$. This means that for every centimetre increase in height, the weight of the kiwi bird increases on average by 0.061kg or 61g."[15]



[15] Note that this doesn't refer to Kiwi birds growing up from babies to adults. It is saying that if one Kiwi is 1cm taller than another, it will be 61g heavier on average.

Predictions

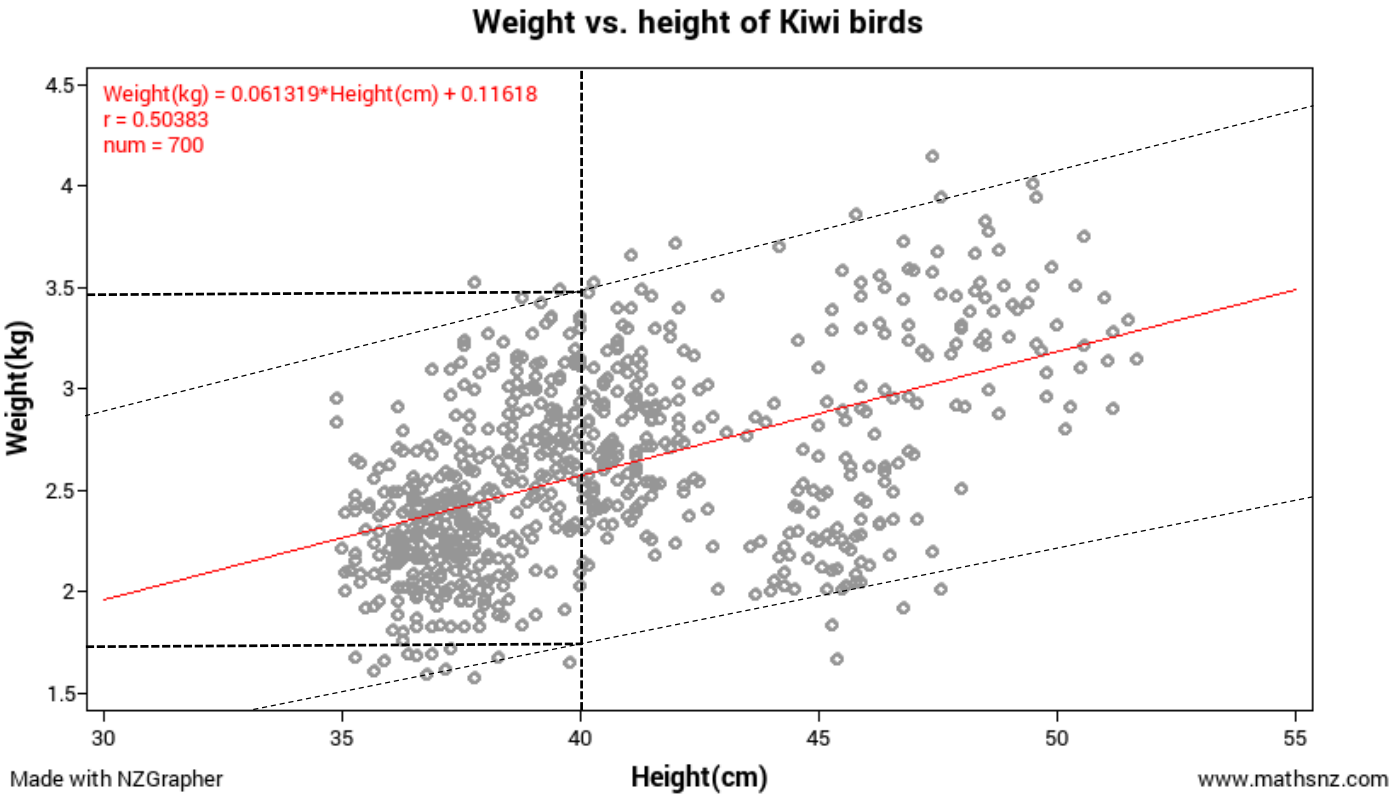
You need to use your regression line to make a prediction of your dependent variable for an arbitrary value of your independent variable, e.g. given a height of 40cm, how heavy do you predict Kiwis to weigh? Your regression line tells you:

$$weight(kg) = 0.061319 \times height(cm) + 0.11618$$
$$weight(kg) = 0.061319 \times 40 + 0.11618$$
$$weight(kg) = 2.568kg$$[16]

We can confirm this value by looking at the graph and regression line. Notice that there are many Kiwis that are 40cm tall and *don't* weigh 2.6kg. Recall that this is only an average.

We can only predict data values inside our data range. In this example, between 35cm and 52cm. The reason for this is the regression line is only valid for the range of x values we used to calculate it.

I suggest you use the graph to comment on the validity of your prediction. Make two lines showing the bounds where most[17] of the data is found and check the two values of the dependent variable at the same point as your prediction. This gives you an upper and lower bound on your prediction.



I could rewrite my prediction as follows: "My regression line predicts a weight of 2.6kg for a Kiwi 40cm tall. But I can see that there is a large spread of weights for Kiwi's weighing 40cm and their weight could realistically be anywhere from 1.75kg and 3.5kg. Given than all of the data collected is between 1.5kg and 4kg, this is not a great prediction."[18]

---

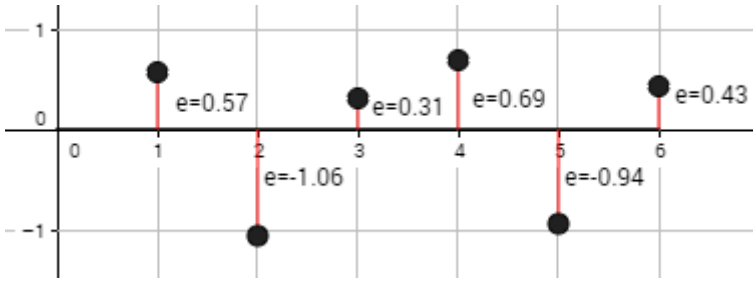
[16] I have rounded this to match the level of accuracy given in the data. Always round in this way.

[17] How you define 'most' is completely up to you.

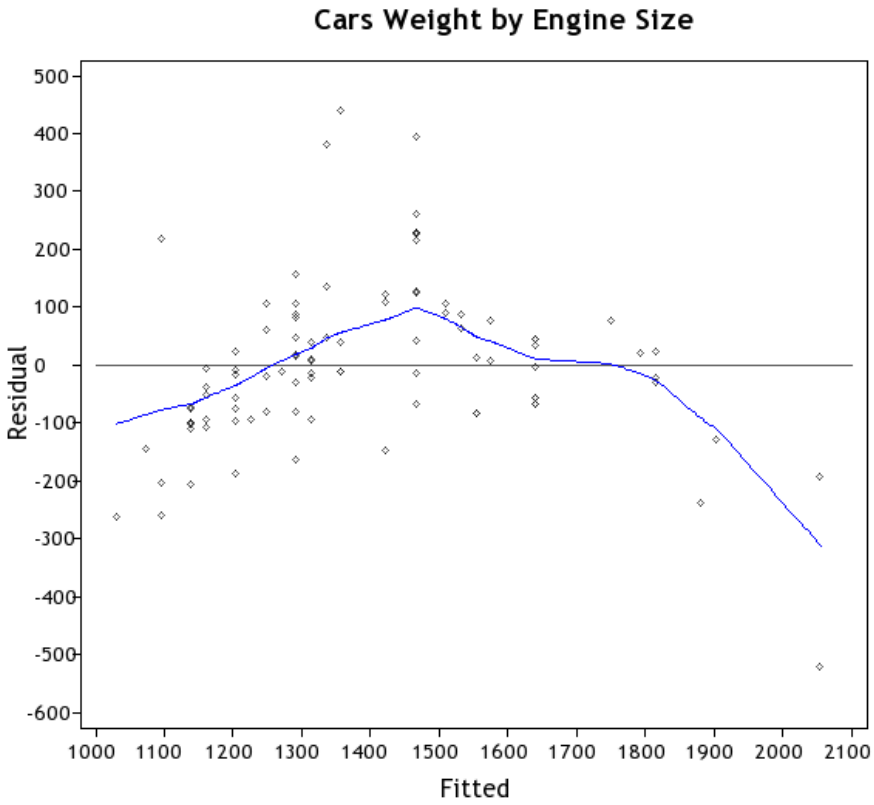
[18] This can be used as evidence for a weak relationship between height and weight. This backs up the evidence of a low correlation coefficient (0.5)

Residuals

You've calculated residuals by hand. You could get a computer to do it for you. This is a great tool to test if a linear model is appropriate or if a non-linear model would be better. Here's a residuals plot for the arbitrary dataset from page 15.

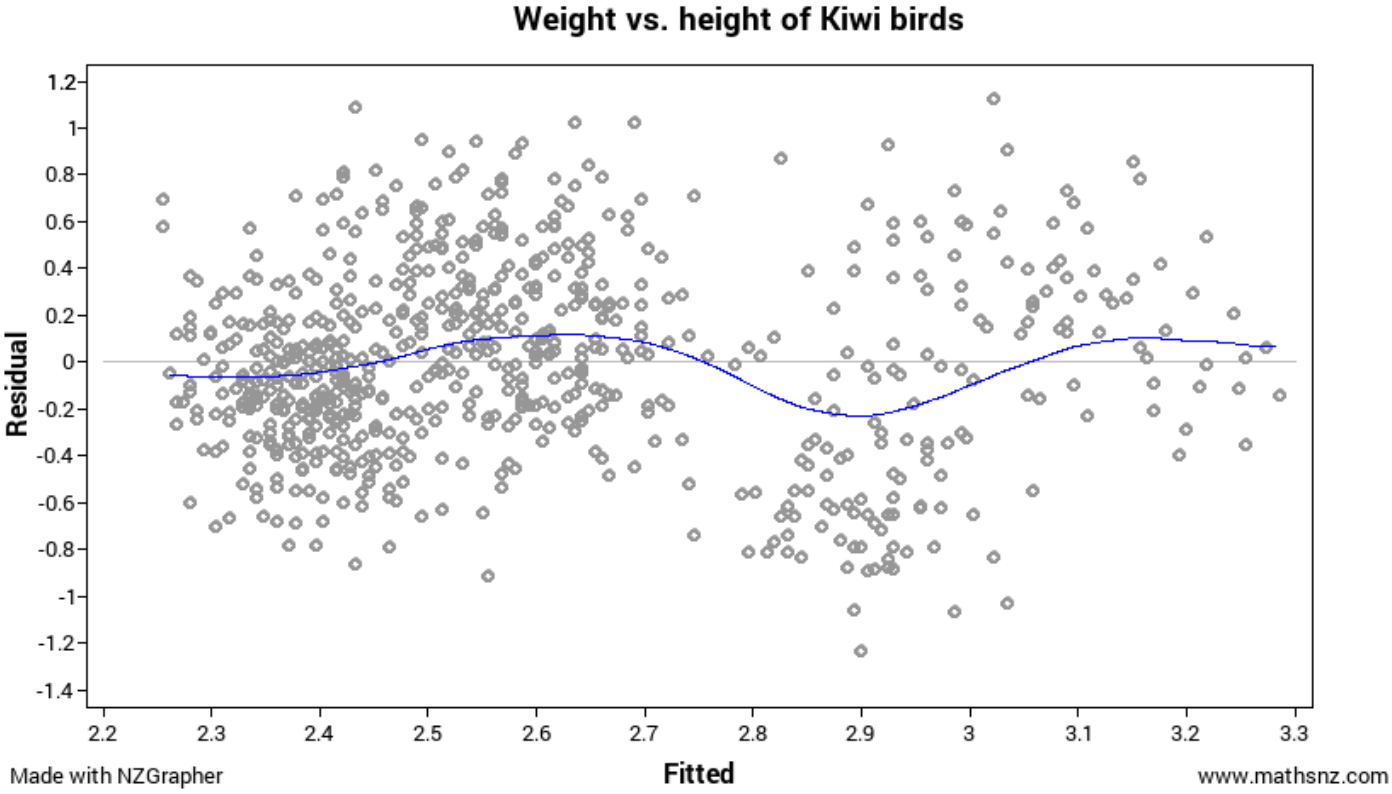


In other words, lie the trend line flat on the x-axis and show the residual values as they should be above or below this flat trend line. In this case, the residuals are seemingly randomly spread above and below the trend line which is a good justification for a linear model. If there was a curve, you would have reason to suspect the relationship was non-linear and therefore a non-linear model would be more appropriate.



You could also use your residuals plot to fine-tune your predication. For example, you can see that where the regression line predicts a weight of 2.6kg, the average of the observed data was around 0.1kg higher. Similarly, if we had predicted a weight of a Kiwi bird to be 2.9kg [e.g. if they had a height of 46cm] you can see that the observed data was much lighter than that, on average it was around 0.2kg lighter.

We can use this to fine-tune our prediction of 2.6kg at 40cm height to 2.7kgkg at 40cm height.



Practice internals
Perform a linear regression analysis on both practice internals. You may choose how in-depth of an analysis you do.

Lesson Seven: Conclusion

You now have a wealth of evidence about your dataset. In your conclusion, summarise the key information and use that to answer your question: "Is there are relationship between…" Don't give a yes or no answer, rather, describe the kind of relationship using the key evidence you've found.

Be aware of opportunities to give evidence as justification for your judgements and to show statistical or contextual insight[19].

Consider what evidence you'd use for the Kiwi heights vs. weights dataset and what judgements you'd make about it. Then proceed to the practice internals.

Practice internals
Conclude the two practice internals.

[19] Contextual insight = prior research

Appendix A: Data Set Information

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

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-North Island 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	SF-South Fiordland	E-East North Island
	CW-Central Westland	N-Northland	W-West North Island
	StI-Stewart Island	EC-Eastern Canterbury	NF-North Fiordland

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.

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.

Appendix B: Calculating the correlation coefficient (r)

The correlation coefficient is actually the "Pearson product-moment correlation coefficient".[20] It has the following formula:

$$r = \frac{S_{xy}}{S_{xx}S_{yy}} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Yes, it's long. But it helps if you complete the following table. Let's take the same arbitrary dataset as on page 15: {(1,2), (2,4), (3,3), (4,3), (5,5), (6,4)}.

	x	y	xy	x^2	y^2
	1	2			
	2	4			
	3	3			
	4	3			
	5	5			
	6	4			
$\sum$					

We can now substitute the appropriate values into the formula.

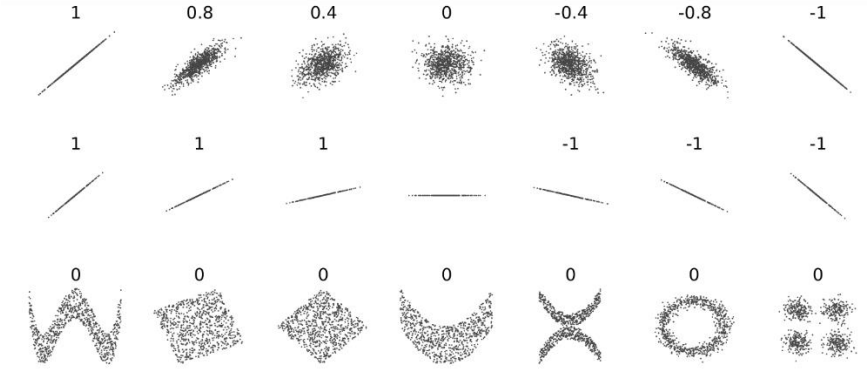
$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{6 \times 80 - 21 \times 21}{\sqrt{[6 \times 91 - 21^2] \times [6 \times 79 - 21^2]}}$$

$$r = \frac{39}{\sqrt{3465}}$$

$$r = 0.6625\ (4d.p.)$$

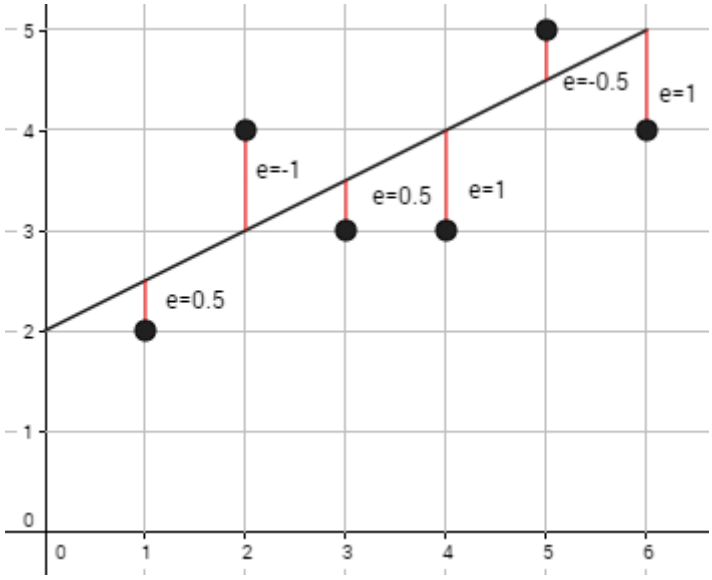
Below are graphs showing the different correlation coefficients (r values) above each one.



[20] https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

Appendix C: Linear regression

In lesson six, I gave you an arbitrary dataset and I guessed where the best line of best fit might be. You calculated the sum of the squared residuals as $\sum e^2 = 3.75$



We could try guessing other lines and calculating their sums of the squared residuals but we're mathematicians. We can do better.

Remind yourself of the equation of lines. There are several different options to choose from, e.g.

$$y = mx + c$$
$$ax + by + c = 0$$
$$y = a + bx$$

We'll use the last one. Remind yourself what a and b mean in this equation. When we have calculated a and b, we have our best line of best fit. We start by calculating b using the following formula:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

You may recall S_{xy} and S_{xx} from appendix B. You need to calculate the following values:

	x	y	xy	x^2
	1	2		
	2	4		
	3	3		
	4	3		
	5	5		
	6	4		
$\sum$				

Now you can substitute your values into the equation:

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$b = \frac{80 - \frac{21 \times 21}{6}}{91 - \frac{21^2}{6}}$$

$$b = \frac{\frac{13}{2}}{\frac{35}{2}}$$

$$b = \frac{13}{35}$$

To calculate a, we use the following formula:

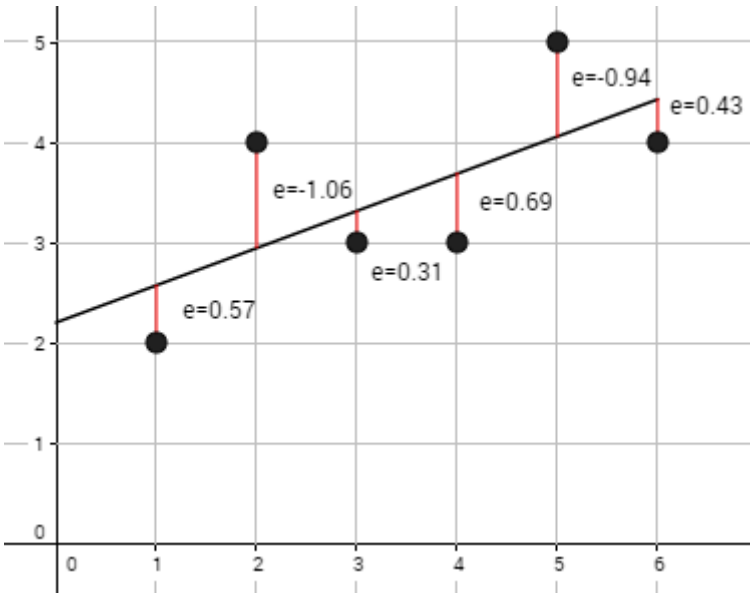
$$a = \bar{y} - b\bar{x}$$

where $\bar{y}$ is the mean of the y-values and $\bar{x}$ is the mean of the x-values

Calculate $\bar{x}$ and $\bar{y}$ and use them to calculate a:

$$a = \bar{y} - b\bar{x}$$
$$a = 3.5 - \frac{13}{35} \times 3.5$$
$$a = 2.2$$

Therefore our best line of best fit is $y = \frac{13}{35}x + 2.2$ and can be shown like this:



Calculate the sum of the squared residuals, knowing that this is the minimum value. Note that the sum of the residuals is 0. This is always true for a least-squares regression line.

Appendix D: Correlation and causation

Correlation is not causation.

There, I said it. Once more for effect:

Correlation is not causation.

What does that mean? I notice that between December and Feburary, spending on raincoats hits rock-bottom while spending on ice-cream goes through the roof. Therefore people aren't buying raincoats because they're spending all of their money on ice-cream.

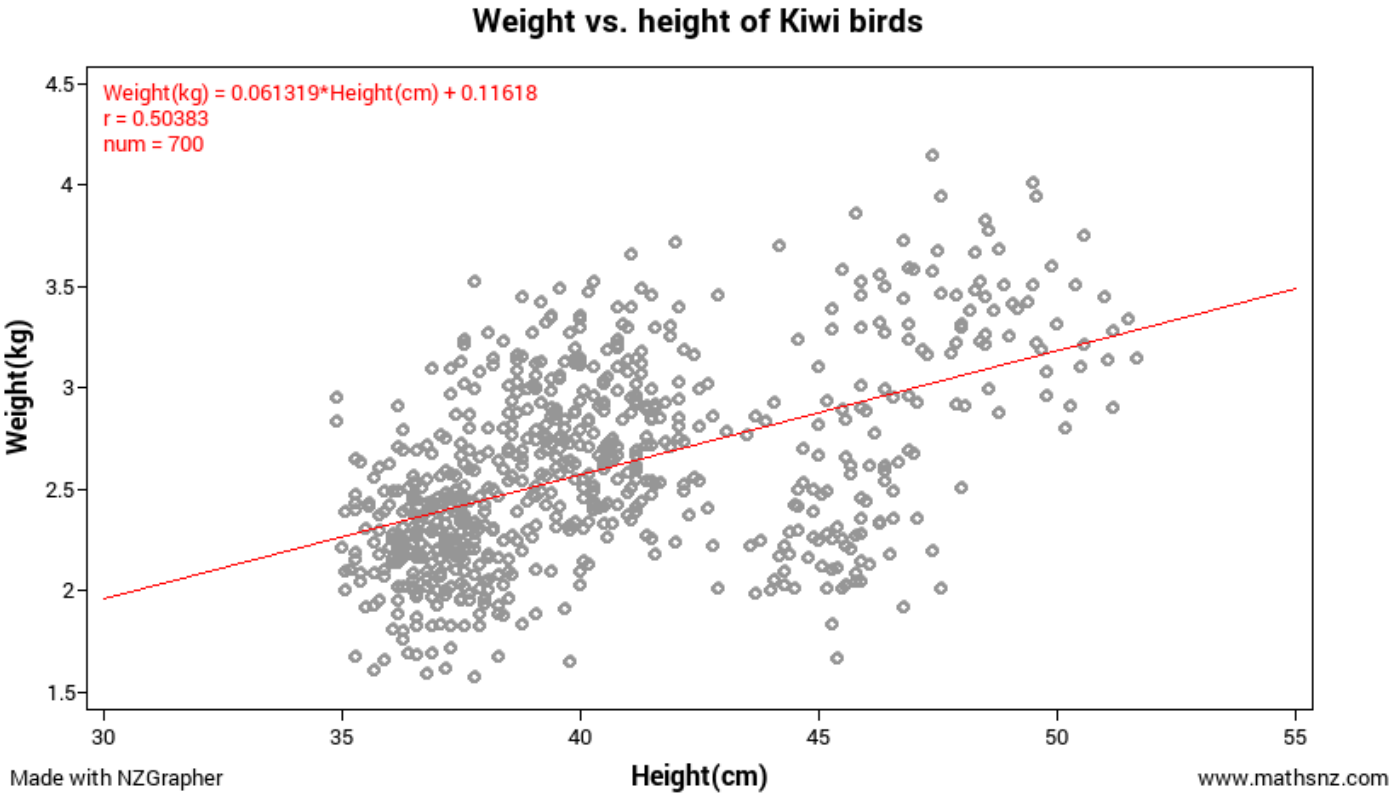
These two variables are correlated [literally: co-related] but one doesn't cause the other. There is what we call a lurking variable.

Lurking variable [noun]: A variable that is neither the explanatory nor response variable but has an effect on one or both of them.

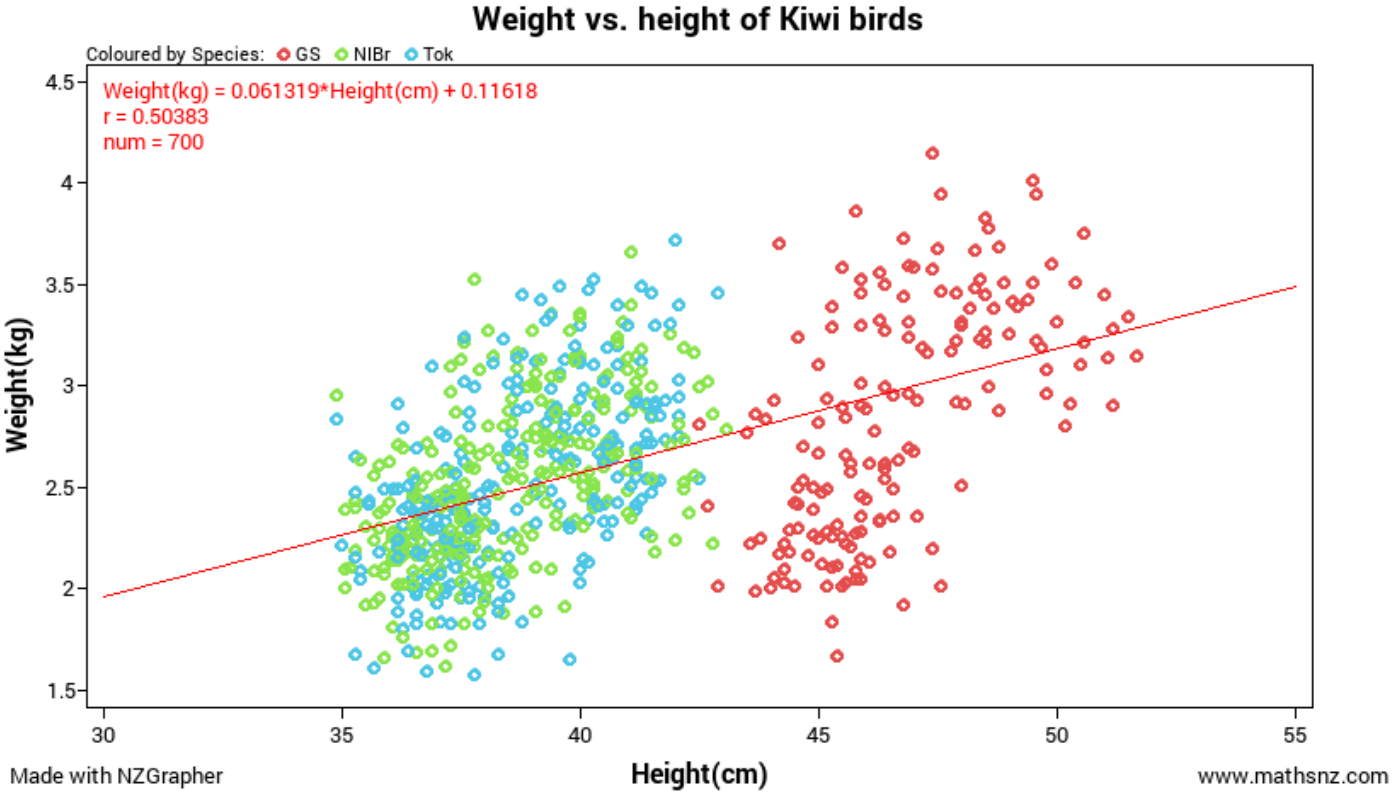
Causal [not casual] relationships can be determined but only through randomised, controlled experiments. Surveys or samples are not these.

Appendix E: Separating Variables

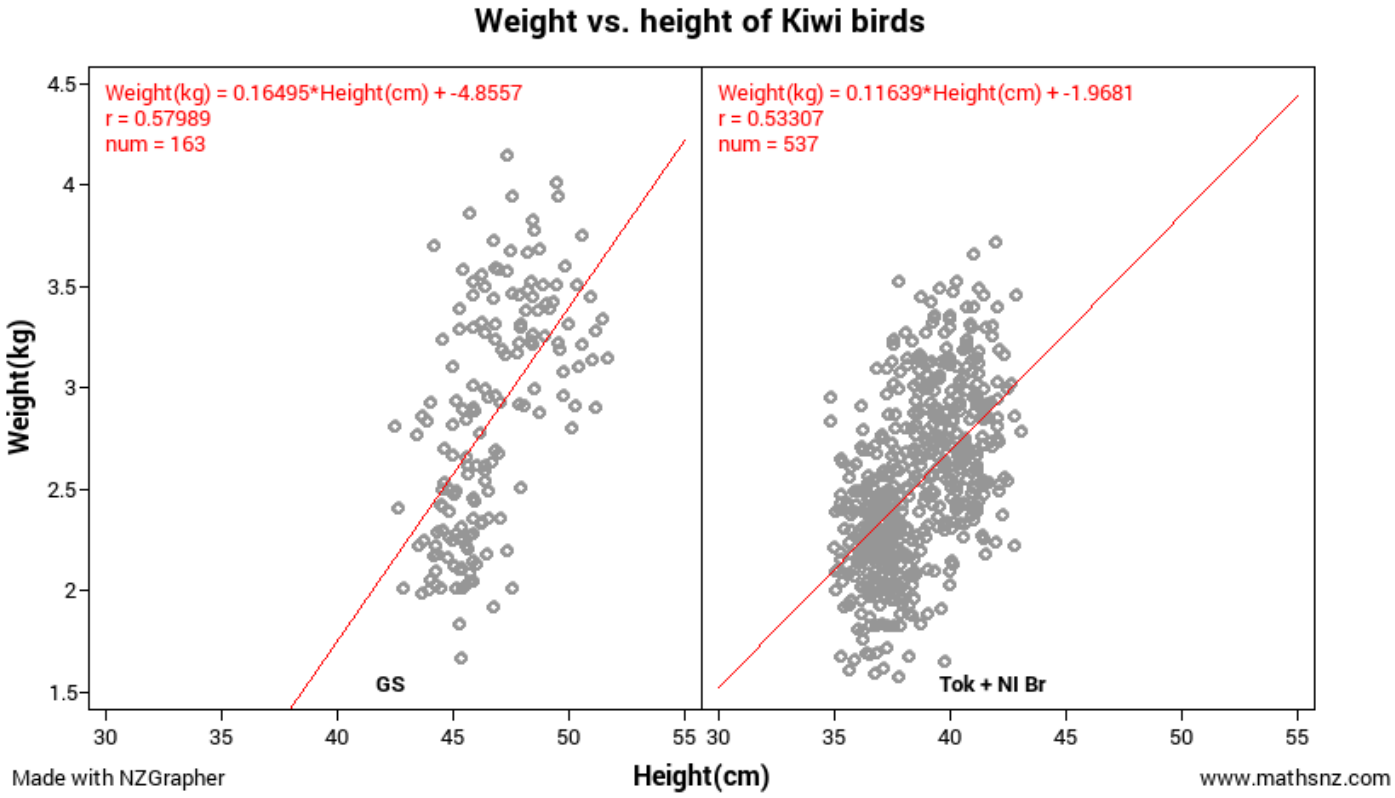
You may suspect that a lurking variable is affecting your response variable. Let's look again at the Kiwi dataset.



There seems to be two groups [or clusters] of data. Let's see if we can colour the data by any other variables. Gender and species are good candidates. I'll colour by species first.



It's clear that Great Spotted Kiwi are almost always bigger than the North Island Brown and the Southern Tokoeka Kiwi. Therefore it makes sense to compare them separately. I could subset by species but you may also choose to have the North Island Brown and Southern Tokoeka together and the Great Spotted kiwi separate. To do the first option in NZGrapher, define variable 3 as species. To do the second option, download the dataset, make the changes in Excel and upload it back to NZGrapher. This is a graph of the second option.



Notice that the correlation coefficients have improved and the graphs look nicer.

Try doing the same thing for gender.

Appendix F: Non-linear regression

There are real-world situations which are non-linear so it would be inappropriate to model them with a linear model. If you suspect you are dealing with a non-linear model, justify your judgement[21] then choose one or several non-linear models. You should compare them to determine which one best fits your data and makes most sense in the context.

For non-linear regression, there is no equivalent of the correlation coefficient, you have to rely on your eyes.

In NZGrapher, you have the following non-linear options:

- Quadratic $y = ax^2 + bx + c$
- Cubic $y = ax^3 + bx^2 + cx + d$
- Exponential $y = Ae^{bx}$
- Logarithmic $y = A\ln(x) + b$
- Power $y = Ax^b$

To help make a comparison, you should understand the behaviour of these functions, e.g. asymptotes, turning points, what happens for very large or very small values of x, etc.

Image credits

Lesson Two: Problem

PPDAC cycle	http://new.censusatschool.org.nz/resource/data-detective-poster/

Lesson Five: Analysis

Relationship strength number line	http://www.mathsnz.com/resources/files/3.9/3.9%20Booklet.pdf

Appendix B: Examples of Correlation Coefficient (r)

Graph diagrams	https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/2000px-Correlation_examples2.svg.png

[21] Presumably by looking at the scatter graph backed up by a residuals plot.