Topic materials:
Dr Raquel Iniesta
Department of Biostatistics and Health Informatics

Narration and contribution:
Zahra Abdula

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Institute of Psychiatry, Psychology and Neuroscience

KING'S College LONDON

Module Title: Introduction to Statistics

Session Title: Scatter Plots

Topic title: Correlation and Linear Regression

Learning Outcomes

- Understand use of scatterplots to investigate associations between two continuous variables.
- Be able to create a scatter plot using statistical software.

Previously on ‘Introduction to Statistics’

Based on the **type** of data, we use different statistical tests for hypothesis testing.

Hypothesis testing	**means**	**proportions**
	Approximately normal (symmetrical data)	**χ^2 assumptions hold**
one group versus a pre-defined value	one sample t-test	one sample χ^2-test
one group versus another group	two independent samples t-test	Pearson’s χ^2-test
one group (twice or) *versus* another matched group	two paired samples t-test	McNemar test

Grouping Variables

Based on the **type** of data, we use different statistical tests for hypothesis testing

Hypotheses testing	Means	Proportions
one group *versus* another group	two independent samples t-test	(two independent samples) Pearson’s χ^2-test
one group (twice or) *versus* another matched group	two paired samples t-test	(two dependent samples) McNemar test

Group: is a binary variable, i.e. a categorical variable with two categories. E.g. Gender ‘female’ and ‘male’. We were partitioning a whole sample in groups based on the levels of a binary variable, producing two samples

Grouping Variables

Based on the **type** of data, we use different statistical tests for hypothesis testing

Hypotheses testing	Means	Proportions
one group *versus* another group	two independent samples t-test	(two independent samples) Pearson's χ^2-test
one group (twice or) *versus* another matched group	two paired samples t-test	(two dependent samples) McNemar test

When we are comparing **continuous variables**, for normally distributed variables we compute means (SDs) and compare them between two groups. E.g. Age

Grouping Variables

Based on the **type** of data, we use different statistical tests for hypothesis testing

Hypotheses testing	Means	Proportions
one group *versus* another group	two independent samples t-test	(two independent samples) Pearson's χ^2-test
one group (twice or) *versus* another matched group	two paired samples t-test	(two dependent samples) McNemar test

When we are comparing **categorical variables**, we compute proportions and compare them between two groups. E.g. Smoking 'yes' or 'no'

Types of Variable

Type of **Outcome** variable

Type of **Predictor** variable

Type of variable	Continuous	Categorical
Categorical	two independent samples t-test	(two independent samples) Pearson's χ^2-test
Categorical	two paired samples t-test	(two dependent samples) McNemar test
Continuous	Correlation & Linear Regression	?

We **do not partition the sample**. We have two continuous variables, measured in every individual. We seek to understand if the variables are related, and what this relationship is.

Scatterplots

Scatterplots are:

- A method of displaying a relationship between two variables (x and y) observed over a number of instances.
- Obtained by plotting points defined by (x, y) pairs (conventionally, **x** represented on the **horizontal** axis, **y** on the **vertical** axis).

Scatterplots are used to:

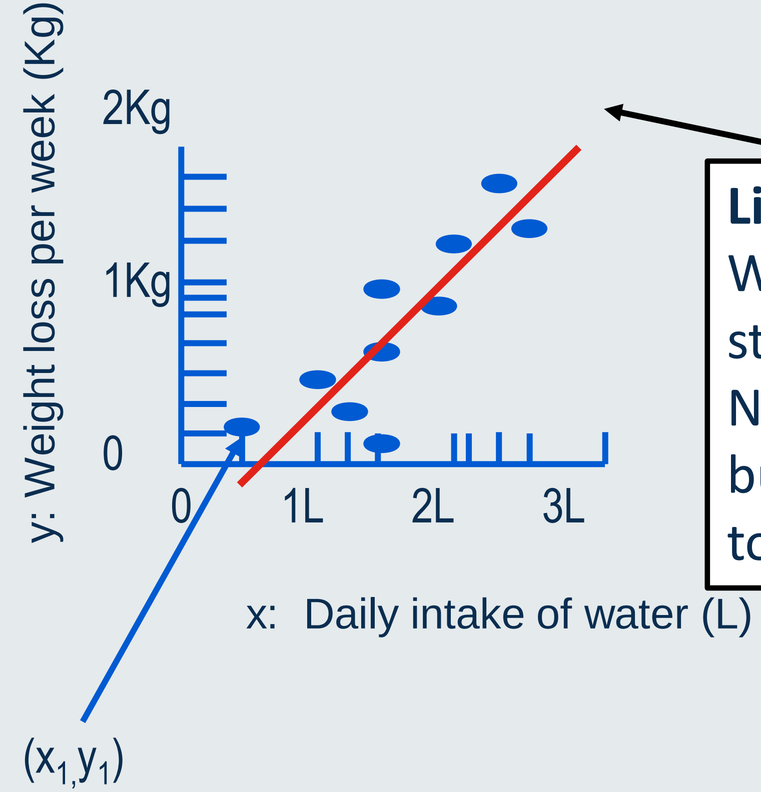
- Investigate an empirical relationship between **x** (the **independent**) and **y** (**dependent** variable).
- Attempt to predict y from x

Example

Let’s imagine we collect data for 10 people to study the Hypothesis ‘The higher the intake of water, the higher the weight loss’.

How do you think a plot of the data approximately would look like?

	x	y
$(x_{1,}y_1)$	0.5	0.10
$(x_{2,}y_2)$	1.0	0.30
$(x_{3,}y_3)$	1.2	0.40
	…	…



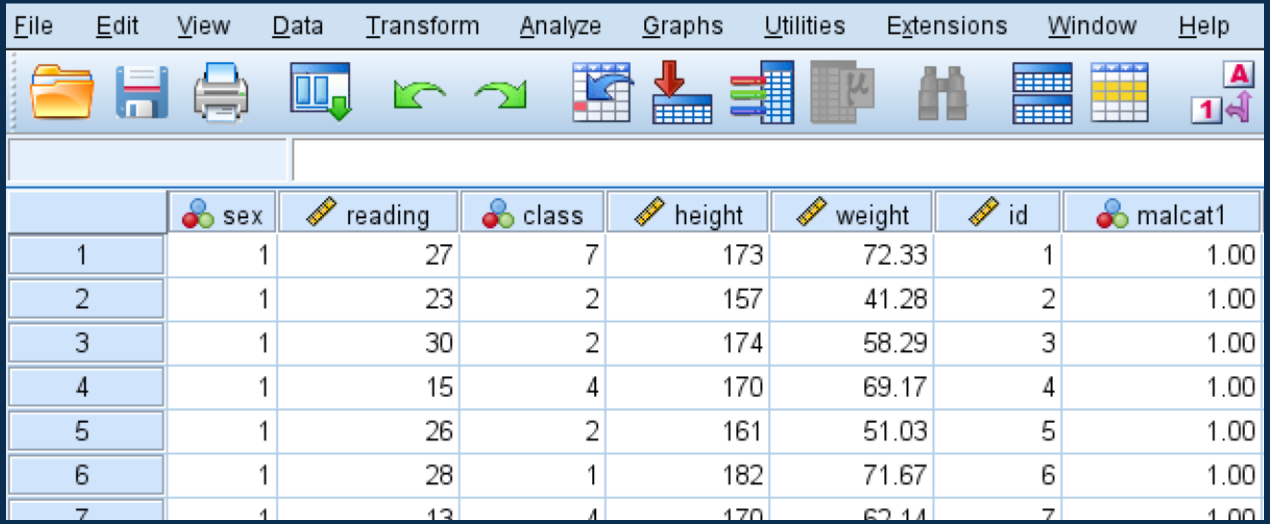
Linear relationship:
We can draw a straight line.
Not perfect fit,
but the line is “close” to the points

Scatterplots

- Plotting the data is essential to understand and visually assess the relationship between pairs of continuous variables
- The plot of data points (x,y) with x and y being continuous is called a **scatterplot**
- Most statistical software is able to generate scatter plots

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_6a_data.sav**



	sex	reading	class	height	weight	id	malcat1
1	1	27	7	173	72.33	1	1.00
2	1	23	2	157	41.28	2	1.00
3	1	30	2	174	58.29	3	1.00
4	1	15	4	170	69.17	4	1.00
5	1	26	2	161	51.03	5	1.00
6	1	28	1	182	71.67	6	1.00
7	1	13	4	170	62.14	7	1.00

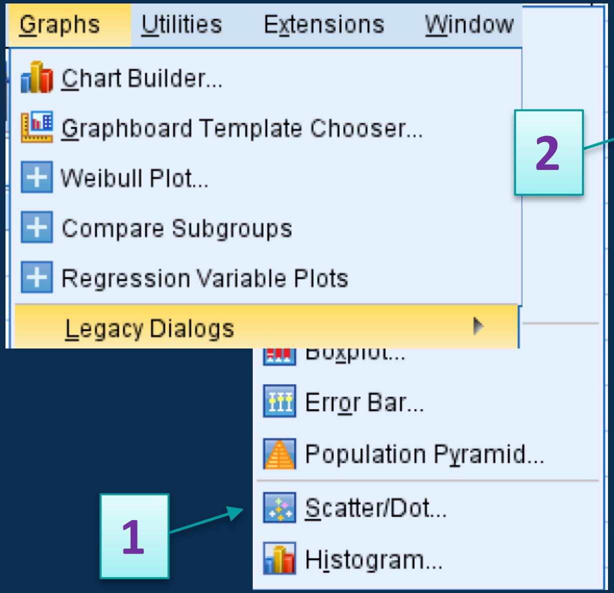
The dataset contains data from 1000 individuals, from the National Child Development Study (NCDS) with respect to their

- **sex**: gender of child (1=male, 2=female)
- **height**: height in cm at age 16
- **weight**: weight in kg at age 16
- **reading**: reading score
- **malcat1**: incidence of malaise at 22 years (0=yes, 1 = No)

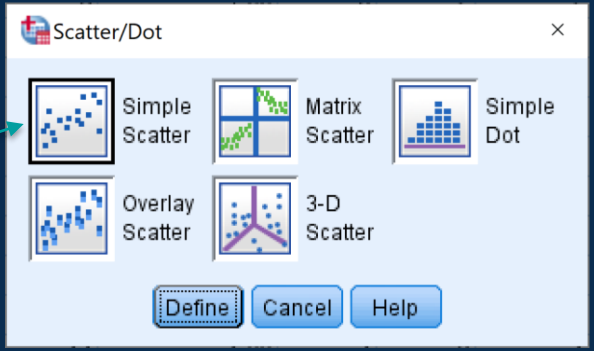
SPSS Slide: ‘how to’

According to the researchers, in the population from which our data came, they believe there is a relationship between weight and height of the 16 year old children.

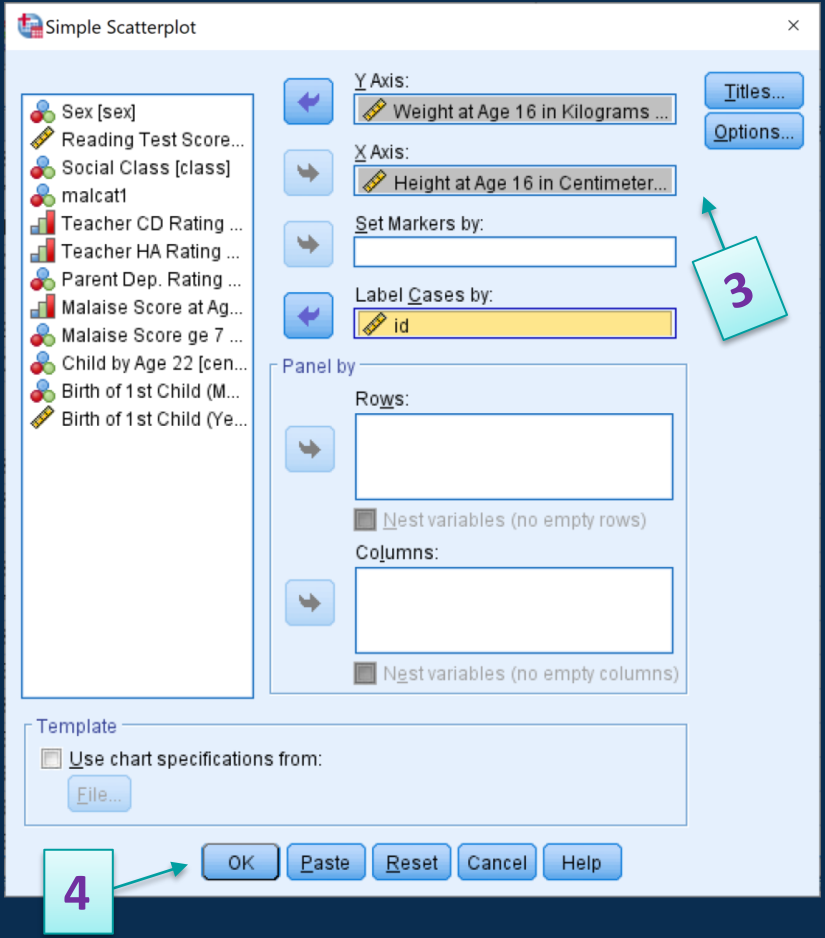
Step 1: Generate a Scatter Plot for variables ‘height’ and ‘weight’ from the data



Use ‘Graphs’ -> ‘Legacy Dialogs’ -> ‘Scatter/Dot’



Click on ‘simple scatter’
Click on ‘define’.
Add the dependent variable (weight) into the ‘y-axis’ box.
Add the independent variable (height) into the ‘x-axis’ box.
‘Label cases by’ ID.
Click ‘Ok’


Simple Scatterplot

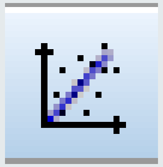
Y Axis: Weight at Age 16 in Kilograms ...

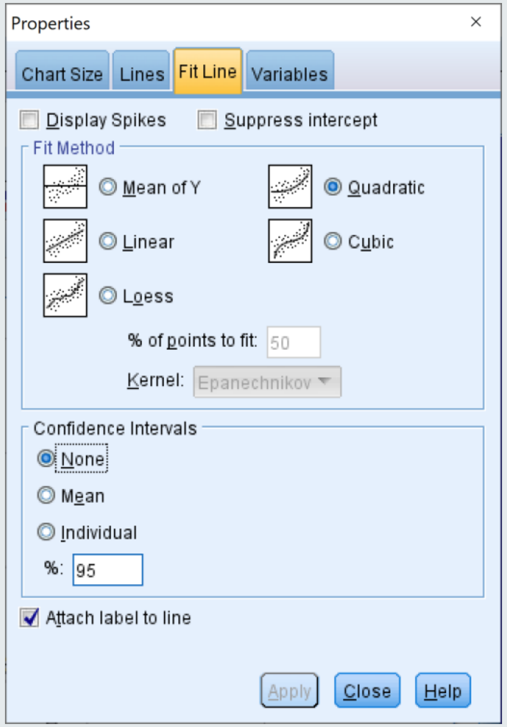
X Axis: Height at Age 16 in Centimeter...

Set Markers by:

Label Cases by: id

Output & Interpretation Slide





Properties

Chart Size | Lines | Fit Line | Variables

□ Display Spikes □ Suppress intercept

Fit Method

○ Mean of Y ● Quadratic

○ Linear ○ Cubic

○ Loess

% of points to fit: 50

Kernel: Epanechnikov

Confidence Intervals

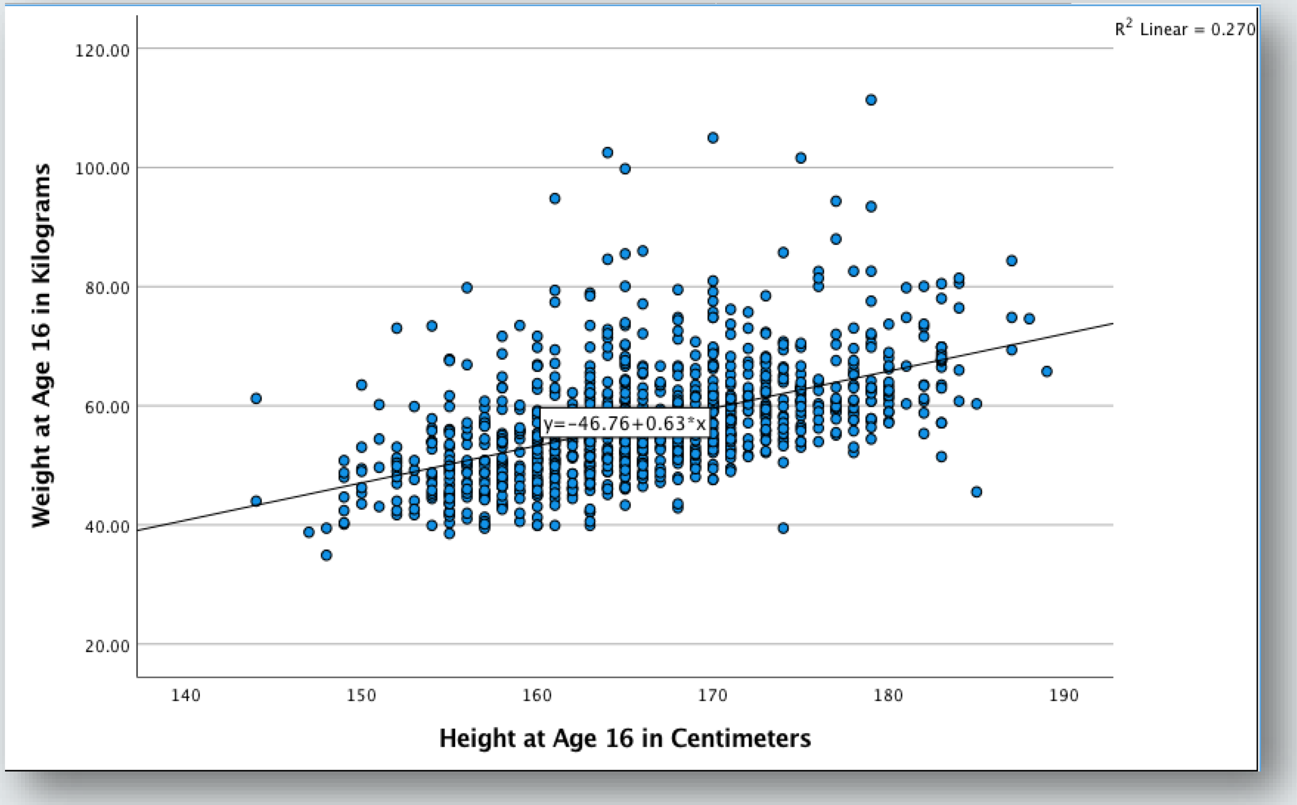
● None

○ Mean

○ Individual

%: 95

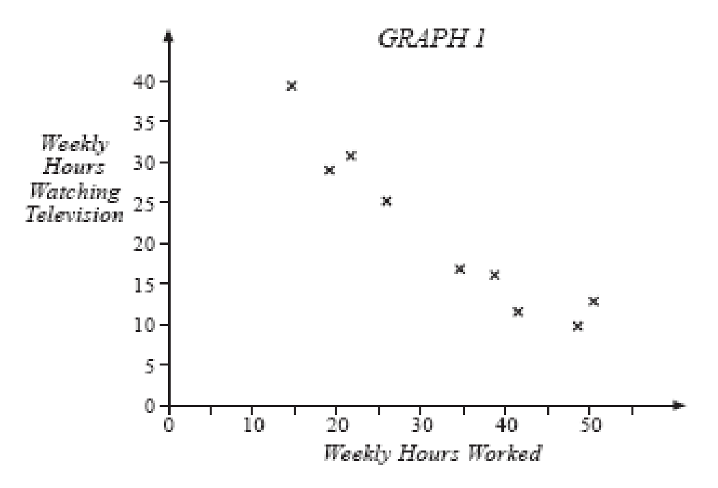
☑ Attach label to line

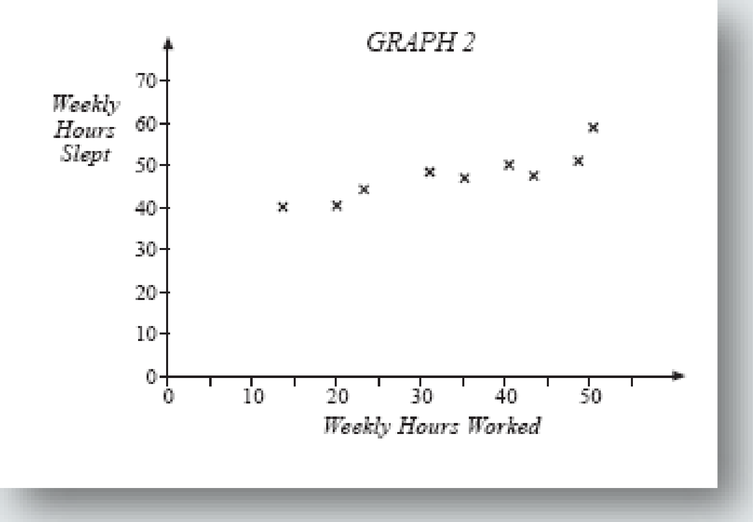


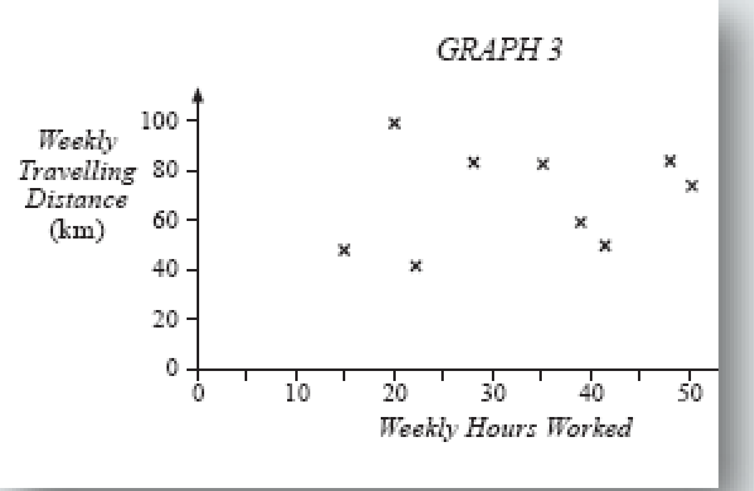
The scatterplot shows a positive linear trend between height of 16 year olds and their weight. As the height of the child increases, so does the weight.

A line of best fit can be added to the figure by double clicking on the graph, clicking on the icon shown and choosing ‘Linear fit line’ Also note that the line is an approximation of the points cloud, but it does not fit the cloud “perfectly”.

Knowledge Test





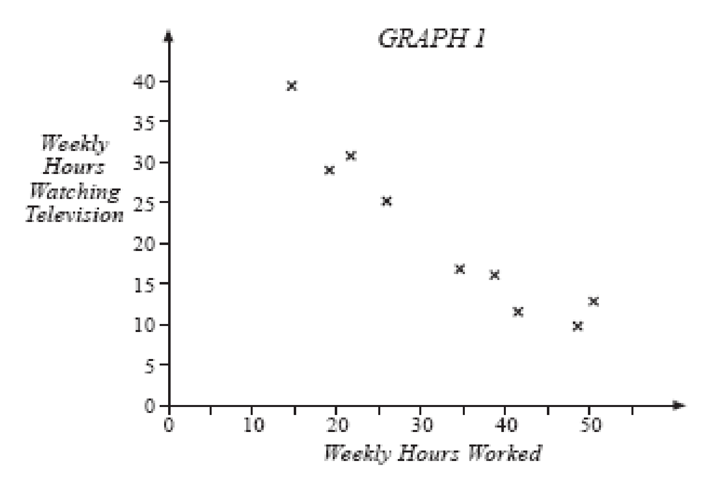


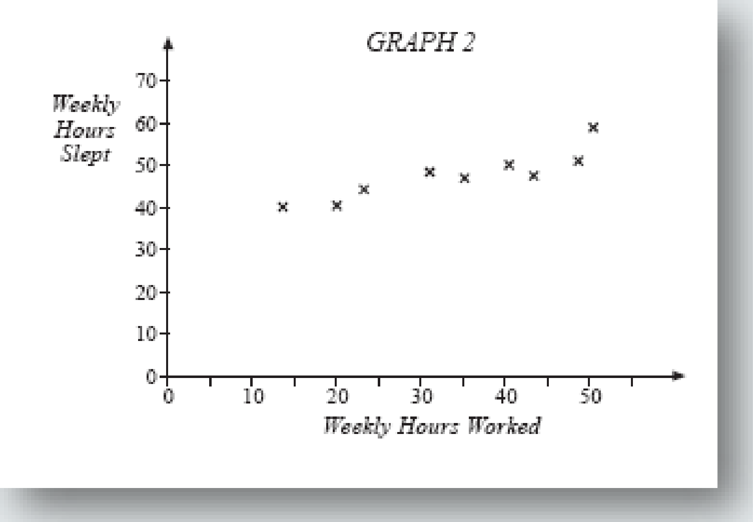
a) What does Graph 1 show about the relationship between the weekly hours spent watching television and the weekly hours worked?

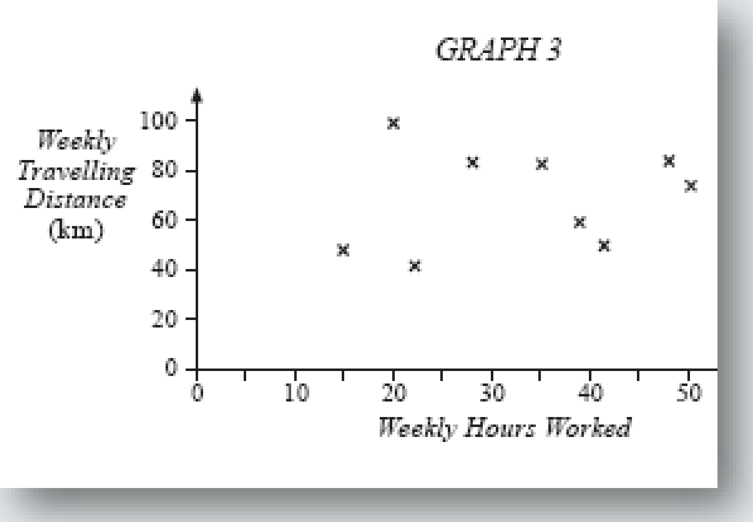
b) What does Graph 2 show about the relationship between the weekly hours slept and the weekly hours worked?

c) What does Graph 3 show about the relationship between the weekly travelling distance and the weekly hours worked?

Knowledge Test Solutions







a) As the weekly hours worked increases the hours watching television decreases. Showing a negative linear relationship between hours worked and hours watching TV

b) As weekly hours worked increases the hours spent sleeping marginally increases, showing a positive linear relationship between hours worked and hours slept

c) There appears to be no linear trend between hours worked and the weekly travel distance.

Reflection

Reflecting on your own field of study.

Write down an example from your research where it would be appropriate to investigate if there is a linear relationship between two continuous variables.

Reference List

- Agresti, A., & Finlay, B. (2009). Statistical Methods for the Social Sciences (4th ed., pp. 255-300) New Jersey, NJ: Pearson Hall.
- Field, A. (2005). Discovering Statistics using SPSS (2nd ed., pp. 116-204). London, England: Sage.

Thank you

Contact details/for more information:

Zahra Abdulla

Department of Biostatistics and Health Informatics (BHI)

IoPPN

+44 (0)20 7848 0847

Zahra.abdulla@kcl.ac.uk

www.kcl.ac.uk/xxxx