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 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	?

Type of **Predictor** variable

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:

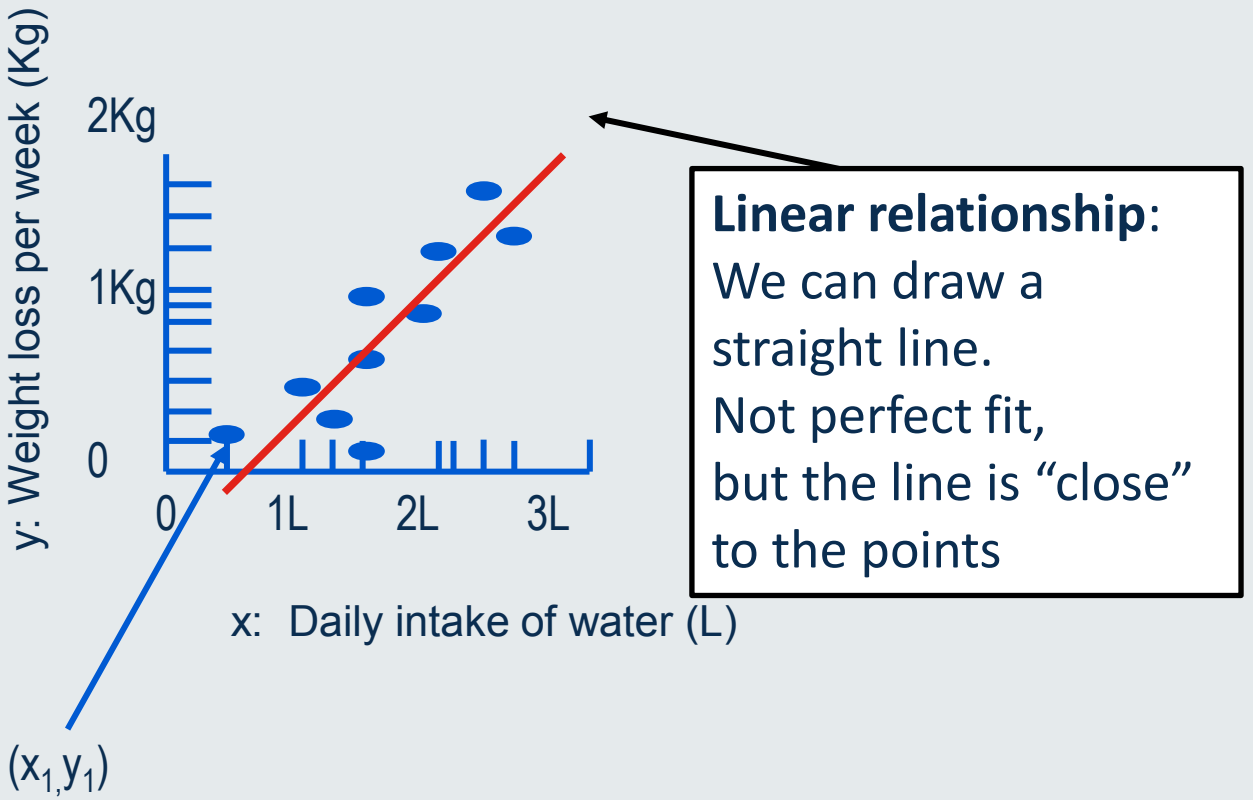
- Investigate an empirical relationship between **x** (the **independent**) and **y** (**dependent** variable).
- Attempt to predict y from x

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
	...	...



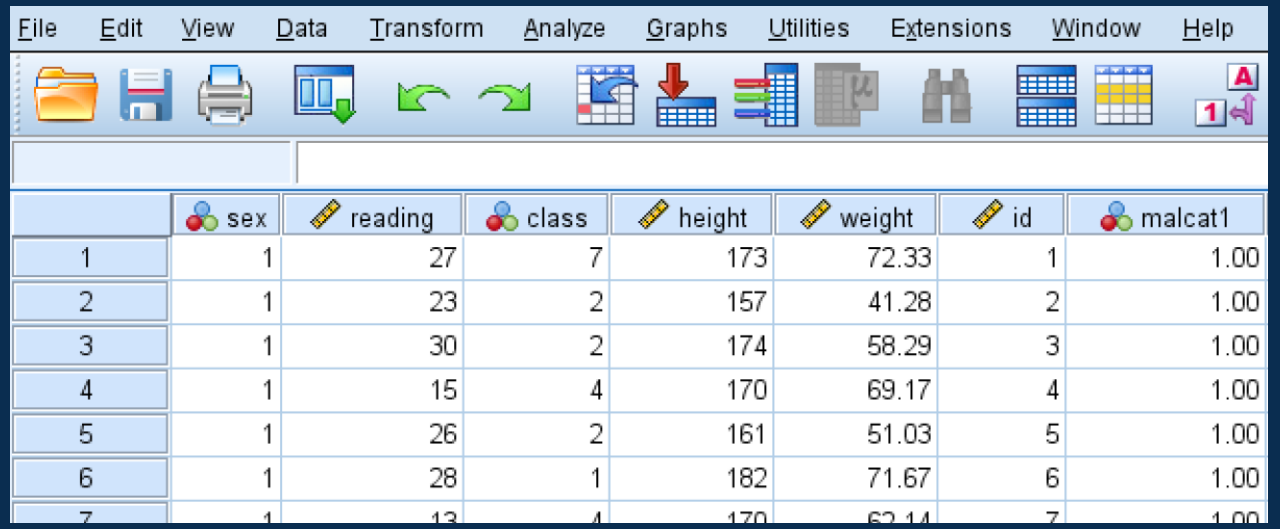
Linear relationship:
We can draw a straight line.
Not perfect fit,
but the line is "close" to the points

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

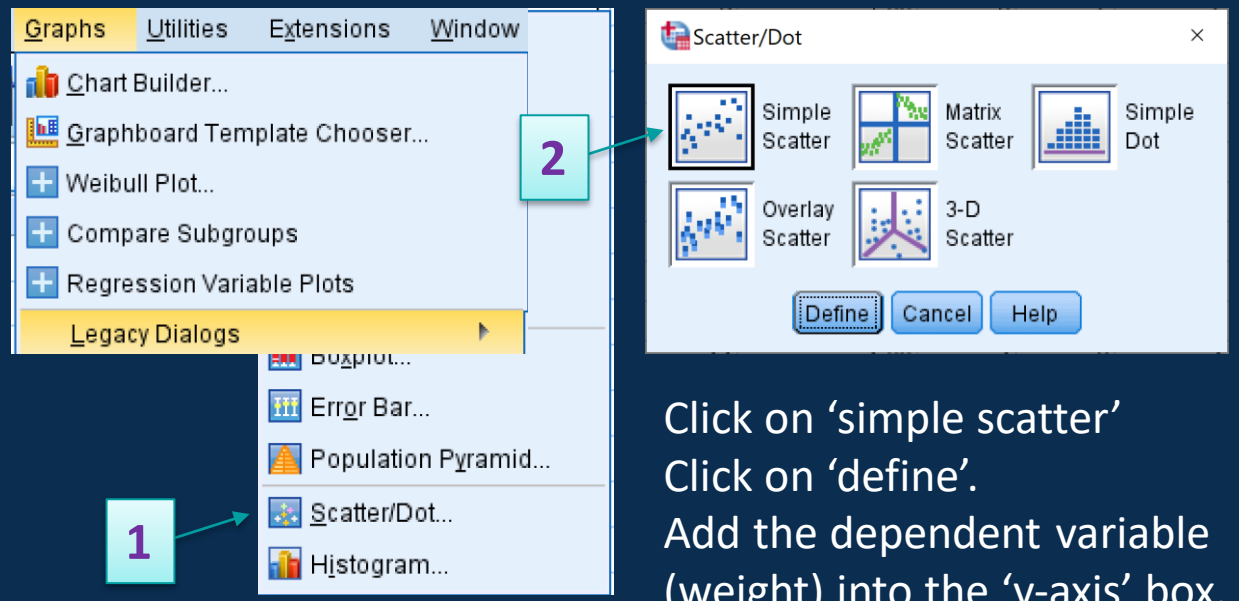
The dataset contains data from 1000 individuals, from the National Child Development Study (NCDS) with respect to their

- **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’

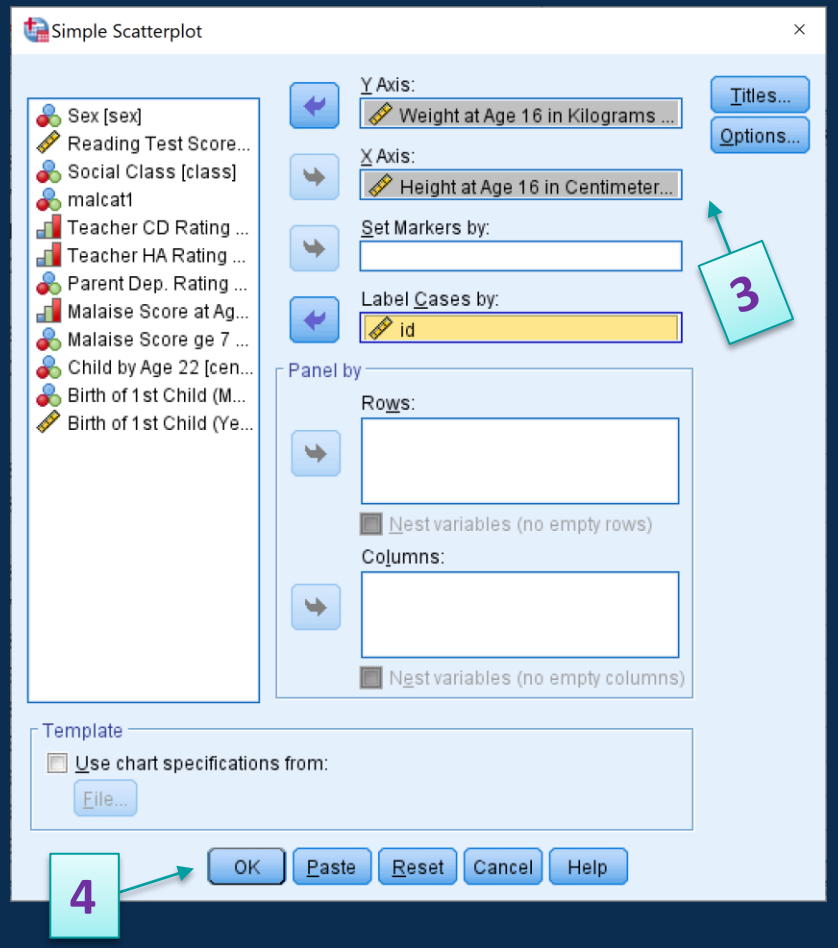
According to the researchers, in the population from which our data came, they believe there is a relationship between weight and height of the 16 year old children.

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

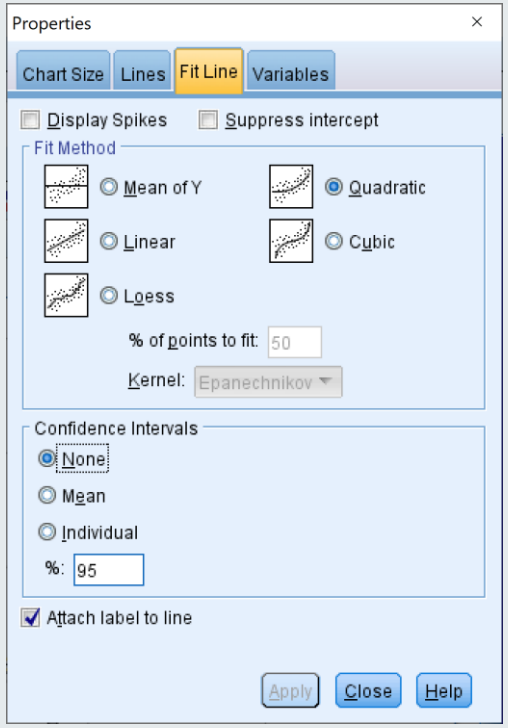


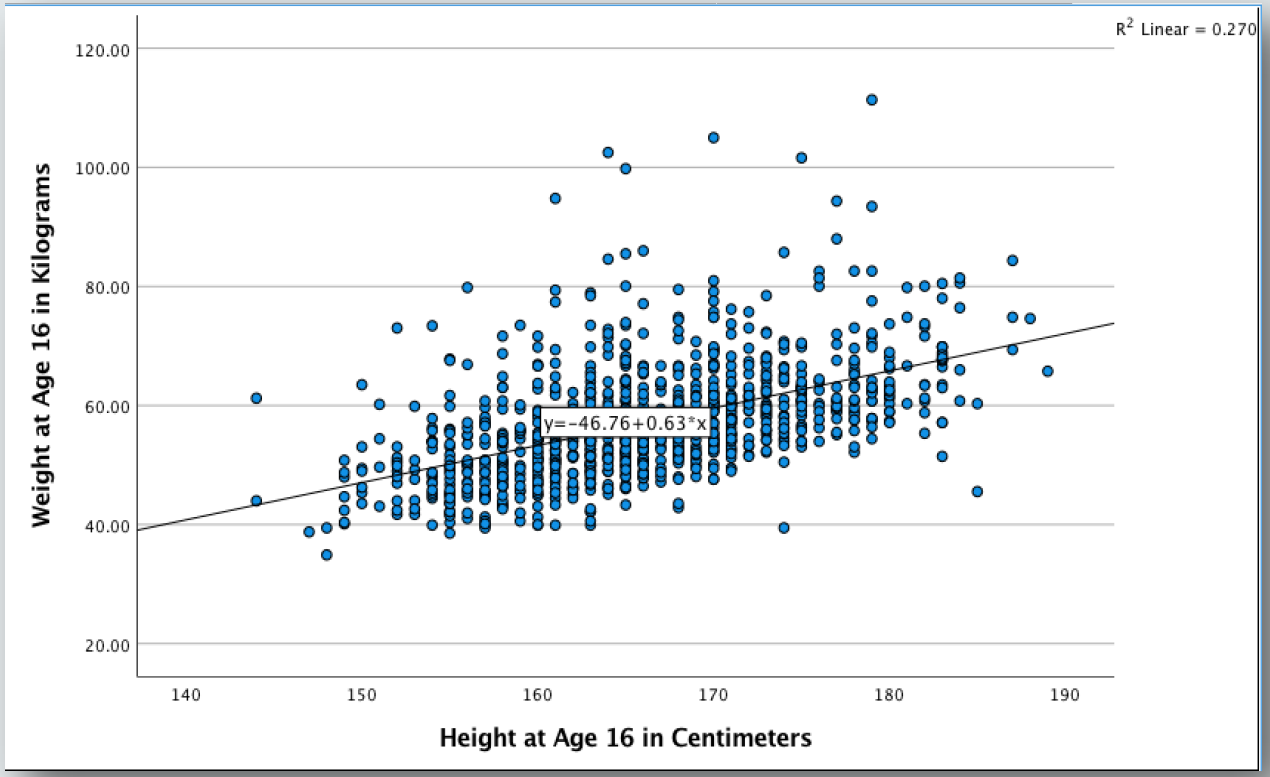
Use ‘Graphs’ -> ‘Legacy Dialogs’ -> ‘Scatter/Dot’

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’



Output & Interpretation Slide

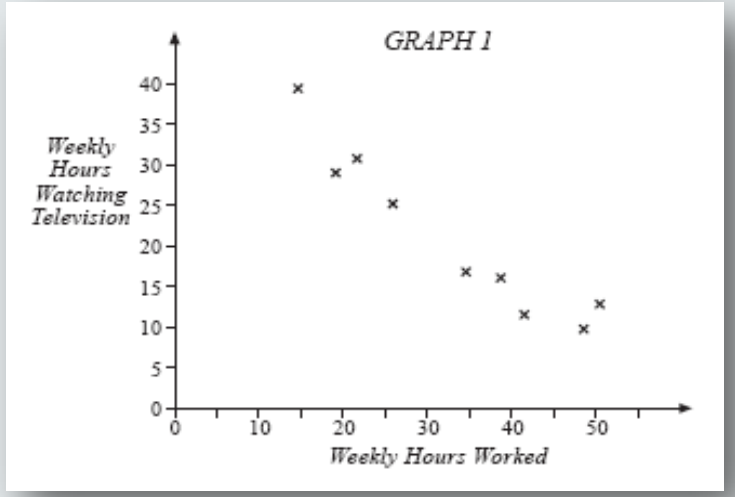


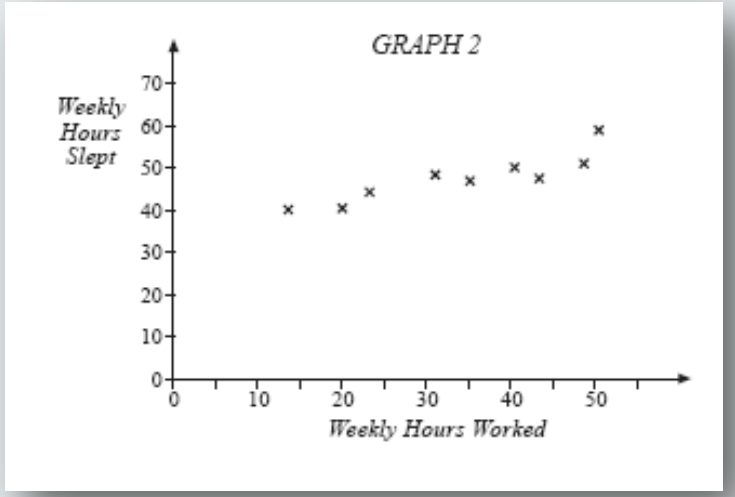


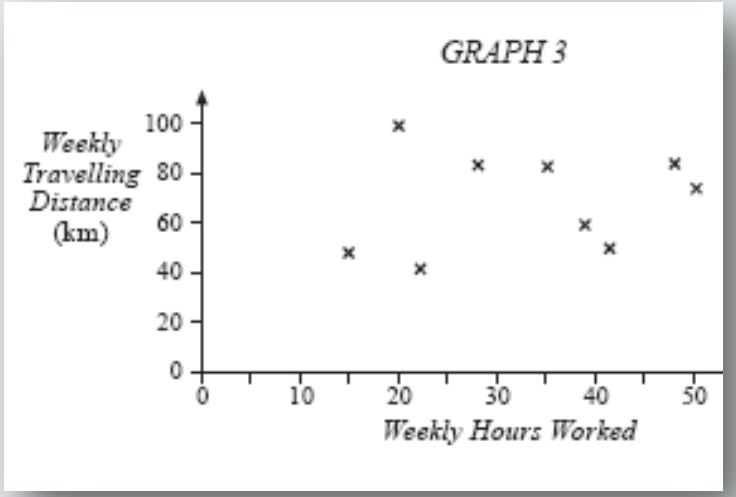
The scatterplot shows a positive linear trend between height of 16 year olds and their weight. As the height of the child increases, so does the weight.

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





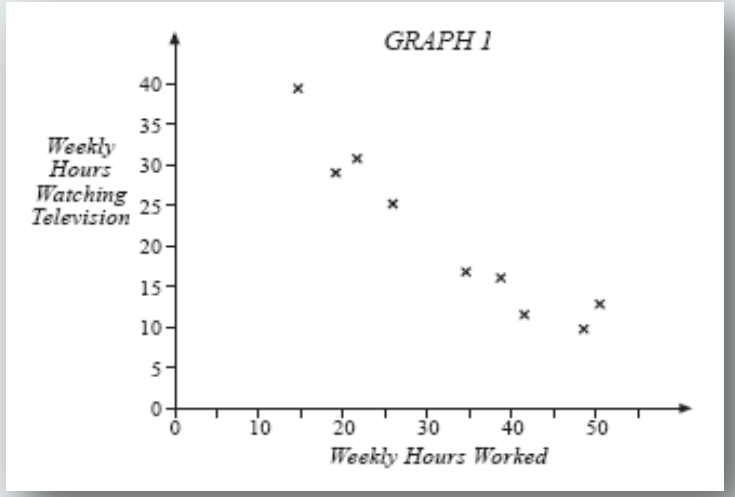


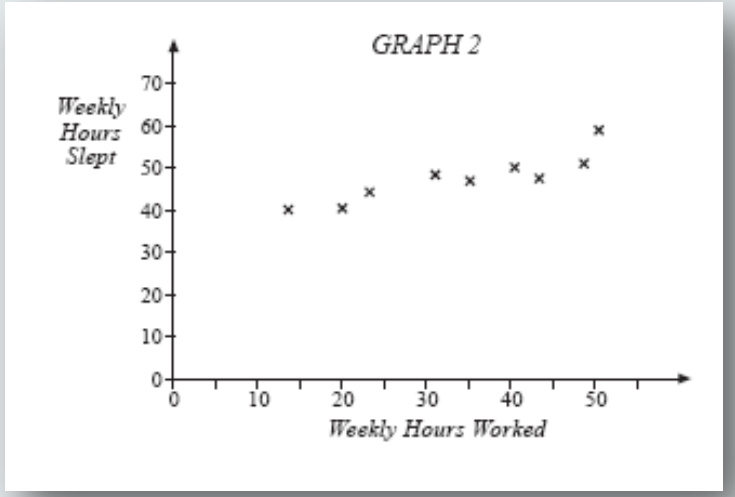
a) What does Graph 1 show about the relationship between the weekly hours spent watching television and the weekly hours worked?

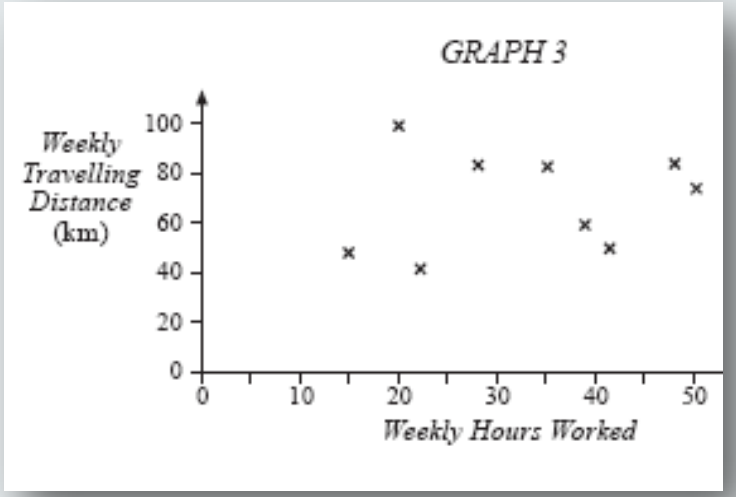
b) What does Graph 2 show about the relationship between the weekly hours slept and the weekly hours worked?

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

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

03/08/2020

Module Title: Introduction to Statistics

Session Title: Correlation

Topic title: Correlation and Linear Regression

Learning Outcomes

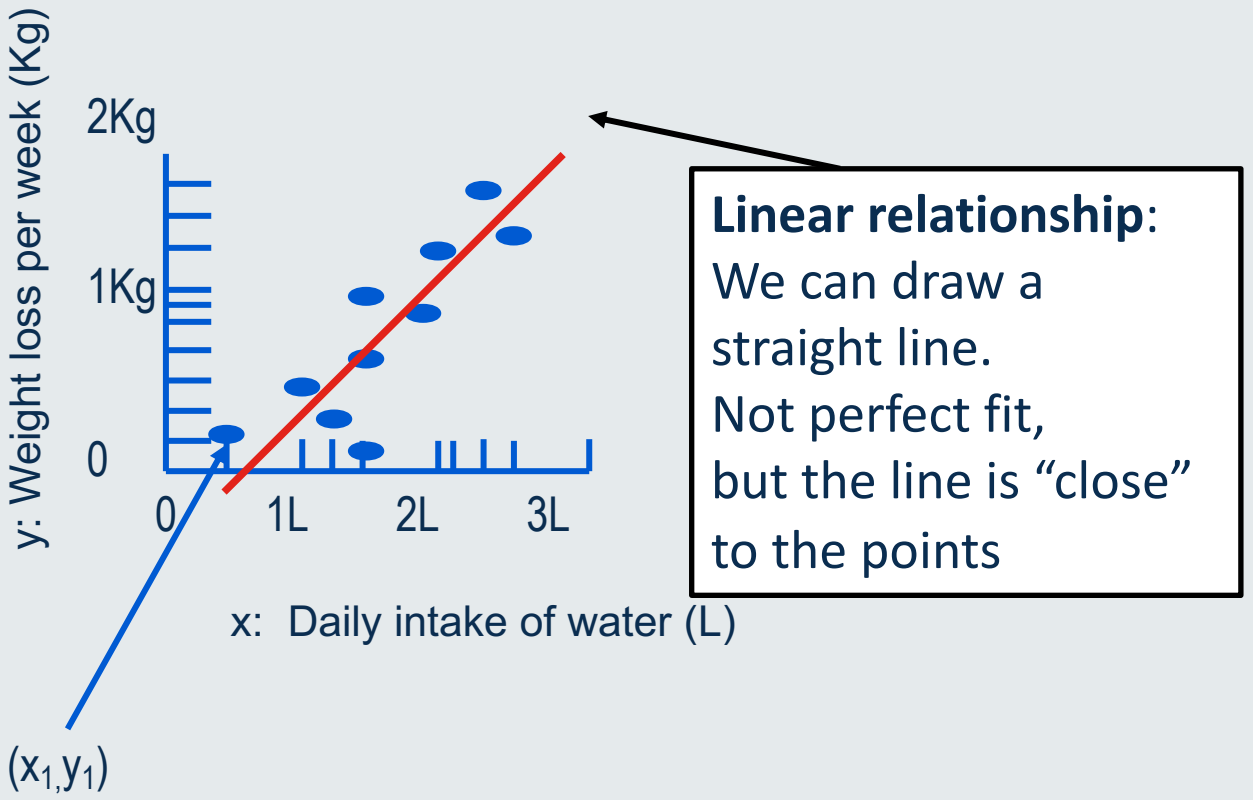
- Understand the features of a Pearson’s correlation coefficient and when to use it
- Understand the features of a Spearman’s Rank correlation coefficient and when to use it
- Understand ‘Spurious’ correlation

Previously on ‘Introduction to Statistics’

10 people were studied for the Hypothesis ‘The higher the intake of water, the higher the weight loss’.

- 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**

	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
	…	…



Correlation

We need an objective measure of strength of a linear relationship.

Correlation '**r**' is a statistical concept that refers to how close two variables are to having a linear relationship with each other, or in other words, the strength of their linear relationship. Correlation '**r**' is a method to quantify the **Direction** and **Magnitude**, of linear association between two continuous variables.

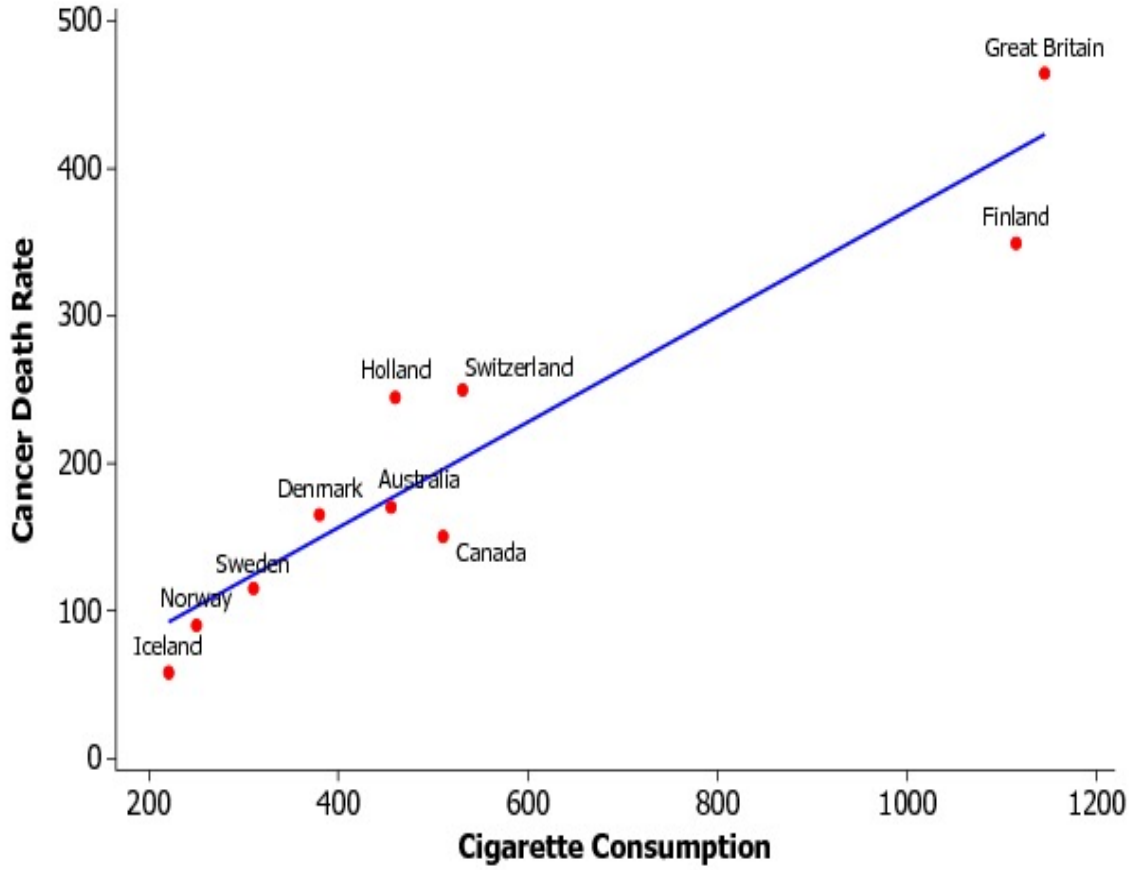
'r' belongs to the range [-1,1]

There are two types of correlation coefficients:

Pearson's Correlation Coefficient (Parametric approach)
Spearman's Correlation Coefficient (non-Parametric approach)

Example

Cigarette consumption (in 1930, average number of packets per year), Lung cancer death rate (in 1950)



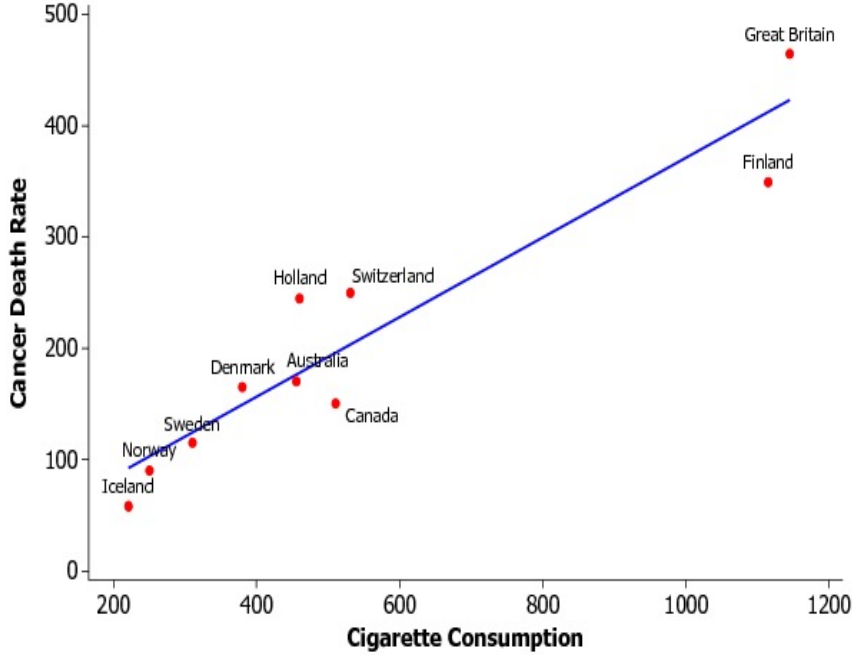
Here we show a scatterplot between country level: Cigarette Consumption and Cancer Death Rate

In addition to the scatterplot we fit a line of best fit through the centre of the data

The line indicates an increase in Cigarette Consumption is associated with an increase in Cancer Death Rate

The line and points are close, but we need a measure of the magnitude of this linear relationship.

Example



Direction of effect

The co-efficient is **positive**, thus a country with increased cigarette smoking is associated with an increased Cancer Death Rate

Magnitude of effect

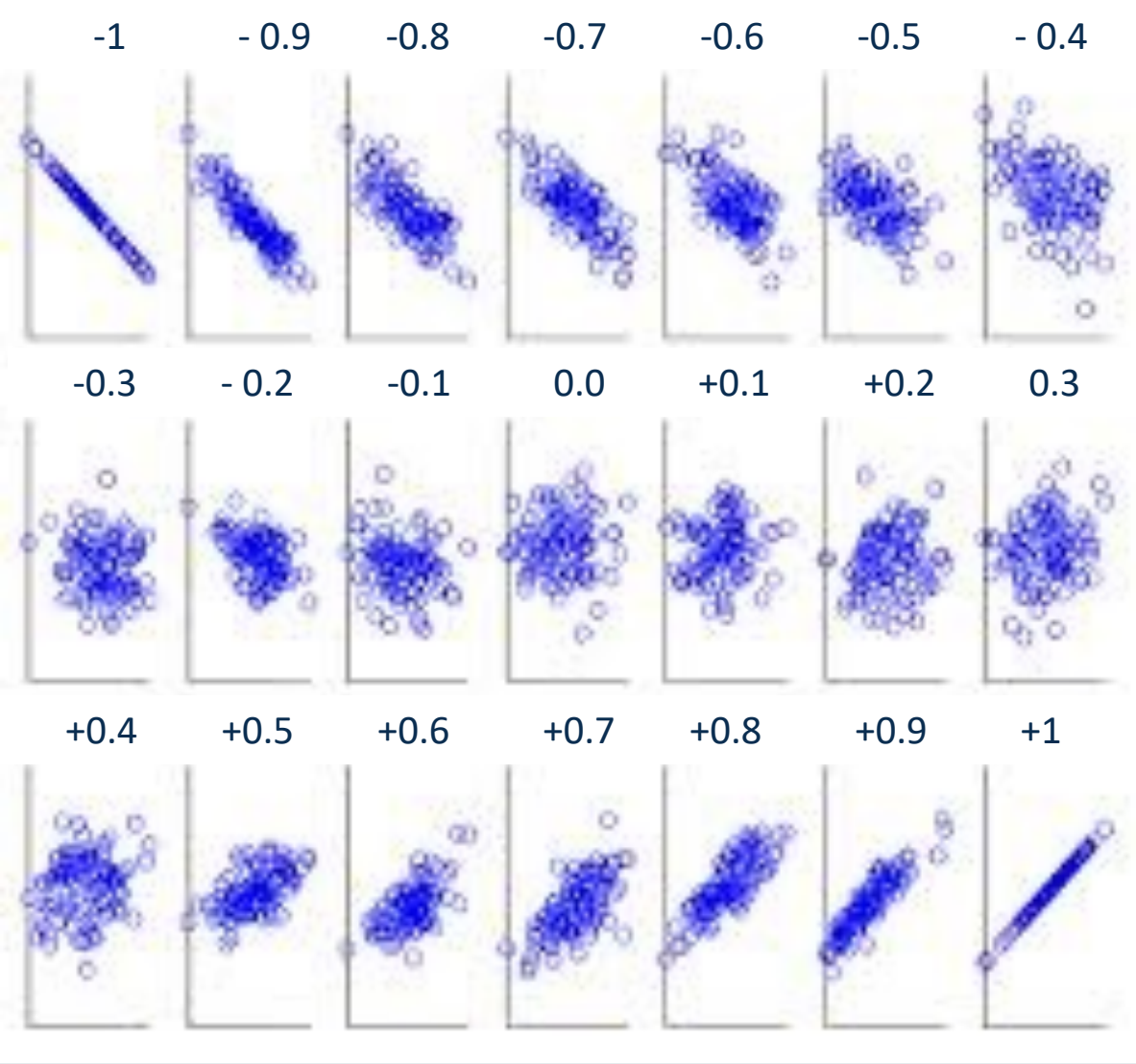
The **magnitude** of the correlation coefficient is **0.73**, thus, there is **strong** correlation.

Linear association

The points follow the line of best fit in a linear manner, thus there is **linear association**

There is strong, positive, linear association between country level cigarette consumption (in 1930) and Cancer Death Rate (in 1950) (r = 0.73).

Direction and Strength of ‘r’



Range of correlation coefficients	Degree of Correlation
0.80 to 1.00	Very strong positive
0.60 to 0.79	Strong positive
0.40 to 0.59	Moderate positive
0.20 to 0.39	Weak positive
0.00 to 0.19	Very weak positive - none
-0.19 to 0.00	Very weak negative - none
-0.39 to -0.20	Weak negative
-0.59 to -0.40	Moderate negative
-0.79 to -0.60	Strong negative
-1.00 to -0.80	Very strong negative

Direction of effect

The co-efficient is **positive** or **negative**

Magnitude of effect

The **magnitude** of the correlation coefficient ranges from -1 to 1, the closer to ±1 the stronger the effect

Pearson's Correlation Coefficient 'r'

When to use it

- To check the magnitude and direction of a linear relationship between two variables.

Hypotheses:

- H_0: the correlation in the population equals to 0
- H_a: the correlation in the population does not equal to 0

Assumptions:

- Variables should be approximately normally distributed.
- Each variable should be continuous.
- Each participant or observation should have a pair of values
- No significant outliers in either variable
- Linearity, a "straight line" relationship between the variable should be formed

Pearson's Correlation Coefficient 'r'

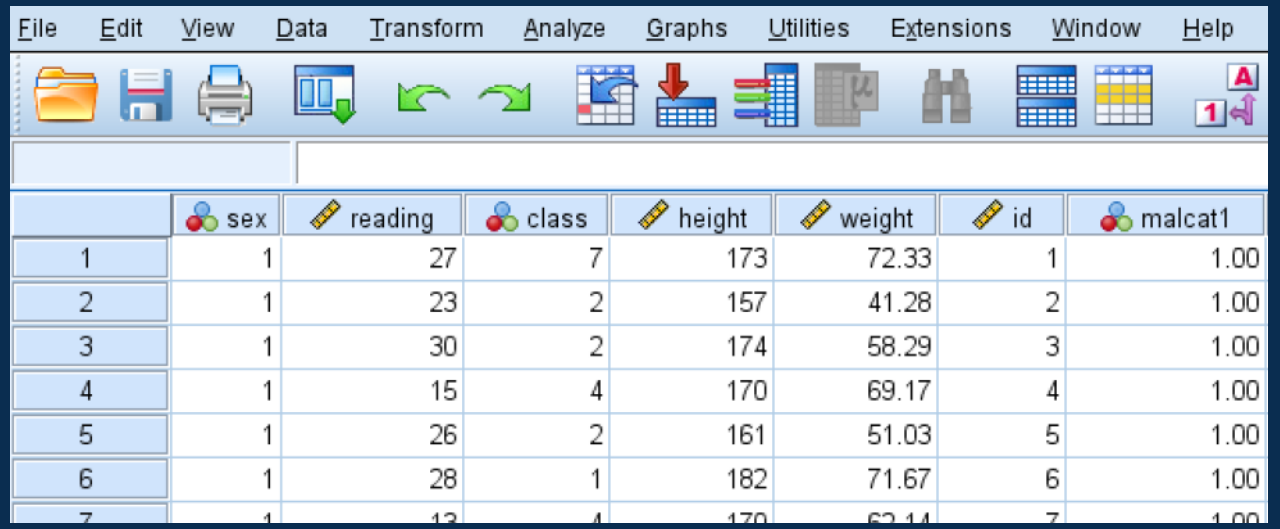
Correlation can be measured using the **Pearson's correlation coefficient 'r',** defined as

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

where s_x is the st. dev, $\bar{x}$ *is the* mean of x_i
and similarly for s_y and $\bar{y}$

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

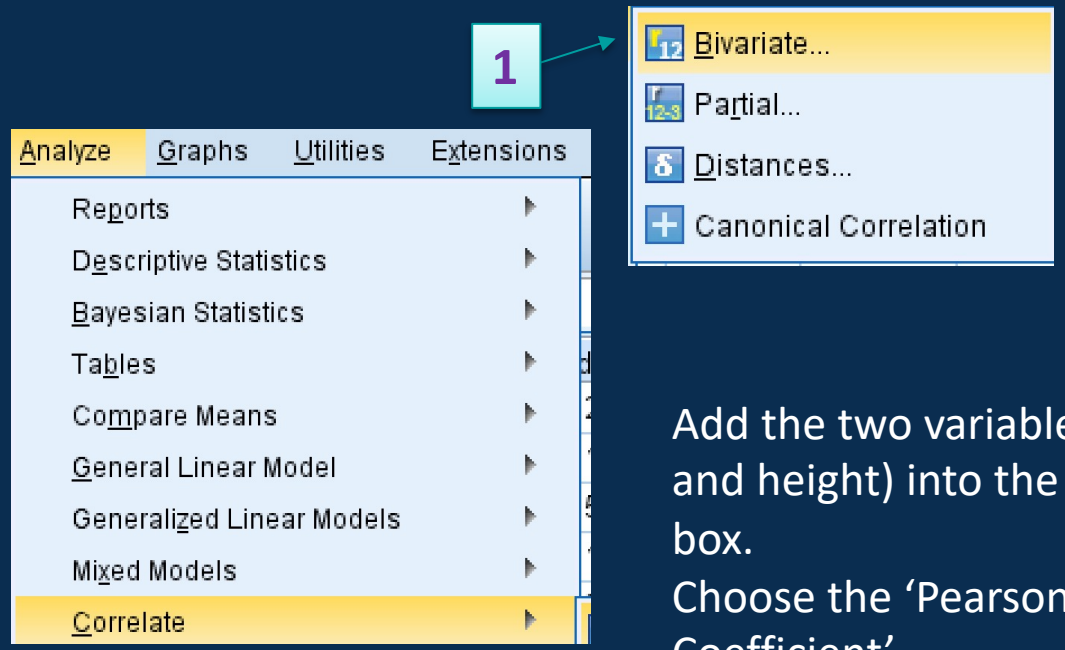
The dataset contains data from 1000 individuals, from the National Child Development Study (NCDS) with respect to their

- **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 and want to understand the direction and magnitude of the relationship.

Step 1: Calculate a correlation coefficient for variables ‘height’ and ‘weight’ from the data

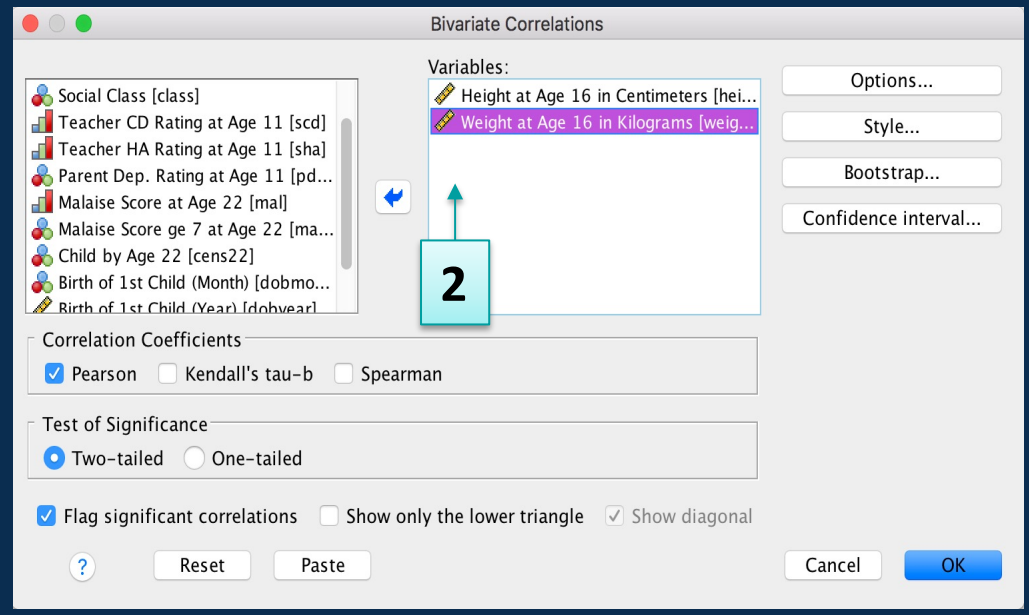


Use ‘Analyse’ -> ‘Correlate’ -> ‘Bivariate’

Add the two variables (weight and height) into the ‘Variables’ box.

Choose the ‘Pearson’ ‘Correlation Coefficient’

Click ‘Ok’



Output and Interpretation Slide

Correlations

		Height at Age 16 in Centimeters	Weight at Age 16 in Kilograms
Height at Age 16 in Centimeters	Pearson Correlation	1	.520**
	Sig. (2-tailed)		.000
	N	1000	1000
Weight at Age 16 in Kilograms	Pearson Correlation	.520**	1
	Sig. (2-tailed)	.000	
	N	1000	1000

**. Correlation is significant at the 0.01 level (2-tailed).

There is a positive moderate correlation (r=0.52) between the height and weight of children aged 16. The correlation coefficient is significantly different from 0 (p<0.001) so we can extrapolate the moderate linear relationship observed in the sample, to the whole population.

Spearman's Correlation Coefficient 'r_S'

When to use it?

When **one or both of the variable**s are not **normally distributed**. This concept of correlation is less sensitive to extreme influential points, so it should be used in the case of non normality.

What it measures?

- The strength and direction of the **monotonic** relationship between two variables.
- **A monotonic** relationship is a relationship varying in such a way that when one variable decreases or increases the other variable also decreases or increases (but not necessarily at a constant rate, as it does a linear relationship for which we use the Pearson correlation)

Hypotheses:

- H_0: the correlation in the population equals to 0
- H_a: the correlation in the population does not equal to 0

$$t = \frac{r\sqrt{n-2}}{1-r^2} \qquad \text{df=N-2}$$

Spearman's Correlation Coefficient 'r_s'

The Spearman's correlation is the **nonparametric** version of the Pearson correlation. Spearman's correlation coefficient, (**ρ**, also signified by $\mathbf{r_s}$) measures the strength and direction of association between two ranked variables.

There are two methods to calculate Spearman's correlation depending on whether: (1) your data **does not have tied ranks** or (2) your data has **tied ranks**.

The formula for when there are no tied ranks is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

where d_i is the difference in paired ranks and
n = number of cases

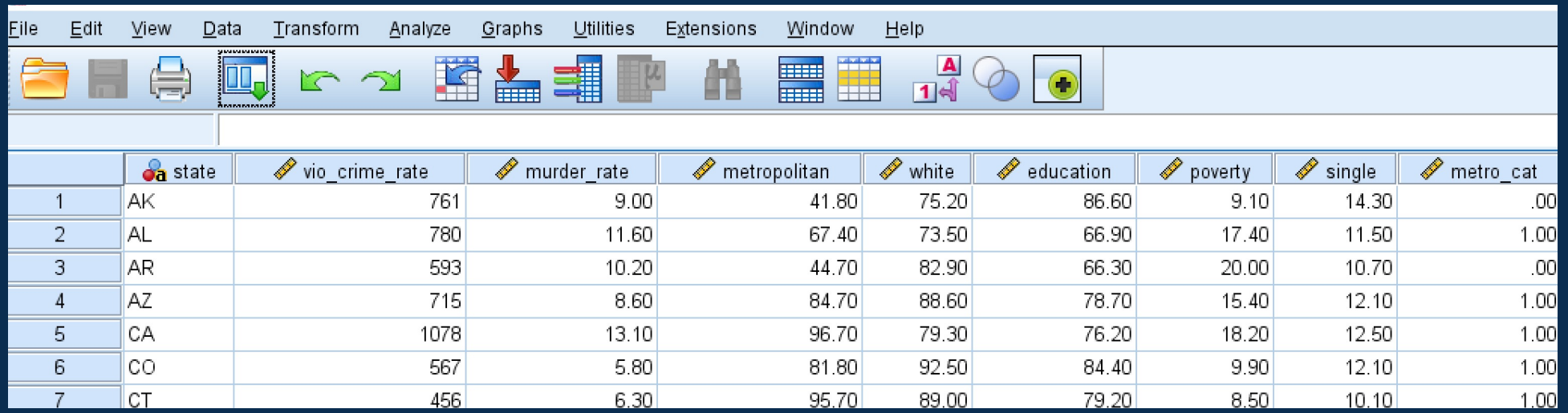
The formula for when there are tied ranks is:

$$\rho = \frac{\sum d_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2 \sum_i (y_i-\bar{y})^2}}$$

where i is the paired score

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_6b_data.sav**.



	state	vio_crime_rate	murder_rate	metropolitan	white	education	poverty	single	metro_cat
1	AK	761	9.00	41.80	75.20	86.60	9.10	14.30	.00
2	AL	780	11.60	67.40	73.50	66.90	17.40	11.50	1.00
3	AR	593	10.20	44.70	82.90	66.30	20.00	10.70	.00
4	AZ	715	8.60	84.70	88.60	78.70	15.40	12.10	1.00
5	CA	1078	13.10	96.70	79.30	76.20	18.20	12.50	1.00
6	CO	567	5.80	81.80	92.50	84.40	9.90	12.10	1.00
7	CT	456	6.30	95.70	89.00	79.20	8.50	10.10	1.00

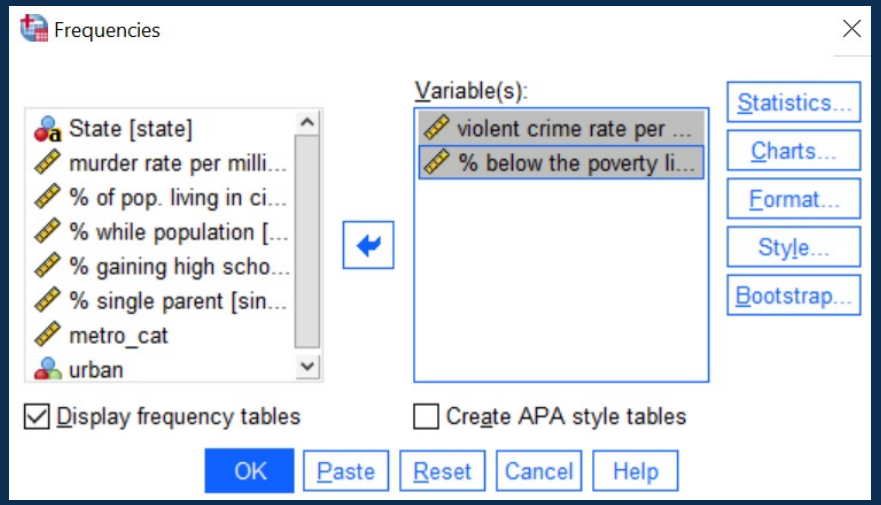
The dataset contains data from 51 US states, measuring the crime rates and background measures for each state with respect to their

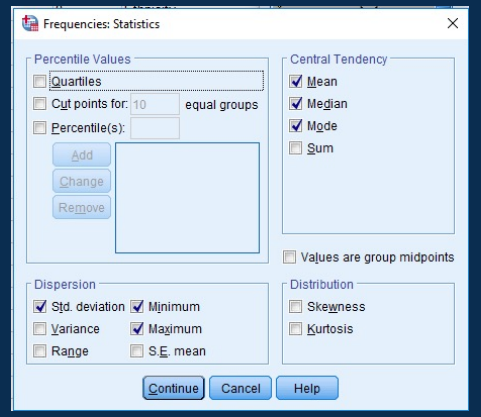
- **violent crime**: per 100,000 population
- **murder**: per 100,000 population
- **poverty**: percent below the poverty line
- **single**: percentage of lone parents

SPSS Slide: ‘how to’

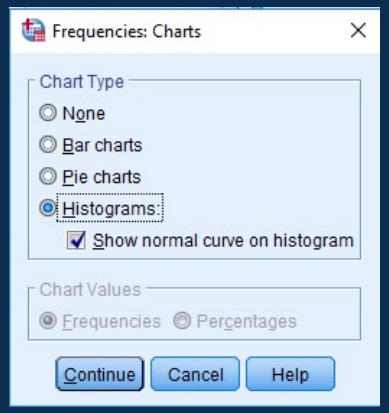
According to the researchers, in the population from which our data came, they believe there is a relationship between violent crime measured per 100,000 and the percentage of people below the poverty line per 100,000.

Step 1: Check the suitability of the data.



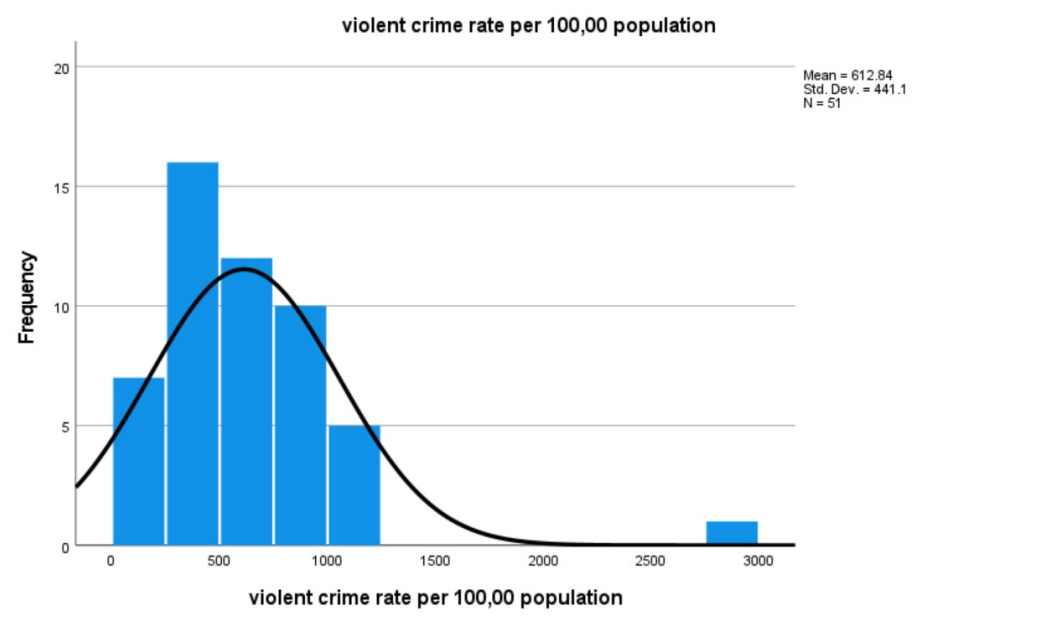


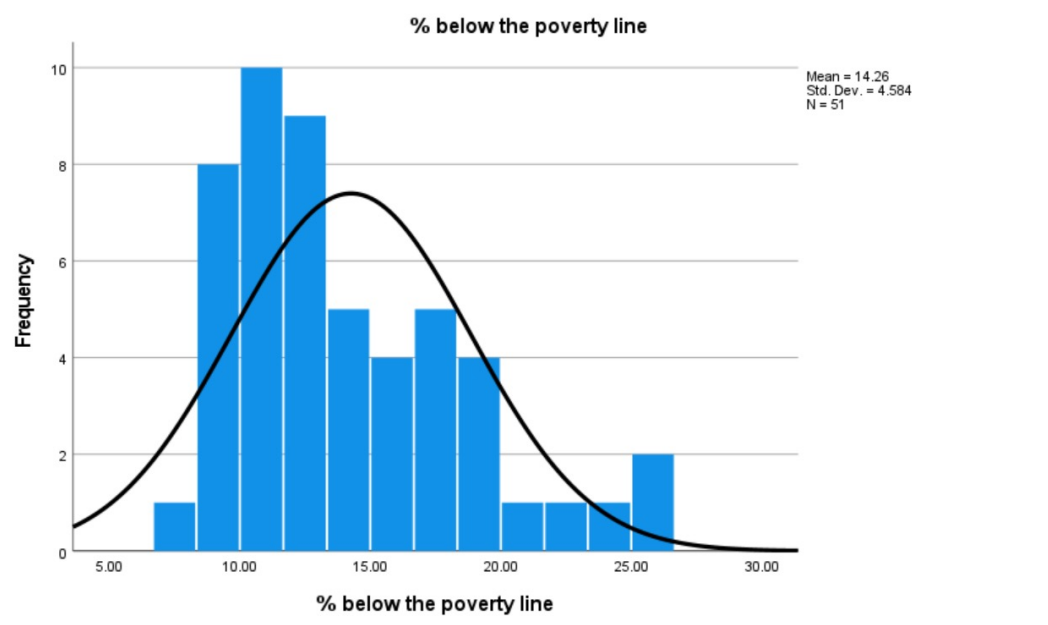
In ‘Statistics’ ask for descriptive statistics



In ‘Charts’ ask for a Histogram

Output and Interpretation Slide



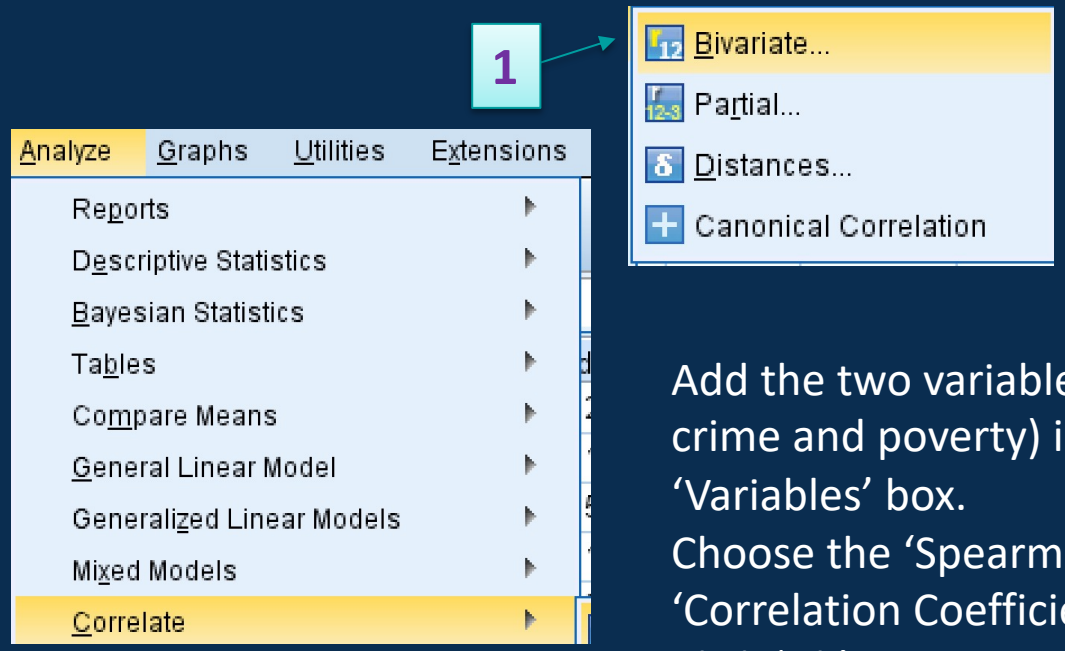


‘Violent Crime’ is a positively skewed variable. ‘Poverty’ is a positively skewed variable. Pearson’s product moment correlation coefficient is unsuitable for this data. Use Spearman’s correlation coefficient instead.

SPSS Slide: ‘how to’

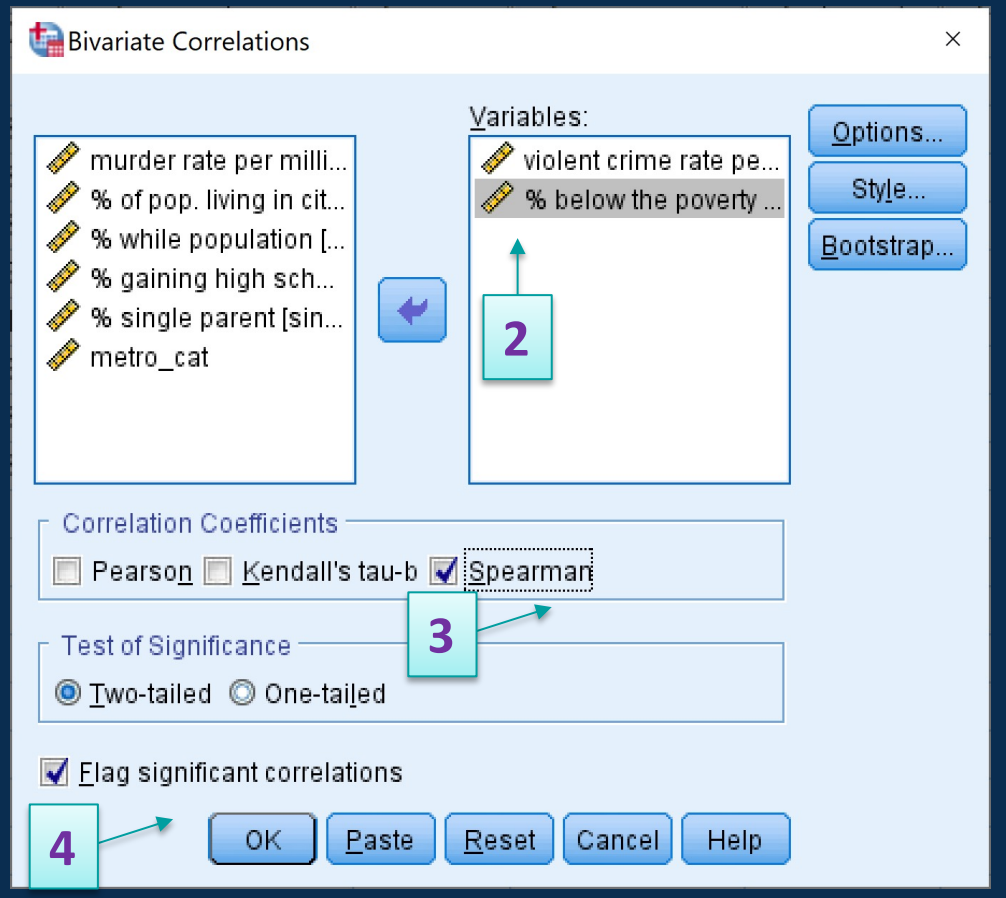
According to the researchers, in the population from which our data came, they believe there is a relationship between Violent crime measured per 100,000 and the percentage of people below the poverty line per 100,000.

Step 2: Calculate a correlation coefficient for variables ‘violent crime’ and ‘poverty’ from the data



Use ‘Analyse’ -> ‘Correlate’ -> ‘Bivariate’

Add the two variables (violent crime and poverty) into the ‘Variables’ box.
Choose the ‘Spearmans’ ‘Correlation Coefficient’
Click ‘Ok’



Output and Interpretation Slide

Correlations

			violent crime rate per 100,00 population	% below the poverty line
Spearman's rho	violent crime rate per 100,00 population	Correlation Coefficient	1.000	.391**
		Sig. (2-tailed)	.	.005
		N	51	51
	% below the poverty line	Correlation Coefficient	.391**	1.000
		Sig. (2-tailed)	.005	.
		N	51	51

**. Correlation is significant at the 0.01 level (2-tailed).

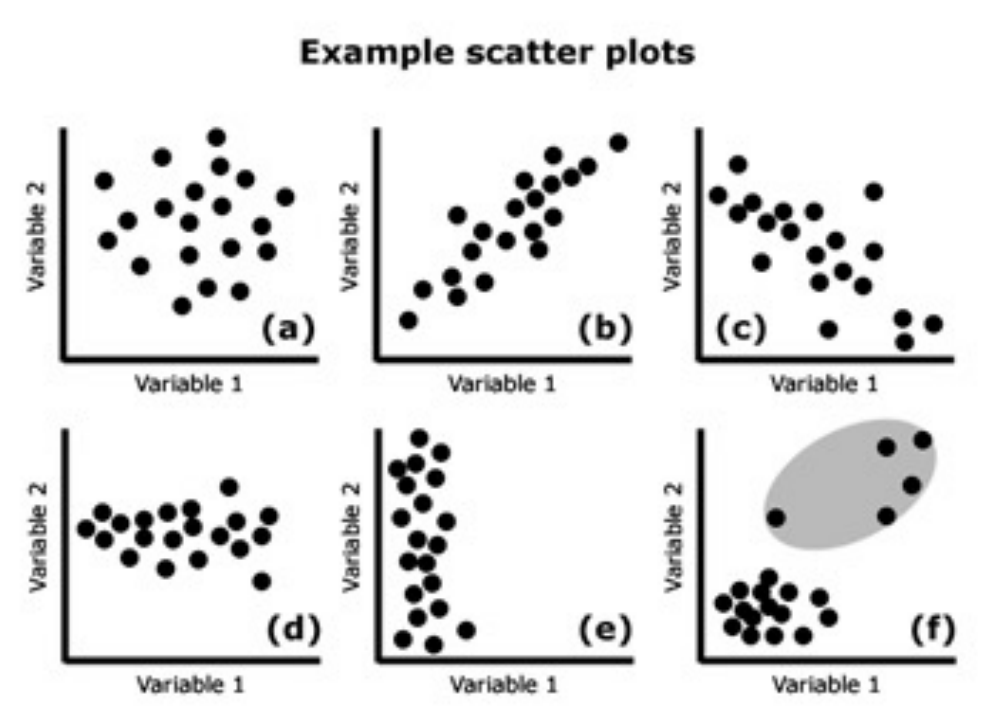
There was a weak positive (r_s=0.39) relationship between 'violent crime per 100,000' and 'percent below the poverty line per 100'000'. The correlation coefficient is significantly different from 0 (p=0.005) so we can extrapolate the weak linear relationship observed in the sample, to the whole population.

Spurious Correlation: A Word of Caution

Just because two variables are correlated, this doesn’t mean there is a causal association between the variables.

The correlation may be due the result of a third unknown variable.

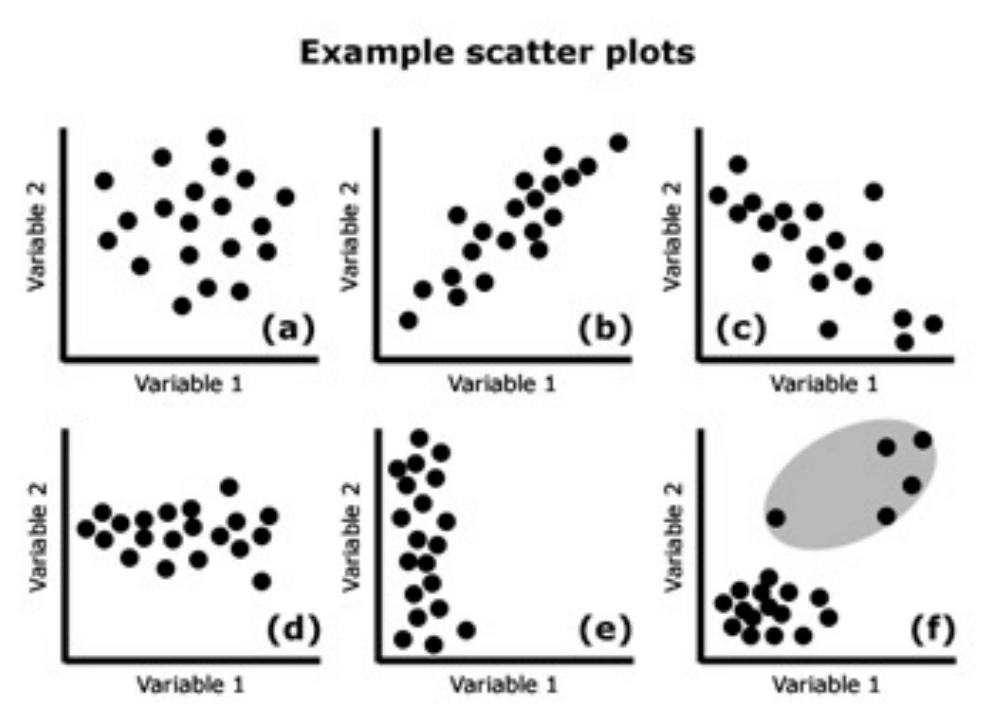
Knowledge Test



1. Quantifying linear relationships

   Looking at the 6 figures to the left. For each figure how would you describe the linear relationship between variable 1 and variable 2?

Knowledge Test Solution



1. Quantifying linear relationships

   Looking at the 6 figures to the left. For each figure how would you describe the linear relationship between variable 1 and variable 2?

a) No linear relationship apparent
b) Positive linear relationship
c) Negative Linear relationship
d) No linear relationship variable 1 is utterly immaterial to the variable 2.
e) No linear relationship variable 1 is utterly immaterial to the variable 2.
f) Two distinct clusters of data showing different relationships between variable 1 and 2. This may indicate there is some other variable modifying the relationship between these variables

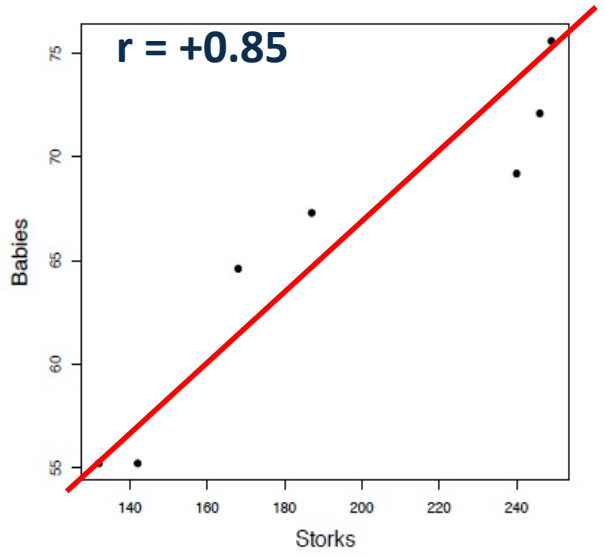
Reflection

Biology: Stork Population vs. Births

If you examine the records of the city of Copenhagen for the ten or twelve years following World War II, you will find a strong positive correlation between (i) the annual number of storks nesting in the city, and (ii) the annual number of human babies born in the city. These data were researched by Dr. Gustav Fischer and subsequently published in *Ornithologische Monatsberichte*, 44 No. 2, Jahrgang, 1936, Berlin *Ornithologische Monatsberichte*, 48 No. 1, Jahrgang, 1940, Berlin *Statistiches Jahrbuch Deutscher Gemeinden*, 27-33, Jahrgang, 1932-1938, Gustav Fischer, Jena.

Can we conclude that storks bring babies? Let's examine the correlation coefficient, just as Dr. Fischer has done in his published work. Visually inspecting the correlation coefficient between the dependent and independent variables, stork population and births, respectively, we can guess that the correlation coefficient is positive and near +0.85.

In this example what you have is a situation where two variables end up as correlated, not because one is influencing the other, but rather because both are influenced by a third variable, Z, that is not being taken into account. That is, the causal relationship here is not $X \rightarrow Y$ or $X \leftarrow Y$,



Read the following article:
What could be this third variable influencing these other variables?

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, what kind of direction and strength might you expect this relationship to have.

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

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

- 03/08/2020

Module Title: Introduction to Statistics

Session Title: Simple Linear Regression

Topic title: Correlation and Linear Regression

Learning Outcomes

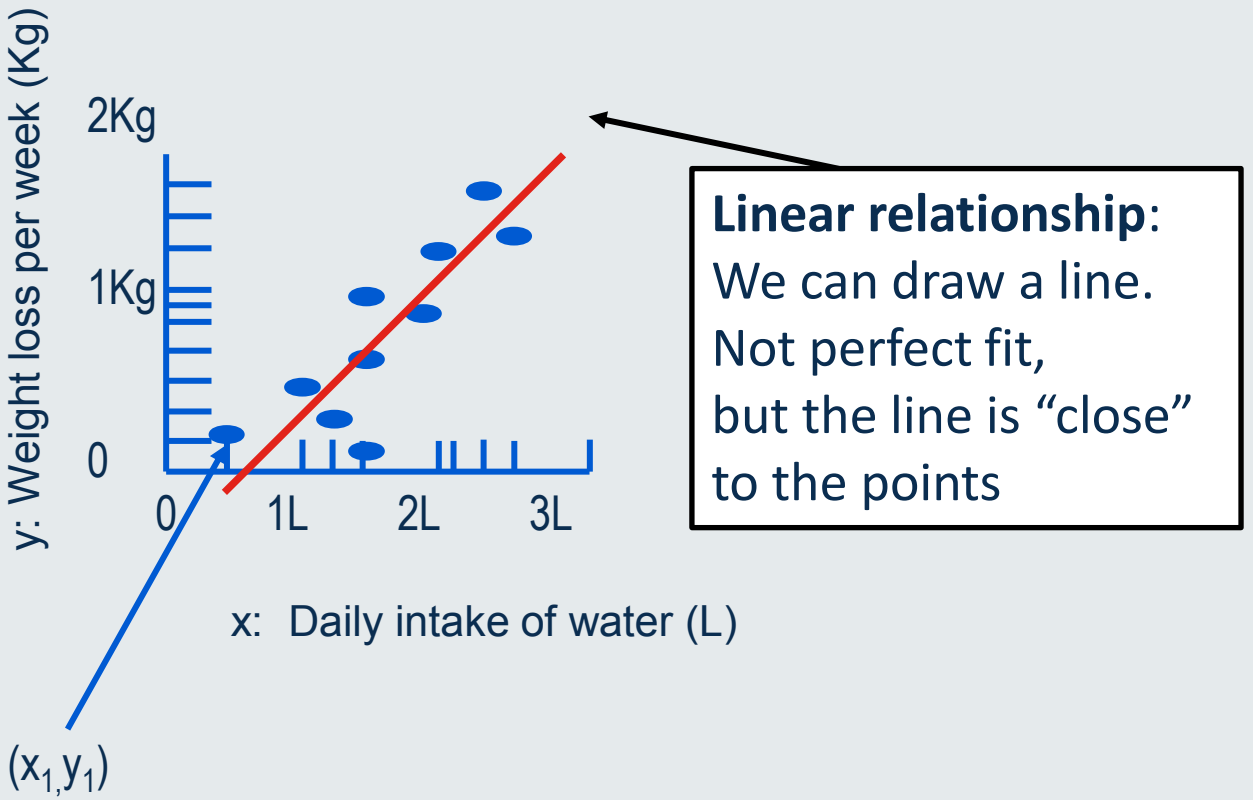
- Understand the difference between an **independent** and **dependent** variable
- Understand the **parameters** of simple linear regression (SLR)
- Interpret the **intercept** and **slope** parameters from a regression equation
- Use the simple linear regression (SLR) parameters to predict future observations
- Understand how to introduce a dummy categorical variable

Previously on ‘Introduction to Statistics’

10 people were studied for the Hypothesis ‘The higher the intake of water, the higher the weight loss’.

- 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**

	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
	…	…



Previously on ‘Introduction to Statistics’

We need an objective measure of strength of a linear relationship

Correlation is a statistical concept that refers to how close two variables are to having a linear relationship with each other, or in other words, the strength of their linear relationship. Correlation is a method to quantify the **Direction** and **Magnitude**, of linear association between two continuous variables.

Range of correlation coefficients	Degree of Correlation
0.80 to 1.00	Very strong positive
0.60 to 0.79	Strong positive
0.40 to 0.59	Moderate positive
0.20 to 0.39	Weak positive
0.00 to 0.19	Very weak positive - none
-0.19 to 0.00	Very weak negative - none
-0.39 to -0.20	Weak negative
-0.59 to -0.40	Moderate negative
-0.79 to -0.60	Strong negative
-1.00 to -0.80	Very strong negative

Direction of effect
The co-efficient is positive or negative

Magnitude of effect
The magnitude of the correlation coefficient ranges from -1 to 1, the close to ±1 the stronger the effect

There are two types of correlation coefficients

- Pearson's Correlation Coefficient (normally distributed data)
- Spearman's Correlation Coefficient (skewed or ordinal data)

Simple Linear Regression

In statistical modelling, a regression model is a set of statistical processes for estimating the relationships among variables. These models describe the relationship between variables by fitting a line to the observed data. The relationship is expressed as an equation.

In this session, we will focus on cases where there is a linear relationship between one continuous outcome and one predictor, for which the equation will look like:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

This model is known as the **simple linear regression model.**

Simple Linear Regression

$$\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon}$$

- $\mathbf{x}$ is called the **independent variable, predictor, explanatory** or **covariate** (continuous or categorical)
- $\mathbf{y}$ is called the **dependent variable, outcome or response**. y **'depends on'** x (always continuous)
- The **intercept** $\boldsymbol{\beta_0}$ is the value that y takes when x is zero.
  - If the intercept is zero then y increases in proportion to x (i.e. double x then y doubles y=x)
- The **slope** $\boldsymbol{\beta_1}$ determines the change in y when x changes by one unit.
  - It is the amount that the dependent variable will increase (or decrease) for each unit increase in the independent variable
- $\boldsymbol{\varepsilon}$ is called the **residual** (distance between the points and the line).
- $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are together known as **regression coefficients.**

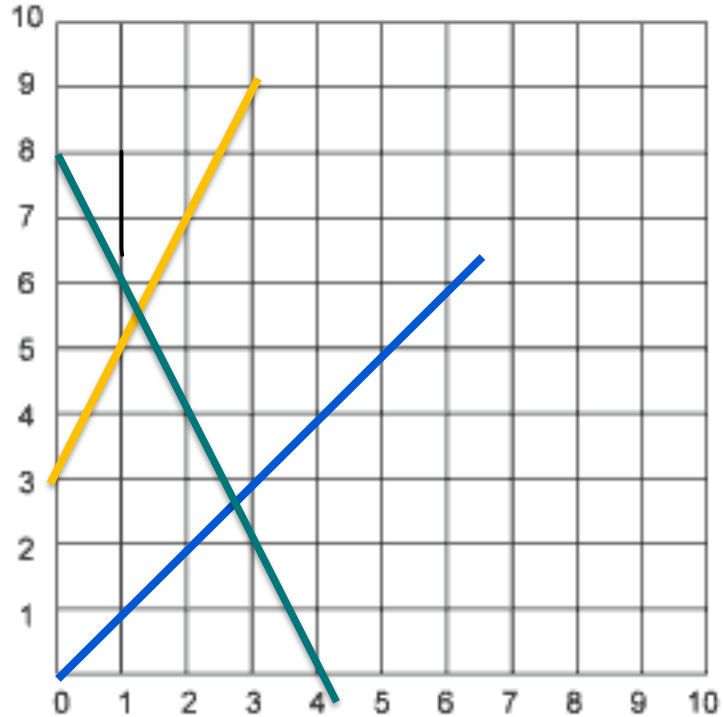
$y=\beta_0 + \beta_1 + \varepsilon$

x	y
1	1
2	2
3	3
…	…
7	7

y = x

x	y
0	3
1	5
2	7
…	…
7	17

y = 3 +2x



x	y
0	8
1	6
2	4
…	…
7	-6

y = 8 -2x

β_0 represents where the line intercepts the y axis.

β_1 represent the slope of the line as x increases by one unit how much does y increase or decrease

Example

You are a social researcher interested in the relationship between income and happiness. You survey 500 people whose incomes range from £15k to £75k and ask them to rank their happiness on a scale from 1 to 10.

Your independent variable (income) and dependent variable (happiness) are both quantitative, so you can do a regression analysis to see if there is a linear relationship between them.

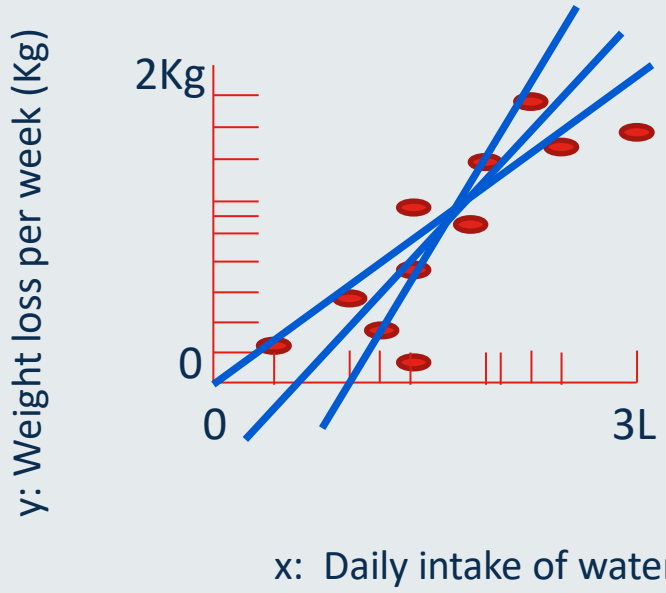
We can ask

- How strong the relationship is between two variables (e.g. the relationship between income and happiness).
- The value of the dependent variable at a certain value of the independent variable (e.g. the amount of happiness at a certain level of income).

Estimation

Independent variable (x): Daily intake of water

Dependent variable (y): Weight loss per week



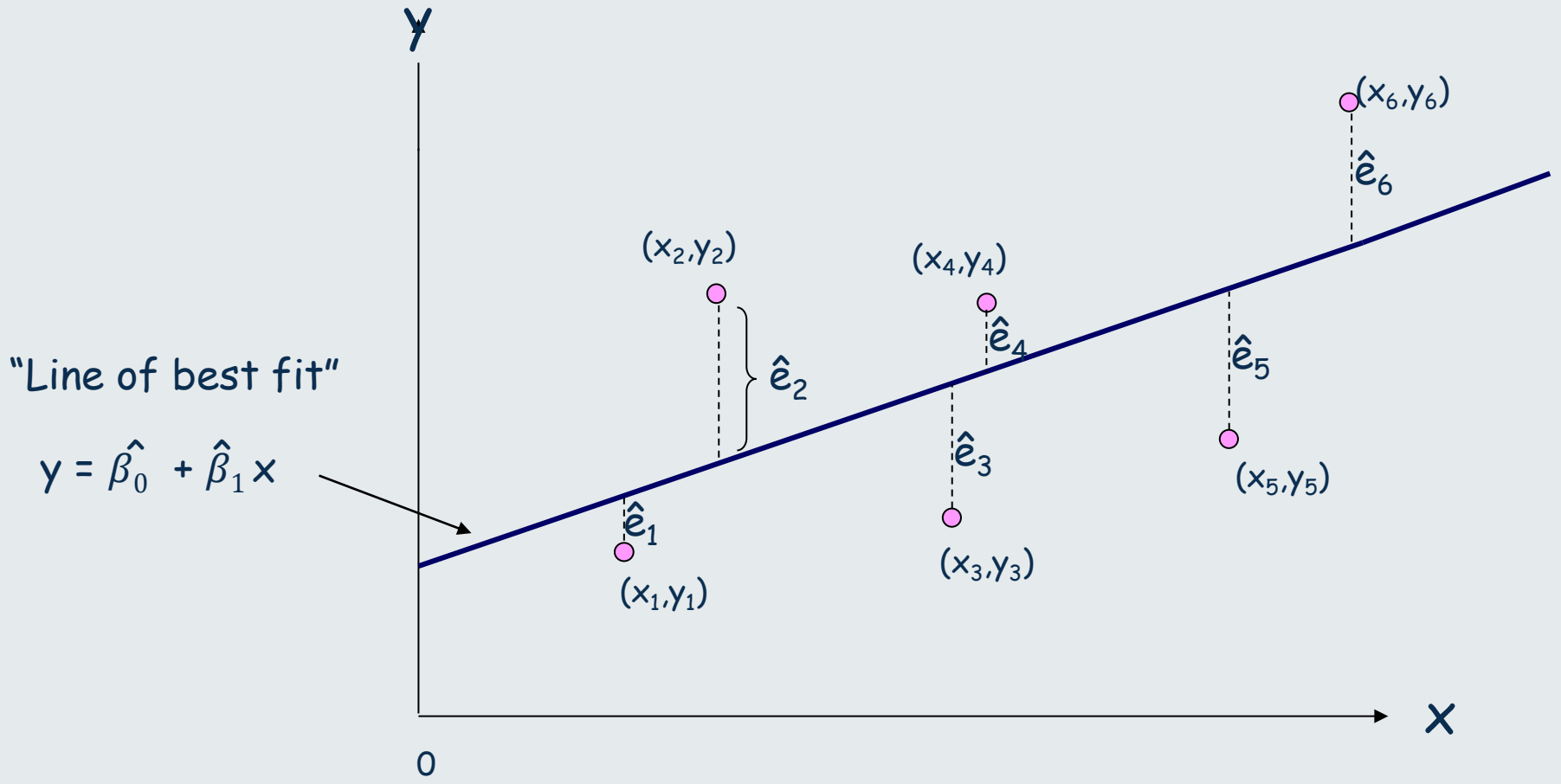
$$y = \beta_0 + \beta_1 x + \varepsilon$$

?

How can we find the best line fitting the cloud of points and therefore find the best estimates for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$?

Estimation

- The best **linear regression line** is the closest to all data points, i.e. the line that makes the **residual** ε as small as possible.
- **Ordinary Least Squares (OLS)** – Is one method that can be used to estimate the regression line that **minimises** the **squared residuals** ($\boldsymbol{\varepsilon_1^2 + \varepsilon_2^2 + \cdots \varepsilon_n^2}$) to give us the estimates for $\boldsymbol{\widehat{\beta_0}}$ and $\boldsymbol{\widehat{\beta_1}}$.



Simple Linear Regression Model

When to use it

- To measure to what extent there is a linear relationship between two variables

Hypotheses:

- H_0: There is no linear association e.g. the slope β_1 in the population equals to 0
- H_a: There is a linear association e.g. the slope β_1 in the population does not equal to 0

Assumptions:

- There is a linear relationship between the dependent and independent variable
- Residuals (or "errors") ε are independent of one another: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations
- Residuals follow a Normal distribution, with mean 0 and constant Standard Deviation σ
- Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.

Formulae – for the curious

The slope is estimated as $\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$

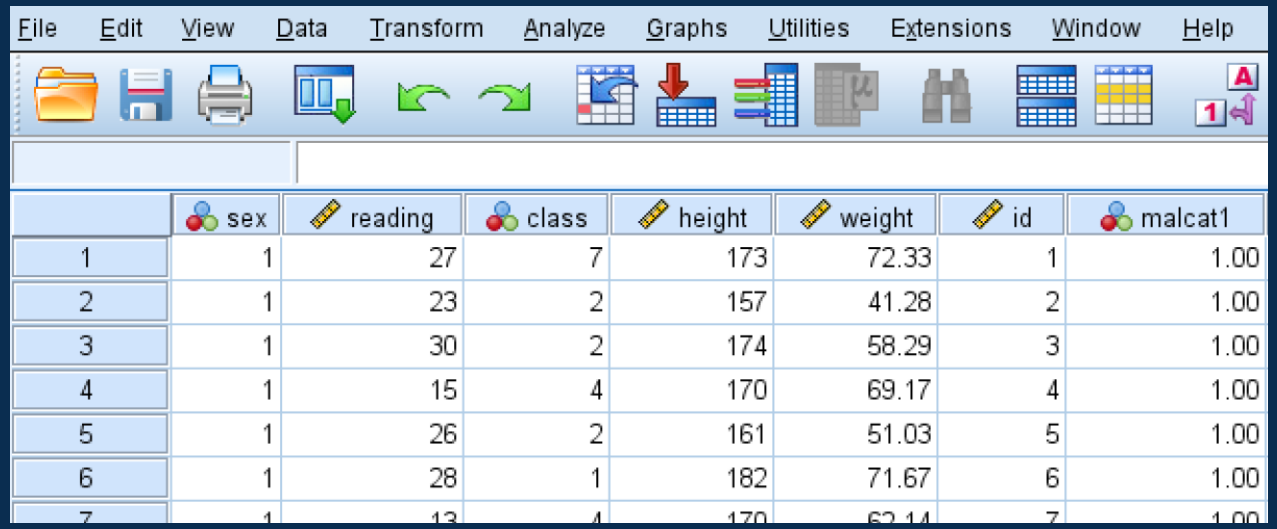
The intercept is estimated as $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$

$$\left[\; \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \;\right]$$

Test Statistic for the hypothesis test $t = \frac{\widehat{\beta_1}}{\widehat{se}(\widehat{\beta_1})}$, df=n-1

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

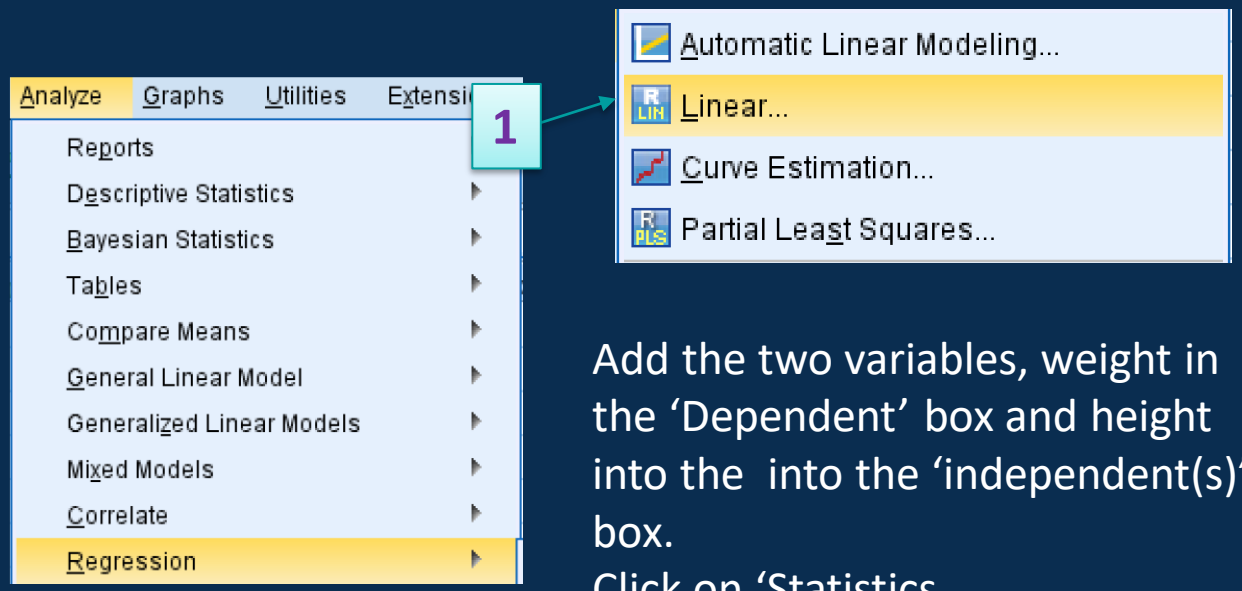
The dataset contains data from 1000 individuals, from the National Child Development Study (NCDS) with respect to their

- **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)

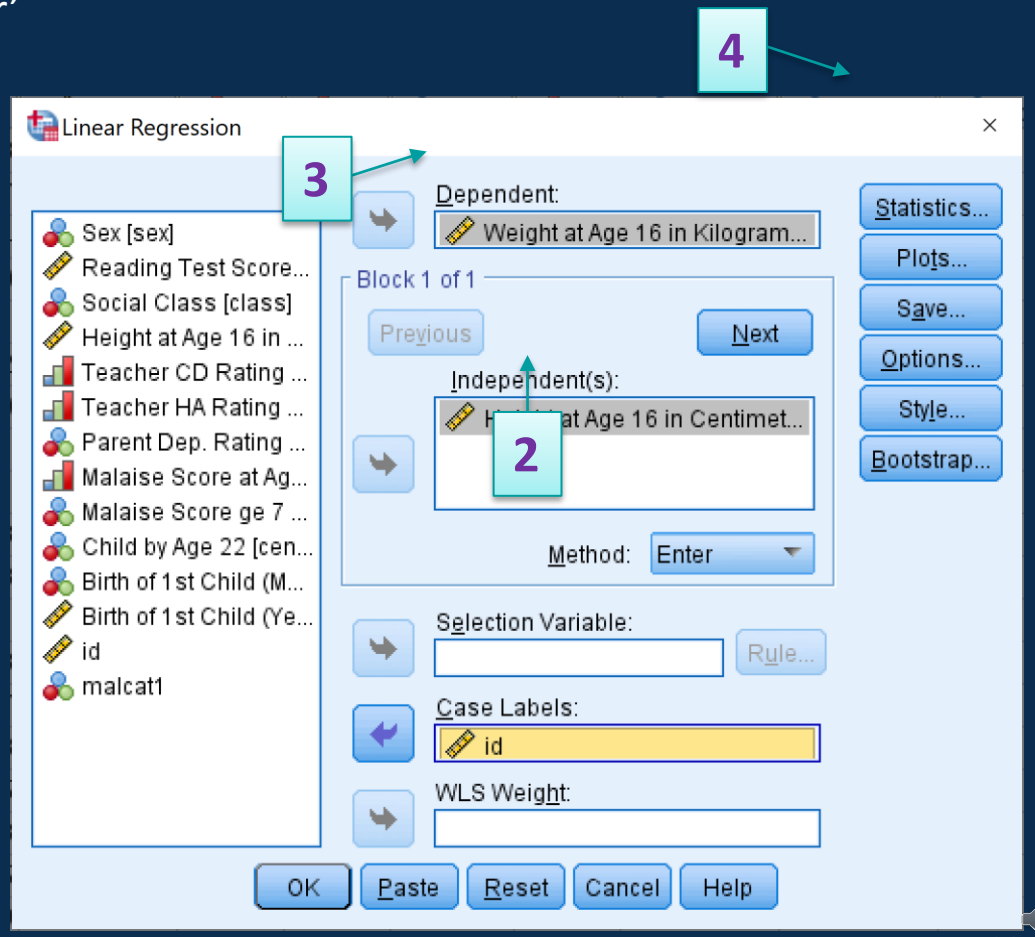
SPSS Slide: ‘how to’

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: Compute a Linear regression model for dependent variable ‘weight’ and independent variable ‘height’ from NCDS data Use ‘Analyse’ -> ‘Regression’ -> ‘Linear’



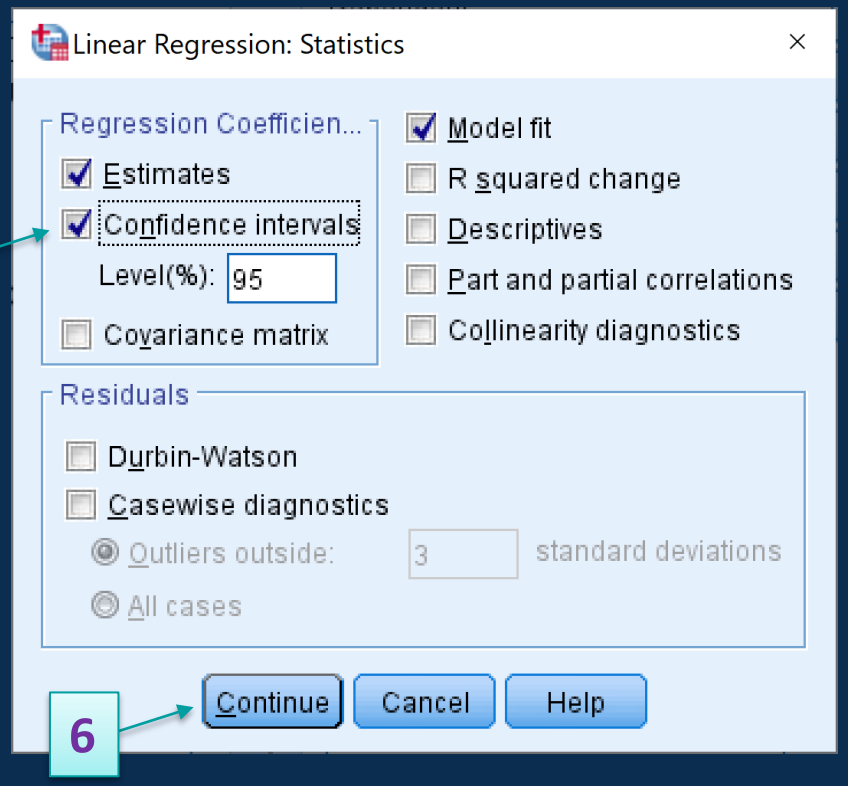
Add the two variables, weight in the ‘Dependent’ box and height into the into the ‘independent(s)’ box.
Click on ‘Statistics



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: Compute a Linear regression model for dependent variable ‘weight’ and independent variable ‘height’ from NCDS data



In the Statistics tab.
Check the ‘Estimates’
Check the ‘Confidence Intervals’
Click on ‘Continue’
Click on ‘OK’

Output and Interpretation Slide

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.520[a]	.270	.270	8.25311

a. Predictors: (Constant), Height at Age 16 in Centimeters

b. Dependent Variable: Weight at Age 16 in Kilograms

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25172.852	1	25172.852	369.570	.000[b]
	Residual	67977.581	998	68.114		
	Total	93150.434	999			

a. Dependent Variable: Weight at Age 16 in Kilograms

b. Predictors: (Constant), Height at Age 16 in Centimeters

This table provides the R and R^2 values. The R value represents the simple correlation and is 0.520 which indicates a moderate degree of correlation.

The R^2 value indicates how much of the total variation in the dependent variable, weight, can be explained by the independent variable, height. In this case, 27.0% can be explained.

The ANOVA table, reports how well the regression equation fits the data (i.e., predicts the dependent variable). This table indicates that the regression model predicts the dependent variable significantly well ($p<0.001$).

This indicates the statistical significance of the regression model that was run and overall, the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data).

Output and Interpretation

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-46.764	5.413		-8.639	.000	-57.386	-36.142
	Height at Age 16 in Centimeters	.626	.033	.520	19.224	.000	.562	.689

a. Dependent Variable: Weight at Age 16 in Kilograms

β_0

β_1

$SE(\beta_1)$

The estimated slope coefficient (β_1), suggests a 1cm increase in height is associated with a 0.626kg increase in weight. The units of the slope is kg/cm.

The intercept (β_0), is the extrapolated weight for a 16 year old of zero height.

In addition to getting point estimation for β1, it is possible to calculate a confidence interval for the slope parameter
The confidence interval formula is:

$$95\%\ CI = [\beta 1 - 1.96 \times SE(\beta 1), \beta 1 + 1.96 \times SE(\beta 1)]$$

E.g. for the NCDS data, a CI for $\beta 1$ can be derived as follows:
Lower limit: 0.626-1.96*0.033=0.562
Upper limit: 0.626+1.96*0.033=0.689
= [0.562, 0.689]

Output and Interpretation

Coefficients[a]

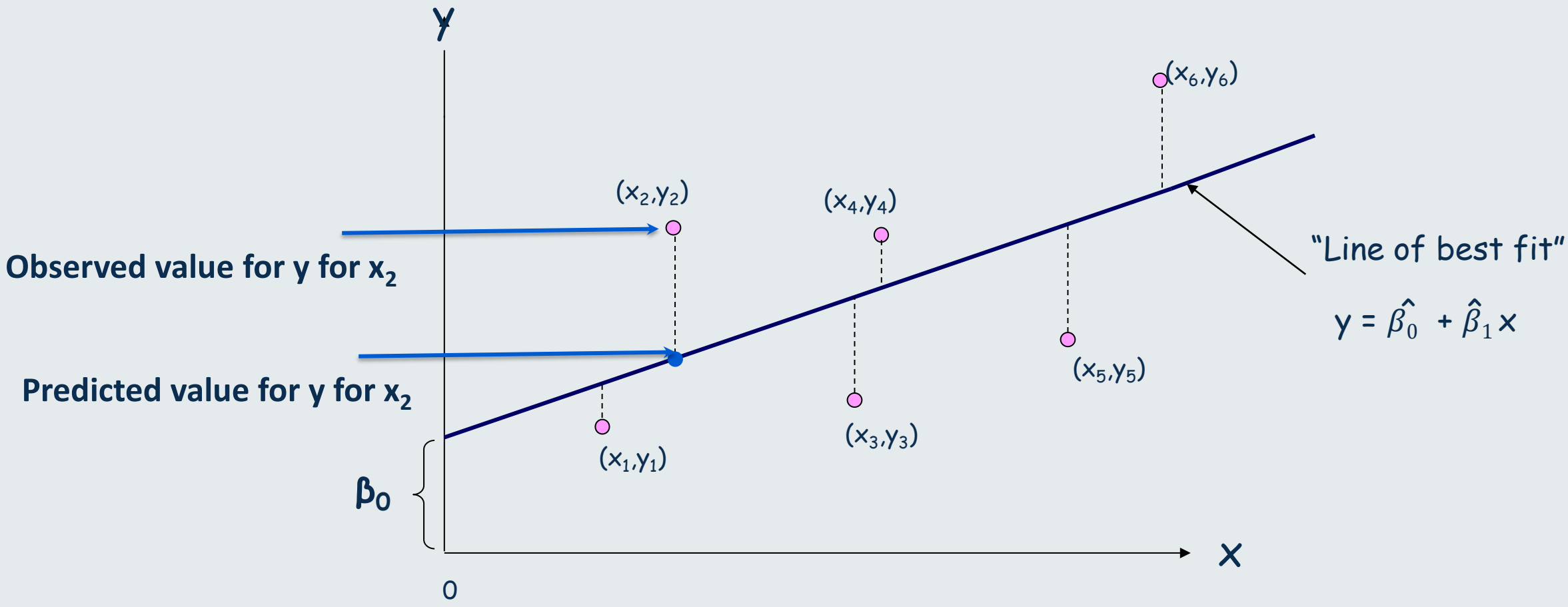
Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-46.764	5.413		-8.639	.000	-57.386	-36.142
	Height at Age 16 in Centimeters	.626	.033	.520	19.224	.000	.562	.689

a. Dependent Variable: Weight at Age 16 in Kilograms

We found a significant relationship between weight and height of 16 year olds with a 1cm increase in height associated with a 0.626kg increase in weight ($\beta 1$=0.626, t=19.224, p<0.001 95%CI (0.562, 0.689))

Prediction

Regression models are used to predict new cases.
The predicted value $\hat{y}$ for a new observation x is its corresponding value on the regression line.



Prediction

We can use the regression equation to predict the weight for new case, added to the sample:
If x=186cm for a given 16 years old new case, and knowing that y=-46.764 + 0.626 x,
What would be the child's weight?

We can estimate:

$$y = -46.764 + 0.626\ x,$$
$$y = -46.764 + 0.626 \times 186$$
$$y = 69.672\ Kg$$

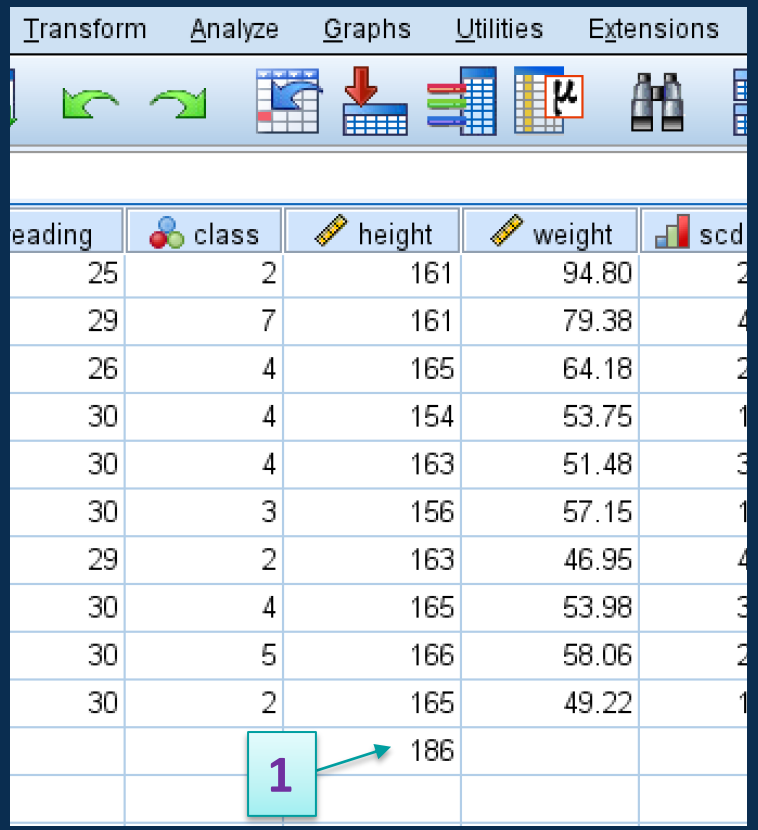
The model predicts a weight of 69.672 Kg for a 16-years old child that is 186 cm tall

In addition to getting predicted values of weight for any given height, it is possible to calculate a confidence interval for that prediction.

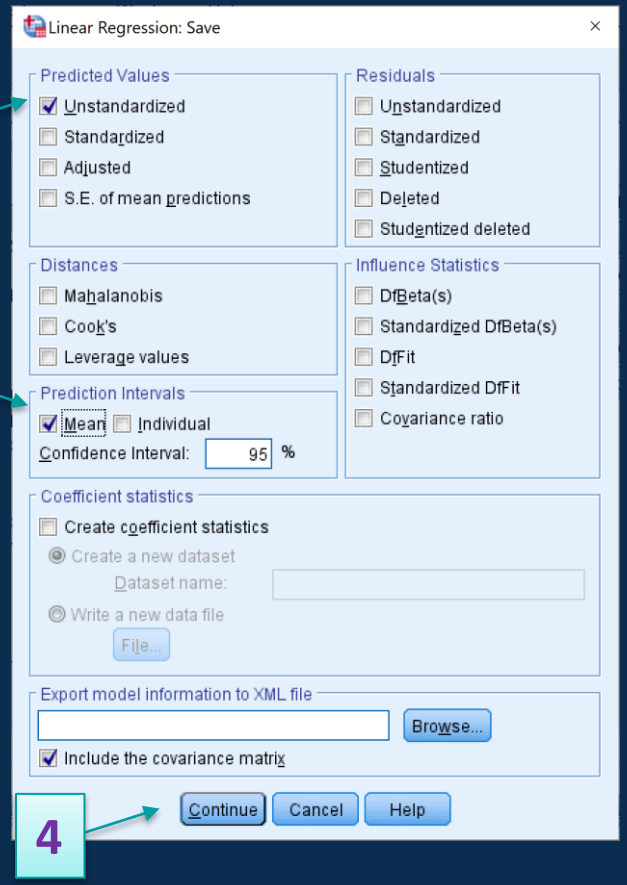
SPSS Slide: 'how to'

If x=186cm for a given 16 years old new case, and knowing that y=-46.764 + 0.626 x,
What would be expect the child's weight to be?

Step 1: Add the x-values at which you want to predict y to the y-variable (here height) in the data. Use height= 186 cm



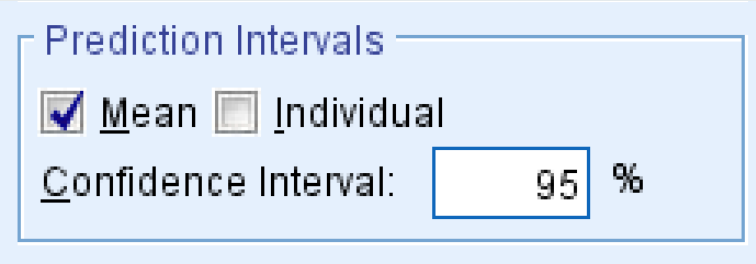
Step 2) Use **Analyse -> Regression -> Linear**
Step 2) Put 'weight' in dependent, and 'height' in independent.
Click '**Save'**, select 'Prediction values' 'Unstandardised' and 'Prediction intervals' 'mean'.
Click on 'Continue'
Click on 'OK'



Output and Interpretation

PRE_1	LMCI_1	UMCI_1
53.94261	53.33354	54.55167
53.94261	53.33354	54.55167
56.44463	55.92713	56.96213
49.56407	48.63380	50.49434
55.19362	54.64309	55.74414
50.81508	49.98842	51.64174
55.19362	54.64309	55.74414
56.44463	55.92713	56.96213
57.07013	56.55788	57.58238
56.44463	55.92713	56.96213
69.58024	68.21403	70.94645

The 'Data View' in SPSS you will see three new columns one for the predicted y (PRE_1) $\hat{y} = \mathbf{69.58\ kg}$ based on the value of 186cm height and the lower (LMCI_1) and upper (UMCI_1) confidence interval limits **95%CI (68.21, 70.95.**



For instance, to predict the average weight of 16 year olds if the height is 186cm use the **confidence interval of the mean**.

To predict the weight of Jasmine, a 16 year old with weight 186cm then use the **confidence interval for the individual.**

Categorical Predictors

What do we do if we have a predictor that is **categorical** ?
Focus on continuous outcome **y** = weight and categorical explanatory variable **x** = gender.

When x is categorical binary then:

- The regression line connects the mean response in one group with the mean response in the other.
- The slope coefficient simply measures the group difference in means (remember: slope measures predicted change in y when x changes by one unit=switches groups)

Coefficients[a]

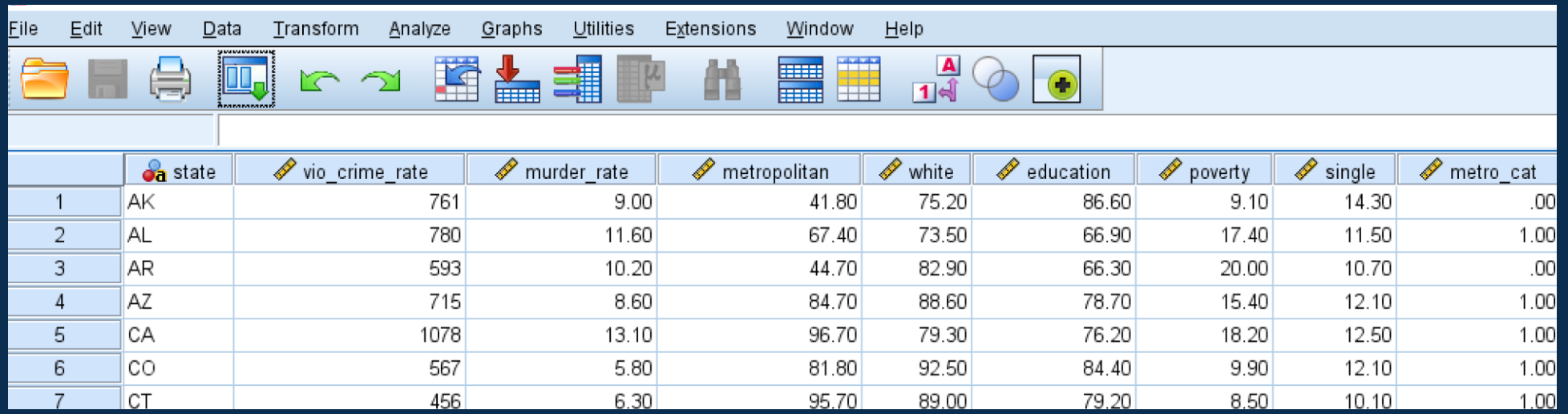
Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	64.124	.943		67.971	.000	62.273	65.975
	Sex	-4.607	.593	-.239	-7.763	.000	-5.772	-3.442

a. Dependent Variable: Weight at Age 16 in Kilograms

Represents the difference in means between males and females, as we change x by one unit (move from male to female), the weight changes by 4.607kg. **On average females weigh 4.607kg less than males (β1 =-4.607. t=-7.763, p<0.001, 95% CI (-5.772, -3.442)**

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_6b_data.sav**.



	state	vio_crime_rate	murder_rate	metropolitan	white	education	poverty	single	metro_cat
1	AK	761	9.00	41.80	75.20	86.60	9.10	14.30	.00
2	AL	780	11.60	67.40	73.50	66.90	17.40	11.50	1.00
3	AR	593	10.20	44.70	82.90	66.30	20.00	10.70	.00
4	AZ	715	8.60	84.70	88.60	78.70	15.40	12.10	1.00
5	CA	1078	13.10	96.70	79.30	76.20	18.20	12.50	1.00
6	CO	567	5.80	81.80	92.50	84.40	9.90	12.10	1.00
7	CT	456	6.30	95.70	89.00	79.20	8.50	10.10	1.00

The dataset contains data from 51 US states, measuring the crime rates and background measures for each state with respect to their

- **violent crime**: per 100,000 population
- **murder** : per 100,000 population
- **poverty**: percent below the poverty line
- **single**: percentage of lone parents

Categorical Predictors

What do we do if we have a predictor that has more than 2 **categories**?
Focus on continuous outcome $\mathbf{y}$ = Violent Crime and categorical explanatory variable $\mathbf{x}$ = Urbanicity.

state	urban
AK	Low
AR	Low
IA	Low
ID	Low
KY	Low
ME	Low
AL	Medium
GA	Medium
KS	Medium
MN	Medium
MO	Medium
NC	Medium
AZ	High
CA	High
CO	High
CT	High
DE	High

The variable **urban** is a categorical variable with three levels "**Low**", "**Medium**" and "**High**"

- Categorical variables which are non binary cannot be included directly in a regression model.
- Need to be recoded into a set of dummy variables
- A dummy (indicator) variable is a binary (0,1) variable indicating a category of the predictor variable.
- A predictor with k levels can be coded as k dummy variables
- Only k-1 dummy variables are necessary to fully represent a categorical predictor.

Categorical Predictors

US crime data. The variable urban is a categorical variable with three levels “Low”, “Medium” and “High”
Let’s consider a linear regression for violent_crime and urban

state	urban
AK	Low
AR	Low
IA	Low
ID	Low
KY	Low
ME	Low
AL	Medium
GA	Medium
KS	Medium
MN	Medium
MO	Medium
NC	Medium
AZ	High
CA	High
CO	High
CT	High
DE	High

The variable **urban** is a categorical variable with three levels “**Low**”, “**Medium**” and “**High**”

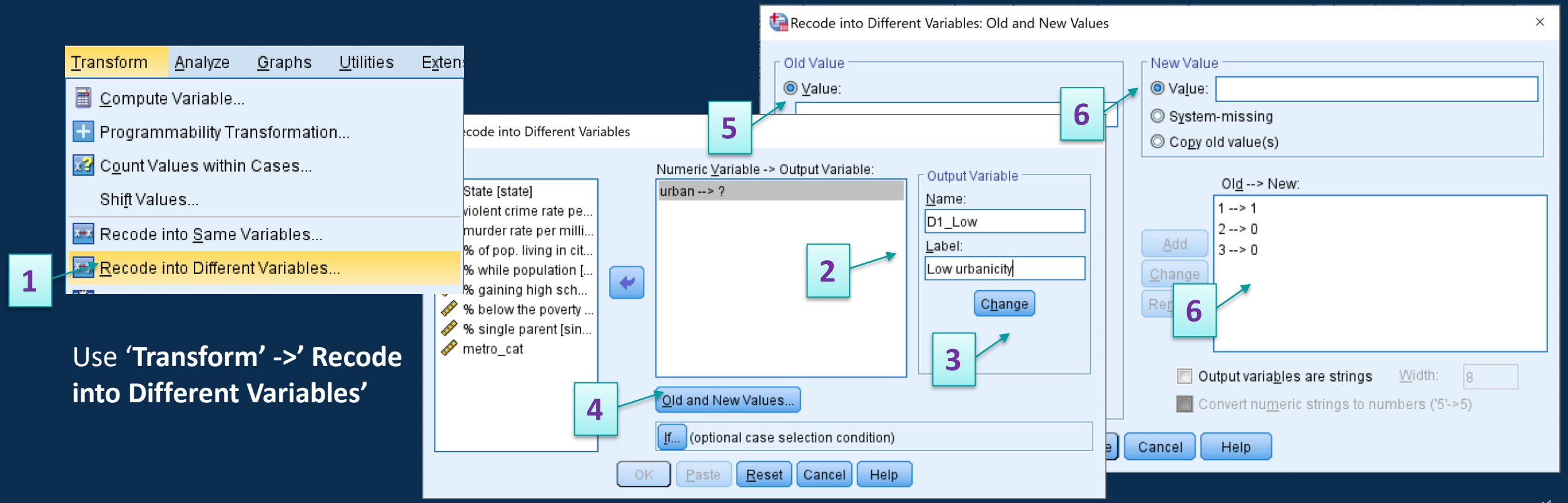
Dummy coding of **`urban`** (*k=3*)

d1	d2	d3
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

SPSS Slide: ‘how to’

Researchers believe there is a relationship between Violent Crime and the level of urbanicity in an area. The variable urban is a categorical variable with three levels “Low”, “Medium” and “High” and needs to be converted to dummy variables to include in the regression.

Step 1: Generating a dummy variable for “Low” urbanicity level in ‘urban’ variable from US crime dataset (We need to repeat this process to create a dummy variable for “Medium” level)



Output and Interpretation Slide

urban	D1_Low	D2_Med	D3_High
2.00	.00	1.00	.00
3.00	.00	.00	1.00
2.00	.00	1.00	.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
2.00	.00	1.00	.00
1.00	1.00	.00	.00

Only 2 dummy variables (e.g. d1 and d2) are needed to represent a variable with 3 levels.

The model will be: $violent\_crime = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \varepsilon$

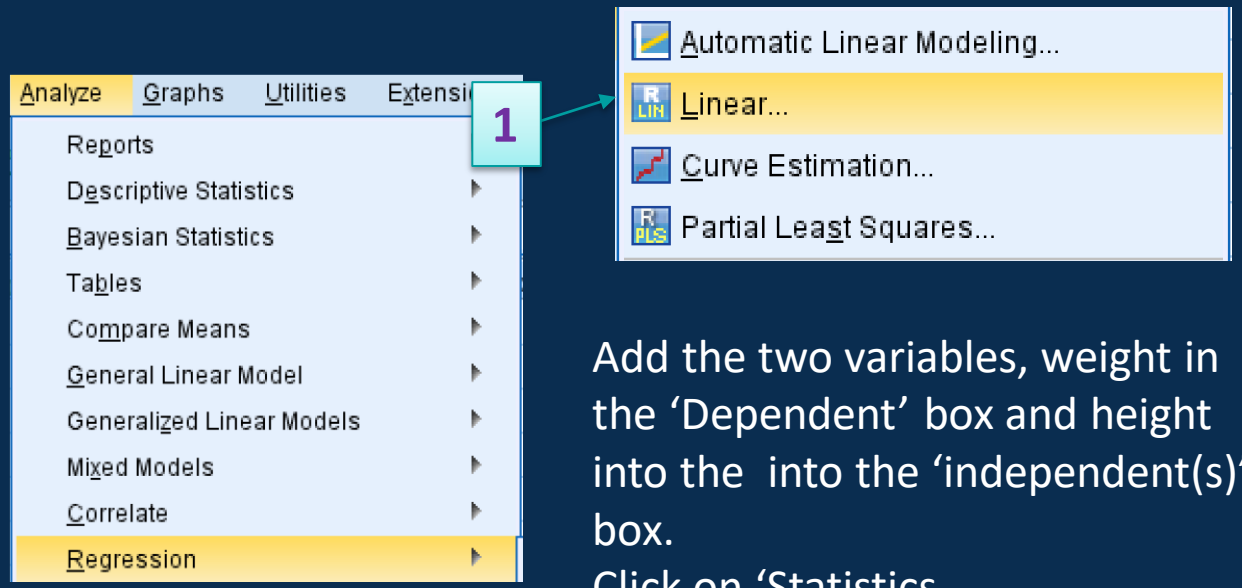
- β_1 will be the difference in mean between "Low" vs. "High" (the latter is called the "reference category")

- β_2 will be the difference in mean between "Medium" vs. "High" (the latter is called the "reference category")

SPSS Slide: ‘how to’

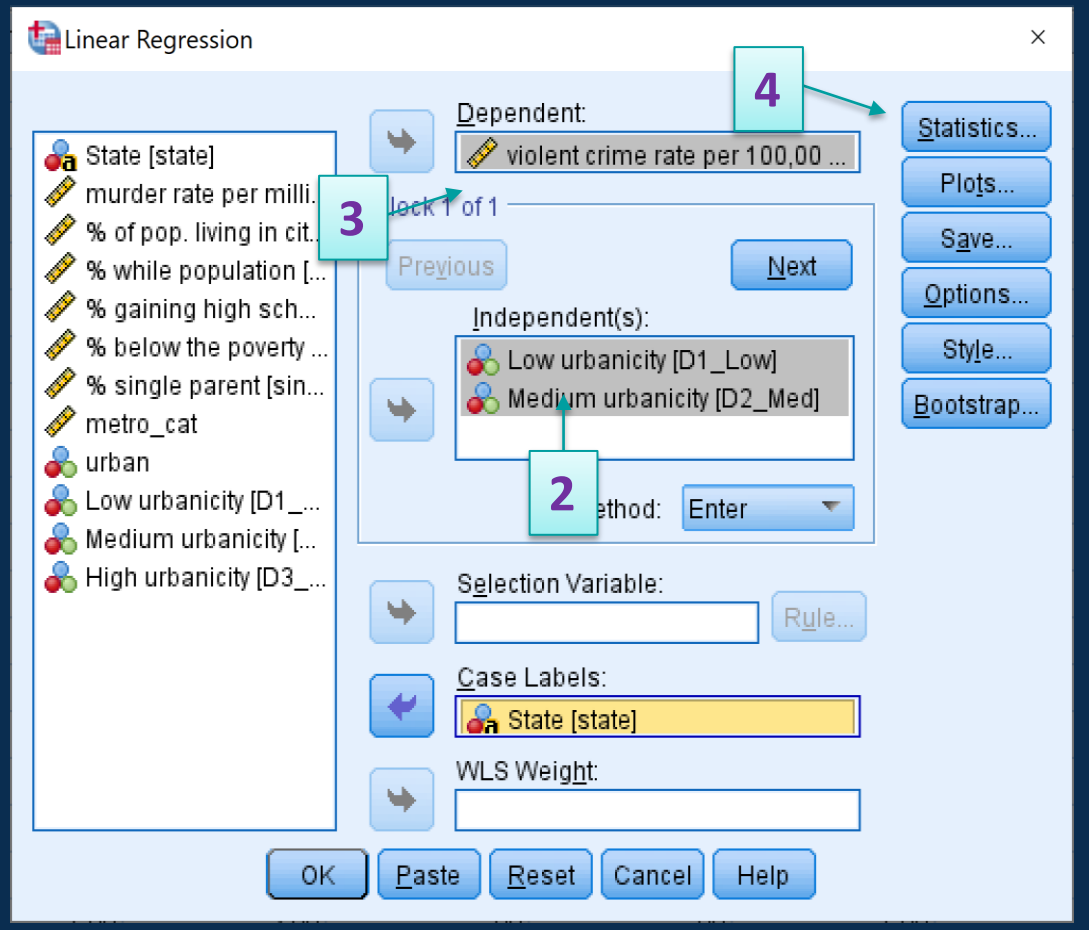
Researchers believe there is a relationship between Violent Crime and the level of urbanicity in an area.

Step 2: Compute a Linear regression model for dependent variable ‘Violent Crime’ and independent variable ‘urban’ using the dummy variables created



Add the two variables, weight in the ‘Dependent’ box and height into the into the ‘independent(s)’ box.
Click on ‘Statistics

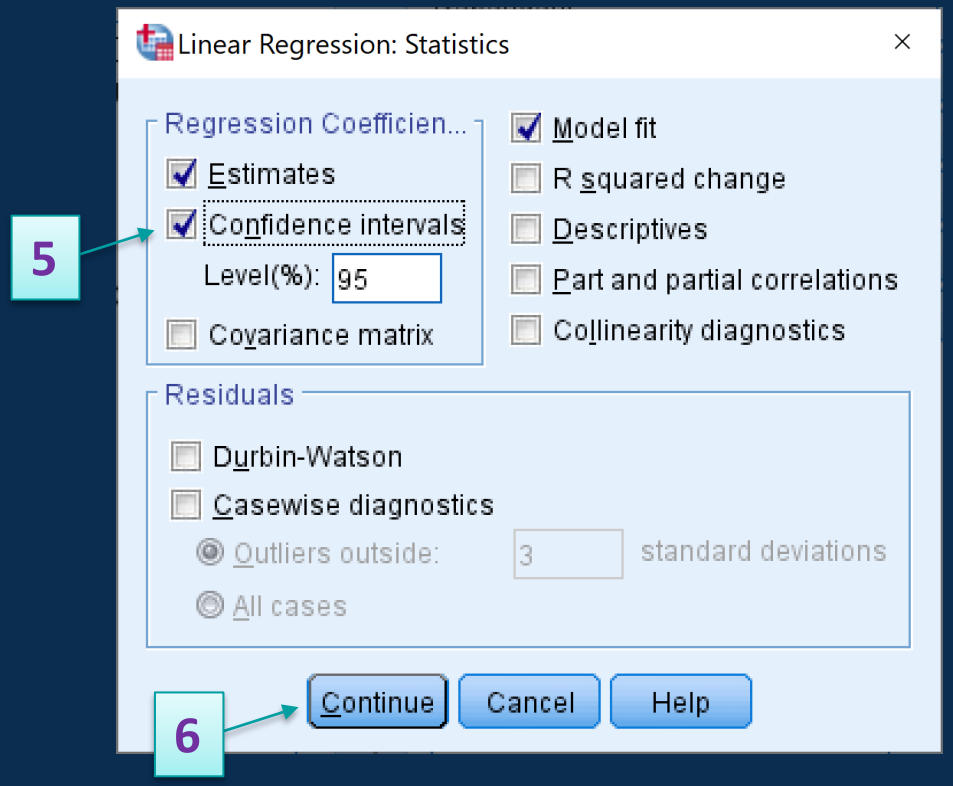
Use ‘Analyse’ -> ‘Regression’ -> ‘Linear’



SPSS Slide: ‘how to’

Researchers believe there is a relationship between Violent Crime and the level of urbanicity in an area

Step 2: Compute a Linear regression model for dependent variable ‘Violent Crime’ and independent variable ‘urban’ using the dummy variables created



In the Statistics tab.
Check the ‘Estimates’
Check the ‘Confidence Intervals’
Click on ‘Continue’
Click on ‘OK’

Output and Interpretation Slide

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.431[a]	.186	.151	410.381

a. Predictors: (Constant), Medium urbanicity, Low urbanicity

b. Dependent Variable: violent crime rate per 100,00 population

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1808632.428	2	904316.214	5.370	.008[b]
	Residual	7915378.052	47	168412.299		
	Total	9724010.480	49			

a. Dependent Variable: violent crime rate per 100,00 population

b. Predictors: (Constant), Medium urbanicity, Low urbanicity

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	749.281	72.546		10.328	.000	603.338	895.224
	Low urbanicity	-498.948	182.569	-.368	-2.733	.009	-866.230	-131.666
	Medium urbanicity	-324.531	138.915	-.314	-2.336	.024	-603.991	-45.071

a. Dependent Variable: violent crime rate per 100,00 population

There is a moderate degree of correlation between Violent Crime and Urbanicity r = 0.431. 18.6% of the variation in Violent crime can be explained by Urbanicity. the regression model statistically significantly predicts the outcome variable i.e., it is a good fit for the data.

On average low urbanised areas have 498.95 less cases of violent crime per 100 000 compared to high urbanised areas (β1 =-498.948. t=-2.733, p<0.009, 95% CI (-866.230, -131.666) , on average med urbanised areas have 324.53 less cases of violent crime per 100 000 compared to high urbanised areas (β2 =-324.531. t=-2.336, p<0.024, 95% CI (-603.991, -45.071)

Knowledge Check

We examined the medical records of participants when they were between 65 and 70 years old, counting the number of health problems they had. Participants were given a questionnaire on how much they've smoked at different times in their life e.g number of cigarettes smoked per day between ages 20 and 50.

The following regression model describes the relationship between health problems (y) and smoking (x)

$$y' = 3.109 + 1.578x \ .$$

- Output from the analysis of the data showed r = 0.77
- Effects of both the intercept and slope show p = .045 and p = 0.049 respectively

1. Write an appropriate Null and Alternative hypothesis for these data.
2. Interpret the coefficients of the regression.
3. How many health problems will a participant be predicted to have if the number of cigarettes they smoke is 10 and 30. Calculate a confidence interval for the prediction given the s.e. is 0.435

Knowledge Check Solutions

We examined the medical records of participants when they were between 65 and 70 years old, counting the number of health problems they had. Participants were given a questionnaire on how much they've smoked at different times in their life e.g number of cigarettes smoked per day between ages 20 and 50.

1. Write an appropriate Null and Alternative hypothesis for these data.
   H0: There is no linear association between health problems and amount of smoking e.g. the slope $\beta 1$ in the population equals to 0
   Ha: There is a linear association between number of health problems and amount of smoking e.g. the slope $\beta 1$ in the population does not equal to 0

2. Interpret the coefficients of the regression.
   β0 = 3.109 The intercept ($\beta 0$), is the extrapolated number of health problems for a participant who does not smoke, this suggests that if a participant is a non-smoker they will have approx. 3 health problems.
   β1 = 1.578 The estimated slope coefficient (β1), suggests a increase of 1 cigarette smoked is associated with a 1.578 increase to number of health problems

3. How many health problems will a participant be predicted to have if the number of cigarettes they smoke is 10 and 30. Calculate a confidence interval for the prediction given the s.e. is 0.435

10 cigarettes will lead to 3.109 + (10 x 1.578) = 18.89 health problems
95% CI (18.89 ± 1.96x0.435) = (18.04,19.74)

30 cigarettes will lead to 3.109 + (30 x 1.578) = 50.45 health problems
95% CI (50.45 ± 1.96x0.435) = (49.60,51.30)

References

Field (2017) Discovering Statistics using SPSS, 5th Ed.
Chapter 8: Correlation
Chapter 9: The Linear Model (Regression)

Agresti and Finlay (2014) Statistical Methods for the Social Sciences, 4th Ed.
Chapter 9: Linear Regression and Correlation

Acock (2018) A Gentle Introduction to Stata, 6th Ed.
Chapter 8: Bivariate correlation and regression

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