🔷 1. Scatterplots

Scatterplots visually represent the relationship between two **continuous variables**.

- **Purpose:** Check direction, strength, and linearity.
- **Axes:**
  - X-axis = **Independent variable**
  - Y-axis = **Dependent variable**

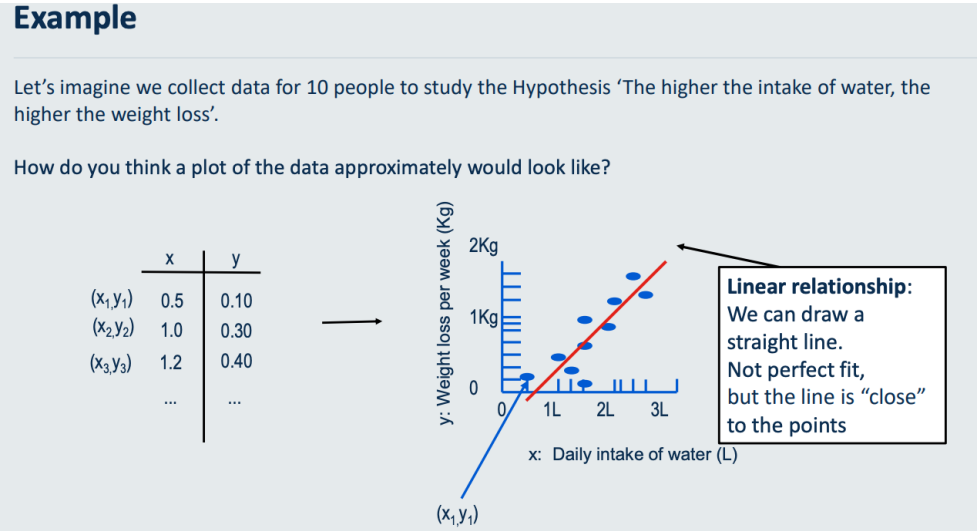
📌 SPSS Command: Create a Scatterplot

```sql
Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter → Define
```

- Put dependent variable (Y) in Y-axis box (e.g., `weight`)
- Put independent variable (X) in X-axis box (e.g., `height`)
- Label cases if needed → Click OK

Also...

Role	Example Question	Variable Example
Explanatory variable	"Does time spent studying affect test scores?"	⏱️ Hours of study (X)
Response variable	"What is the outcome?"	📝 Test score (Y)



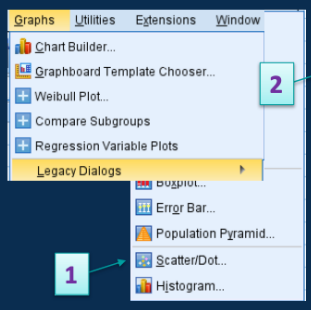
• Plotting the data is essential to understand and visually assess the relationship between pairs of continuous variables

• The plot of data points (x, y) with x and y being continuous is called a scatterplot

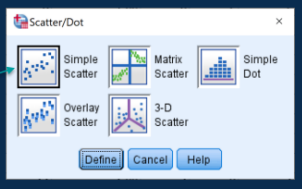
SPSS Slide: 'how to'

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.

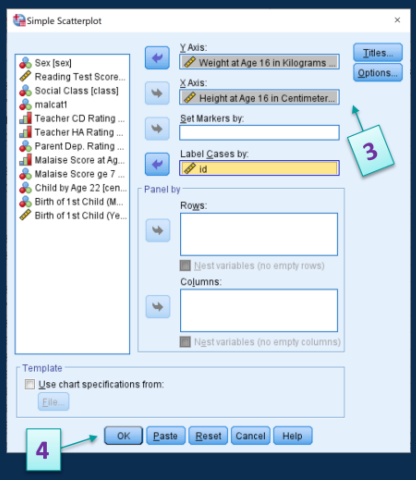
Step 1: Generate a Scatter Plot for variables 'height' and 'weight' from the data



Use 'Graphs' -> 'Legacy Dialogs' -> 'Scatter/Dot'

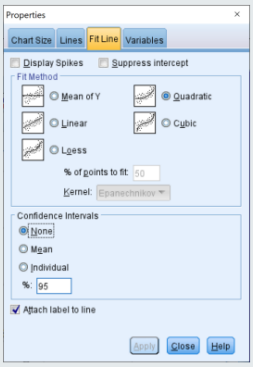


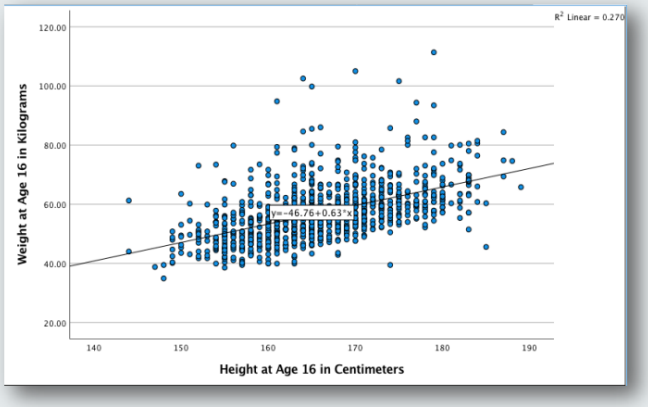
Click on 'simple scatter'
Click on 'define'.
Add the dependent variable (weight) into the 'y-axis' box.
Add the independent variable (height) into the 'x-axis' box.
'Label cases by' ID.
Click 'Ok'



13

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".

14

🔷 2. Correlation Analysis

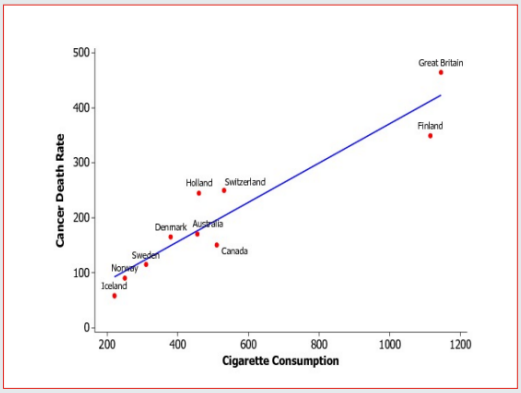
🔶 Pearson's Correlation (r)

- **Use when:** Both variables are continuous and **normally distributed**
- **Hypotheses:**
  - H_0: r = 0 (no correlation)
  - H_1: r ≠ 0 (some 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]



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.

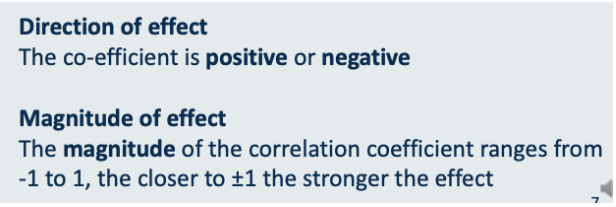
Linear association
The points follow the line of best fit in a linear manner, thus there is **linear association**

When to use it

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

Range	Interpretation
0.80 – 1.00	Very strong +ve
0.60 – 0.79	Strong +ve
0.40 – 0.59	Moderate +ve
0.20 – 0.39	Weak +ve
0.00 – 0.19	Very weak/none

... negative values = same strength but in –ve direction

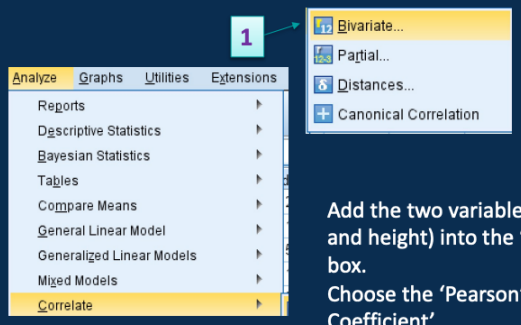


- **Assumptions:**
  - Both variables continuous
  - Linear relationship
  - No significant outliers
  - Normal distribution

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

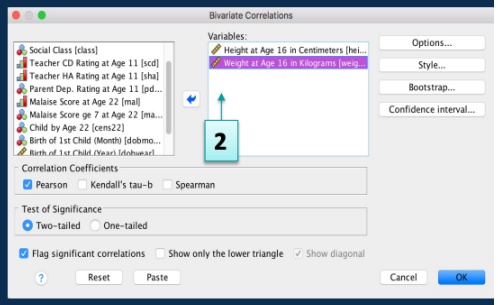


Use 'Analyse' -> 'Correlate' -> 'Bivariate'

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 (r_s)

- **Use when:** At least one variable is **ordinal** or **not normally distributed**
- **Hypotheses:**
  - H_0: $\rho = 0$
  - H_1: $\rho \neq 0$

📌 SPSS Command: Compute Pearson or Spearman Correlation

```nginx
Analyze → Correlate → Bivariate
```

- Add both variables
- Tick: Pearson (default) or Spearman
- Click OK

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:

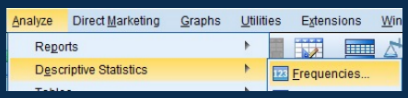
- H_0: the correlation in the population equals to 0
- H_a: the correlation in the population does not equal to 0

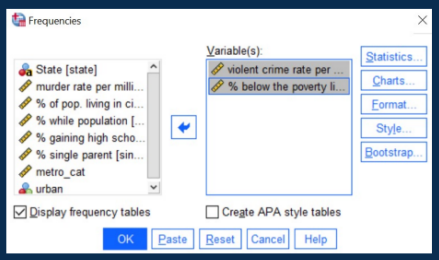
$$t = \frac{r\sqrt{n-2}}{1-r^2} \qquad \text{df=N-2}$$

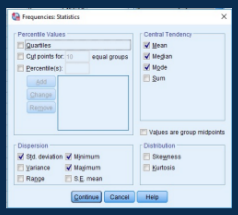
SPSS Slide: 'how to'

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.

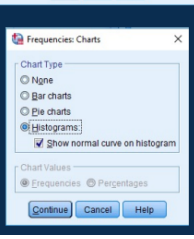
Step 1: Check the suitability of the data.





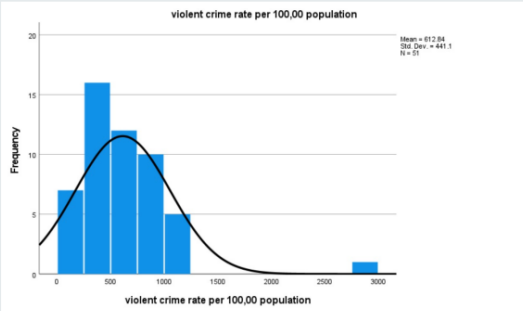


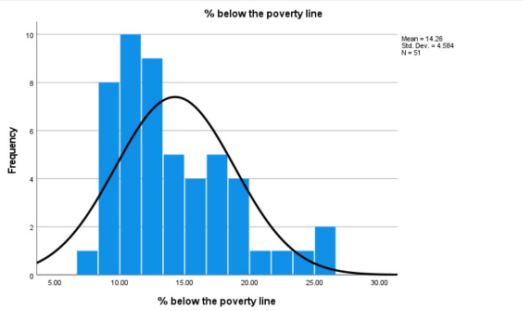
In 'Statistics' ask for descriptive statistics



In 'Charts' ask for a Histogram

Output and Interpretation Slide

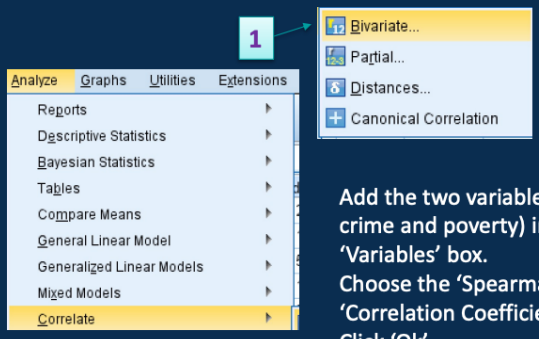




'Violent Crime' is a positively skewed variable. 'Poverty' is a positively skewed variable. Pearson's product moment correlation coefficient is unsuitable for this data. Use Spearman's correlation coefficient instead.

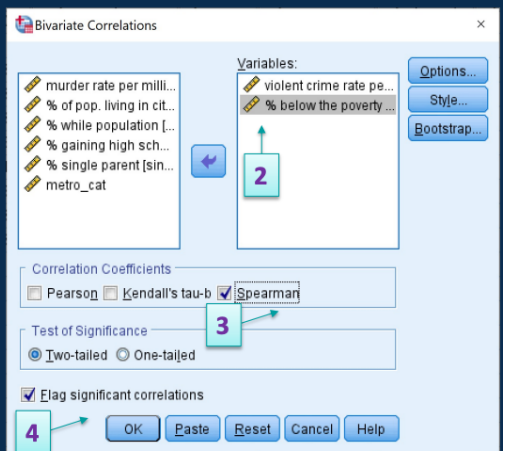
According to the researchers, in the population from which our data came, they believe there is a relationship between Violent crime measured per 100,000 and the percentage of people below the poverty line per 100,000.

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.

🔷 3. Spurious Correlation

> Just because two variables are correlated doesn't mean one causes the other.

- A third variable may explain both — **correlation ≠ causation**

🔷 4. Simple Linear Regression

Model:

$Y = \beta_0 + \beta_1 X + \varepsilon$

- **Y**: Dependent (continuous)
- **X**: Independent (continuous or dummy coded)
- **β_0**: Intercept (Y when X = 0)
- **β_1**: Slope (change in Y for 1 unit change in X)
- **ε**: Error term

Hypotheses (for slope β_1):

- H_0: $\beta_1 = 0$ → No linear association
- H_1: $\beta_1 \neq 0$ → Linear association

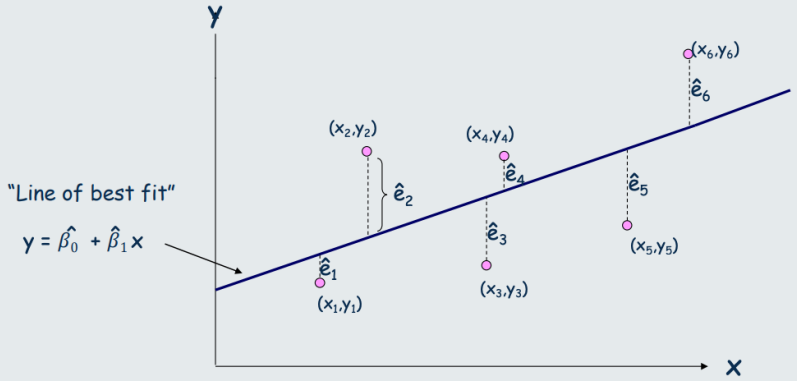
When to use it

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

- ε is called the **residual** (distance between the points and the line).
- $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are together known as **regression coefficients.**

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 $\boldsymbol{\widehat{\beta_1}}$.



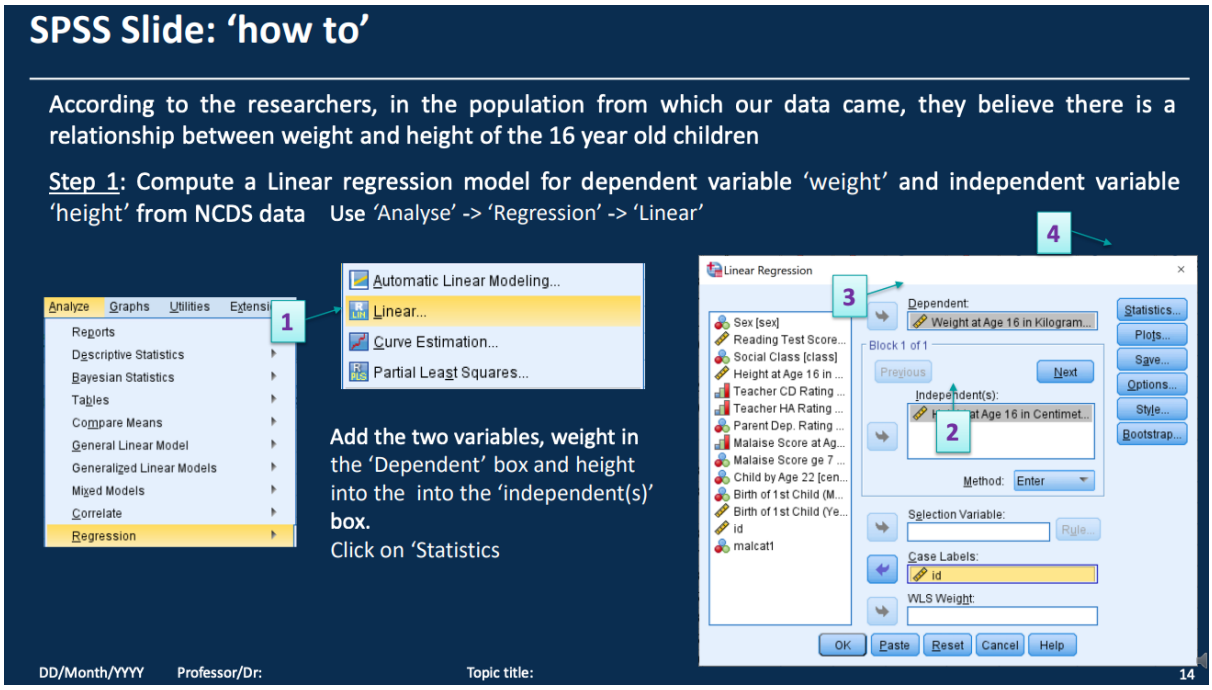
Assumptions:

- Linearity
- Homoscedasticity
- Normality of residuals
- Independent errors

📌 SPSS Command: Linear Regression

```nginx
Analyze → Regression → Linear
```

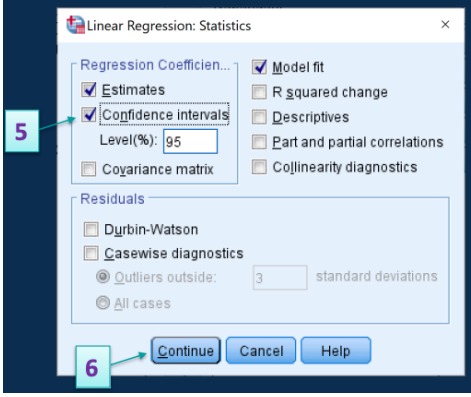
- Dependent = outcome (e.g., weight)
- Independent = predictor (e.g., height)
- Click 'Statistics' → tick “Estimates” and “Confidence Intervals”
- Click OK



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'

5. Interpretation of Output

- **R** = correlation between actual Y and predicted Y
- **R^2** = proportion of variance in Y explained by X
- **β_1 (slope):** For each 1 unit increase in X, Y changes by β_1
- **t-value and p-value:** Test whether β_1 is significantly different from 0
- **Confidence Interval for β_1:**

$$95\%\ \text{CI} = [\beta_1 \pm 1.96 \times SE(\beta_1)]$$

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

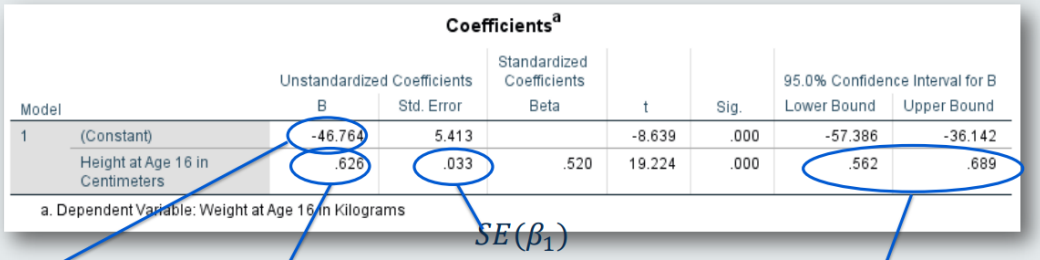
This table provides the R and R^2 values. The R value represents the simple correlation and is 0.520 which indicates a moderate degree of correlation.

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 B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B 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

$SE(\beta_1)$

β_0 β_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.

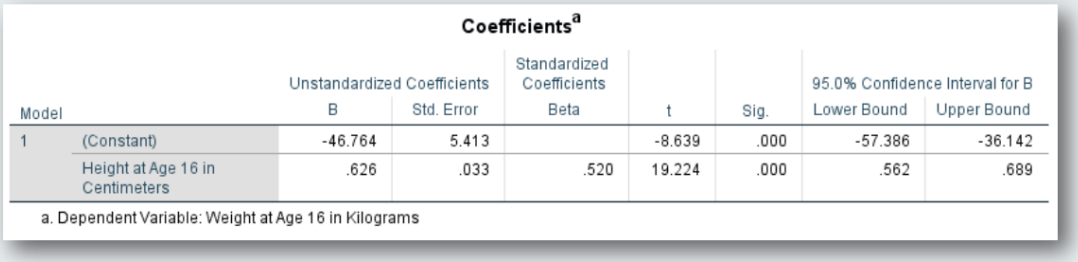
The intercept (β_0), is the extrapolated weight for a 16 year old of zero height.

In addition to getting point estimation for $\beta 1$, it is possible to calculate a confidence interval for the slope parameter
The confidence interval formula is:

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

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 B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B 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))

🔷 6. Predictions Using Regression

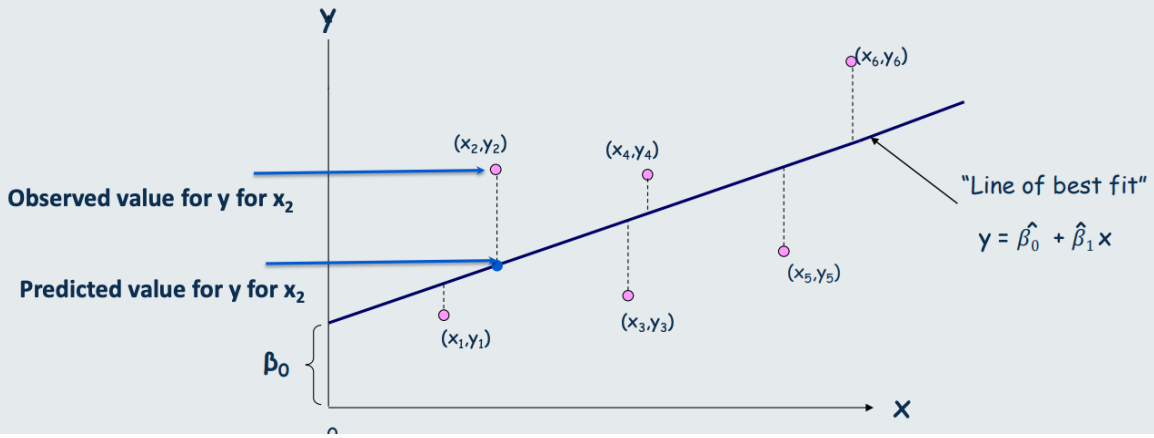
Formula:

$$\hat{y} = \beta_0 + \beta_1 \cdot x$$

📌 SPSS Command: Save Predictions

```java
Regression Dialog → Click 'Save' → Tick:
✓ Unstandardized predicted values
✓ Prediction intervals (mean)
```

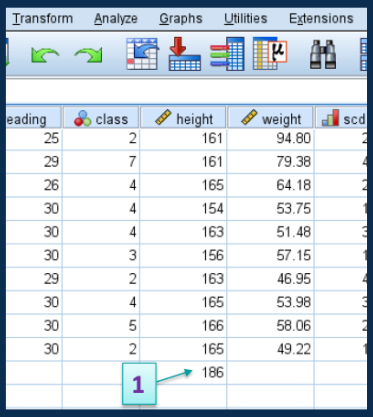
Regression models are used to predict new cases.
The predicted value $\hat{y}$ for a new observation x is its corresponding value on the regression line.



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

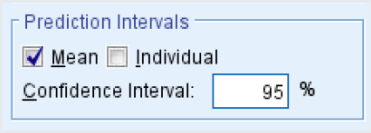


Step 2) Use **Analyse -> Regression -> Linear**
Step 2) Put 'weight' in dependent, and 'height' in independent.
Click '**Save'**, select 'Prediction values' 'Unstandardised' and 'Prediction intervals' 'mean'.
Click on 'Continue'
Click on 'OK'

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) $\hat{y} = 69.58\ \mathbf{kg}$ 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.**

🔷 7. Categorical Predictors in Regression

A. Binary Predictor

- E.g. Gender (Male = 1, Female = 0)
- β_1 = difference in group means

B. >2 Categories: Create Dummy Variables

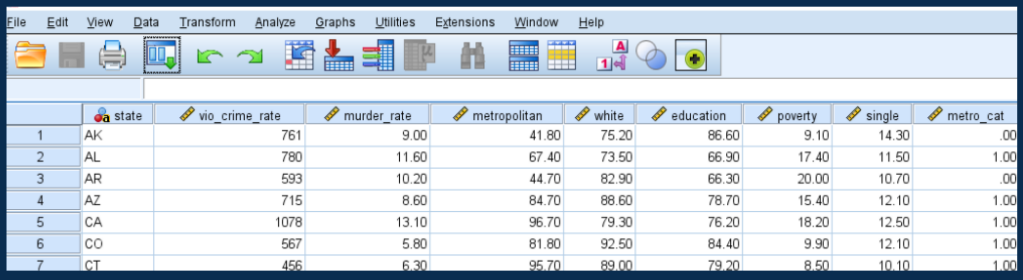
- k categories → create k–1 dummy variables

📌 SPSS Command: Recode to Dummy

```pgsql
Transform → Recode into Different Variables
```

- For each category, set 1 = present, 0 = all others

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 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 B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B 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)**

Categorical Predictors

What do we do if we have a predictor that has more than 2 **categories**?
Focus on continuous outcome **y** = Violent Crime and categorical explanatory variable **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**"

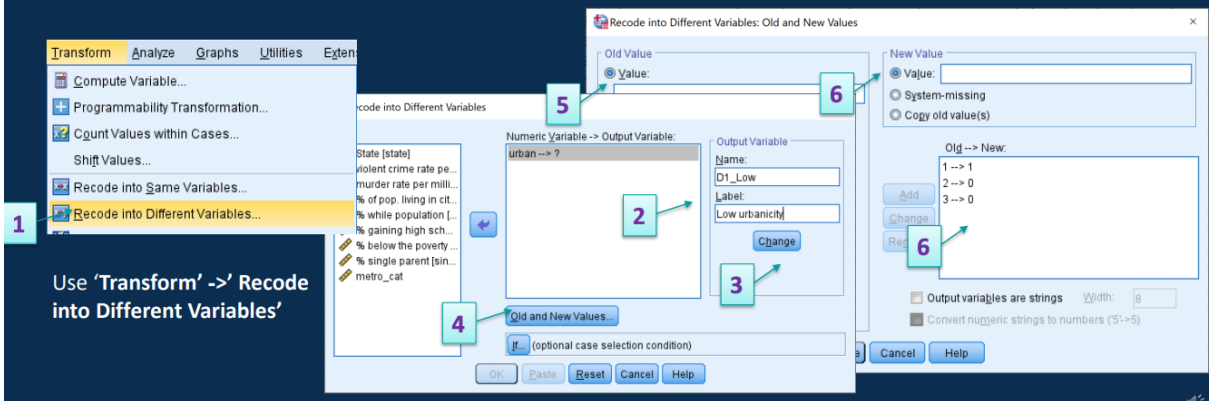
Dummy coding of **urban** (*k=3*)

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.

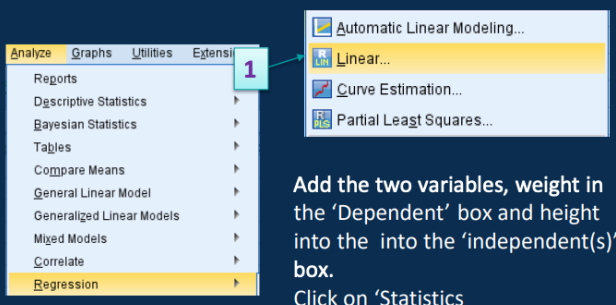
The model will be: $violent\_crime = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \varepsilon$

- 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'

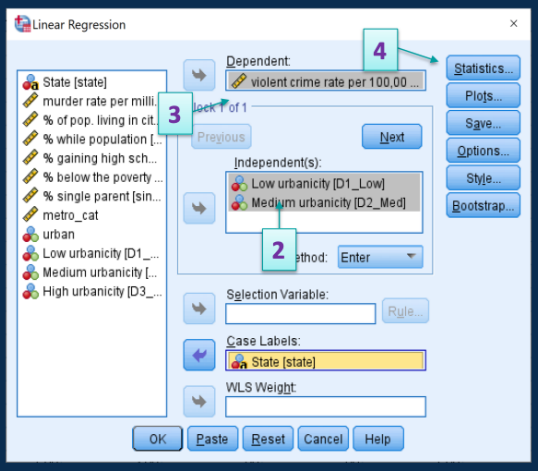
Researchers believe there is a relationship between Violent Crime and the level of urbanicity in an area.

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



Use 'Analyse' -> 'Regression' -> 'Linear'

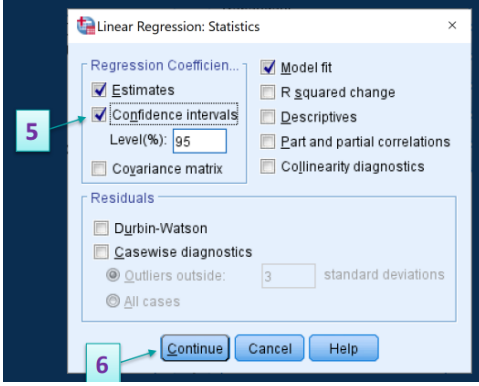
Add the two variables, weight in the 'Dependent' box and height into the into the 'independent(s)' **box.**
Click on 'Statistics



SPSS Slide: 'how to'

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 B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B 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)

Quiz:

A simple linear regression model is useful for:

Select one:

- a. Estimating the association between a continuous outcome and a continous explanatory variable.
- b. Predicting a value of an explanatory variable, given a value of the dependent variable.
- c. Predicting a value for the independent variable, given a value for the dependent variable.
- d. Predicting the outcome of any dependent variable with continuous predictor variables. ×

Your answer is incorrect.

The correct answer is: Estimating the association between a continuous outcome and a continous explanatory variable.

Role	Example Question	Variable Example
Explanatory variable	"Does time spent studying affect test scores?"	⏱️ Hours of study (X)
Response variable	"What is the outcome?"	📝 Test score (Y)

🧠 What is an Explanatory Variable?

An **explanatory variable** (also called an **independent variable**, **predictor**, or **X variable**) is:

> A variable that you think might **explain**, **influence**, or **predict** changes in another variable (called the **response** or **dependent variable**).

✅ Why Option A is correct:

Simple linear regression is used when you want to:

- **Quantify the relationship** between one continuous outcome (dependent variable, Y)
  and one continuous predictor (independent variable, X).
- It estimates **how much Y changes for a one-unit change in X**, which is **β_1 (the slope)** in the regression equation.

So it's **not just about prediction** — it's also about **estimating the nature of the association** between two variables.

📌 Example:
If you're trying to understand **how weight (Y)** changes depending on **height (X)**, linear regression helps **estimate that association** — not just blindly predict weight.

Concept	Explanation
Simple linear regression	Models the **relationship** between 1 continuous predictor and 1 continuous outcome
Main use	Estimate **how much Y changes** when X changes
Also used for	Making predictions (but that's secondary)
Predictor can be	Continuous or categorical (with dummy coding)
Outcome must be	Continuous

The coefficients of the least squares regression line are determined by minimising the sum of the squares of the:

Select one:

- a. Differences ✕
- b. Residuals
- c. y-coordinates
- d. x-coordinates

Your answer is incorrect.

The correct answer is: Residuals

Topic Knowledge Check
Quiz 6 Solutions

Q1. A [scatter plot] is the appropriate plot to visualise the relationship between two continuous variables.

Q2. The Pearson coefficient is the appropriate measure to estimate the linear correlation between any two continuous variables.

False, only when those two continuous variables are normally distributed. If one or either are not normally distributed we would use Spearman's Correlation coefficient which measures the monotonic relationship between variables.

Monotonic relationships are where: One variable increases and the other increases. Or, One variable decreases and the other decreases

Q3. A simple linear regression model is useful for ...

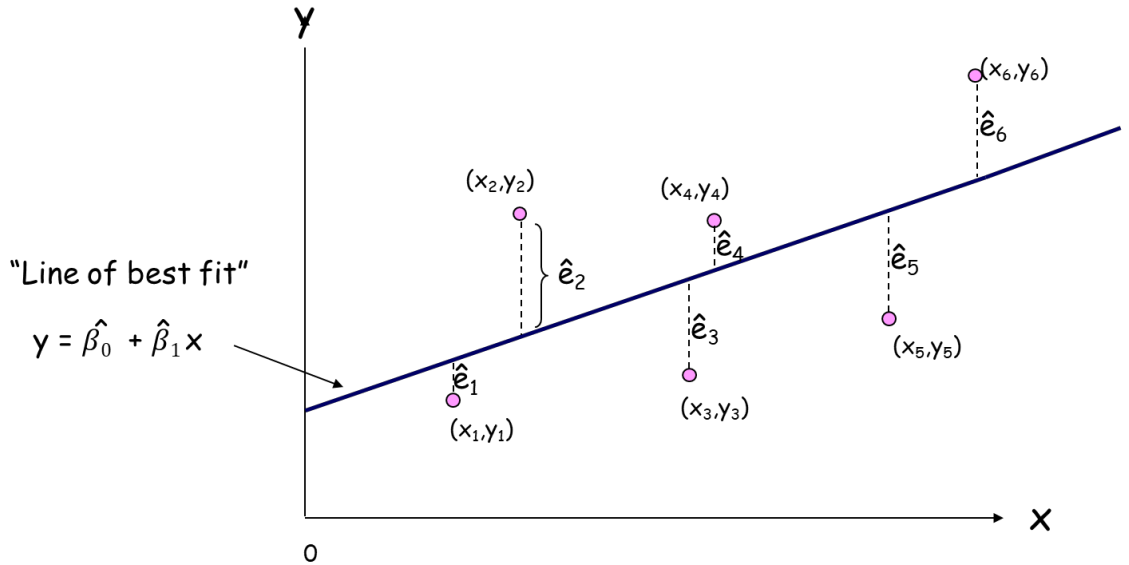
Estimate the association between an outcome and a explanatory variable.

The resulting regression equation can be used to make a prediction of the dependent variable based on a value of the independent variable. Simple linear regression only estimates association and cannot prove causation.

Q4.The coefficients of the least squares regression line are determined by minimising the sum of the squares of the...

Residuals

- 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 we used to estimate the regression line that **minimises** the **squared residuals** $(\varepsilon_1^2 + \varepsilon_2^2 + \cdots \varepsilon_n^2)$ to give us the estimates for $\widehat{\beta_0}$ and $\widehat{\beta_1}$.



Q5. The equation of the regression line is equation. Predict y when x = 5.

$$\hat{y} = 1.2x - 3.4$$

Est. y = 1.2 x5 – 3.4

Est. y = 6 – 3.4

Est. y = 2.6