Statistical Analysis 6: Simple Linear Regression

Research question type: When wanting to predict or explain one variable in terms of another

What kind of variables? Continuous (**scale/interval/ratio**)

Common Applications: Numerous applications in finance, biology, epidemiology, medicine etc.

Example 1:

A dietetics student wants to look at the relationship between calcium intake and knowledge about calcium in sports science students. Further she wants to know if knowledge about calcium can be used to predict calcium intake of the students. Table 1 shows the data she collected.

Table 1: Dietetics study data

Respondent number	Knowledge score (Out of 50)	Calcium intake (mg/day)	Respondent number	Knowledge score (Out of 50)	Calcium intake (mg/day)
1	10	450	11	38	940
2	42	1050	12	25	733
3	38	900	13	48	985
4	15	525	14	28	763
5	22	710	15	22	583
6	32	854	16	45	850
7	40	800	17	18	798
8	14	493	18	24	754
9	26	730	19	30	805
10	32	894	20	43	1085

Research question: Does knowledge about calcium **predict** calcium intake in sports science students?

In this example there is a single predictor variable (knowledge about calcium) for one response variable (calcium intake).

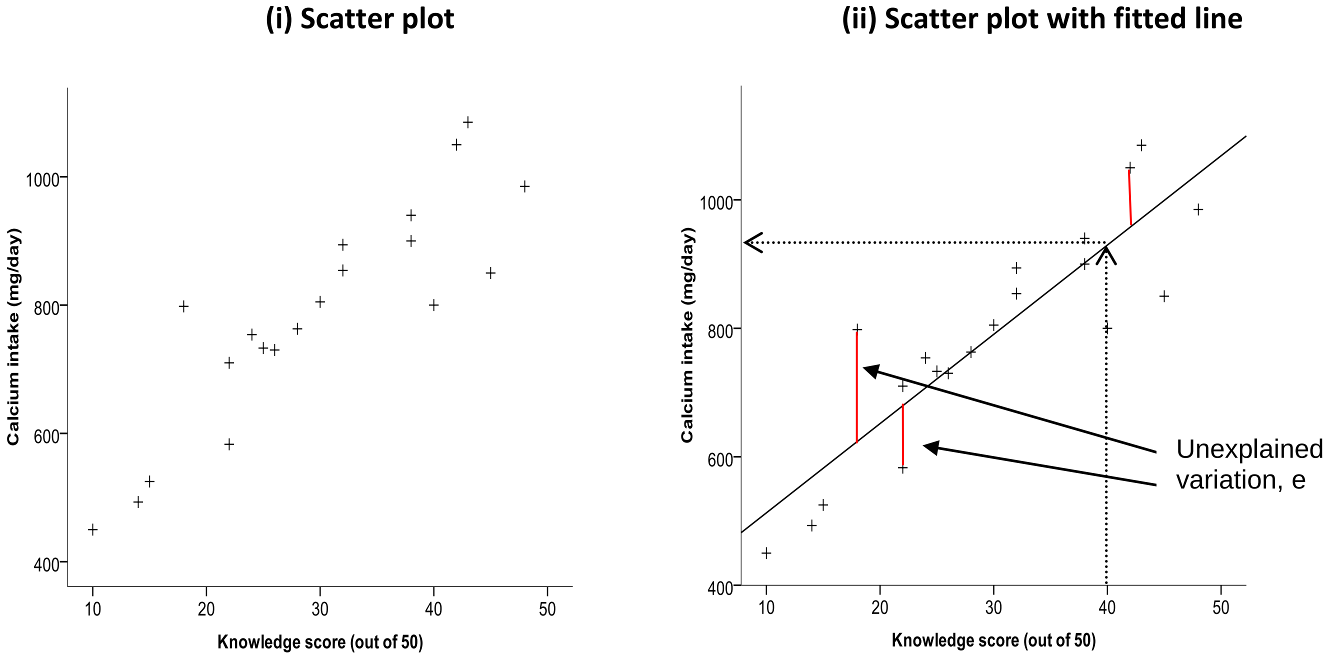
It can be seen from the scatter plot in Figure 1(i) that the calcium intake seems to increase as the knowledge scores increase, and that, although there is some variation, the relationship roughly follows a straight line (described as **linear**).

A straight line can be written in terms of its two variables, its gradient (**slope**) and where it crosses the vertical axis (**intercept**), plus an allowance for the unexplained variation (e). See Figure 1(ii), which shows a fitted line, sometimes termed the line of best fit, calculated from all the given data.

Data can be found in W:\EC\STUDENT\ MATHS SUPPORT CENTRE STATS WORKSHEETS\calcium.sav

Steps in SPSS (PASW):

Step 1: Draw a **scatter plot** of the data to check for an underlying straight line relationship. If there is NO underlying straight line apparent, ask for advice on how to proceed.



(i) Scatter plot

(ii) Scatter plot with fitted line

Figure 1: Calcium intake against Knowledge score

A scatter plot can be drawn in SPSS, using the **Graphs> Chart Builder** options

- choose **Scatter/Dot**
- drag the Simple Scatter plot into the plotting region
- drag the Predictor (independent) variable (in this case Knowledge score) into the X-axis box
- drag the Response (dependent) variable (in this case Calcium intake) into the Y-axis box
- click OK

Step 2: The fitted line can be added to the chart from the SPSS Chart Editor:

- double-click on the chart in the SPSS Output window to open the Chart Editor [The above charts have also had changes made to the marker and text size]
- choose Fit Line at Total from the Elements menu – make sure Linear Fit Method is selected in the Properties window
- close the Chart Editor window to update the chart in the Output window.

The fitted line gives an idea of how much variation there is in the observed values compared to the line. The variation is assumed to be just in the response variable, so all the unexplained variation is shown by the difference between the plotted point and point on the fitted line from the value on the horizontal axis.

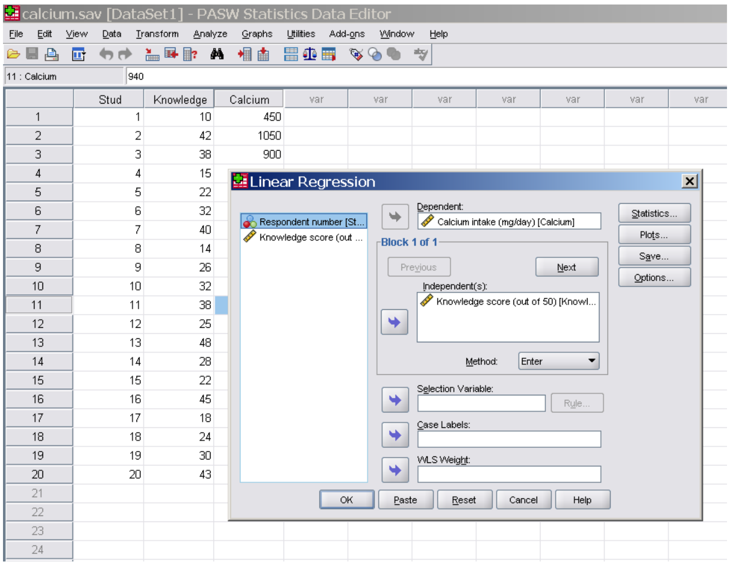
Look at Figure 1(ii) and follow a Knowledge score of 40, up to the observed value at 800, and compare this to value on the line (following the arrow) at approximately 930. So there is about 130 mg/day discrepancy (**residual**). Other discrepancies are smaller or larger.

Further exploration of these residuals can be carried-out to check the validity of the regression model – ask for more details.

Step 3: Estimating the model

Estimates of the intercept and slope can be made from the observed data using a technique called Simple Linear Regression, which aims to get a model such that the line fits the data in the 'best' way.

- choose **Analyse>Regression>Linear** – see right
- move the response variable (Calcium intake) into the **Dependent** box, and
- move the predictor variable (Knowledge score) into the **Independent(s)** box
- **OK**



calcium.sav [DataSet1] - PASW Statistics Data Editor

11 : Calcium 940

	Stud	Knowledge	Calcium	var	var	var	var	var	var	var
1	1	10	450							
2	2	42	1050							
3	3	38	900							
4	4	15								
5	5	22								
6	6	32								
7	7	40								
8	8	14								
9	9	26								
10	10	32								
11	11	38								
12	12	25								
13	13	48								
14	14	28								
15	15	22								
16	16	45								
17	17	18								
18	18	24								
19	19	30								
20	20	43								
21										
22										
23										
24										

Linear Regression

Dependent: Calcium intake (mg/day) [Calcium]

Independent(s): Knowledge score (out of 50) [Knowl...

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

Results:

Table 2 shows some of the output from the regression analysis

Table 2: Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	373.743	55.067	6.787	.000	258.051	489.435
	Knowledge score (out of 50)	13.897	1.748	7.951	.000	10.225	17.569

The **estimates** of the intercept and slope are given in the 'B' column:

Intercept = Constant = **373.743**, and

Slope = Knowledge score = **13.897**

The p-value (< 0.001) next to the knowledge score implies that this variable is significant in explaining calcium consumption.

Conclusion:

Our estimated model is (on average):

(Calcium intake) = 373.7 + 13.90 x (Knowledge score)

This 'model' can be used as follows:

A student with knowledge of calcium score equal to 30 (out of 50) could be predicted to have an approximate calcium intake of

= 373.7 + 13.90 x 30
= 373.7 + 417.0
= 790.7
i.e. approximately 791 mg/day

For prediction, knowledge scores should lie in the range of the data collected on knowledge.

Confidence Intervals:

Obviously this model is subject to uncertainty, as the observed points did not all lie on a perfect straight line, so the coefficients for the intercept and slope are only estimates of the 'true' value.

Confidence intervals (CIs) can be calculated for these values to give a range of possible values; choose the Statistics button in the Linear Regression dialog box, and check the Confidence Intervals box, Level(%) = 95. This means that If we were to do this experiment 100 times, 95 times the true value for the intercept and slope would lie in the 95% CI.
95% CIs were calculated for the two coefficients in this example (Table 2 above) as:

intercept: CI=(260, 490), slope: CI=(10, 18), both corrected to 2 sig fig.

In other words, we expect that the true estimates are in the intervals (260, 490) and (10, 18). These CIs should be inspected in context by the reader for practical importance/usefulness.

Validity of simple linear regression:

This is based on several assumptions:

- both sets of data are measured at continuous (scale/interval/ratio) level
- data values are independent of each other; ie, only one pair of readings per participant is used
- there is a linear relationship between the two variables
- the residuals are normally distributed
- there should be some rationale for assuming one variable depends on another in a linear way

Further exploration of variation can be performed – ask for details if required.

More than one predictor variable can be used in a model with the same provisos as above – see Multiple Linear Regression. There are also other regression modelling techniques for data not considered to be at continuous/interval/ratio level.

Comments:

Table 3: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.882[a]	.778	.766	84.348

a. Predictors: (Constant), Knowledge score (out of 50)

Another SPSS output table – see Table 3 – gives a useful value 'R square', or the '**coefficient of determination**'. In this example R^2 = 0.778 (or a value 0.766, adjusted for generalisation).

This value can be interpreted as 78% (or 77%) of variation in calcium intake can be explained by knowledge score. A 95% CI can be calculated for this value, too. [But not in SPSS]

The remaining 22% (or 23%) arises from other variables not taken into account in the analysis.

Note we CANNOT assume that knowledge about calcium CAUSES the increase in calcium intake.

Pearson's coefficient of linear correlation R = 0.882 is also given in the output. It is a measure of the strength of the linear relationship between the predictor and response variable.