statstutor

community project

encouraging academics to share statistics support resources

All stcp resources are released under a Creative Commons licence

stcp-gilchristsamuels-4

The following resources are associated:

Pearson Correlation worksheet

Simple Linear Regression – Additional Information worksheet

Simple Linear Regression

Research question type: When using one variable to predict or explain another variable in terms of a linear relationship; looking for a significant linear relationship between a dependent variable and an independent variable.

What kind of variables: Continuous (scale)

Common Applications: Simple linear regression is the simplest model for predicting the value of one variable in terms of another

Definition

Simple linear regression estimates the coefficients b_0 and b_1 of a linear model which predicts the value of a single dependent variable (y) against a single independent variable (x) in the form:

$$y = b_0 + b_1x$$

b_0 is the intercept of the straight line (the value of y when it crosses the Y-axis) whilst b_1 is its slope.

Example: Dietetics

A dietetics student wants to look at the relationship between calcium intake and knowledge about calcium in sports science students. Also, 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.

Research question: How can 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).

© Mollie Gilchrist and Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Based on material provided by Loughborough University Mathematics Learning Support Centre and Coventry University Mathematics Support Centre

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

Steps in SPSS

Step 1: Draw a **scatterplot** 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. The scattering of points also lie approximately within an ellipse of cigar shape tilted in the direction of the fitted line for the assumptions of linear regression to be satisfied.

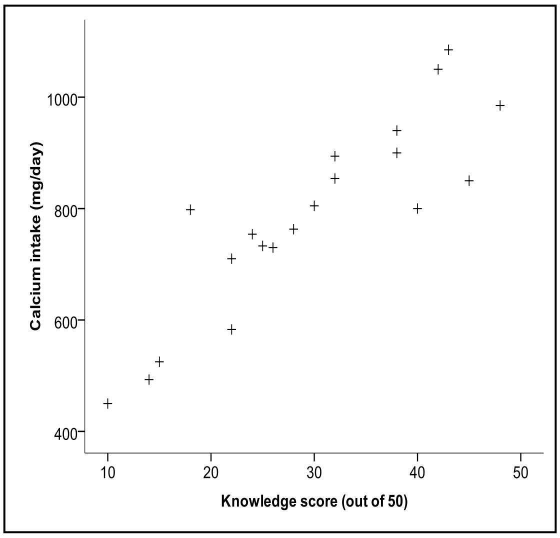
A scatterplot can be drawn in SPSS, using the Graphs – Chart Builder option (note: alternatively, the legacy chart builder can be used.):

- 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 – this create a scatter plot like the one shown in Figure 1

It can be seen from this scatterplot 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 a **linear relationship**). The value of each plotted point also includes an allowance for the unexplained variation, also known as a residual or **error**.

The scatter plot in Figure 2 shows a fitted line, sometimes called the **line of best fit**. This is the line given by the coefficients b_0 and b_1.

Figure 1



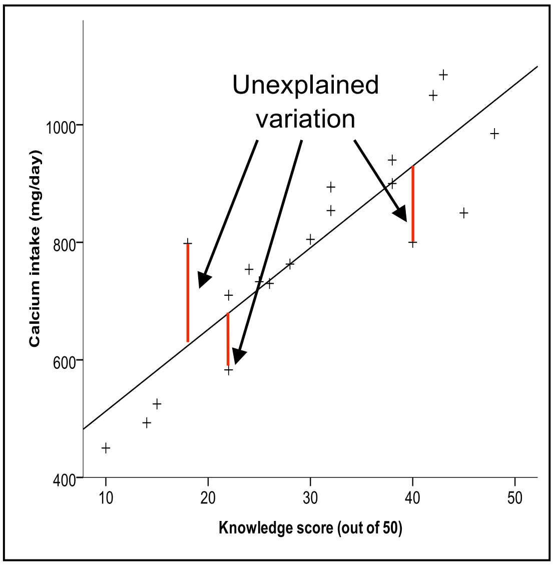
Step 2: The fitted line in Figure 2 can be added to the chart from the SPSS Chart Editor:

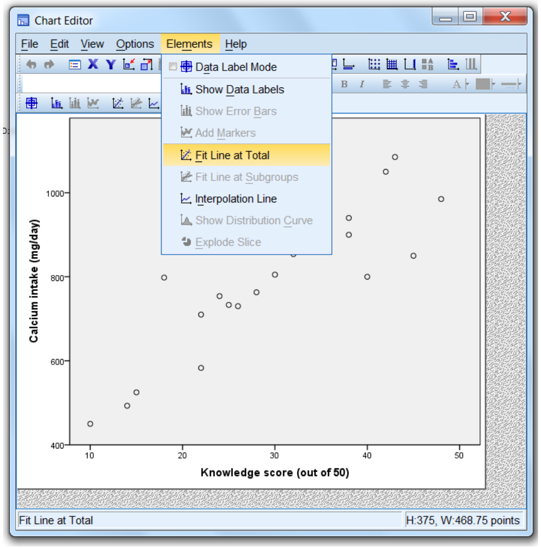
- Double-click on the chart in the SPSS Output window to open the Chart Editor (The above charts have also had changes made to the marker and text size)
- Choose Fit Line at Total from the Elements menu – make sure Linear Fit Method is selected in the Properties window
- Close the Chart Editor window 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 vertical distance between the plotted point and points on the fitted line. For example, in the figure on the right, the observation with a Knowledge score of 40 has an observed value of about 800 and a projected value on the regression line of approximately 930. So the residual for this observation is about -130.

Further exploration of these residuals can be carried out to check the validity of the regression model – ask a tutor for more details. SPSS can also calculate residuals.

Figure 2





Step 3: Estimate the model using simple linear regression:

- Choose Analyze – Regression – Linear
- Put the response variable (*CalciumIntake*) into the Dependent box, and the predictor variable (*KnowledgeScore*) into the Independent(s) box
- Choose the Statistics button in the Linear Regression dialog box, and check the Confidence Intervals box, keeping the Level(%) at 95 and select OK

Results

The regression analysis should create the following output:

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)	373.743	55.067		6.787	.000	258.051	489.435
	Knowledge score (out of 50)	13.897	1.748	.882	7.951	.000	10.225	17.569

a. Dependent Variable: Calcium intake (mg/day)

The estimates of the intercept and slope coefficient are given in the *B* column. The intercept (constant, called b_0 above) is 373.743 and the slope (*KnowledgeScore* out of 50, called b_1 above) is 13.897.

The p-values for the two coefficients are given in the Sig. column and are both shown as 0.000, meaning 0.000 rounded to 3 decimal places (never write 0.000 in your write-up); the null hypotheses are that these coefficients are zero. They should be interpreted as **less than 0.001**, indicating **very strong evidence** that both the Constant and *KnowledgeScore* explain the variation in *CalciumConsumption*

Note: It is not advisable to report on a model with a non-significant slope estimate (the b_1 coefficient). However, if the slope estimate is significant it is normal practice to include the constant in the model even if its p-value is not significant.

Conclusion

Our estimated model is:

$$\textit{CalciumIntake} = 373.7 + (13.90 \times \textit{KnowledgeScore})$$

[Note: Both coefficients have been rounded to 4 significant figures.]

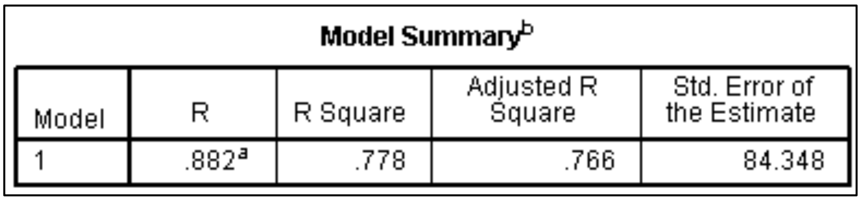
For example, a student with a knowledge of calcium score of 30 is predicted to have an approximate calcium intake of:

$$\textit{CalciumIntake} = 373.7 + (13.90 \times 30) = 373.7 + 417.0 = 790.7 \text{ mg/day}$$

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

Comments

SPSS also gives another table which provides the adjusted R square (R^2) value of the model, also called the **coefficient of determination** adjusted for generalisation. In this example the value of Adjusted R^2 is 0.766.



Model Summary[b]

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

This can be interpreted as 76.6% of the variation in *CalciumIntake* can be explained by *KnowledgeScore*. A 95% confidence interval can be also calculated for this value (but not in SPSS). The remaining 23.4% arises from other variation not taken into account in this analysis.

Note: Fitting a simple linear regression model does <u>not</u> allow us to conclude that a change in the independent variable <u>causes</u> a change in the dependent variable.

The 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 – see the **Pearson Correlation** worksheet.

For information on confidence intervals and the validity of simple linear regression see the **Additional Information** worksheet.