Chapter Seven

Multiple regression

An introduction to multiple regression
Performing a multiple regression on SPSS

Section 1: An introduction to multiple regression

WHAT IS MULTIPLE REGRESSION?

Multiple regression is a statistical technique that allows us to predict someone's score on one variable on the basis of their scores on several other variables. An example might help. Suppose we were interested in predicting how much an individual enjoys their job. Variables such as salary, extent of academic qualifications, age, sex, number of years in full-time employment and socio-economic status might all contribute towards job satisfaction. If we collected data on all of these variables, perhaps by surveying a few hundred members of the public, we would be able to see how many and which of these variables gave rise to the most accurate prediction of job satisfaction. We might find that job satisfaction is most accurately predicted by type of occupation, salary and years in full-time employment, with the other variables not helping us to predict job satisfaction.

When using multiple regression in psychology, many researchers use the term "independent variables" to identify those variables that they think will influence some other "dependent variable". We prefer to use the term "predictor variables" for those variables that may be useful in predicting the scores on another variable that we call the "criterion variable". Thus, in our example above, type of occupation, salary and years in full-time employment would emerge as significant predictor variables, which allow us to estimate the criterion variable – how satisfied someone is likely to be with their job. As we have pointed out before, human behaviour is inherently noisy and therefore it is not possible to produce totally accurate predictions, but multiple regression allows us to identify a set of predictor variables which together provide a useful estimate of a participant's likely score on a criterion variable.

HOW DOES MULTIPLE REGRESSION RELATE TO CORRELATION AND ANALYSIS OF VARIANCE?

In a previous section (Chapter 4, Section 2), we introduced you to correlation and the regression line. If two variables are correlated, then knowing the score on one variable will allow you to predict the score on the other variable. The stronger the correlation, the closer the scores will fall to the regression line and therefore the more accurate the prediction. Multiple regression is simply an extension of this principle, where we predict one variable on the basis of several other variables. Having more than one predictor variable is useful when predicting human behaviour, as our actions, thoughts and emotions are all likely to be influenced by some combination of several factors. Using multiple regression we can test theories (or models) about precisely which set of variables is influencing our behaviour.

As we discussed in Chapter 6, Section 1, on Analysis of Variance, human behaviour is rather variable and therefore difficult to predict. What we are doing in both ANOVA and multiple regression is seeking to account for the variance in the scores we observe. Thus, in the example above, people might vary greatly in their levels of job satisfaction. Some of this variance will be accounted for by the variables we have identified. For example, we might be able to say that salary accounts for a fairly large percentage of the variance in job satisfaction, and hence it is very useful to know someone's salary when trying to predict their job satisfaction. You might now be able to see that the ideas here are rather similar to those underlying ANOVA. In ANOVA we are trying to determine how much of the variance is accounted for by our manipulation of the independent variables (relative to the percentage of the variance we cannot account for). In multiple regression we do not directly manipulate the IVs but instead just measure the naturally occurring levels of the variables and see if this helps us predict the score on the dependent variable (or criterion variable). Thus, ANOVA is actually a rather specific and restricted example of the general approach adopted in multiple regression.

To put this another way, in ANOVA we can directly manipulate the factors and measure the resulting change in the dependent variable. In multiple regression we simply measure the naturally occurring scores on a number of predictor variables and try to establish which set of the observed variables gives rise to the best prediction of the criterion variable.

A current trend in statistics is to emphasise the similarity between multiple regression and ANOVA, and between correlation and the *t*-test. All of these statistical techniques are basically seeking to do the same thing – explain the variance in the level of one variable on the basis of the level of one or more other variables. These other variables might be manipulated directly in the case of controlled experiments, or be observed in the case of surveys or observational studies, but the underlying principle is the same. Thus, although we have given separate chapters to each of these procedures they are fundamentally all the same procedure. This underlying single approach is called the General Linear Model – a term you first encountered when we were undertaking ANOVA in Chapter 6, Section 1.

WHEN SHOULD I USE MULTIPLE REGRESSION?

1. You can use this statistical technique when exploring linear relationships between the predictor and criterion variables – that is, when the relationship follows a straight line. (To examine non-linear relationships, special techniques can be used.)
2. The criterion variable that you are seeking to predict should be measured on a continuous scale (such as interval or ratio scale). There is a separate regression method called logistic regression that can be used for dichotomous dependent variables (not covered here).
3. The predictor variables that you select should be measured on a ratio, interval, or ordinal scale. A nominal predictor variable is legitimate but only if it is dichotomous, i.e. there are no more that two categories. For example, sex is acceptable (where male is coded as 1 and female as 0) but gender identity (masculine, feminine and androgynous) could not be coded as a single variable. Instead, you would create three different variables each with two categories (masculine/not masculine; feminine/not feminine and androgynous/not androgynous). The term dummy variable is used to describe this type of dichotomous variable.
4. Multiple regression requires a large number of observations. The number of cases (participants) must substantially exceed the number of predictor variables you are using in your regression. The absolute minimum is that you have five times as many participants as predictor variables. A more acceptable ratio is 10:1, but some people argue that this should be as high as 40:1 for some statistical selection methods (see page 210).

TERMINOLOGY

There are certain terms we need to clarify to allow you to understand the results of this statistical technique.

Beta (standardised regression coefficients)

The beta value is a measure of how strongly each predictor variable influences the criterion variable. The beta is measured in units of standard deviation. For example, a beta value of 2.5 indicates that a change of one standard deviation in the predictor variable will result in a change of 2.5 standard deviations in the criterion variable. Thus, the higher the beta value the greater the impact of the predictor variable on the criterion variable.

When you have only one predictor variable in your model, then beta is equivalent to the correlation coefficient between the predictor and the criterion variable. This equivalence makes sense, as this situation is a correlation between two variables. When you have more than one predictor variable, you cannot compare the contribution of each predictor variable by simply comparing the correlation coefficients. The beta regression coefficient is computed to allow you to make such comparisons and to assess the strength of the relationship between each predictor variable to the criterion variable.

R, R Square, Adjusted R Square

R is a measure of the correlation between the observed value and the predicted value of the criterion variable. In our example this would be the correlation between the levels of job satisfaction reported by our participants and the levels predicted for them by our predictor variables. R Square (R^2) is the square of this measure of correlation and indicates the proportion of the variance in the criterion variable which is accounted for by our model – in our example the proportion of the variance in the job satisfaction scores accounted for by our set of predictor variables (salary, etc.). In essence, this is a measure of how good a prediction of the criterion variable we can make by knowing the predictor variables. However, R square tends to somewhat over-estimate the success of the model when applied to the real world, so an Adjusted R Square value is calculated which takes into account the number of variables in the model and the number of observations (participants) our model is based on. This Adjusted R Square value gives the most useful measure of the success of our model. If, for example we have an Adjusted R Square value of 0.75 we can say that our model has accounted for 75% of the variance in the criterion variable.

DESIGN CONSIDERATIONS

Multicollinearity

When choosing a predictor variable you should select one that might be correlated with the criterion variable, but that is not strongly correlated with the other predictor variables. However, correlations amongst the predictor variables are not unusual. The term multicollinearity (or collinearity) is used to describe the situation when a high correlation is detected between two or more predictor variables. Such high correlations cause problems when trying to draw inferences about the relative contribution of each predictor variable to the success of the model. SPSS provides you with a means of checking for this and we describe this below.

Selection methods

There are different ways that the relative contribution of each predictor variable can be assessed. In the "simultaneous" method (which SPSS calls the **Enter** method), the researcher specifies the set of predictor variables that make up the model. The success of this model in predicting the criterion variable is then assessed.

In contrast, "hierarchical" methods enter the variables into the model in a specified order. The order specified should reflect some theoretical consideration or previous findings. If you have no reason to believe that one variable is likely to be more important than another you should not use this method. As each variable is entered into the model its contribution is assessed. If adding the variable does not significantly increase the predictive power of the model then the variable is dropped.

In "statistical" methods, the order in which the predictor variables are entered into (or taken out of) the model is determined according to the strength of their correlation with the criterion variable. Actually there are several versions of this method, called forward selection, backward selection and stepwise selection. In **Forward** selection, SPSS enters the variables into the model one at a time in an order determined by the strength of their correlation with the criterion variable. The effect of adding each is assessed as it is entered, and variables that do not significantly add to the success of the model are excluded.

In **Backward** selection, SPSS enters all the predictor variables into the model. The weakest predictor variable is then removed and the regression re-calculated. If this significantly weakens the model then the predictor variable is re-entered – otherwise it is deleted. This procedure is then repeated until only useful predictor variables remain in the model.

Stepwise is the most sophisticated of these statistical methods. Each variable is entered in sequence and its value assessed. If adding the variable contributes to the model then it is retained, but all other variables in the model are then re-tested to see if they are still contributing to the success of the model. If they no longer contribute significantly they are removed. Thus, this method should ensure that you end up with the smallest possible set of predictor variables included in your model.

In addition to the **Enter**, **Stepwise**, **Forward** and **Backward** methods, SPSS also offers the **Remove** method in which variables are removed from the model in a block – the use of this method will not be described here.

How to choose the appropriate method?

If you have no theoretical model in mind, and/or you have relatively low numbers of cases, then it is probably safest to use **Enter**, the simultaneous method. Statistical procedures should be used with caution and only when you have a large number of cases. This is because minor variations in the data due to sampling errors can have a large effect on the order in which variables are entered and therefore the likelihood of them being retained. However, one advantage of the **Stepwise** method is that it should always result in the most parsimonious model. This could be important if you wanted to know the minimum number of variables you would need to measure to predict the criterion variable. If for this, or some other reason, you decide to select a statistical method, then you should really attempt to validate your results with a second independent set of data. The can be done either by conducting a second study, or by randomly splitting your data set into two halves (see Chapter 5, Section 3). Only results that are common to both analyses should be reported.

Section 2: Performing a multiple regression on SPSS

EXAMPLE STUDY

In an investigation of children's spelling, a colleague of ours, Corriene Reed, decided to look at the importance of several psycholinguistic variables on spelling performance. Previous research has shown that age of acquisition has an effect on children's reading and also on object naming. A total of 64 children, aged between 7 and 9 years, completed standardised reading and spelling tests and were then asked to spell 48 words that varied systematically according to certain features such as age of acquisition, word frequency, word length, and imageability. Word length and age of acquisition emerged as significant predictors of whether the word was likely to be spelt correctly.

Further analysis was conducted on the data to determine whether the spelling performance on this list of 48 words accurately reflected the children's spelling ability as estimated by a standardised spelling test. Children's chronological age, their reading age, their standardised reading score and their standardised spelling score were chosen as the predictor variables. The criterion variable was the percentage correct spelling score attained by each child using the list of 48 words.

For the purposes of this book, we have created a data file that will reproduce some of the findings from this second analysis. As you will see, the standardised spelling score derived from a validated test emerged as a strong predictor of the spelling score achieved on the word list. The data file contains only a subset of the data collected and is used here to demonstrate multiple regression. (These data are available in the Appendix.)

HOW TO PERFORM THE TEST

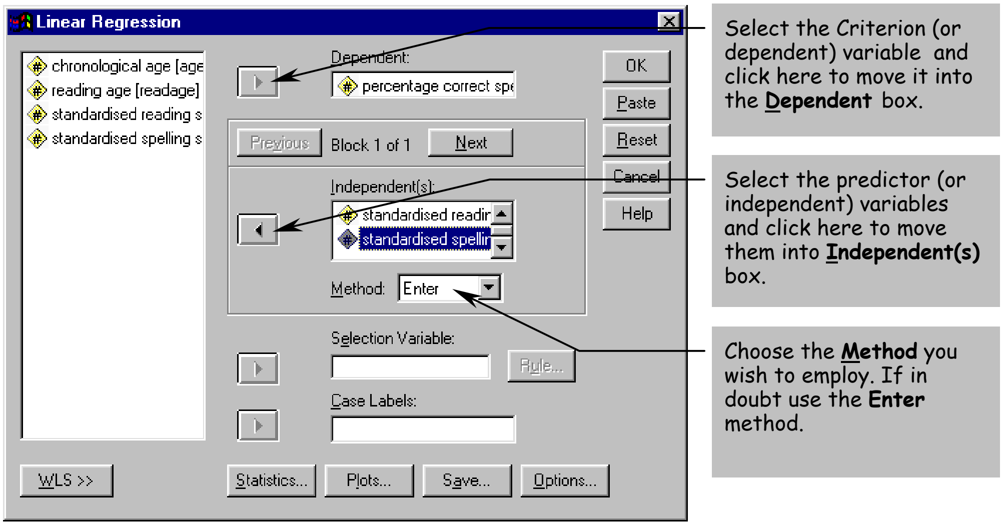
For SPSS Versions 9 and 10, click on **Analyze ⇒ Regression ⇒ Linear**
For SPSS Version 8, click on **Statistics ⇒ Regression ⇒ Linear**

You will then be presented with the **Linear Regression** dialogue box shown below. You now need to select the criterion (dependent) and the predictor (independent) variables.

We have chosen to use the percentage correct spelling score ("spelperc") as our criterion variable. As our predictor variables we have used chronological age

(“age”), reading age (“readage”), standardised reading score (“standsc”), and standardised spelling score (“spellsc”).

As we have a relatively small number of cases and do not have any strong theoretical predictions, we recommend you select **Enter** (the simultaneous method). This is usually the safest to adopt.



Linear Regression

- chronological age [age
- reading age [readage]
- standardised reading s
- standardised spelling s

Dependent: percentage correct spe

Block 1 of 1

Independent(s):
- standardised readir
- standardised spellin

Method: Enter

Selection Variable:

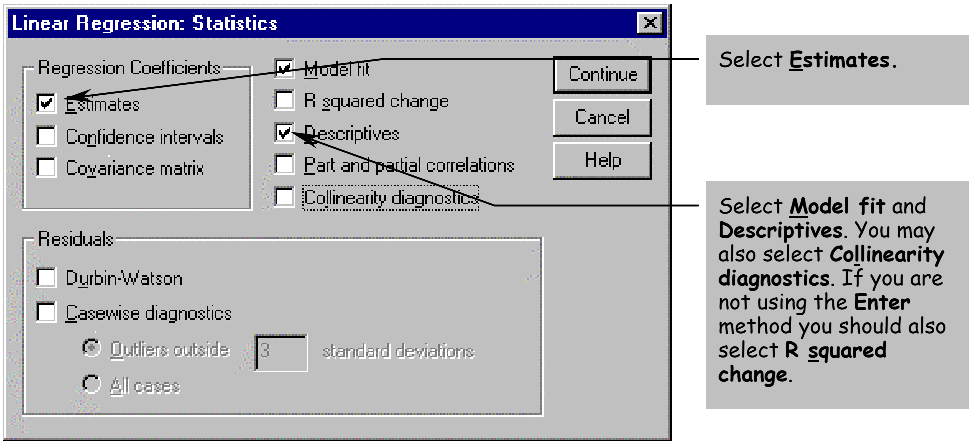
Case Labels:

Select the Criterion (or dependent) variable and click here to move it into the **Dependent** box.

Select the predictor (or independent) variables and click here to move them into **Independent(s)** box.

Choose the **Method** you wish to employ. If in doubt use the **Enter** method.

Now click on the Statistics... button. This will bring up the **Linear Regression: Statistics** dialogue box shown below



Linear Regression: Statistics

Regression Coefficients
- ☑ Estimates
- □ Confidence intervals
- □ Covariance matrix

- ☑ Model fit
- □ R squared change
- ☑ Descriptives
- □ Part and partial correlations
- □ Collinearity diagnostics

Residuals
- □ Durbin-Watson
- □ Casewise diagnostics
  - ● Outliers outside 3 standard deviations
  - ○ All cases

Select **Estimates**.

Select **Model fit** and **Descriptives**. You may also select **Collinearity diagnostics**. If you are not using the **Enter** method you should also select **R squared change**.

The **Collinearity diagnostics** option gives some useful additional output that allows you to assess whether you have a problem with collinearity in your data. The **R squared change** option is useful if you have selected a statistical method such as stepwise as it makes clear how the power of the model changes with the addition or removal of a predictor variable from the model.

When you have selected the statistics options you require, click on the **Continue** button. This will return you to the **Linear Regression** dialogue box. Now click on the OK button. The output that will be produced is illustrated on the following pages.

Tip The SPSS multiple regression option was set to **Exclude cases listwise**. Hence, although the researcher collected data from 52 participants, SPSS analysed the data from only the 47 participants who had no missing values.

SPSS Output for multiple regression USING ENTER METHOD

Obtained Using Menu Items: **Regression** > **Linear** (Method = Enter)

This first table is produced by the **Descriptives** option.

Descriptive Statistics

	Mean	Std. Deviation	N
percentage correct spelling	59.7660	23.9331	47
chronological age	93.4043	7.4910	47
reading age	89.0213	21.3648	47
standardised reading score	95.5745	17.7834	47
standardised spelling score	107.0851	14.9882	47

This second table gives details of the correlation between each pair of variables. We do not want strong correlations between the criterion and the predictor variables. The values here are acceptable.

Correlations

		percentage correct spelling	chronological age	reading age	standardised reading score	standardised spelling score
Pearson Correlation	percentage correct spelling	1.000	-.074	.623	.778	.847
	chronological age	-.074	1.000	.124	-.344	-.416
	reading age	.623	.124	1.000	.683	.570
	standardised reading score	.778	-.344	.683	1.000	.793
	standardised spelling score	.847	-.416	.570	.793	1.000
Sig. (1-tailed)	percentage correct spelling	.	.311	.000	.000	.000
	chronological age	.311	.	.203	.009	.002
	reading age	.000	.203	.	.000	.000
	standardised reading score	.000	.009	.000	.	.000
	standardised spelling score	.000	.002	.000	.000	.
N	percentage correct spelling	47	47	47	47	47
	chronological age	47	47	47	47	47
	reading age	47	47	47	47	47
	standardised reading score	47	47	47	47	47
	standardised spelling score	47	47	47	47	47

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	standardised spelling score, chronological age, reading age, standardised reading score[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: percentage correct spelling

This third table tells us about the predictor variables and the method used. Here we can see that all of our predictor variables were entered simultaneously (because we selected the Enter method.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.923[a]	.852	.838	9.6377

a. Predictors: (Constant), standardised spelling score, chronological age, reading age, standardised reading score

This table is important. The Adjusted R Square value tells us that our model accounts for 83.8% of variance in the spelling scores – a very good model!

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22447.277	4	5611.819	60.417	.000[a]
	Residual	3901.149	42	92.884		
	Total	26348.426	46			

a. Predictors: (Constant), standardised spelling score, chronological age, reading age, standardised reading score

b. Dependent Variable: percentage correct spelling

This table reports an ANOVA, which assesses the overall significance of our model. As $p < 0.05$ our model is significant.

Coefficients[a]

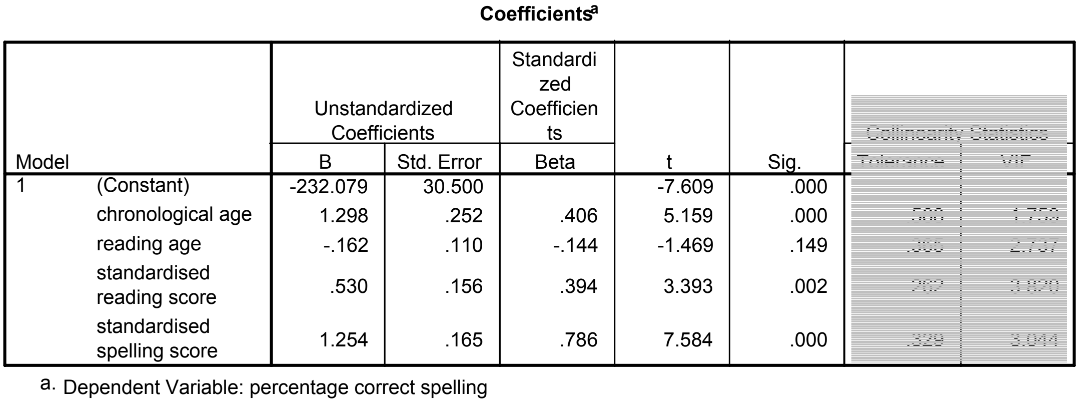
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-232.079	30.500		-7.609	.000
	chronological age	1.298	.252	.406	5.159	.000
	reading age	-.162	.110	-.144	-1.469	.149
	standardised reading score	.530	.156	.394	3.393	.002
	standardised spelling score	1.254	.165	.786	7.584	.000

a. Dependent Variable: percentage correct spelling

The Standardized Beta Coefficients give a measure of the contribution of each variable to the model. A large value indicates that a unit change in this predictor variable has a large effect on the criterion variable. The *t* and Sig (*p*) values give a rough indication of the impact of each predictor variable – a big absolute *t* value and small *p* value suggests that a predictor variable is having a large impact on the criterion variable. If you requested **Collinearity diagnostics** these will also be included in this table – see next page.

Collinearity diagnostics

If you requested the optional **Collinearity diagnostics**, these will be shown in an additional two columns of the Coefficients table (the last table shown above) and a further table (titled Collinearity diagnostics) that is not shown here. Ignore this extra table and simply look at the two new columns.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-232.079	30.500		-7.609	.000		
	chronological age	1.298	.252	.406	5.159	.000	.568	1.759
	reading age	-.162	.110	-.144	-1.469	.149	.365	2.737
	standardised reading score	.530	.156	.394	3.393	.002	.262	3.820
	standardised spelling score	1.254	.165	.786	7.584	.000	.329	3.044

[a] Dependent Variable: percentage correct spelling

The tolerance values are a measure of the correlation between the predictor variables and can vary between 0 and 1. The closer to zero the tolerance value is for a variable, the stronger the relationship between this and the other predictor variables. You should worry about variables that have a very low tolerance. SPSS will not include a predictor variable in a model if it has a tolerance of less that 0.0001. However, you may want to set your own criteria rather higher – perhaps excluding any variable that has a tolerance level of less than 0.01. VIF is an alternative measure of collinearity (in fact it is the reciprocal of tolerance) in which a large value indicates a strong relationship between predictor variables.

Reporting the results

When reporting the results of a multiple regression analysis, you want to inform the reader about the proportion of the variance accounted for by your model, the significance of your model and the significance of the predictor variables. In the results section, we would write:

Using the enter method, a significant model emerged ($F_{4,42}$=60.417, p < 0.0005. Adjusted R square = .838. Significant variables are shown below:

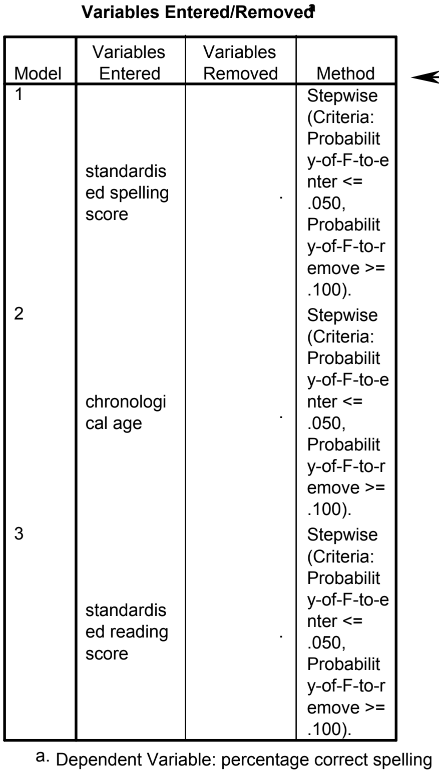
Predictor Variable	**Beta**	***p***
Chronological age	.406	$p < 0.0005$
Standardised reading score	.394	$p = 0.002$
Standardised spelling score	.786	$p < 0.0005$

(Reading age was not a significant predictor in this model.)

OUTPUT FROM MULTIPLE REGRESSION USING STEPWISE METHOD

Obtained Using Menu Items: **Regression** > **Linear** (Method = Stepwise)

Reproduced below are the key parts of the output produced when you the **Stepwise** method is selected. When using this method you should also select the **R Squared Change** option in the **Linear Regression: Statistics** dialogue box (see page 213).



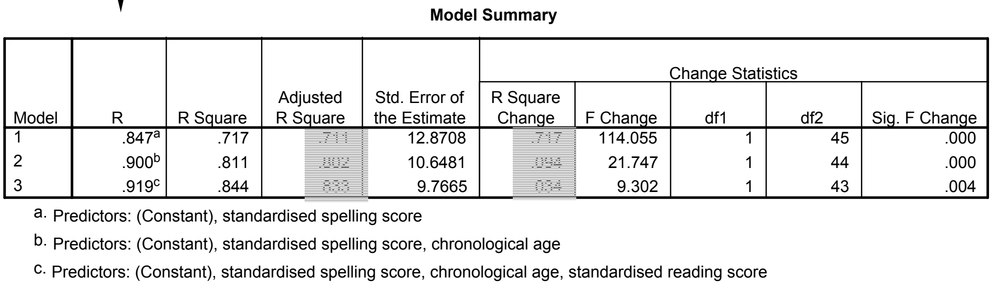
Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	standardised spelling score	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	chronological age	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	standardised reading score	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: percentage correct spelling

This table shows us the order in which the variables were entered and removed form our model. We can see that in this case three variables were added and none were removed.

Here we can see that model 1, which included only standardised spelling score accounted for 71% of the variance (Adjusted R^2=0.711). The inclusion of chronological age into model 2 resulted in an additional 9% of the variance being explained (R^2 change = 0.094). The final model 3 also included standardised reading score, and this model accounted for 83% of the variance (Adjusted R^2=0.833).



Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.847[a]	.717	.711	12.8708	.717	114.055	1	45	.000
2	.900[b]	.811	.802	10.6481	.094	21.747	1	44	.000
3	.919[c]	.844	.833	9.7665	.034	9.302	1	43	.004

a. Predictors: (Constant), standardised spelling score

b. Predictors: (Constant), standardised spelling score, chronological age

c. Predictors: (Constant), standardised spelling score, chronological age, standardised reading score

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18893.882	1	18893.882	114.055	.000[a]
	Residual	7454.543	45	165.657		
	Total	26348.426	46			
2	Regression	21359.610	2	10679.805	94.193	.000[b]
	Residual	4988.815	44	113.382		
	Total	26348.426	46			
3	Regression	22246.870	3	7415.623	77.744	.000[c]
	Residual	4101.556	43	95.385		
	Total	26348.426	46			

a. Predictors: (Constant), standardised spelling score

b. Predictors: (Constant), standardised spelling score, chronological age

c. Predictors: (Constant), standardised spelling score, chronological age, standardised reading score

d. Dependent Variable: percentage correct spelling

This table reports the ANOVA result for the three models.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-85.032	13.688		-6.212	.000		
	standardised spelling score	1.352	.127	.847	10.680	.000	1.000	1.000
2	(Constant)	-209.328	28.959		-7.228	.000		
	standardised spelling score	1.576	.115	.987	13.679	.000	.827	1.209
	chronological age	1.075	.230	.336	4.663	.000	.827	1.209
3	(Constant)	-209.171	26.562		-7.875	.000		
	standardised spelling score	1.197	.163	.750	7.349	.000	.348	2.875
	chronological age	1.092	.211	.342	5.162	.000	.827	1.210
	standardised reading score	.406	.133	.301	3.050	.004	.371	2.698

a. Dependent Variable: percentage correct spelling

Here SPSS reports the Beta, *t* and sig (*p*) values for each of the models. These were explained in the output from the **Enter** method.

Excluded Variables[d]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	chronological age	.336[a]	4.663	.000	.575	.827	1.209	.827
	reading age	.208[a]	2.249	.030	.321	.675	1.481	.675
	standardised reading score	.288[a]	2.317	.025	.330	.371	2.696	.371
2	reading age	.036[b]	.395	.695	.060	.517	1.933	.435
	standardised reading score	.301[b]	3.050	.004	.422	.371	2.698	.348
3	reading age	-.144[c]	-1.469	.149	-.221	.365	2.737	.262

a. Predictors in the Model: (Constant), standardised spelling score

b. Predictors in the Model: (Constant), standardised spelling score, chronological age

c. Predictors in the Model: (Constant), standardised spelling score, chronological age, standardised reading score

d. Dependent Variable: percentage correct spelling

This table gives statistics for the variables that were excluded from each model.

Thus, the final model to emerge from the **Stepwise** analysis contains only three predictor variables. The predictor variable reading age, which was not significant in the **Enter** analysis, was also not included in the **Stepwise** analysis as it did not significantly strengthen the model.

REPORTING THE RESULTS

In your results section, you would report the significance of the model by citing the *F* and the associated *p* value, along with the adjusted R square, which indicates the strength of the model. So, for the final model reported above, we would write:

Adjusted R square = .833; $F_{3,43}$ = 77.7, $p < 0.0005$ (using the stepwise method). Significant variables are shown below.

Predictor Variable	**Beta**	***p***
Standardised spelling score:	.750	p <0.0005
Chronological age	.342	p <0.0005
Standardised reading score	.301	p =0.004

(Reading age was not a significant predictor in this model.)