

Multiple Regression

Regression allows you to investigate the relationship between variables. But more than that, it allows you to model the relationship between variables, which enables you to make predictions about what one variable will do based on the scores of some other variables.

- The variable you want to predict is called the outcome variable (or DV)
- The variables you base your prediction on are called the predictor variables (or IVs)

While simple linear regression only enables you to predict the value of one variable based on the value of a single predictor variable; multiple regression allows you to use multiple predictors.

Worked Example

For this tutorial, we will use an example based on a fictional study attempting to model students' exam performance.

Imagine you are a psychology research methods tutor interested in predicting how well your students will do in their exam. You think that revision intensity and enjoyment of the subject are variables that may allow you to do this.

To investigate this you could measure how many hours of revision your students did in the weeks preceding their exam and ask them to rate their enjoyment of the material on a scale from 0-100. You could then see how well they do in their exam, which would allow you to model how well future students are likely to do based on these predictors. This is what we will explore in this tutorial.

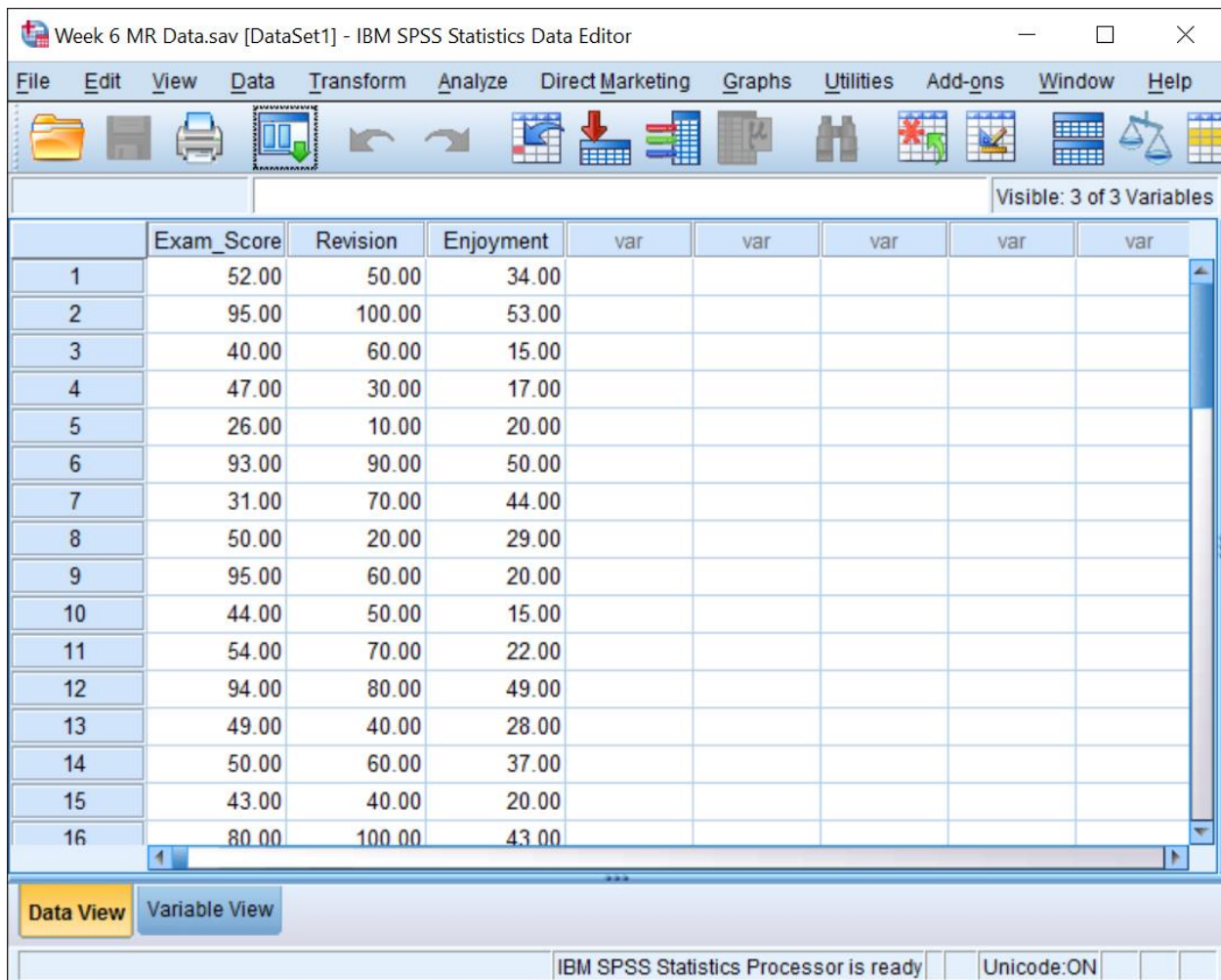
Multiple regression allows you to include multiple predictors (IVs) into your predictive model, however this tutorial will concentrate on the simplest type: when you have only *two* predictors and a single outcome (DV) variable.

In this example our three variables are:

- Exam Score - the outcome variable (DV)
- Revision Intensity - a predictor variable (IV1)
- Subject Enjoyment - a predictor variable (IV2)

As with ANOVA there are a number of assumptions that must be met for multiple regression to be reliable, however this tutorial *only* covers how to run the analysis. If you plan on running a multiple regression as part of your own research project, make sure you also check out the assumptions tutorial.

This what the data looks like in SPSS. It can also be found in the SPSS file: 'Week 6 MR Data.sav'.



Week 6 MR Data.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Visible: 3 of 3 Variables

	Exam_Score	Revision	Enjoyment	var	var	var	var	var
1	52.00	50.00	34.00					
2	95.00	100.00	53.00					
3	40.00	60.00	15.00					
4	47.00	30.00	17.00					
5	26.00	10.00	20.00					
6	93.00	90.00	50.00					
7	31.00	70.00	44.00					
8	50.00	20.00	29.00					
9	95.00	60.00	20.00					
10	44.00	50.00	15.00					
11	54.00	70.00	22.00					
12	94.00	80.00	49.00					
13	49.00	40.00	28.00					
14	50.00	60.00	37.00					
15	43.00	40.00	20.00					
16	80.00	100.00	43.00					

Data View Variable View

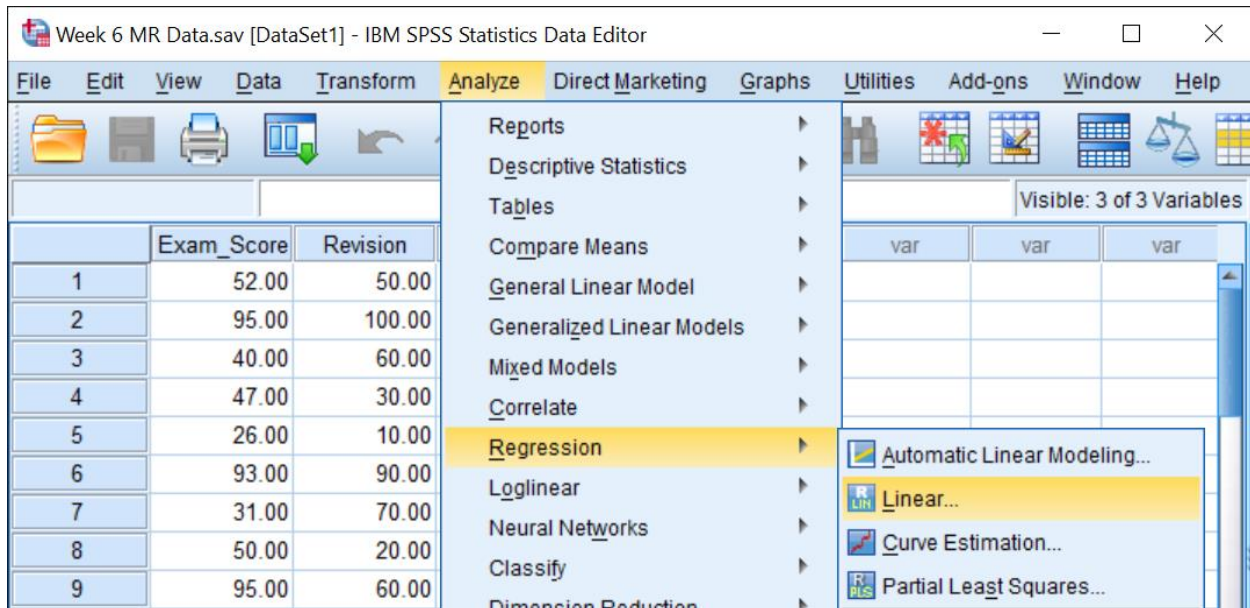
IBM SPSS Statistics Processor is ready Unicode:ON

In multiple regression, each participant provides a score for all of the variables. As each row should contain all of the information provided by *one* participant, there needs to be a separate column for each variable.

In this example, the different columns display the following data:

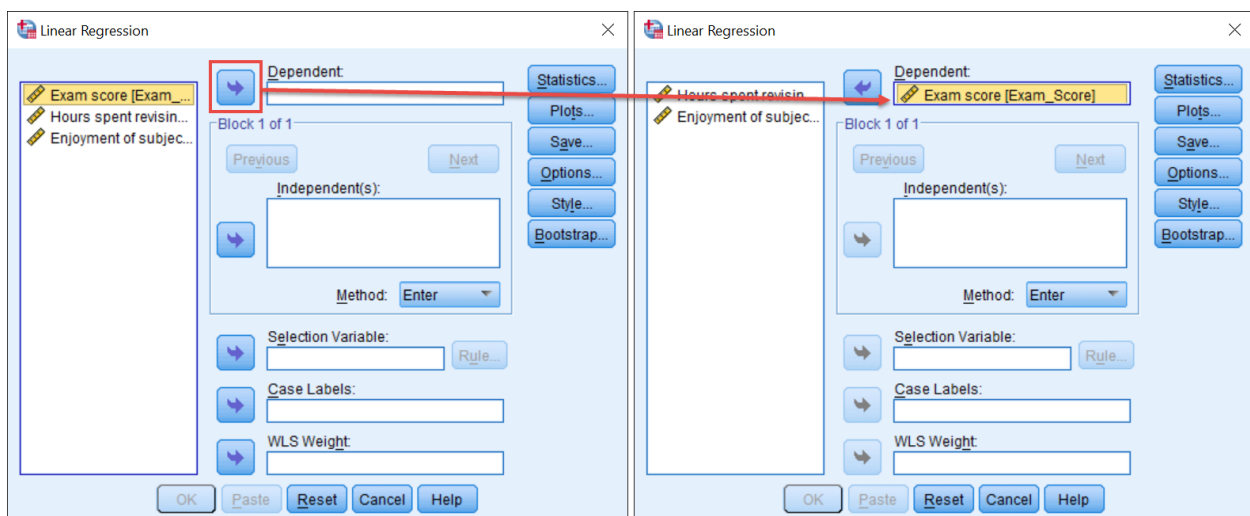
- **Exam_Score:** This is our **outcome variable**.
- **Revision:** This is our first **predictor variable** (IV1) 'Revision Intensity'. It represents how many hours of revision participants did in the weeks leading up to the exam.
- **Enjoyment:** This is our second **predictor variable** (IV2) 'Subject Enjoyment'. It represents how much participants enjoyed the subject they were studying on a scale of 0-100.

To start the analysis, begin by **CLICKING** on the **Analyze** menu, select **Regression**, and then the **Linear...** sub-option.

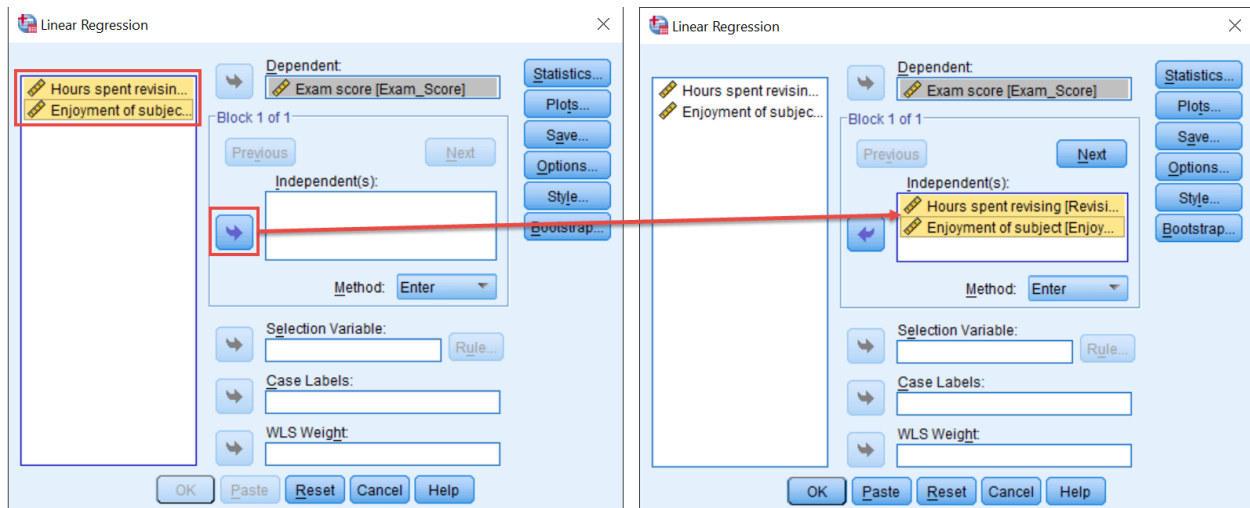


This opens the **Linear Regression** dialog box. Here you will see all of the variables recorded in the data file displayed in the box in the left. To tell SPSS what we want to analyse we need to move our variables to the correct boxes on the right.

Exam_Score is already selected. As this is our *Outcome Variable*, move it across to the **Dependent** box by **CLICKING** the arrow to the left of it.



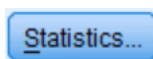
Next, **SELECT** the two predictor variables (**Revision Intensity** and **Subject Enjoyment**) as shown below. When doing this yourself, remember that if you hold down the Ctrl key so you can highlight them all in one go.



Add them to the analysis by **CLICKING** on the blue arrow to the left of the **Independent(s)** box.

Now we have told SPSS which variables are which, we need to tell it what statistics we want it to produce.

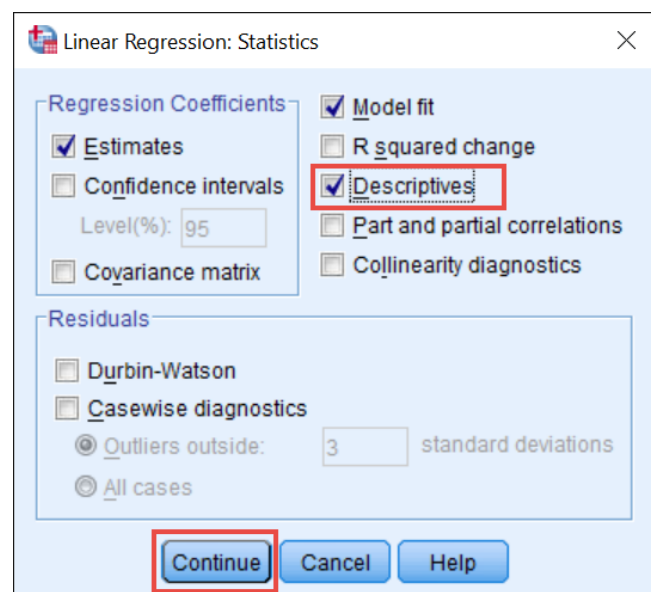
To do this, **CLICK** on **Statistics** button.



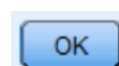
This opens the **Statistics** dialogue box. **Estimates** and **Model Fit** are already selected by default. In addition to this, **SELECT** the **Descriptives** option to see how the different variables are correlated with one another.

We can also use this box to test several of the assumptions of regression, however we will not cover this in this tutorial. Remember to check out the assumptions tutorial if you are going to carry out a multiple regression yourself.

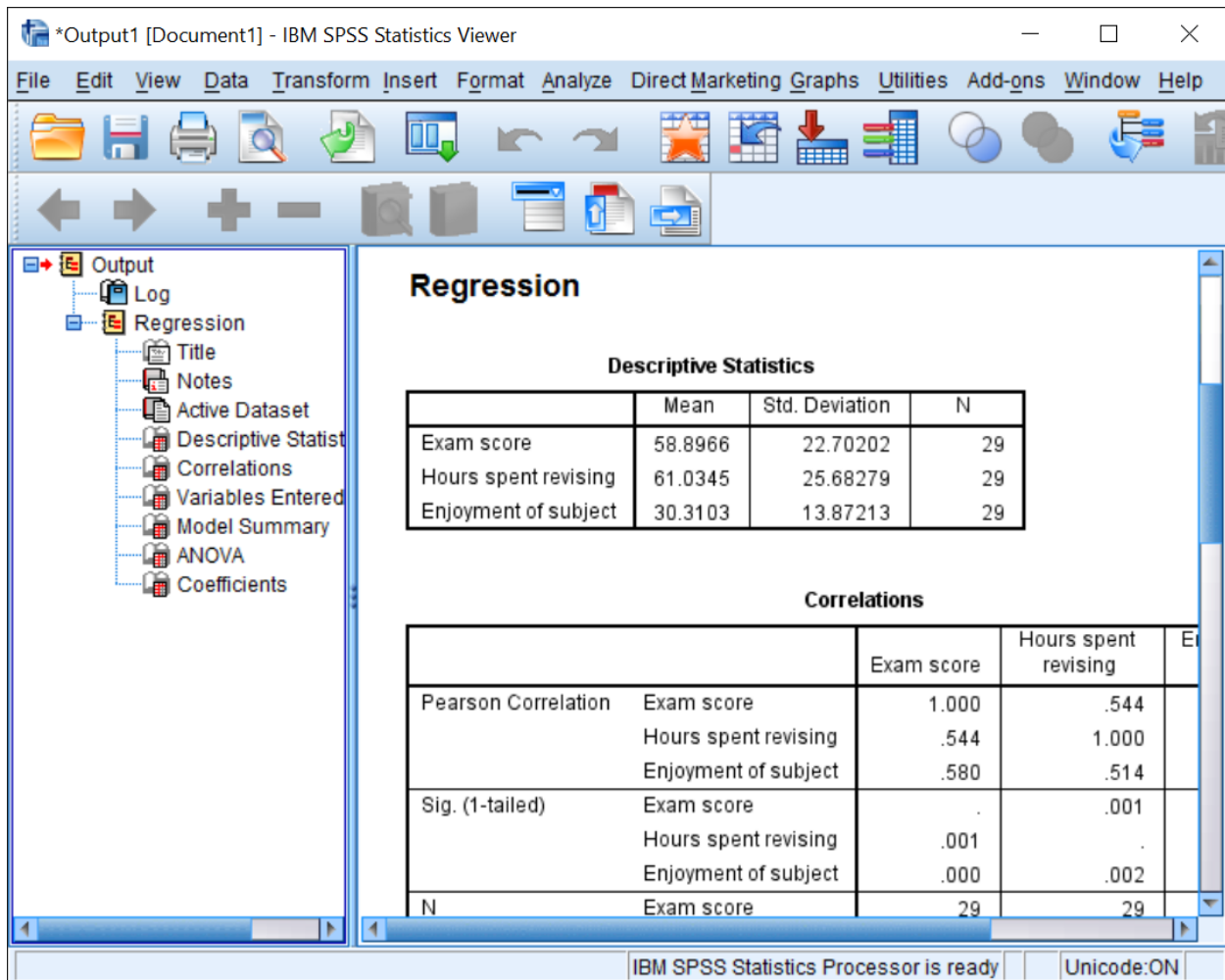
Now **CLICK** on **Continue**



CLICK on **OK** in the main Regression dialog box to proceed.



The output window gives you the results of the regression.



This tutorial will now take you through the results, box-by-box.

Descriptive Statistics

The first box simply gives you the means and standard deviations for each of your variables. You don't really need this information to interpret the multiple regression, it's just for your interest.

Descriptive Statistics

	Mean	Std. Deviation	N
Exam score	58.8966	22.70202	29
Hours spent revising	61.0345	25.68279	29
Enjoyment of subject	30.3103	13.87213	29

Correlations

		Correlations		
		Exam score	Hours spent revising	Enjoyment of subject
Pearson Correlation	Exam score	1.000	.544	.580
	Hours spent revising	.544	1.000	.514
	Enjoyment of subject	.580	.514	1.000
Sig. (1-tailed)	Exam score	.	.001	.000
	Hours spent revising	.001	.	.002
	Enjoyment of subject	.000	.002	.
N	Exam score	29	29	29
	Hours spent revising	29	29	29
	Enjoyment of subject	29	29	29

The next box gives you the correlations between each of the variables. The first row shows the correlation coefficients (r), while the second tells you their statistical significance. To establish which values are associated with which correlations you can find the name of the first variable at the top of each column, and the name of the correlated variable at the start of the intersecting row.

In multiple regression, you want the predictor variables to be related to your outcome variable (otherwise, there is no point in including them in the predictive model). In contrast, you don't want your predictors to be too strongly related to one another, as this can make your analysis unreliable. When predictors correlate at more than $r = .8$, you have multicollinearity which is a problem for multiple regression, so you may want to remove one of the variables. You can learn more about this in the separate tutorials on Assumptions of Multiple Regression. In this case there are several correlations of around $r = .5$, suggesting multiple regression is appropriate.

Variables Entered/Removed

The third box simply tells you which variables you have included into the model. That is, which variables are acting as predictor variables (or IVs).

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	Enjoyment of subject, Hours spent revising ^b	.	Enter

In this case we have included two predictors:

- Subject Enjoyment ('Enjoyment of subject')
- Revision Intensity ('Hours spent revising')

Model Summary

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.647 ^a	.418	.373	17.97120

a. Predictors: (Constant), Enjoyment of subject, Hours spent revising

The next box displays information about how the two variables relate to one another. In this case, the term 'model' is used because we are trying to build a model of the relationship between our variables. The model consists of the predictor variables we are using to try to predict the outcome variable (Exam Score). In this case, we have two predictor variables in the model: Revision Intensity and Subject Enjoyment.

The key sections of the table are:

- **R** The value in the *R* column is a very similar statistic to *r*, and can be interpreted like any regular correlation coefficient. But instead of telling you the relationship between two variables, it tells you the strength of the relationship between the outcome variable (DV) and *all* of the predictor variables (or IVs) combined.

In this case $R = 0.65$, which is a strong relationship. This suggests our model is a relatively good predictor of the outcome.

- **R Square** The *R Square* column contains the value we are most interested in. Usually written as R^2 , this value indicates the proportion of variation in the outcome variable (Exam Score) that can be explained by the model (i.e. by Revision Intensity and Subject Enjoyment).

You can either report this as $R^2 = .418$, or you can multiply it by 100 to give a proportion. In this case we could say that 41.8% of the variance in the data can be explained by the predictor variables.

ANOVA

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6033.628	2	3016.814	9.341	.001 ^b
	Residual	8397.062	26	322.964		
	Total	14430.690	28			

a. Dependent Variable: Exam score

b. Predictors: (Constant), Enjoyment of subject, Hours spent revising

The next box in the output tells us whether or not our model (which includes Revision Intensity and Subject Enjoyment) is a significant predictor of the outcome variable. This is tested using Analysis of Variance. As the significance value is less than $p=0.05$, we can say that the regression model significantly predicts Exam Score.

How do we write up our findings?

So we know that the model is significant, but how do we write up the numbers? To report your findings in APA format, you report your results as:

$$F(\text{Regression df, Residual df}) = F\text{-Ratio}, p = \text{Sig}$$

You need to report these statistics along with a sentence describing the results. In this case we could say:

The results indicated that the model was a significant predictor of exam performance, $F(2,26) = 9.34, p = .001$.

Coefficients

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	20.647	9.530		2.167	.040
	Hours spent revising	.295	.154	.333	1.911	.067
	Enjoyment of subject	.668	.285	.408	2.342	.027

a. Dependent Variable: Exam score

While the ANOVA table tells us whether the *overall* model is a significant predictor of the outcome variable, this table tells us the extent to which the individual predictor variables contribute to the model.

There are two sections of the table that you need to look at to interpret your multiple regression. The first part of the table that we need to look at is the **Sig** column. This tells us whether the predictors significantly contributed to the model or not.

By reading across the rows for each of the predictor variables, we can see that:

- Subject Enjoyment significantly contributed to the model (**p=.03**)
- However, Revision Intensity did not (**p=.07**)

Remember... this example is fictional!! Please don't go away thinking that revision does not lead to better marks!!! 😊

The next column we need to look at contains the unstandardized beta coefficients for the model (the B values). These values tell us about the relationships between the outcome and both predictor variables. As both values are positive, so are the relationships. That is, as time spent revising increases (or as subject enjoyment scores go up), exam scores also get higher. In addition, these B values give us an idea of the influence each predictor has on the outcome *if the effects of the other variables are held constant*.

- Revision Intensity (**B₁ = .295**): as revision intensity increases by one unit (i.e. by one hour), exam scores increase by 0.295 units.
- Subject Enjoyment (**B₂ = .668**): as people's enjoyment of the subject increased by one unit on the enjoyment scale, exam scores went up by 0.668 units.

In regression, we can produce a statistical model that allows us to predict values of our outcome variable based on our predictor variable. This table also gives us all of the information we need to do that. This model takes the form of a statistical equation where:

$$Y = B_0 + B_1X_1 + B_2X_2$$

- Where Y represents the outcome variable
- X₁ represents the first predictor variable
- X₂ represents the second predictor variable

In this case, we can say:

$$\text{Exam score} = B_0 + B_1\text{Revision Intensity} + B_2\text{Subject Enjoyment}$$

Creating the Model

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	20.647	9.530		2.167	.040
	Hours spent revising	.295	.154	.333	1.911	.067
	Enjoyment of subject	.668	.285	.408	2.342	.027

a. Dependent Variable: Exam score

We can also use the B column to create our predictive model. To do this, we need to replace the Bs with their actual numerical values. B_0 is found in the row labelled (Constant); B_1 in the row for the first predictor (Revision Intensity); B_2 in the row for the second predictor (Subject Enjoyment).

Replacing the Bs with the correct values gives us the following predictive model:

$$\text{Exam score} = 20.657 + (.295 * \text{Revision Intensity}) + (.668 * \text{Subject Enjoyment})$$

By inserting the hours participants spent revising and their subject enjoyment score, we can now predict how well somebody is likely to do on their exam.

How do we write up our results?

When writing up your results there are certain statistics that you need to report:

- First, you need to state the proportion of variance that can be explained by your model. This is represented by the statistic R^2 and is a number between 0 and 1. It can either be reported in this format (e.g. $R^2 = .418$) or it can be multiplied by 100 to represent the percentage of variance your model explains (e.g. 41.8%).
- Second, you need to report whether or not your model was a significant predictor of the outcome variable using the results of the ANOVA.
- Finally, you need to include information about your predictor variables. In this case, you need to include your B values for both variables and the significance of their contribution to the model. It is also a good idea to include your final model here.

Writing up the results

Now you have seen what information you need to include in a results section, for this example you might write something like this:

A multiple regression was carried out to investigate whether Revision Intensity and Subject Enjoyment could significantly predict participants' exam scores. The results of the regression indicated that the model explained 41.8% of the variance and that the model was a significant predictor of exam performance, $F(2,26) = 9.34$, $p = .001$. While Subject Enjoyment contributed significantly to the model ($B = .668$, $p < .05$), Revision Intensity did not ($B = .295$, $p = .07$). The final predictive model was:

$$\text{Exam score} = 20.657 + (.295 * \text{Revision Intensity}) + (.668 * \text{Subject Enjoyment})$$

This brings us to the end of the tutorial. Why not download the dataset used in this tutorial and see if you can produce the same output on your own. Remember, practice makes perfect!