Session 8

Using the Statistical Add-Ins (Part II)

SESSION 8: Statistical Add-Ins Part II

One-Way ANOVA – Single Factor

For some data sets, we want to compare several independent groups. For this we use a **One-Way ANOVA** (ANALYSIS OF VARIANCE).

This analysis tool performs simple analysis of variance (ANOVA) to test the hypothesis that means from two or more samples are equal (drawn from populations with the same mean). This technique expands on the tests for two means, such as the *t*-test. To be able to perform ANOVA, the different datasets must be given in different columns.

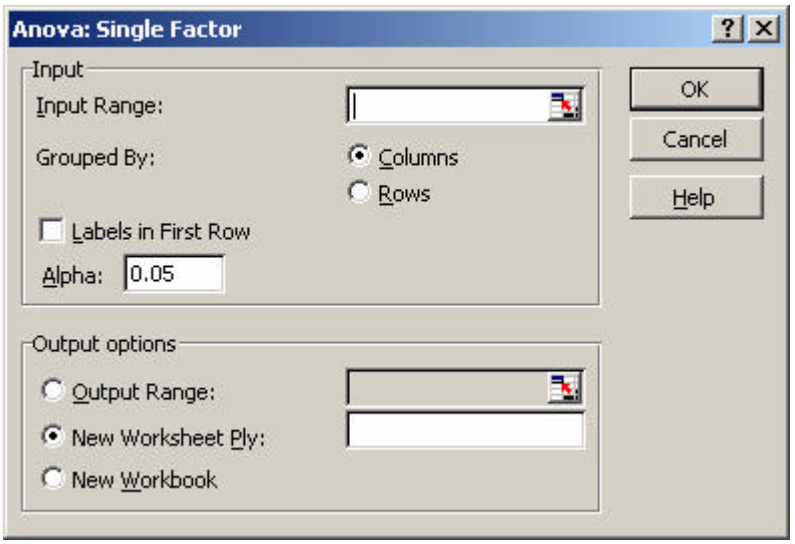
In this case, the null hypothesis (H_0) and the alternative hypothesis (H_1) are:

> H_0 : all the groups have same mean
>
> H_1 : at least 1 group has a different mean

We are going to use the **gss91tanova.xls** data set. This is a subset of **gss91t.xls**, suitably arranged so that an **ANOVA** procedure can be applied. We can split the respondents into three groups according to which category of the variable **LIFE** they fall into; EXCITING (1), ROUTINE (2) or DULL (3). The other values are all missing data. We want to know if there is any difference in the average years of education of these groups. Our **Null Hypothesis** is that there is no difference between them in terms of education.

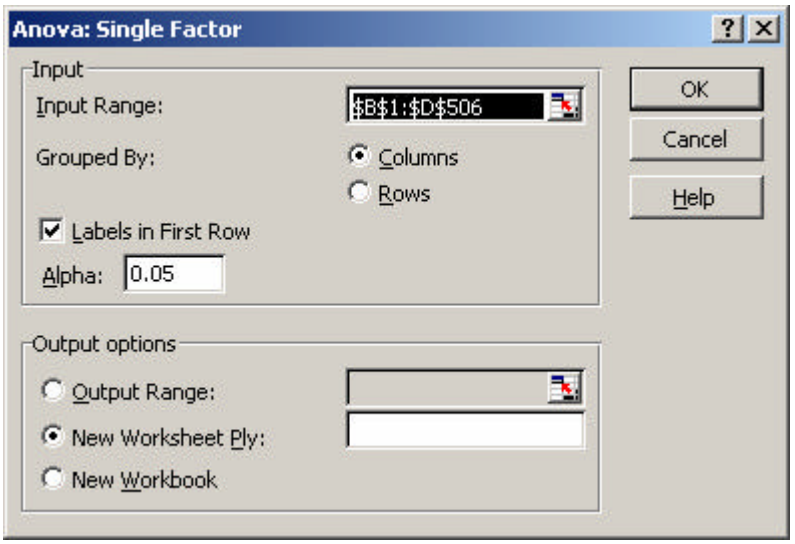
Click on **Tools** ➢ **Data Analysis** ➢ **Random Number Generation** ➢ **OK**.

Excel displays the following dialog box.



You can then select the input range so that it reflects the 3 columns that you want the check the mean of. You have to enter the row number of the group having the most cases; **Excel** will count the data points in each group.

The completed form looks like this.



Press **OK**, the following output is obtained.

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	EDUC1	434	6006	13.83871	26.42658		
6	EDUC2	505	6456	12.78416	37.01483		
7	EDUC3	41	430	10.4878	10.6061		
8							
9							
10	ANOVA						
11	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12	Between Groups	563.9976	2	281.9988	9.026571	0.00013	3.004942
13	Within Groups	30522.43	977	31.24097			
14							
15	Total	31086.42	979				
16							

Other statistical packages would allow you to test on which group differs from the rest. However, this feature is not present in **Excel**. We can tell from the **P-value** obtained that the groups are different; however we cannot really see which differ.

The 1st table obtained gives some descriptive statistics, such as the count in each group, as well as mean and average.

The 2nd table gives the results of the **One-Way ANOVA**. A measure of the variability found between the groups is shown in the **Between Groups**

line, while the **Within Groups** line gives a measure of how much the observations within each group vary. These are used to perform the **F-Test** which we use to test our **Null Hypothesis** that there is no difference between the three groups in terms of their years in education.

We interpret the **F-Test** in the same way as we did the **T-Test**; if the significance (in the **P-value** column) is less than 0.05, we have evidence, at the 5% level; to reject the Null Hypothesis, and say that there **is** some difference between the groups. Otherwise, we accept our Null Hypothesis.

We can see that the **F-value** of 9.026 has a significance of less than 0.0005, and therefore we reject the Null Hypothesis.

The following table provides a brief description of the elements that Excel calculates for each row/column in the input range when you execute the **ANOVA: single factor** macro:

Option	Description
Descriptive Statistics	For each group, Excel starts by giving the number of observations (count), the sum of the data points, the mean and the variance.
Source of Variation	There are 2 estimates of variability, namely the between group variation, and the within group variation.
Sum of squares (SS)	The 2 estimates of variability
df	Degrees of freedom needed by the *F* distribution.
Mean Square	Evaluated by SS divided by df.
F	The F-Score worked out by dividing the 2 MS.
P-Value	The probability of accepting H_0. If this value is greater than alpha, we accept H_0, otherwise we accept H_1.
F Critical	The critical point for the acceptance of the hypothesis.

Types of Relationships, Scatter Plots and Lines of Best Fit

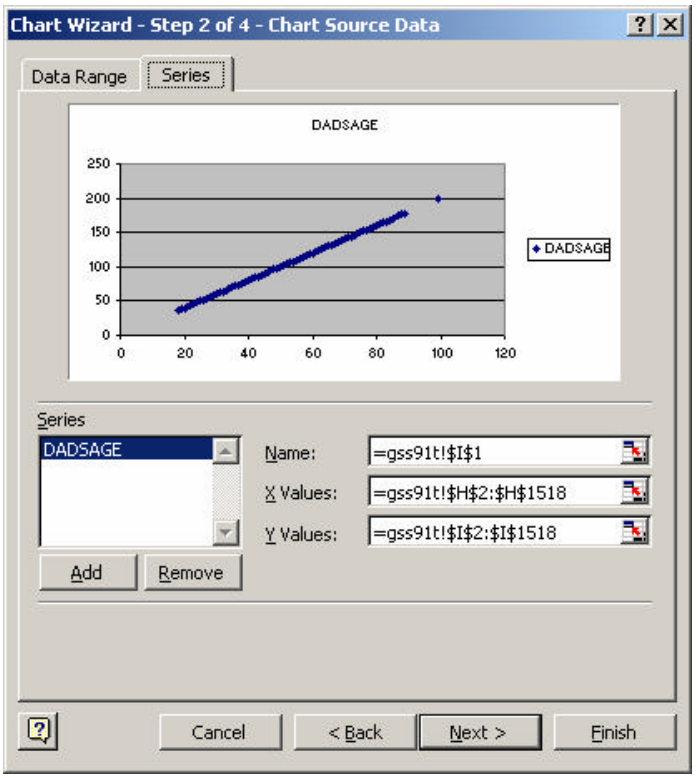
What do we mean when we say that two variables are **related**? Nothing complicated; simply that knowing the value of one variable tells us something about the other.

In Session 5 we produced some **Scatter Plots**. A Scatter Plot of two variables that are unrelated produces what appears to be a random pattern. The other extreme is a **Perfect Relationship**, where knowing the

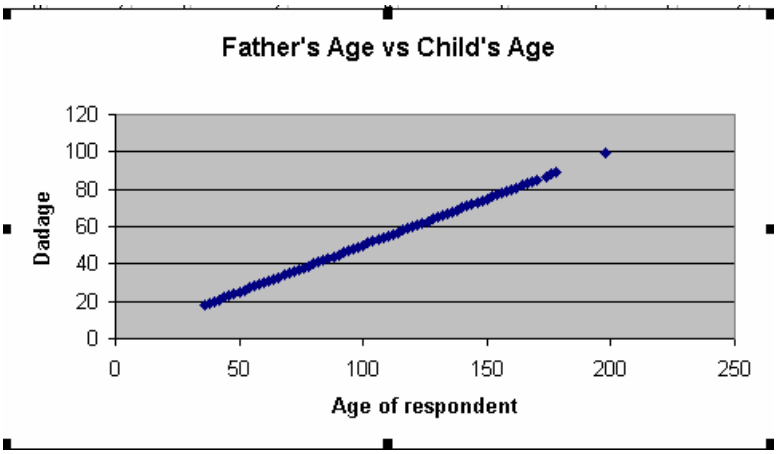
value of one variable can tell you the exact value of the other. In these cases, the points on the Scatter Plot can be joined to from a smooth line. We will be interested in **Linear Relationships**; that is, where the line would be straight.

Perfect relationships are rare, so we will create some from the **gss91t.xls** data. If we imagine that all the fathers are exactly twice the age of their children, we can create a new variable **DADSAGE**.

To plot the new variable against **AGE**, we click on **Chart Wizard ➢ XY (Scatter)**.



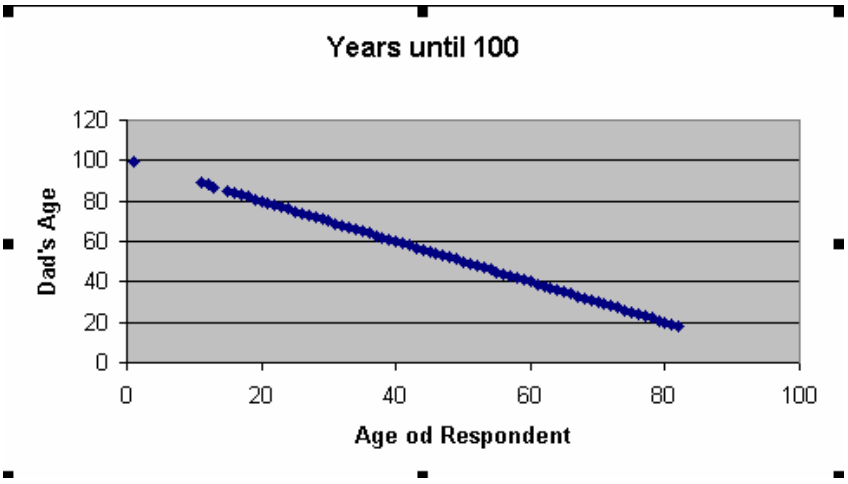
Modify the selections until you obtain the following scatter diagram.



DADSAGE is selected for the **Y Axis**, and **AGE** for the **X Axis.**

This is a **positive** relationship; that is, as the value of one variable increases, so does the value of the other.

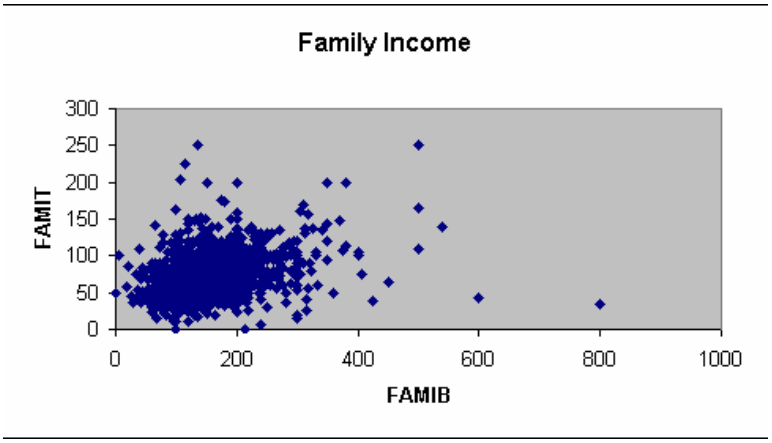
In a **negative** relationship the opposite is true: as the value of one variable *increases*, the value of the other *decreases*. As an example, we can create a new variable, **HUNAGE**, which is the number of years each respondent has to go before they reach 100 years of age. We do this using the expression **100 – Age**. A scatter plot now looks like the following.



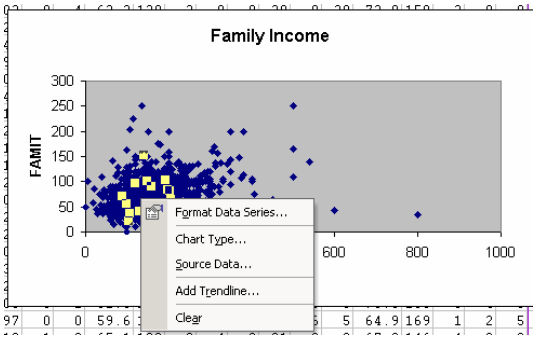
But what of the middle ground, where there is some relationship between two variables, but it is not a perfect relationship? To describe a non-perfect linear relationship we use a **Line of Best Fit**. This takes the form $y = a + bx$, where a is the *intercept* (where the line crosses the vertical y-axis) or *constant*, and b is the *gradient* or *slope* of the line (i.e. how steep it is). The sign of b indicates a positive or negative relationship. If b is zero, this indicates the absence of any linear relationship between the two variables x and y. If b is large (either positively or negatively), this indicates that a small change in x would lead to a large change in y.

As an example from the **STATLAB** data (**stalaba.xls**), we will look at the relationship between the Family Income at the point the child was aged10 (**FAMIT**) and when the child was born (**FAMIB**). Does the family income at the later time depend on what it was 10 years earlier?

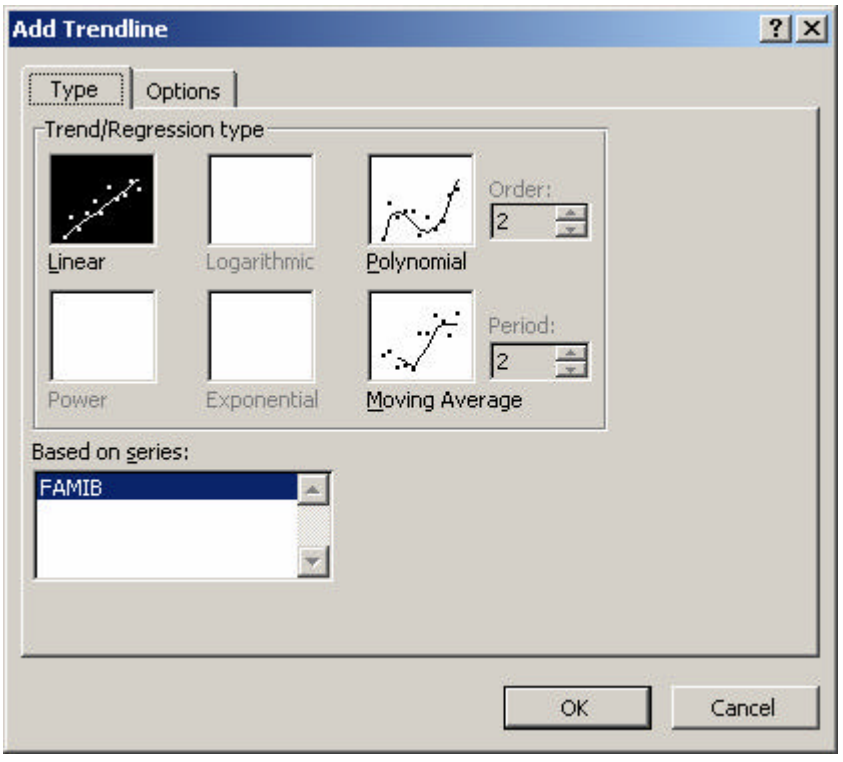
Firstly, we produce a Scatter Plot in the normal way, with **FAMIT** as the variable on the Y Axis, and **FAMIB** on the X Axis.



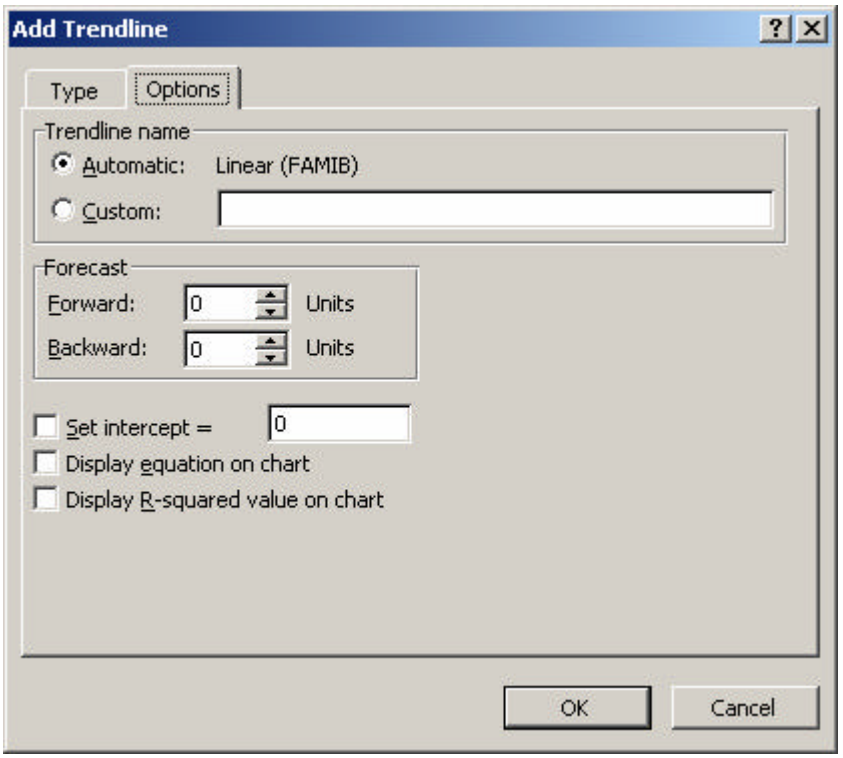
If you click on any data point in the graph, and then right click, you obtain the following menu.



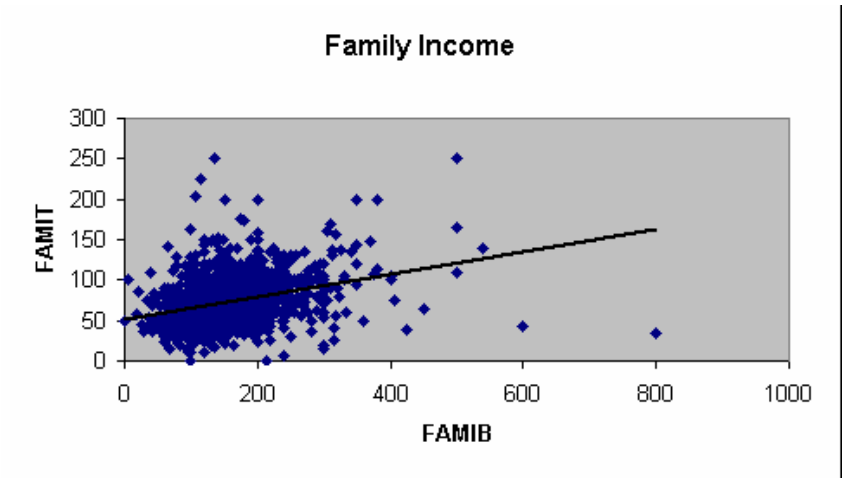
Click on **Add Trendline**.



To obtain a regression line, you have to choose the **Linear Option** and press **OK**. If you click on the **Options** tab, you obtain a series of options, such as displaying **R squared** or the equation.



The graph with the trendline is as follows.



The Scatter Plot and Line of Best Fit do not tell us the values of a and b; nor do they tell us if b is zero (or close enough to be taken as zero). It certainly seems that there is a positive relationship between the family income at the two points, but is this a **significant** relationship?

Simple Linear Regression

Linear **Regression** estimates the equation of the **Line of Best Fit**, using a technique called **Least Squares**. The **Least Squares Line** is the line that has the smallest sum of squared vertical distances from the observed points to the line.

In the above figure, imagine you have measured the distance in a vertical line from every point to the Line of Best Fit. Square these distances and add them together to get a total, T say. If you draw a different line through the points, and go through the same measuring procedure, you won't get a smaller value than T no matter which line you draw. The Line of Best Fit is just that – the best fit to all the points on the plot.

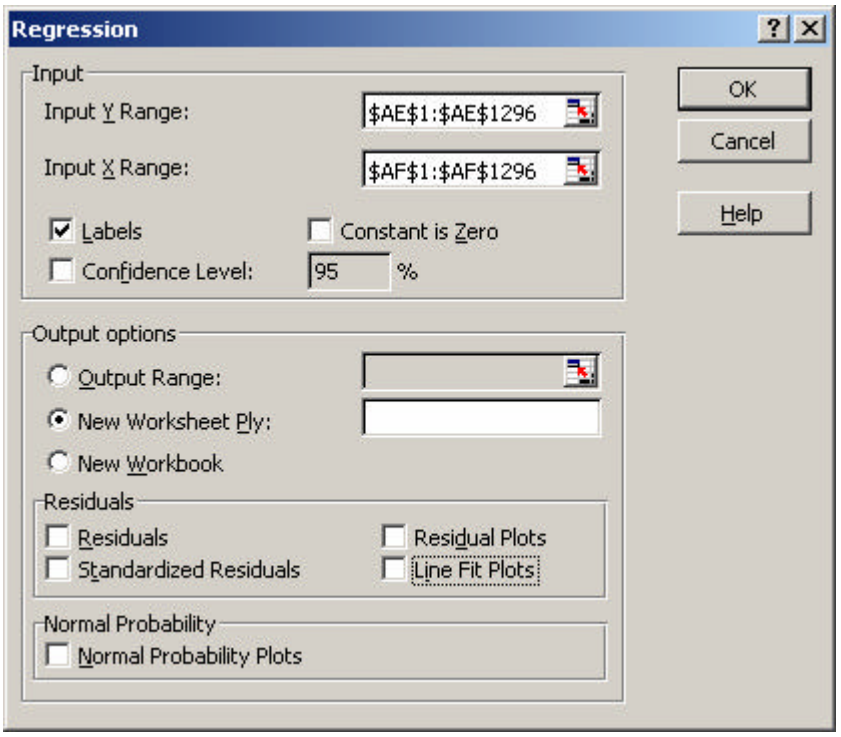
To perform a **Linear Regression** in Excel, we click on **Data Analysis** ➢ **Regression** and this gives the **Linear Regression** dialogue box.

The following table describes the options in the **Regression** dialog box:

Option	Description
Input Y Range	Enter the reference for the range of dependent data. The range must consist of a single column of data.
Input X Range	Enter the reference for the range of independent data. Excel orders independent variables from this range in ascending order from left to right. The maximum number of independent variables is 16.
Confidence Level	Select to include an additional level in the summary output table. In the box, enter the confidence level you want applied in addition to the default 95 percent level.
Constant is zero	Select to force the regression line to pass through the origin.
Residuals	Select to include residuals in the residuals output table
Standardised Residuals	Select to include standardized residuals in the residuals output table.
Residual Plots	Select to generate a chart for each independent variable versus the residual.
Line Fit Plots	Select to generate a chart for predicted values versus the observed values.
Normal Probability Plots	Select to generate a chart that plots normal probability.

The **Dependent** variable (or the **Y range**) is the Family Income at the time the child was 10 (**FAMIT**), and the **Independent** variable (or the **X range**)

is **FAMIB**, the income when the child was born. The filled in **regression form** should look like the following.



Click on **OK**, **Excel** executes the **Regression** macro according to your specifications. The following screen displays the output reflecting the data range selected.

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.323955				
5	R Square	0.104947				
6	Adjusted R Square	0.104254				
7	Standard Error	64.60484				
8	Observations	1295				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1	632773.8836	632773.9	151.6067098	5.04742E-33
13	Residual	1293	5396704.621	4173.785		
14	Total	1294	6029478.505			

	Coeff	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	99.5459	4.8652	20.4607	0.0000	90.0014	109.0905
FAMIB	0.7536	0.0612	12.3129	0.0000	0.6335	0.8736

The following table provides a brief description of the elements that Excel calculates for each row/column in the input range when you execute the **Regression** macro:

Option	Description
Multiple R	The absolute value of the correlation coefficient between life expectancy and birth rate is 0.97. That is the value labelled *r* in the table. This is a measure of how much the two variables are linearly related.
R Square	This is the square of the correlation. It tells you what proportion of variability of the dependent variable is explained by the regression model. In this example, close to 94% of the variability in observed female life expectancies is explained by birth-rate.
Adjusted R Square	It is an estimate of how well your model would fit another data set from the same population.
Standard Error	It is a measure of the variation given by the sample size.
ANOVA Table	Gives a measure of whether your model fits the data. The null hypothesis is that there is no linear relationship between the dependent and independent. In this case, we can see that there is a linear relationship.
Coefficients	Give the coefficients that Excel estimated both for the slope and for the intercept.
P-Value	When you calculate linear regression, you want to test whether there is a linear relationship between the two variables in the population. This is equivalent to testing the null hypothesis that the population slope is 0. In this example, the sample slope is –0.7 and its standard error is 0.05. Since the significance level is very small, we can reject the null hypothesis. Therefore there appears to be a linear relationship between 1992 female life expectancy and birth-rate.

The estimated values of a and b are displayed in the **Coefficients** column. This tells us that the equation of the Line of Best Fit ($y = a + bx$) is:

FAMIT = 99.546 + (0.754 * FAMIB)

This tells us that the Family Income when the child was aged 10 can be estimated by multiplying the Family Income at the time of the child's birth by 0.754 and adding 99.546.

Is this relationship between the two variables a significant one? In other words, is the coefficient of **FAMIB**, 0.754, significantly different from zero?

The Linear Regression procedure performs a test for this, and the results are produced in the final two columns.

Our **Null Hypothesis** is that the coefficient is zero (or not significantly different from zero). On the evidence of the **T-Test** in the **FAMIB** row of the table, we reject this hypothesis, since the **P-value** is less than 0.05.

Therefore we say that, at the 5% level, there is evidence that the Family Income at the child's birth has a significant effect on the Family Income 10 years later.

Strengths of Linear Relationships

We have just looked at relationships between variables and the Line of Best Fit through the points on a plot. Linear Regression can tell us whether any perceived relationship between the variables is a significant one. But what about the strength of a relationship? How tightly are the points clustered around the line? The strength of a linear relationship can be measured using the **Pearson Correlation Coefficient**.

The values of the **Correlation Coefficient** can range from –1 to +1. The following table provides a summary of the types of relationship and their Correlation Coefficients:

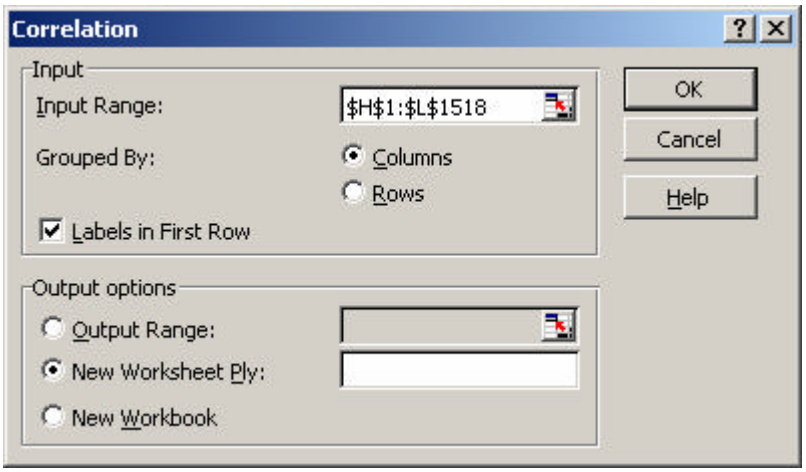
Linear Relationship	Correlation Coefficient
Perfect Negative	-1
Negative	-1 to 0
None	0
Positive	0 to +1
Perfect Positive	+1

The higher the Correlation Coefficient, regardless of sign, the stronger the linear relationship between the two variables.

From the **gss91t.xls** data set, we can look at the linear correlation between the education of the respondent (**EDUC**), that of the parents (**MAEDUC** and **PAEDUC**), the age of the respondent (**AGE**), and the Occupational Prestige Score (**PRESTG80**). Note that **Excel** does not do any test on the correlation, it simply calculates the value. You need to do further testing if you want to know whether the 2 variables are significantly correlated.

To obtain correlations, click on **Tools** ➢ **Data Analysis** ➢ **Correlation**. Fill in the Correlation form to reflect the columns of the 5 variables chosen. All possible pairs of variables from your chosen list will have the Correlation Coefficient calculated.

The completed **Correlation** form is as follows.



Click on **OK** and **Excel** executes the **Correlation** macro according to your specifications. The output obtained is as follows.

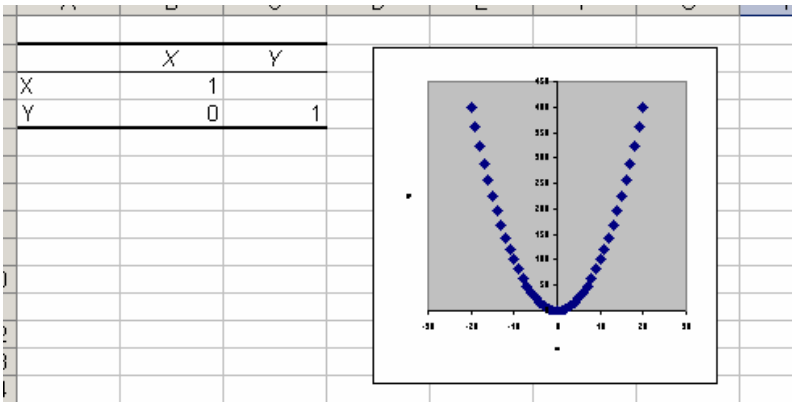
	AGE	EDUC	PAEDUC	MAEDUC	PRESTG80
AGE	1				
EDUC	−0.06872	1			
PAEDUC	0.150804	−0.09254	1		
MAEDUC	0.191532	−0.00993	0.378869	1	
PRESTG80	0.027144	0.128797	−0.12087	−0.12251	1

The strongest correlation is between **paeduc** and **maeduc**, however, we cannot really tell whether the other correlations are significant or not. We have to try to use multiple regressions and hence determine which variables can be excluded from the equation.

A Note on Non-Linear Relationships

It must be emphasised that we are dealing with **Linear** Relationships. You may find that the Correlation Coefficient is very close to 0, indicating no significant Linear Relationship between two variables, but they may have a **Non-Linear Relationship** which we are not testing for.

The following contains the results of the Correlation and Scatter Plot procedures performed on some hypothetical data. You can see that although the correlation between the 2 variables is 0, there is a very strong relationship (quadratic) between the 2 variables. It is always a good idea to check for relationships visually using graphics as well as using formal statistical methods!



	X	Y
X	1	
Y	0	1

Multiple Linear Regression

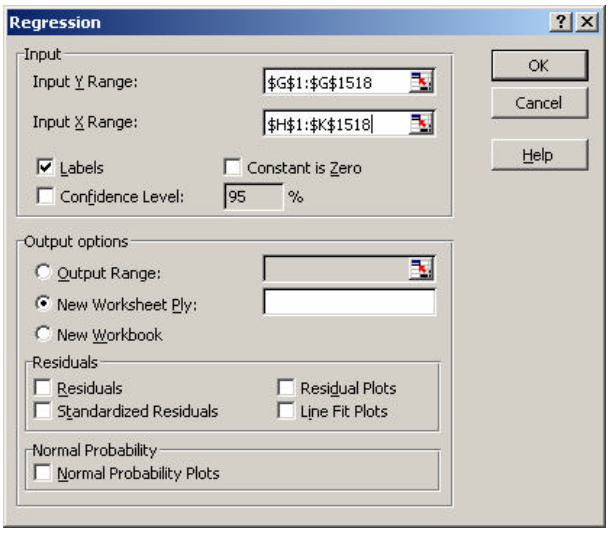
Simple Linear Regression looks at one **Dependent** variable in terms of one **Independent** (or **Explanatory**) variable. When we want to 'explain' a **Dependent** variable in terms of two or more **Independent** variables we use **Multiple Linear Regression**.

Just as in Simple Linear Regression, **Least Squares** is used to estimate the **Coefficients** (the constant and the *Bs*) of the Independent variables in the now more general equation:

$$Y = b_0 + b_1(x_1) + b_2(x_2) + \ldots$$

where b_0, b_1, b_2 are constants, and x_1, x_2, x_3,… are the independent variables.

We are going to use the **gss91t.xls** data set. We will investigate the effect of the respondent's age (**AGE**), sex (**SEX**), education (**EDUC**) and spouse's education (**SPEDUC**) on the Occupational Prestige score (**PRESTG80**).
Follow the same procedure as with simple regression; however choose all the independent variables in the **Input X Range**.



In the **Linear Regression** dialogue box, we choose **PRESTG80** as our **Dependent** variable, and **EDUC**, **SPEDUC**, **AGE** and **SEX** (not a continuous variable, but, as it is a binary variable, we can use it if we interpret the results with care) as the **Independent** variables.

	A	B	C	D	E	F
	SUMMARY OUTPUT					
	Regression Statistics					
	Multiple R	0.225834				
	R Square	0.051001				
	Adjusted R Square	0.04849				
	Standard Error	16.09133				
	Observations	1517				
	ANOVA					
		df	*SS*	*MS*	*F*	*Significance F*
	Regression	4	21040.14667	5260.037	20.31444	2.57113E-16
	Residual	1512	391503.4855	258.9309		
	Total	1516	412543.6322			

5							
6		*Coeff*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
7	Intercept	43.2090	1.9861	21.7554	0.0000	39.3131	47.1048
8	AGE	0.0376	0.0232	1.6244	0.1045	-0.0078	0.0830
9	SEX	-4.1645	0.8439	-4.9346	0.0000	-5.8199	-2.5091
0	EDUC	0.3261	0.0633	5.1514	0.0000	0.2019	0.4502
1	SPEDUC	-0.0478	0.0099	-4.8468	0.0000	-0.0672	-0.0285
2							

The 3rd table could be used to write out the equation for **PRESTG80** in terms of a **constant**, **EDUC**, **SPEDUC**, **AGE** and **SEX**, just as we did for Simple Linear Regression.

If we then estimate **PRESTG80** from this equation for each respondent, using their values of the **Explanatory** variables, and perform a **Correlation** calculation between this **Predicted** variable and the actual variable **PRESTG80**, we will have a measure of how strongly related they are. This is the value in row **Multiple R** of the **Regression Statistics** table.

Simply put, the closer this value is to 1, the more accurate is our prediction; if it is close to zero, our model is not explaining what is happening very well.

The figure in the row headed **R Square** is the proportion of variability in the **Dependent** variable that can be explained by changes in the values of the **Independent** variables. This is calculated from the figures in the 2nd table, **ANOVA** (the Analysis Of Variance); it is the **Regression Sum of Squares** divided by the **Total Sum of Squares**. The higher this proportion, the better the model is fitting to the data.

The **ANOVA** table also indicates whether there is a significant Linear Relationship between the **Dependent** variable and the combination of the **Explanatory** variables; an **F-Test** is used to test the **Null Hypothesis** that there is no Linear Relationship. We can see in our example that, with a

Significance value (**Sig.**) of less than 0.05, we have evidence that there **is** a significant Linear Relationship

In the 3rd table we have the figures that will be used in our equation. All four **Explanatory** variables have been entered, but should they all be there? Looking at columns, headed **t-Stat** and **P-value.**, we can see that the Significance level for the variable **age** is more than 0.05. This indicates that, when the other variables (a **constant**, **EDUC**, **SPEDUC** and **SEX**) are used to explain the variability in **PRESTG80**, using **AGE** as well doesn't help to explain it any better; the Coefficient of **AGE** is **not significantly different from zero**. It is not needed in the model.

If you want to remove this variable from the model, simply run the procedure again, removing the range of the variable **AGE**.

Practical Session 8

1. The Scores of the Offspring of Different Paternal Occupational Groups

Open the data file **statlaba.xls**. At the age of ten, the children in the sample were given two tests; the **Peabody Picture Vocabulary Test** and the **Raven Progressive Matrices Test**. Their scores are stored in the variables **CTP** and **CTR**.

Create a new variable called **TESTS** which is the sum of the two tests; this new variable will be used in Questions 1 and 3.

State your **Null** and **Alternative Hypotheses**, and which of the two you accept on the evidence of the relevant test, and the **Significance Level**.

The fathers' occupation is stored in the variable **FTO**, with the following categories:

0 Professional
1 Teacher / Counsellor
2 Manager / Official
3 Self-employed
4 Sales
5 Clerical
6 Craftsman / Operator
7 Labourer
8 Service worker

Recode **FTO** into a new variable, **OCCGRP**, with categories:

1 Self-employed
2 Professional/ Manager / Official
3 Teacher / Counsellor
4 Sales/ Clerical/ Service worker
5 Craftsman / Operator
6 Labourer

Modify the data file so that a One-Way ANOVA could be applied. Using a **One-Way ANOVA**, test whether there is any difference between the occupation groups in terms of the test scores of their children.

2. Weight against Height

Use the **statlaba.xls** dataset. Produce a **Scatter Plot** of the child's weight at age 10 (**CTW**) against the child's height at that time (**CTH**). *(NB – 'Y against X')*

Superimpose the **Line of Best Fit** on the plot.

Perform a Linear Regression to estimate the Line of Best Fit.

At the 5% level, is there evidence that the child's height significantly affects how much the child weighs? If so, what happens as the child grows taller?

Estimate the weight of a 10 year old child who is 52 inches tall.

3. The Years of Education of Mother and Child

Use the **gss91t.xls** data set. Using the variables **EDUC** and **MAEDUC**, investigate whether the mother's education has a significant effect on her child's.

4. Age and Occupational Prestige

Is the Occupational Prestige Score (**PRESTG80**) significantly affected by the age of the respondent (**AGE**)?

Produce some **Descriptive Statistics** of **PRESTG80** and compare them to your **Linear Regression**.

How would you estimate the Occupational Prestige Score for a respondent aged 32? How about one aged 67?

5. A Child's Weight and Physical Characteristics

Use the **statlaba.xls** data file. Use the methods from this session to investigate the relationship between the weight of the child at age 10 (**CTW**) and some physical characteristics:

CBW	child's weight at birth
CTH	child's height at age 10
SEX	child's gender (coded 1 for girls, 2 for boys)

6. A Child's Weight and Hereditary Characteristics

Repeat Question 1, but instead use the following explanatory variables:

FTH	Father's height
FTW	Father's weight
MTH	Mother's height
MTW	Mother's weight

7. Educational Relationships

Use the **gss91t.xls** data set. Investigate the Linear Relationships between the following variables using Correlations:

EDUC	Education of respondent
MAEDUC	Education of respondent's mother
PAEDUC	Education of respondent's father
SPEDUC	Education of respondent's spouse

Using Linear Regression, investigate the influence of education and parental education on the choice of marriage partner (Dependent variable **SPEDUC**). Use the variable **SEX** to distinguish between any gender effects.

It is thought that the size of the family might affect educational attainment. Investigate this using **EDUC** and **SIBS** (the number of siblings) in a Linear Regression.

Also investigate whether the education of the parents (**MAEDUC** and **PAEDUC**) affects the family size (**SIBS**).

How does this result influence your interpretation of the model? Are you perhaps finding a spurious effect? Test whether **SIBS** still has a significant effect on **EDUC** when **MAEDUC** and **PAEDUC** are included in the model.

8. Average Years of Education

Compute a new variable **PARED** = (**MAEDUC** + **PAEDUC**) / 2, being the average years of education of the parents. By including **PARED**, **MAEDUC** and **PAEDUC** in a Linear Regression, investigate which is the better predictor of **EDUC**; the separate measures or the combined measure.