How to get started.

This is a tutorial for you to work through at the computer at your own pace. Please make sure that you have a floppy disk which will also be supplied by Andrew. You can now proceed following the tutorial. If you have any questions or problems let Andrew know and he will help out.

1. Switch on the PC.

The power button is on the right hand side of the PC System Unit at the front. "Booting Up" takes about 60 seconds.

After booting up the PC will display a log on box. Simply leave the password blank and hit enter.

2. Access Excel from Windows 95.

Excel is accessed either by double clicking on the Excel Icon or through the Start button / Programs options. Let me know if you have not used Windows 95 or Windows 3.1 before and I will happily give you a personalized introduction.

3. Disk Drives.

The LAN has several "logical disk drives" . A: is the floppy disk as on your PC, but there are also C: through Z:

For this session, you will use the floppy disk supplied which contains the data files that you will need to follow the tutorials. When opening or saving files make sure that drive A: is selected.

Using EXCEL for Statistical Analysis, a brief Tutorial.

I am making the assumption that you followed the general instructions and that you are now sitting in front of Windows 95 or Windows NT

If not, get me to come over.

So in this session we are going to:-

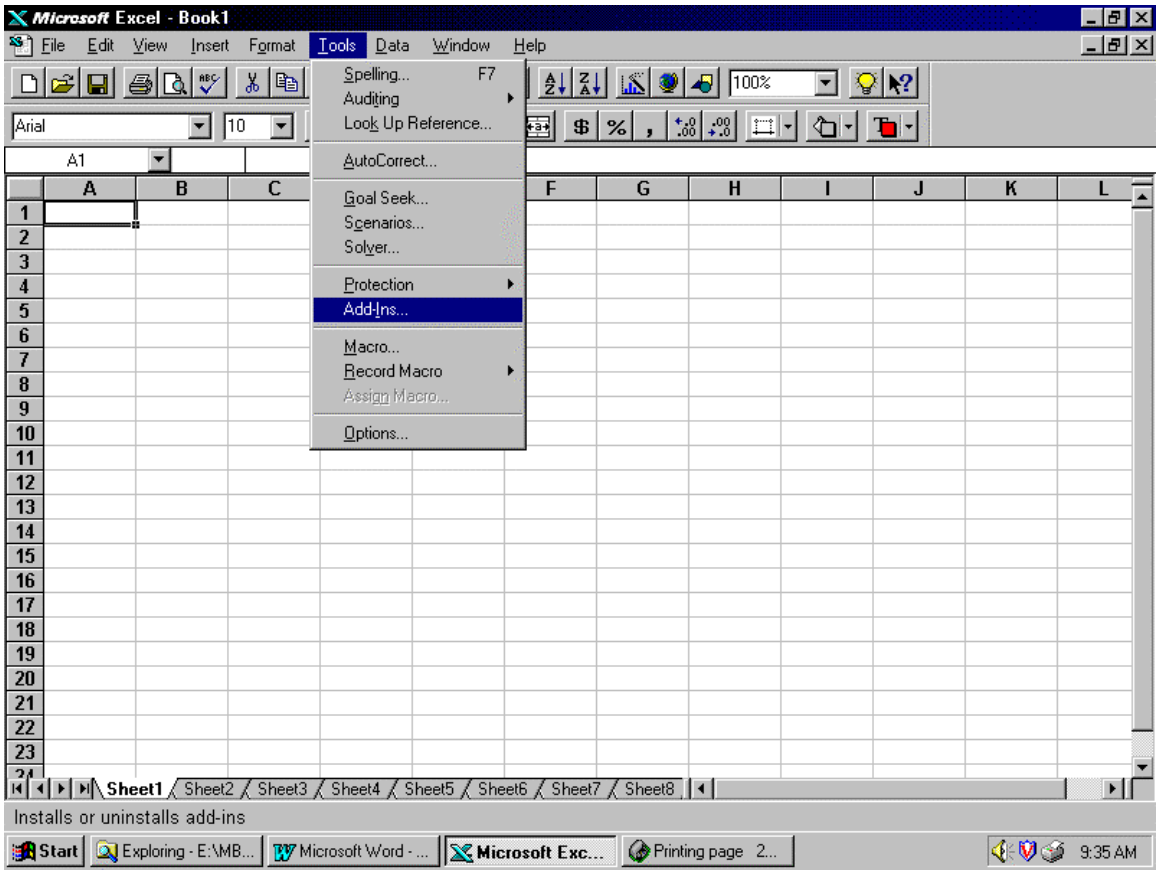
1. Learn how to install the Data Analysis Tools available in Excel.
2. Learn how to import data in other formats.
3. Learn how to download data from the World Wide Web.
4. Do correlation analysis on data on your floppy disk.
5. Perform a linear regression on data on your floppy disk.
6. Perform a multiple regression on data on your floppy disk.
7. If time allows, consider how to use the other statistical functions that are available.

If Excel is completely new to you, let me know and I will come over and give you an introduction.

1. Install the Data Analysis Tools available in Excel.

The following works for Excel version 5 and version 7. If you have an earlier version then the following may work, but I cannot guarantee it. On the SOM LAN Excel has the Tools installed, but for your own machine these are the procedures. Try them on the SOM LAN, anyway.

Select the "Tools" Menu and the "Add-Ins" Option as shown below.

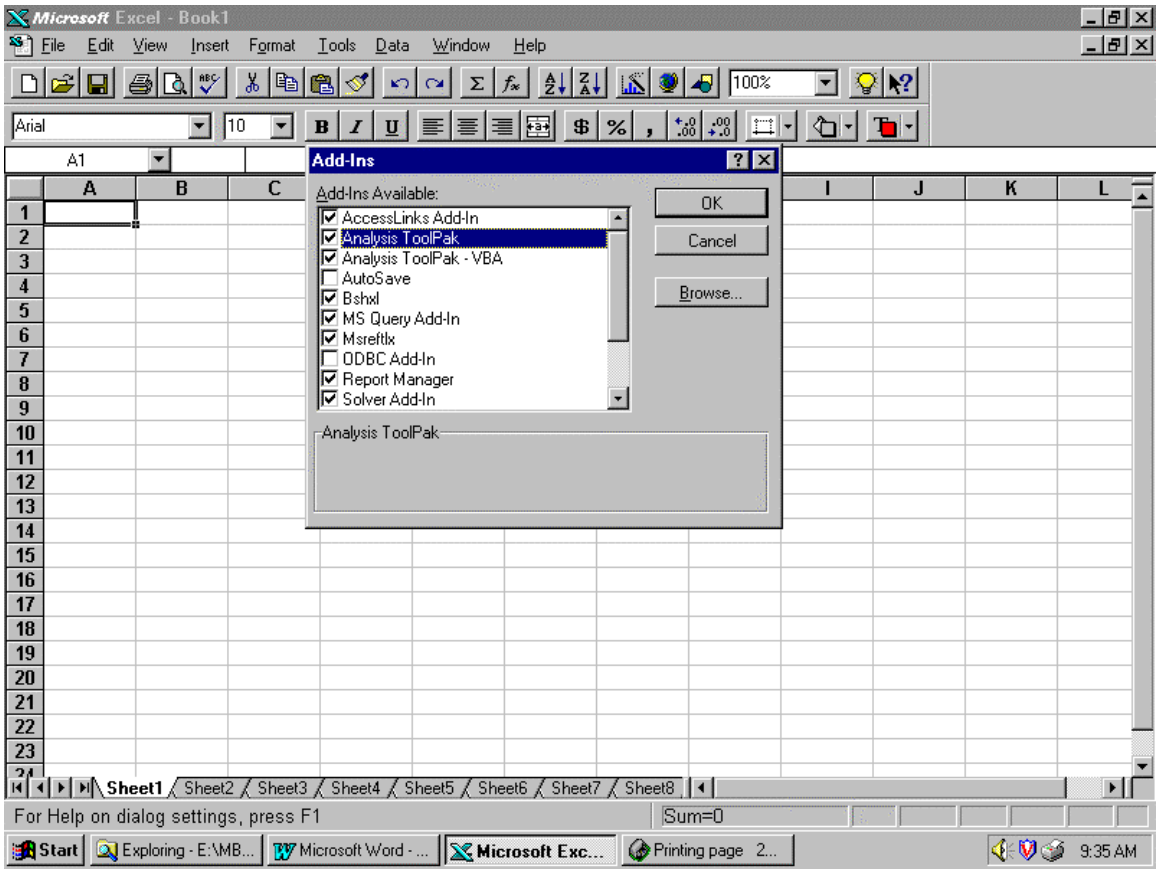


If the last item on the tools menu is "Data Analysis", then the tools are already installed.

Set the check boxes

- Analysis ToolPak,
- Analysis ToolPak - VBA

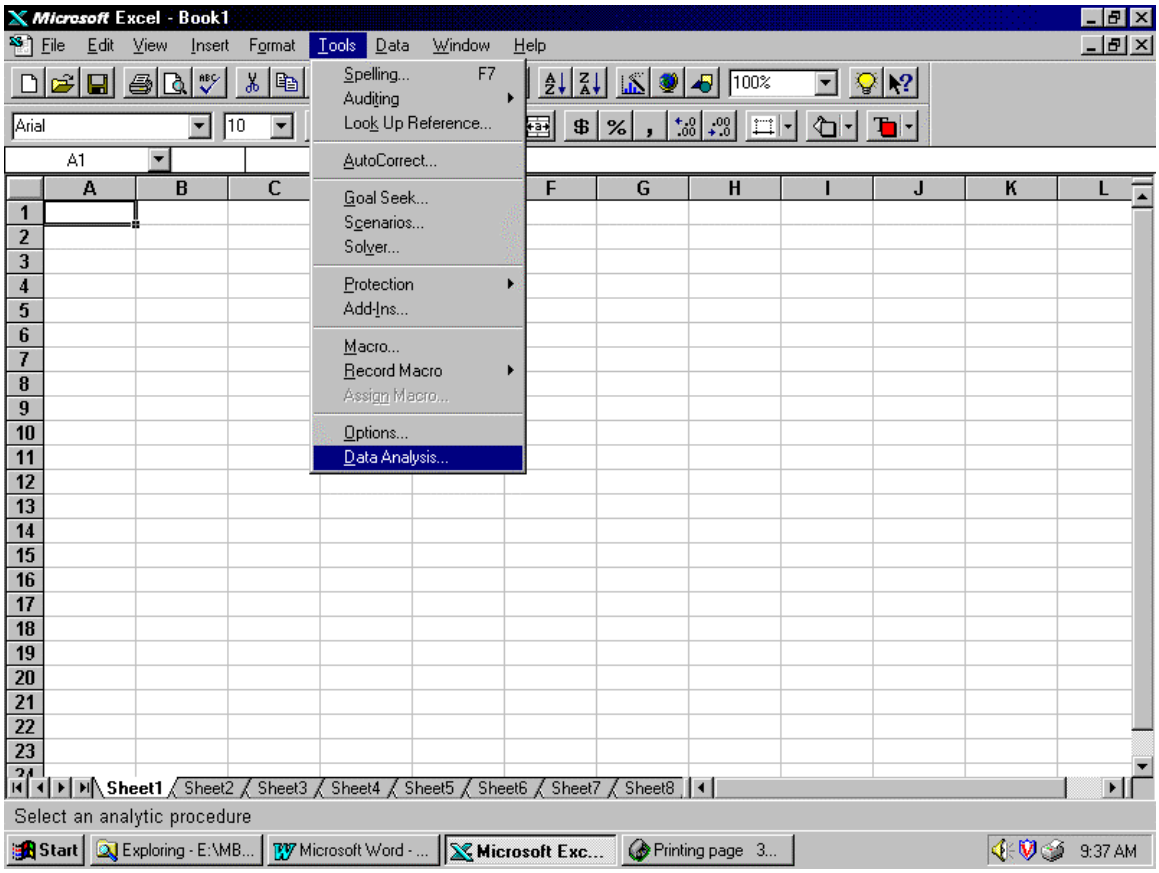
as shown on the following page.



Then press OK.

At home, Excel may ask you to insert disks from the Excel or Office disks that you used to load Excel in the first place, so have these handy.

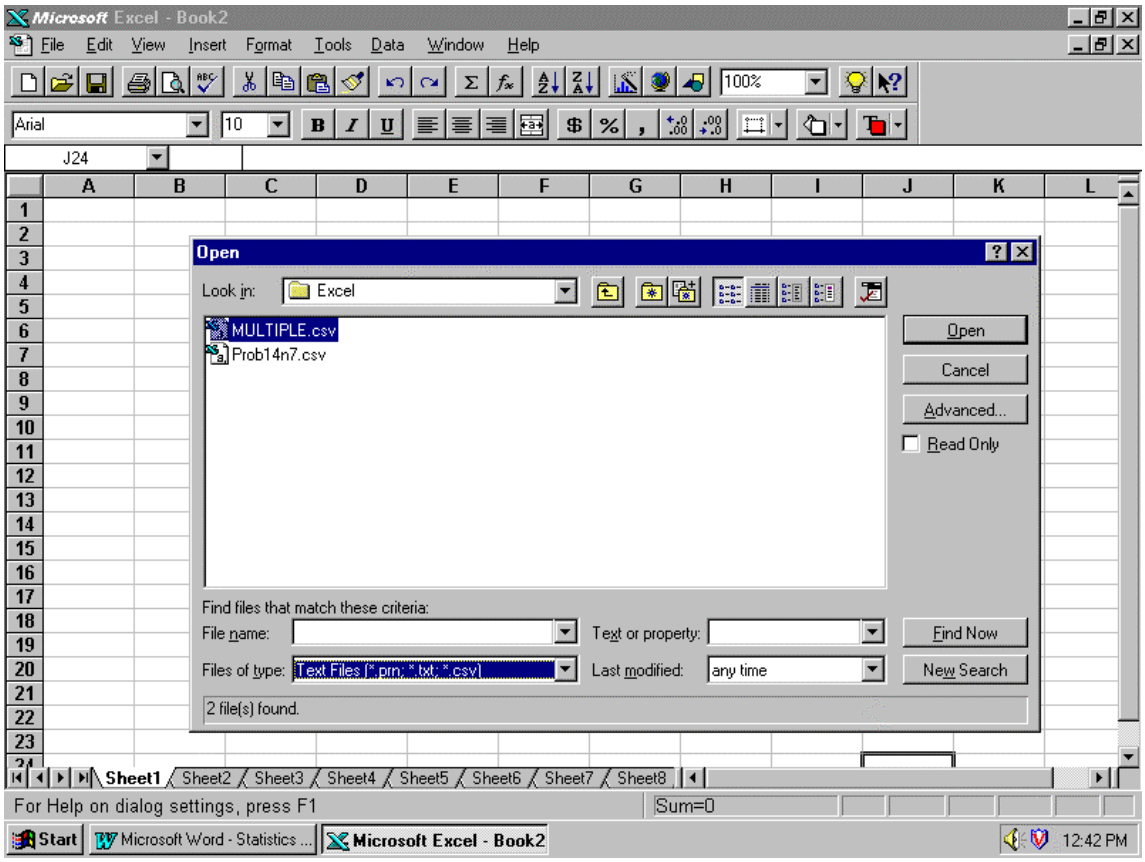
Now when you select the Tools pull down menu, the last item in the menu will be Data Analysis as you can see on the following page.


Microsoft Excel - Book1
File Edit View Insert Format Tools Data Window Help
Spelling... F7
Auditing
Look Up Reference...
AutoCorrect...
Goal Seek...
Scenarios...
Solver...
Protection
Add-Ins...
Macro...
Record Macro
Assign Macro...
Options...
Data Analysis...
Arial
10
A1
100%
Sheet1 Sheet2 Sheet3 Sheet4 Sheet5 Sheet6 Sheet7 Sheet8
Select an analytic procedure
Start
Exploring - E:\MB...
Microsoft Word - ...
Microsoft Exc...
Printing page 3...
9:37 AM

2. Import Data for analysis from files on the floppy disk.

The data that you are trying to analyze could come from a number of different sources. You may down load data from a government Web Site or from your company's accounting system or from a database package. Whichever source you select you will want to be able to import the data into Excel.

To import a file you simply open the file from within Excel using the File Open options on the pull down menus. You will then have the "open" window in front of you. Change the "Files of Type" to the type of file you want to import and then select your file … in this case MULTIPLE.csv. "csv" stands for comma separated variables which describes the way in which the data is stored.



Click on Open when you have highlighted the file that you want or double click on the filename. In the case of the "multiple.csv" file the data is read without any problems by Excel. In other cases a "Wizard" may prompt you with questions which will allow it to put the data into sensible columns.

There are a number of different formats which are commonly used to store data and pass it around and you may encounter:

- Text files where columns of data are separated by commas or tabs or surrounded by quotation marks or with fixed column widths (fixed record length).
- Dbase files from DBaseII, III, or IV or from FoxPro or Clipper. (Usually named as *.dbf).

To practice, import multiple.csv; multiple.dbf and multiple.txt from the floppy disk.

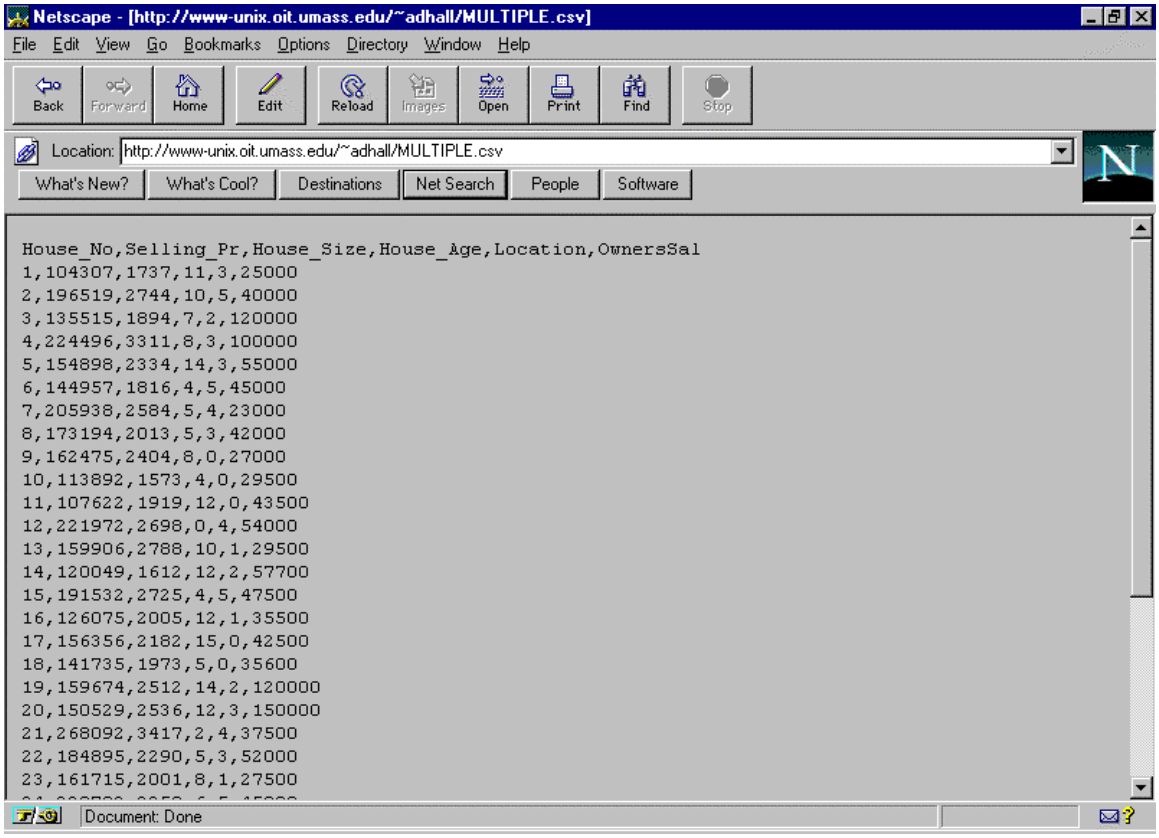
3. Getting Data From a Web Site.

The SOM PCLabs have been undergoing a major upgrade this summer and when I tried this yesterday Netscape was not working on the PC that I tried so you may have to try this from at home.

Invoke Netscape or your favorite browser and call up the following page:

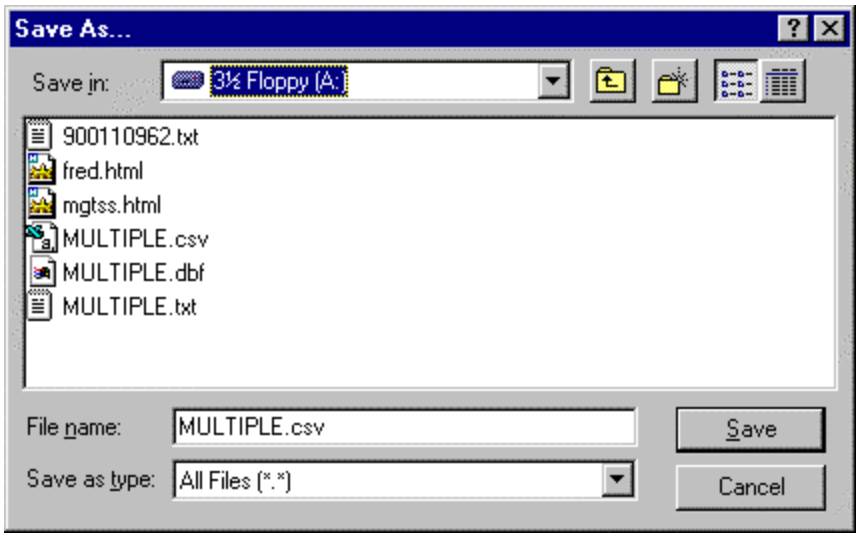
http://www-unix.oit.umass.edu/~adhall/andrew.htm

Double click on “Multiple.csv” and you should see the following:



Go to the File pull down and select Save as.

Fill in the save as options as follows:



and select save using the Save button.

Netscape will save a copy of the file on your floppy disk.

Even if what you saw on the screen was meaningless, when you save it, it may nevertheless be meaningful. A good example is the Multiple.dbf file which cannot be read by Netscape's editor but when saved will be readable by Excel.

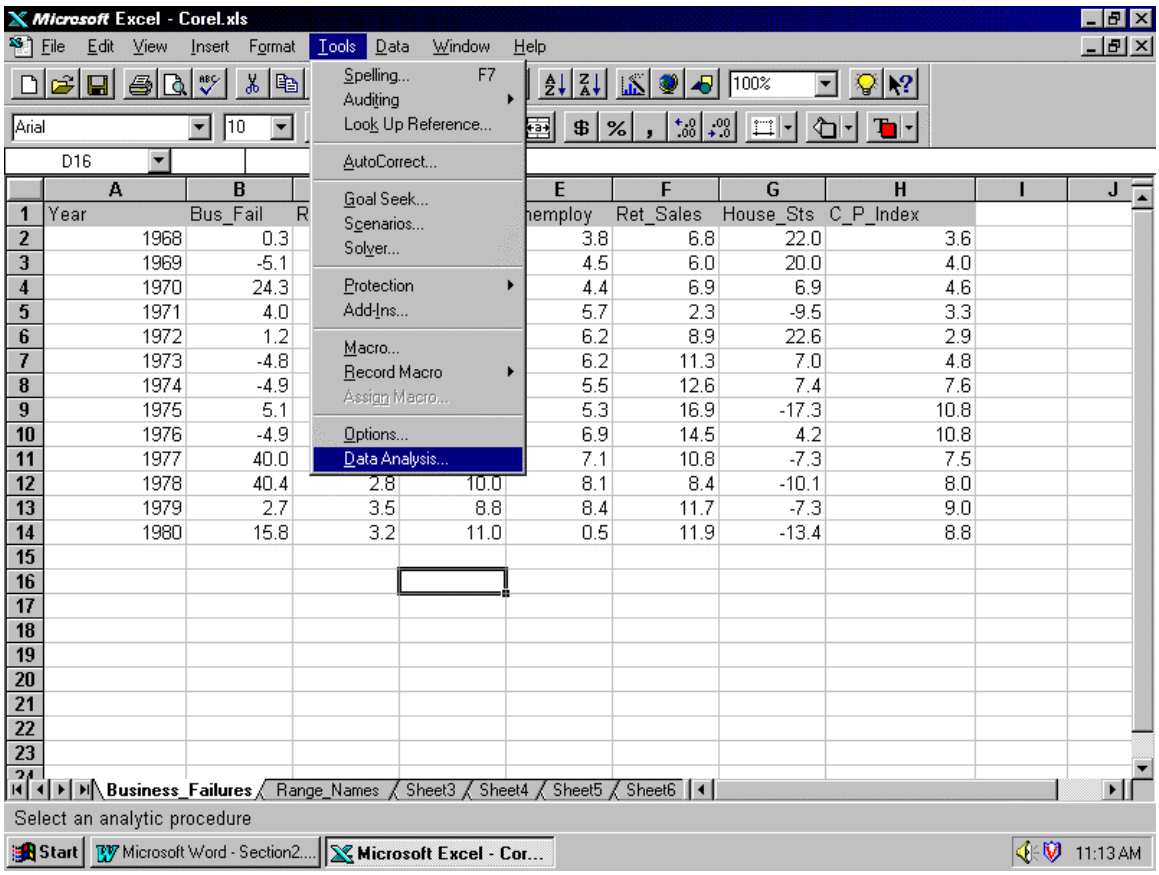
4. Do correlation analysis on data on the floppy disk.

First load up the Worksheet *corel.wks* from the floppy disk in the data directory *A:\data*

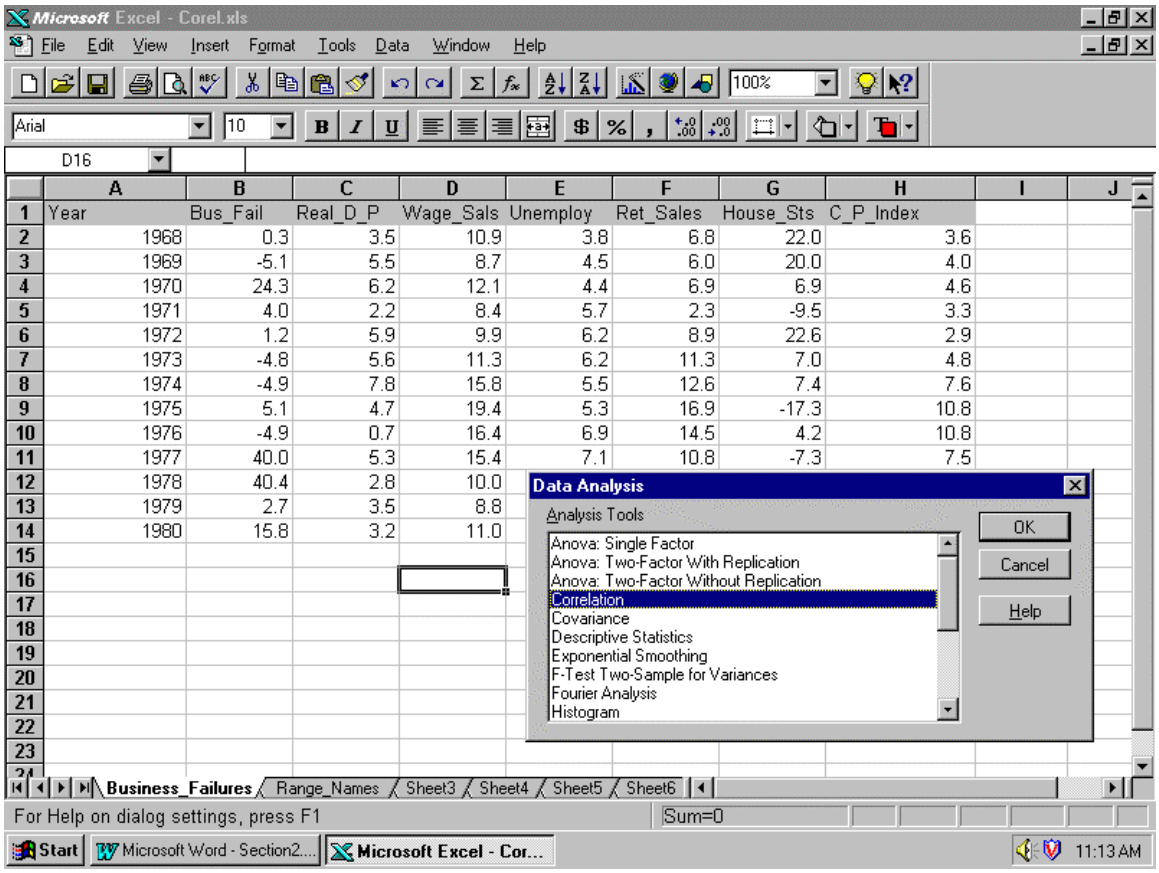
I have named two ranges to make selection of ranges easier. These are:-

Named Ranges	The Range
All_Data	=Business_Failures!$A$1:$H$14
All_Numeric_Data	=Business_Failures!$B$1:$H$14

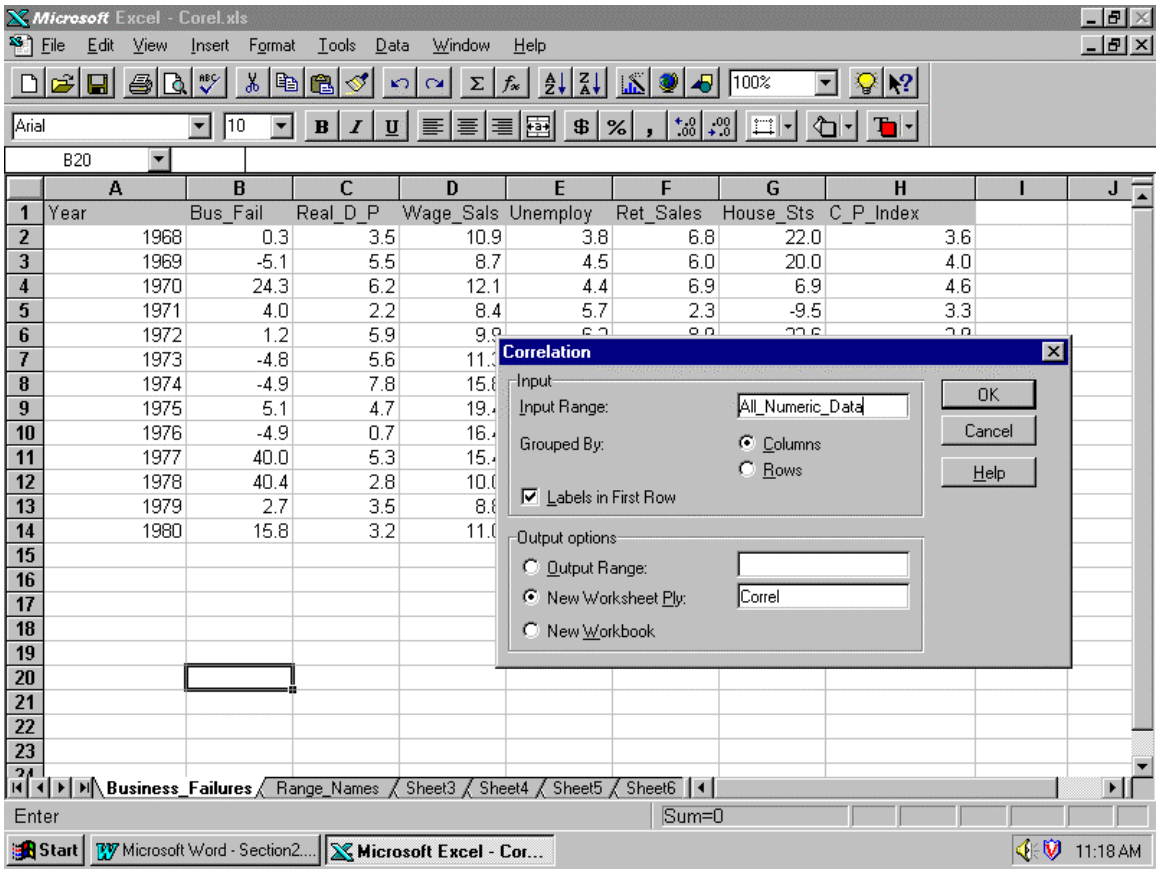
Then you need to access the Data Analysis tools which you do using Tools and the last option in the pull down menu. As you can see in the following panel.



Now select Correlation as on the following screen:-



	A	B	C	D	E	F	G	H	I	J
1	Year	Bus_Fail	Real_D_P	Wage_Sals	Unemploy	Ret_Sales	House_Sts	C_P_Index		
2	1968	0.3	3.5	10.9	3.8	6.8	22.0	3.6		
3	1969	-5.1	5.5	8.7	4.5	6.0	20.0	4.0		
4	1970	24.3	6.2	12.1	4.4	6.9	6.9	4.6		
5	1971	4.0	2.2	8.4	5.7	2.3	-9.5	3.3		
6	1972	1.2	5.9	9.9	6.2	8.9	22.6	2.9		
7	1973	-4.8	5.6	11.3	6.2	11.3	7.0	4.8		
8	1974	-4.9	7.8	15.8	5.5	12.6	7.4	7.6		
9	1975	5.1	4.7	19.4	5.3	16.9	-17.3	10.8		
10	1976	-4.9	0.7	16.4	6.9	14.5	4.2	10.8		
11	1977	40.0	5.3	15.4	7.1	10.8	-7.3	7.5		
12	1978	40.4	2.8	10.0						
13	1979	2.7	3.5	8.8						
14	1980	15.8	3.2	11.0						
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										



I suggest that you complete the options as shown above. Using named ranges makes things easier. If you do more correlations in this file and you want to keep the results then you should give a different "New Worksheet Ply" each time.

If you want to look at a couple of columns, say Bus_Fail and Unemploy, then you have to move them so that they are side by side as Excel will not allow you to define a split range in this instance.

<u>The following is from Excel's help.</u>

Correlation measures the relationship between two data sets that are scaled to be independent of the unit of measurement. The population correlation calculation returns the covariance of two data sets divided by the product of their standard deviations.

You can use the Correlation tool to determine whether two data sets move together; that is, whether large values of one set are associated with large values of the other (positive correlation), whether small values of one set are associated with large values of the other (negative correlation), or whether values in both sets are unrelated (correlation near zero).

Input Range

Type the reference for the range of worksheet data you want to analyze. The input range should contain two or more blocks of data arranged in columns or rows. If your input range includes row or column labels, you must select the Labels In First Column or Labels In First Row check box, or Microsoft Excel displays a message.

Output Range

Type the reference for the upper-left cell of the output table. Only half of the table is complete because correlation between two data sets is independent of the order in which the data sets are processed. Cells in the output table with matching row and column coordinates contain the value 1 because each data set correlates perfectly with itself. The output table is a square whose width and height are one larger than the number of cell ranges of data. If existing data is about to be overwritten in the output range, Microsoft Excel displays a message.

New Worksheet Ply

This option inserts a new ply into the workbook the current ply resides in and pastes the results into cell A1 of the new ply. Use the text box next to New Worksheet Ply to optionally name the new ply.

New Workbook

This option creates a new workbook, adds a new workbook ply in the new workbook, and pastes the results into cell A1 of the new ply.

Grouped By

Select either the Rows or Columns option button to indicate whether the data in the input range is arranged in rows or in columns.

Labels In First Row/Labels In First Column

If you select the Columns option button under Grouped By and the first row of your input range contains labels, select the Labels In First Row check box.

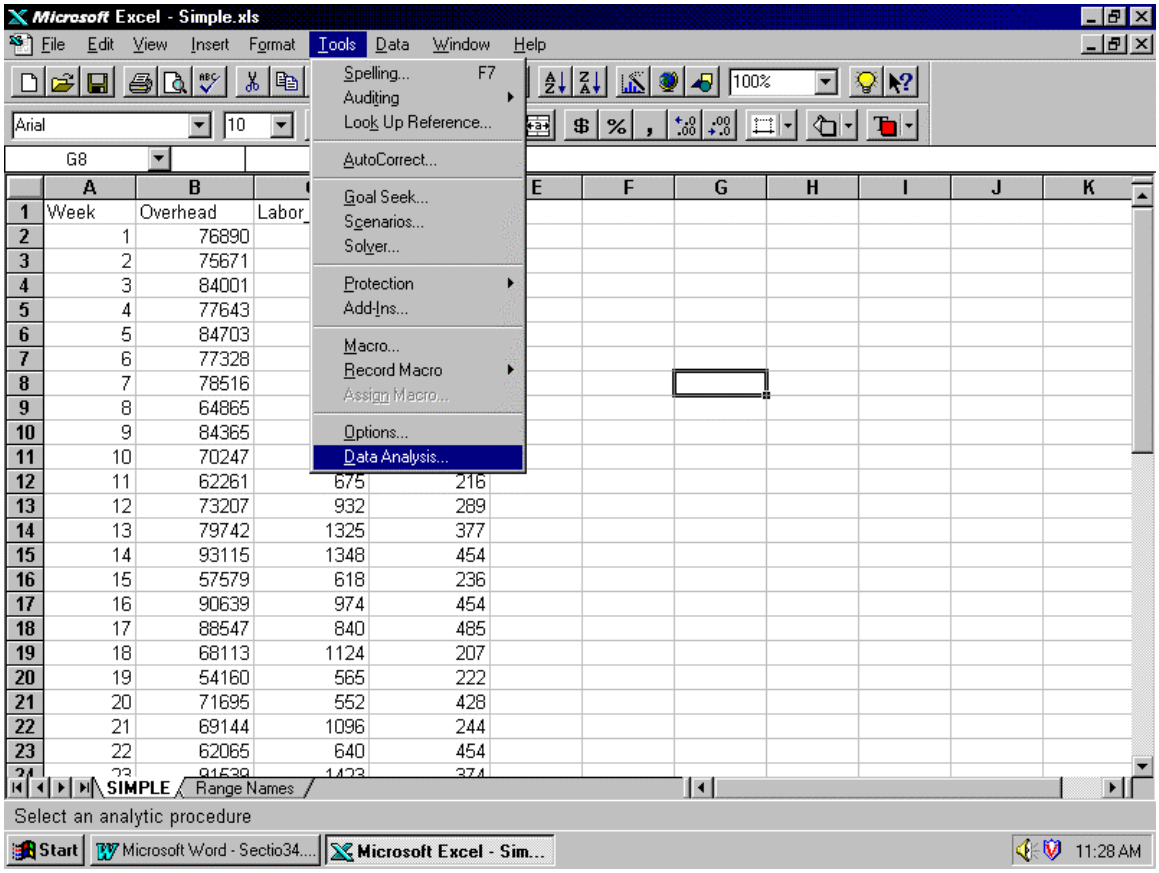
If you select the Rows options button under Grouped By and the first column of your input range contains labels, select the Labels In First Columns check box.

If your input range does not include labels, clear the Labels In First Row or Labels In First Column check box. Microsoft Excel then generates the appropriate data labels for the output table (Row 1, Row 2, Row 3, and so on; or Column 1, Column 2, Column 3, and so on).

5. Perform a linear regression on data on the floppy disk.

First load up the Worksheet *simple.xls* from the floppy disk in the data directory *A:\data*

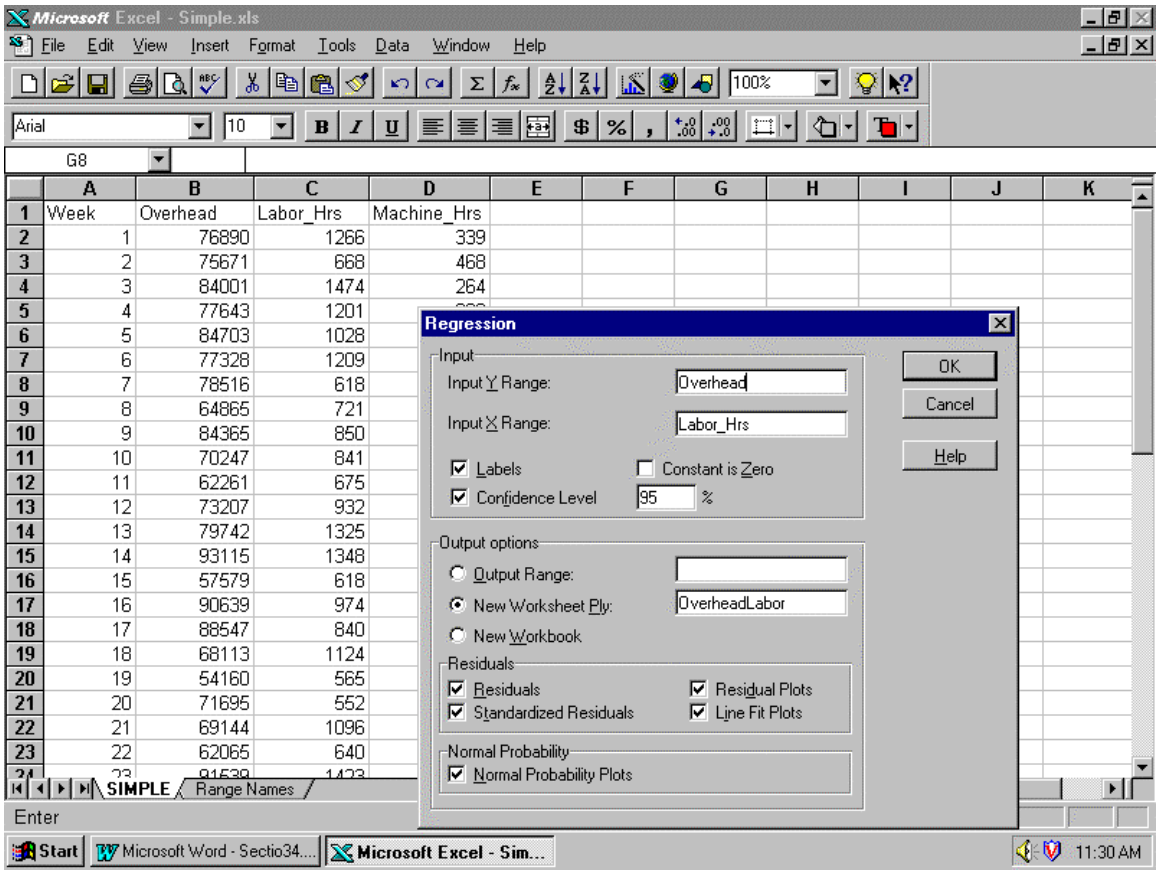
Then you need to access the Data Analysis tools which you do using Tools and the last option in the pull down menu. As you can see in the following panel.



	A	B	C	D
1	Week	Overhead	Labor_	
2	1	76890		
3	2	75671		
4	3	84001		
5	4	77643		
6	5	84703		
7	6	77328		
8	7	78516		
9	8	64865		
10	9	84365		
11	10	70247		
12	11	62261	675	216
13	12	73207	932	289
14	13	79742	1325	377
15	14	93115	1348	454
16	15	57579	618	236
17	16	90639	974	454
18	17	88547	840	485
19	18	68113	1124	207
20	19	54160	565	222
21	20	71695	552	428
22	21	69144	1096	244
23	22	62065	640	454

Now select a Regression Analysis ...

Now complete the form.



I suggest that you complete the options as shown above. I have named ranges Overhead, Labor_Hrs and Machine_Hrs which makes it easier to enter ranges. When you do the second regression of Overhead against Machine_Hrs remember to use a different Worksheet name to the one given above.

You will need to change column widths and so on to really see what you have as a result.

The resulting equation is Total Overhead = $ 49,615 + $ 28.60 for every labor hour.

If you run Overhead against Machine_Hrs you should get:

Total Overhead = $48,783 + $82.36 for every machine hour.

The following is from Excel's help.

Input Y Range

Type the reference for the range of dependent data you want to analyze. The dependent data should be typed in a single column. If you include a report label in the first row of the input y range, you must select the Labels check box, or Microsoft Excel displays a message.

Input X Range

Type the reference for the range of independent data you want to analyze. The data should be typed in a single column for a simple analysis.

Constant Is Zero

Select the Constant Is Zero check box to force the regression line to pass through the origin.

Labels

If the first row in the input range includes report labels, select the Labels check box.
If your input range does not contain labels, clear the Labels check box. Microsoft Excel then generates the appropriate data labels for the summary output table.

Confidence Level

Select the Confidence Level check box if you want an additional level included in the summary output table. In the Confidence Level box, type an additional confidence level that you want Microsoft Excel to apply to the regression in addition to the default 95% confidence level.

Output Range

Type the reference for the upper-left cell of the range where you want the summary output table to appear. Allow at least seven columns for the summary output table.
The summary output table includes the following:

- Anova table
- Coefficients
- Standard error of y estimate
- r2 values
- Number of observations
- Standard error of coefficients

New Worksheet Ply

This option inserts a new ply into the workbook the current ply resides in and pastes the results into cell A1 of the new ply. Use the box next to New Worksheet Ply to optionally name the new ply.

New Workbook

This option creates a new workbook, adds a new workbook ply in the new workbook, and pastes the results into cell A1 of the new ply.

Residuals

Select the Residuals check box if you want to include residuals in the residuals output table.

Standardized Residuals

Select the Standardized Residuals check box if you want to include standardized residuals in the output table.

Residual Plots

Select the Residual Plots check box if you want Microsoft Excel to generate a chart for each independent variable versus the residual.

Line Fit Plot

Select the Line Fit Plot check box if you want Microsoft Excel to generate a chart for predicted values versus the observed values.

Normal Probability Plot

Select the Normal Probability Plot check box if you want Microsoft Excel to generate a chart plotting normal probability plots.

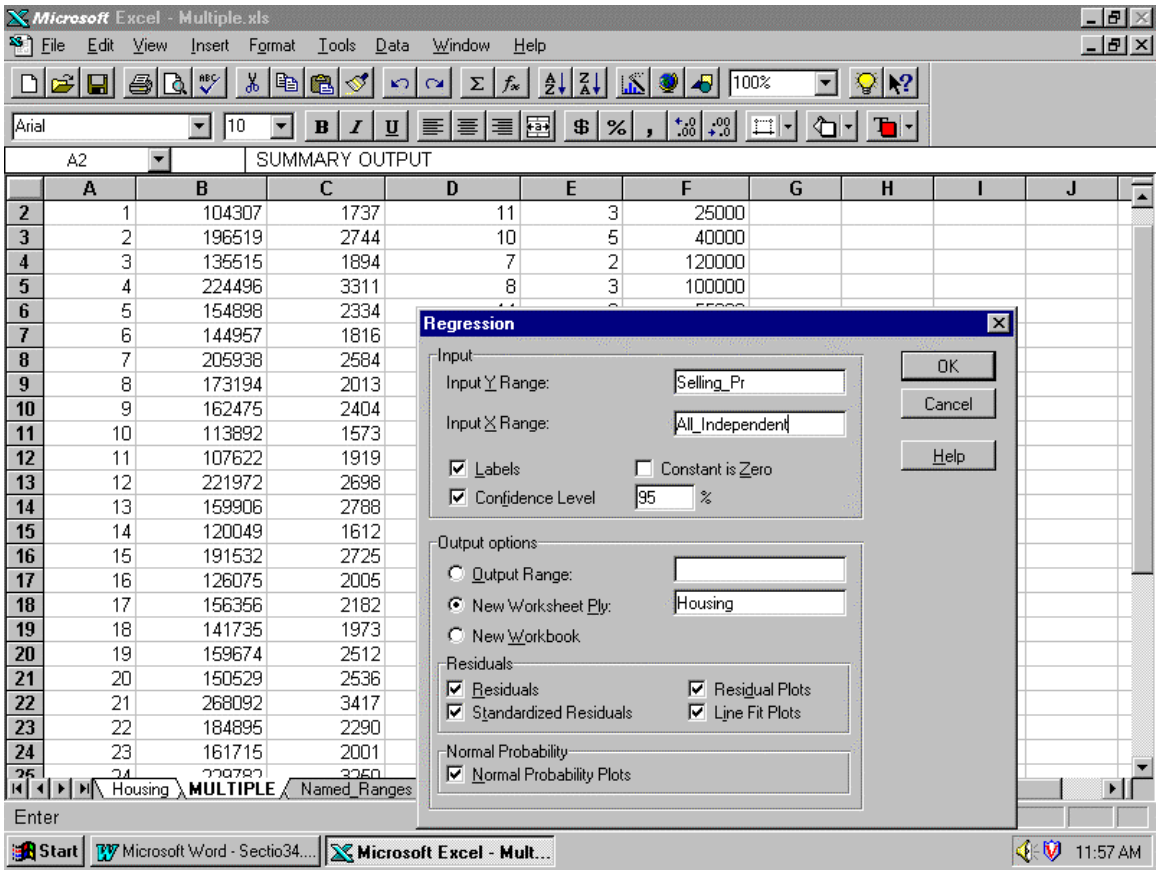
6. Perform a multiple regression on data on the floppy disk.

First load up the Worksheet *multiple.xls* from the floppy disk in the data directory *A:\data*

In Multiple.xls, I have also named ranges as follows:-

All_Independent	=MULTIPLE!$C$1:$F$36
Selling_Pr	=MULTIPLE!$B$1:$B$36

The main difference between simple and multiple regression analysis is that multiple regression involves more than one Independent Variable and hence you can specify more columns for the Independent Variable range. Up to 16 are available.



The resulting equation is:

Selling Price = $ 39,555 + $60 per unit of House Size - $2,568 per unit of House Age + $3,401 per unit of House Location - 5 cents per unit of Owner's Salary.

Unfortunately if you want to perform a regression analysis of Selling_Pr against both House_Age and OwnersSal you have to move the columns around to make the columns House_Age and OwnersSal contiguous (next to one another!).

7. Other statistical functions that are available.

The following table is a list of Excel's statistical functions (You should use the Help system in Excel to get descriptions of these functions and how to use them.).

For a simple linear regression you can establish an equation using:

INTERCEPT() and SLOPE().
RSQ() will give you the R-squared value.
CORREL() will give you the Correlation Coefficient.
STEYX() will give you the Standard Error of Estimation.

Statistical	Functions
AVEDEV	Returns the average of the absolute deviations of data points from their mean.
AVERAGE	Returns the average of its arguments.
BETADIST	Returns the cumulative beta probability density function.
BETAINV	Returns the inverse of the cumulative beta probability density function.
BINOMDIST	Returns the individual term binomial distribution probability.
CHIDIST	Returns the one-tailed probability of the chi-squared distribution.
CHIINV	Returns the inverse of the one-tailed probability of the chi-squared distribution.
CHITES	Returns the test for independence.
CONFIDENCE	Returns the confidence interval for a population mean.
CORREL	Returns the correlation coefficient between two data sets.
COUNT	Counts how many numbers are in the list of arguments.
COUNTA	Counts how many values are in the list of arguments.
COVAR	Returns covariance, the average of the products of paired deviations.
CRITBINOM	Returns the smallest value for which the cumulative binomial distribution is less than or equal to a criterion value.
DEVSQ	Returns the sum of squares of deviations.
EXPONDIST	Returns the exponential distribution.
FDIST	Returns the F probability distribution.
FINV	Returns the inverse of the F probability distribution.
FISHER	Returns the Fisher transformation.
FISHERINV	Returns the inverse of the Fisher transformation.
FORECAST	Returns a value along a linear trend.
FREQUENCY	Returns a frequency distribution as a vertical array.
FTEST	Returns the result of an F-test.
GAMMADIST	Returns the gamma distribution.
GAMMAINV	Returns the inverse of the gamma cumulative distribution.
GAMMALN	Returns the natural logarithm of the gamma function, G(x).
GEOMEAN	Returns the geometric mean.
GROWTH	Returns values along an exponential trend.
HARMEAN	Returns the harmonic mean.
HYPGEOMDIST	Returns the hypergeometric distribution.
INTERCEPT	Returns the intercept of the linear regression line.
KURT	Returns the kurtosis of a data set.
LARGE	Returns the k-th largest value in a data set.
LINEST	Returns the parameters of a linear trend.
LOGEST	Returns the parameters of an exponential trend.
LOGINV	Returns the inverse of the lognormal distribution.

LOGNORMDIST	Returns the cumulative lognormal distribution.
MAX	Returns the maximum value in a list of arguments.
MEDIAN	Returns the median of the given numbers.
MIN	Returns the minimum value in a list of arguments.
MODE	Returns the most common value in a data set.
NEGBINOMDIST	Returns the negative binomial distribution.
NORMDIST	Returns the normal cumulative distribution.
NORMINV	Returns the inverse of the normal cumulative distribution.
NORMSDIST	Returns the standard normal cumulative distribution.
NORMSINV	Returns the inverse of the standard normal cumulative distribution.
PEARSON	Returns the Pearson product moment correlation coefficient.
PERCENTILE	Returns the k-th percentile of values in a range.
PERCENTRANK	Returns the percentage rank of a value in a data set.
PERMUT	Returns the number of permutations for a given number of objects.
POISSON	Returns the Poisson distribution.
PROB	Returns the probability that values in a range are between two limits.
QUARTILE	Returns the quartile of a data set
RANK	Returns the rank of a number in a list of numbers.
RSQ	Returns the square of the Pearson product moment correlation coefficient.
SKEW	Returns the skewness of a distribution.
SLOPE	Returns the slope of the linear regression line.
SMALL	Returns the k-th smallest value in a data set.
STANDARDIZE	Returns a normalized value.
STDEV	Estimates standard deviation based on a sample.
STDEVP	Calculates standard deviation based on the entire population.
STEYX	Returns the standard error of the predicted y-value for each x in the regression
TDIST	Returns the Student's t-distribution.
TINV	Returns the inverse of the Student's t-distribution.
TREND	Returns values along a linear trend
TRIMMEAN	Returns the mean of the interior of a data set.
TTEST	Returns the probability associated with a Student's t-Test.
VAR	Estimates variance based on a sample.
VARP	Calculates variance based on the entire population.
WEIBULL	Returns the Weibull distribution.
ZTEST	Returns the two-tailed P-value of a z-test.