GENSTAT

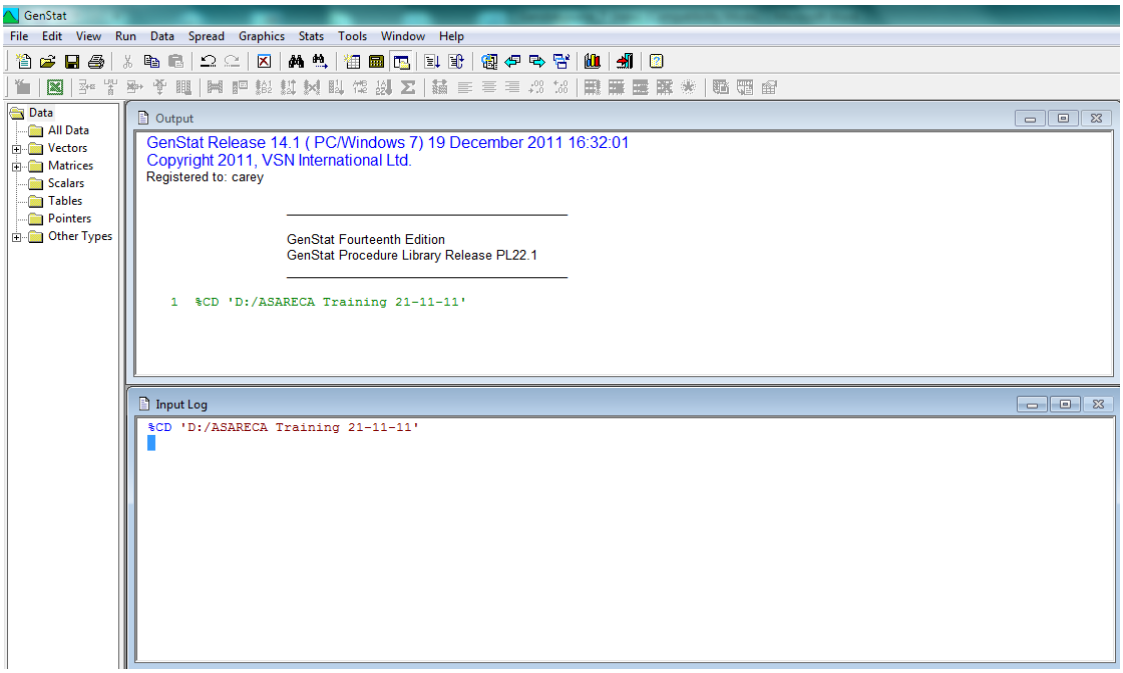
January 2012

Introduction to Genstat

1. Introduction

GENSTAT is a powerful general-purpose statistics package. It includes comprehensive facilities for tabulation, regression, the analysis of designed experiments, multivariate analysis and time series. It also includes good 'language' facilities which permit the user to extend the range of analyses. The version you will be using is GENSTAT 14 for Windows.

The first thing you will see when you go into this version of GENSTAT is a screen looking like the following.



This shows the **Output Window**, where analysis results will appear, and an **Input Log** where you can type in Genstat commands.

Two other windows are accessible: the **Fault Log**, where a record of any faults which occur are kept, and the **Graphics** window. This will appear only when graphical output is requested.

The menu bar at the top of the screen gives you a range of menus such as **Data, Graphics**, **Stats**, **Windows**.

Clicking on the **Windows** menu reveals the list of the four windows just discussed above.
Clicking on the **Stats** menu reveals a list of the statistical areas you may wish to use.

At the top of the screen, you will see the following toolbar

The three buttons to the extreme left are for opening files, saving files and printing the current window. The next three are cut-and-paste buttons (using the clipboard). Other buttons worth mentioning here are**:**  this allows you to submit the current window of commands.

 : these allow you to move backwards and forwards to different windows in Genstat

 : this moves you to the graphics window

 : to exit Genstat

 : to Genstat help; very useful

To use Genstat for Windows fully you will need to be familiar with both the Genstat command structure and the facilities which are available from its menu system. The menu system will be sufficient for a user's use a lot of the time, but not always.

Most of the time in practicals we will use the menu system first, and then look at the syntax.

6. The GENSTAT Command Structure

GENSTAT has a powerful command structure. It also has a system of defaults to make life easy for the user.

The general structure of a command is

Command name [**Option** list] **Parameter** list

Examples of some of the commands:

READ
PRINT

FACTOR
TREAT
ANOVA

The Option List is enclosed in [ ] and is specific to each command.

The final part of the command is the **Parameter List**. For example, the ANOVA command, can be written

anova[print=aovtable,residuals,means] **y**=yields;**residuals**=resexp1

Here there are 2 parameters, Y=yields and RESIDUALS=resexp1

Now notice the structure in all these examples.

a) Each Option or Parameter has the structure

name = list of things e.g. print=aovtable,residuals,means

and different items within the list are separated by a **comma**.

b) There can be more than one option or more than one parameter, e.g.

y=yields;residuals=resexp1

then each option or parameter is separated by a **semicolon**.

Note: Genstat refers to commands as **directives**.

7. Getting HELP and a few other useful features

GENSTAT has extensive on-line HELP facilities which can be accessed via the ? button on the toolbar.

Note also.

(a) Genstat is fussy about spelling and UPPER/lower case. Therefore it is best to use all UPPER or all lower case letters as names for variables and factors.

(b) The backslash (\) is the continuation symbol in Genstat syntax..

(c) A command can be shortened to the first four letters of the command name. It is usual to put each command on a separate line but it is possible to put more than one command on the same line by separating commands with a colon (:).

Getting Started with Genstat for Windows

Topics covered in this session include:

1. Familiarisation with Genstat
2. Data entry via Genstat's spreadsheet facility
3. Reading ascii files
4. Some simple descriptive analyses

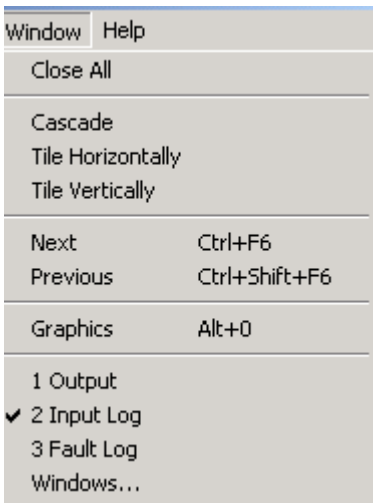
First of all, go into Genstat. Your screen should look like the one demonstrated earlier.

Section A: Familiarisation

Before we do any analyses of data, you should spend some time getting familiar with the layout of Genstat for Windows. There are two ways that you can do statistical analyses in Genstat - one is by entering commands directly (at an Input window). The other is using Genstat's statistics menu. Whichever way you use, output is always set to the **Output** window. There are also different ways of entering data: for example directly or via an ascii file. We are therefore going to start off by looking at some of the options on the menu bar, in particular the **Windows, Stats, Run, Data** and **Spread**.

(a) Windows

Click on **Windows**. This brings up a pull down menu which looks like:

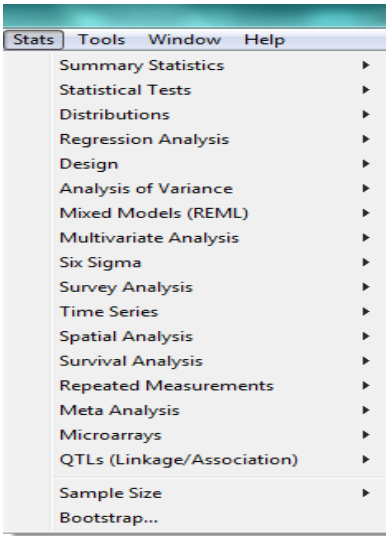


This tells you that there are four main windows in Genstat: **Input log**, **Output**, **Graphics** and **Fault Log.** Clicking on any one of these will move you to that window. There are also shortcut keys that you can use to move between windows. Try clicking on **Graphics** What happens? Now try moving to **Input Log** and **Output** in that order. The **Fault Log** window

is currently empty because you have not yet carried out any analyses. If you click on **Fault Log** here, you may want to close it afterwards.

(b) **Stats**

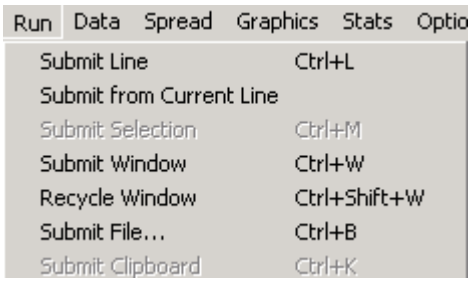
Click on **Stats**. Your screen should look like:



Click on some of the choices which are available with this pull-down menu to see what things you can do via the menu system. Spend no more than a few minutes on it.

(c) **Run**

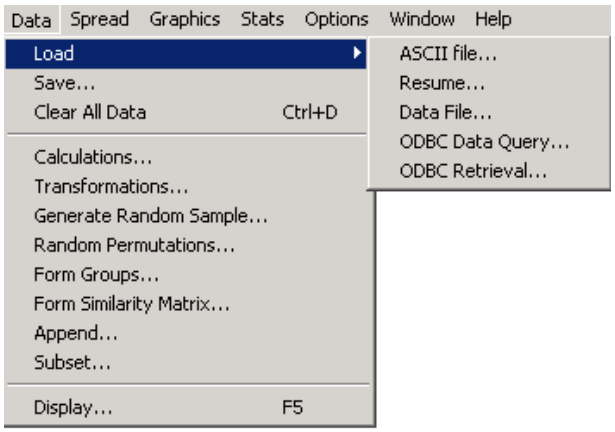
The pull-down menu should look like:

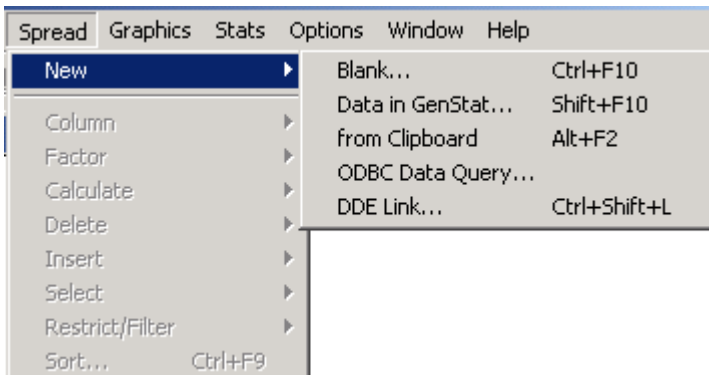


These are the choices available to you if you are entering commands *directly into Genstat* from an input window. You can for instance submit the whole window of Genstat commands, or just a section or even just a single line. Alternatively if you had previously prepared a set of commands and saved them into a file, you can submit that file. You will see much more of these options later.

(d) **Data** and **Spread**

You can read data into Genstat directly or from a previously saved file, such as an ASCII file. Click on **Data** and then **Load** to see the choices which are available to you.





To enter data directly you need to create a new spreadsheet. Click on **Spread** and then **New.** Which of the choices do you think you would choose to enter data directly?

Section B: Entering data directly via a spreadsheet

Example: We will use the following example to illustrate entering data into a Genstat spreadsheet. The total rainfall amounts from 1 - 10 May at two sites in the 10 years from 1966 to 1975 were:

Year	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
Site 1	13.5	10.0	29.7	69.3	38.6	15.8	46.2	4.8	30.2	39.1
Site 2	16.8	9.9	34.1	68.1	32.8	18.1	37.1	6.3	31.0	31.1

Later we will use these data to do some simple analyses. First of all, though, we need to enter them into Genstat, and we will do this using the Spreadsheet approach.

Note that Genstat uses separate columns for each variate. Here we want to have three variates: the data for year, the rainfall data at site1 and the rainfall data at site2.

There are other ways you might want the data; for instance you might want three variables, one indicating year, another indicating site (1 or 2) and a third indicating rainfall. However that is **NOT** what we want here.

Choose **Spread** from the menu bar, **New** and then **Create...**

Highlight the icon for Vector and then specify the size of your spreadsheet. You want to have three columns of data, one for year, one with the rainfall from site 1 and another with the rainfall from site 2 and you want each row to contain data from one single year. How many rows and columns do you think your spreadsheet must have? Enter this information.

You are now given a spreadsheet with this number of rows and columns. By default the columns are named C1, C2 and C3. Before entering your data, you may wish to give your columns meaningful names - say *year, site1* and *site2*.

Click on **Spread**, **Column** and **Edit Attributes...**

Name the first column (currently called C1) *year*. **Apply.**

Repeat the processing, naming the second and third columns *site1* and *site2*.

You can now go ahead and enter your data. Start by clicking on row1, column1, and entering the years from 1966 to 1975. If you make any mistakes they are easy to correct. When you have finished your spreadsheet should look like:

Spreadsheet [Loaded Data]

Row	year	site1	site2
1	1966	13.5	16.8
2	1967	10	9.9
3	1967	29.7	34.1
4	1969	69.3	68.1
5	1970	38.6	32.8
6	1971	15.8	18.1
7	1972	46.2	37.1
8	1973	4.8	6.3
9	1974	30.2	31
10	1975	39.1	31.1

Since we are going to use these data later, you should save them. To save the data as a Gensat spreadsheet, click on **Spread**, **Save As** and then **Genstat GHS File**. Supply a name for the file, using the extension.gsh. How about *rainfall.gsh*. Click Save.

If you move to the **Output** window you will see a summary for each of these three variables you have just entered. The number of values, mean, minimum and maximum are given. Genstat will always give you a summary of the data you have just entered or read in.

Note that because you have entered the data in numeric form, Genstat automatically assumes that your data are variates. To change a variable from a variate to a factor, click on **Spread, Column** and **Convert…..** Change the **Column type:** from Variate to Factor and click on **Apply**. The variable name changes to *Italics* and has a *red exclamation mark* on the spreadsheet. Any variable that shows as Italics with an exclamation on your spreadsheet are Factors.

Section C: Simple Descriptive Analyses

Example: These data below are the weights of a sample of 144 carrots (grams) taken from a crop of carrots grown at one particular farm. They are saved in an ASCII file called *carrots.txt.*

405	109	549	72	32	248
221	234	287	15	388	166
106	257	62	325	673	610
709	494	498	194	499	144
93	203	179	70	717	185
373	380	353	640	509	74
238	149	36	95	529	147
376	662	86	217	380	589
113	496	78	172	318	120
296	1017	197	202	116	129
66	201	292	59	341	606
97	284	915	512	290	87
39	429	643	191	305	258
252	385	216	66	288	670
390	100	579	106	245	426
599	874	597	656	80	175
106	543	340	134	325	55
274	18	71	211	331	254
87	77	749	21	446	88
283	212	31	286	343	91
556	93	132	54	657	223
325	121	436	124	771	505
251	173	258	484	28	813
663	318	393	157	325	87

We are now going to use these data to illustrate reading an ascii file into Genstat. We are going to use the data to illustrate some simple descriptive analyses.

First of all though, you might want to clear the data in Genstat.

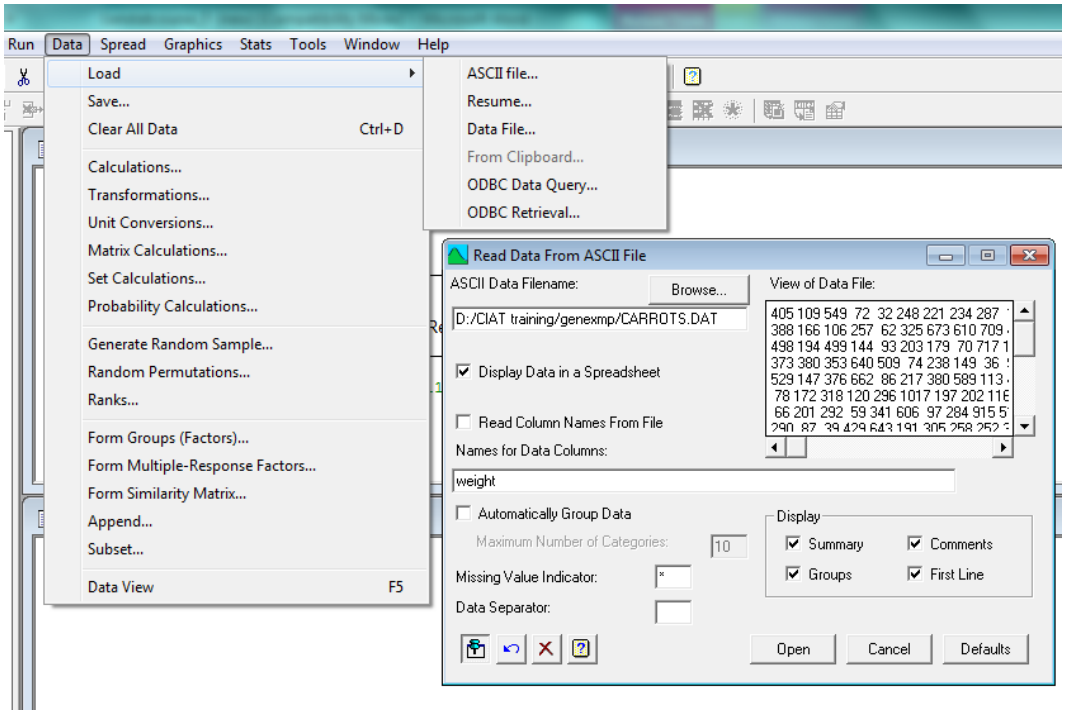
Click on **Data**, then **Clear All Data**.

We are now ready to start a new job. Close the **Data Display** dialogue box.

Reading an Ascii file

To read in the carrot weight data, use the **Data** pull down menu. Choose **Load** and then **Ascii file**. At the **Read Data from ASCII File** dialogue box, enter the name of the file-*carrots.dat*. Note that if you cannot remember the name of the file you have the opportunity to find it using the **Browse...** button. Enter the name of the datafile as *carrots.dat*. The box in the top right-hand corner of the dialogue box has now filled up with the first few lines of the datafile. We have one variable here - carrot weight - so we want the data in Genstat to appear as one column. Name the data column *weight*. Note, too, that if you want to open a spreadsheet with the data you can click on **Display Data in a Spreadsheet** and Genstat will do this for you. Click on **Open**. If you choose the spreadsheet option, you should now have a spreadsheet window open to you.

Select **Data→ Load → ASCII file…**

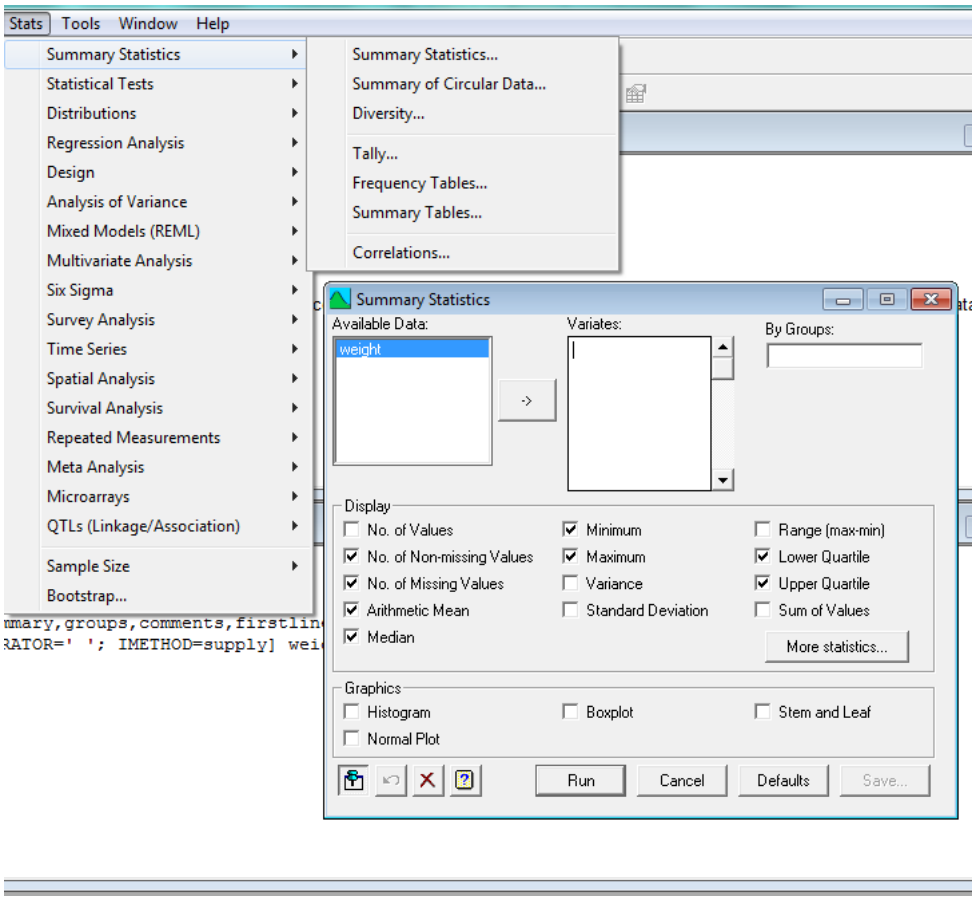


Move to the **Output** window and you will see a summary of the data presented. Check that the number of observations entered is correct. If it is, you are now ready to try out some simple descriptive analyses. We will first of all do some numerical summaries and then look at some graphical presentations.

Descriptive Statistics

To get some descriptive statistics of the carrot weight data we can use the Statistics pull-down menu. Click on **Stats**, **Summary Statistics** and **Summary Statistics...** This will bring up a **Summary of Statistics** dialogue box. Here you can choose from the available data which variate you are interested in. In this case there is only one, *weight*. Double click on this to

transfer it to the variate box. In the lower part of the screen you can see what summary statistics are printed by default by Genstat. Click on **Run**, then move to the **Output** window.



Record, below, some summary statistics about the carrot weight. If you do not already have the information available in the output then go back to the dialogue box, alter your options and resubmit your request**.**

Carrot weight: no. of obsns. _____

mean _____ minimum _____

maximum _____ variance _____

standard deviation _____ s.e.m. _____

Before moving on to look at graphical summaries of the data, take a look at the Genstat command which produced your output. Move to the **Input Log** window.

What is the command which produced your summary statistics of carrot weight?

And can you identify which part told Genstat to produce each summary statistic.

To see how submitting a command works directly as opposed to using the menus, edit the command so that it will produce only the number of observations and the mean; then submit it.

There are different ways that you can submit the command.

1. With the cursor on the command line, click on **Run** at the menu bar and **Submit Line.**

2. Highlight the whole line, click on **Run** at the menu bar and **Submit Selection.**

You might like to try both. Check your output.

Note that if you wanted a print-out of the data there is no facility within the menu system to do this. The Genstat command is PRINT; so at the **Input Log** window type

PRINT weight

and submit it. At the **Output** window you will see the data displayed as one long column of data. You can add options to your PRINT statement. For instance try

PRINT[ORIENTATION=across]weight; DECIMALS = 0

What do you get when you submit this?

To find out more about the structure of any command, and how to customise it, you can use the Genstat help system. This is available with the ? button on the tool bar. We will look at this later.

Graphical Summaries of data

Descriptive analyses of data often include graphical summaries such as histograms and boxplots. These are both available from the **Graphics** pull-down menu. Click on **Graphics** and **Histogram...** , and then at the dialogue box specify that the data are in the variable called *weight*. Click on **Run**. (We will look at the other options in a minute). A histogram of the carrot weights will appear in the Genstat graphics window. Does the distribution seem symmetric or skewed?

Now see if you can now improve the graphical presentation of this histogram. Go back to the histogram dialogue box. Try adding a title, change the number of groups and add labels to the two axes. Does your graph now look better? To do this you will first have to click on **Options.**

And finally, let us look at another graphical summary of data, the boxplot. You can also do this from the **Graphics** pull down menu and **Boxplot...** At the dialogue **Box Plot** box enter *weight* as your data variable, and leave the display at Box and Whisker. **OK**. What does your plot look like? Does it indicate that the data are symmetric or skewed? Now try plotting the data again; this time using the Schematic display. What is the difference between the two plots; ask if you are unsure. Which do you prefer?

That is the end of this practical. Write down below five things you think you now know how to do in Genstat.

1.

2.

3.

4.

5.

Two Sample Tests using Genstat for Windows

Topics covered in this session are:

1. t-tests paired and unpaired in the Windows version of Genstat
2. More practice with spreadsheets.

In Genstat for Windows two sample tests can be carried out using the **Statistical Tests** on the **Stats** pull down menu, where you would choose **One and two sample t-tests ... .** We will illustrate how using the below example.

Two rubber plantations (A and B) are supplying rubber to a factory. The factory is interested in testing whether the mean tensile strength of the rubber is the same for the two plantations and collects eight specimens from each plantation. The following tensile strengths were recorded.

A	B
201	188
181	170
193	179
159	149
179	161
188	181
197	183
185	177

(a) Are these data paired or independent? You may want to carry out a suitable test in Genstat to determine whether the average tensile strength of rubber from the two plantations is the same.

First of all enter these data into a Genstat spreadsheet using variable names *plant_a* and *plant_b*. To do this you first choose **Spread** from the menu bar, then **New** and then **Blank...** Specify that the spreadsheet will have two columns and eight rows, and then enter the data.

When you have entered the data, and transferred it to Genstat, you are ready to do your analysis.

Before doing any formal analysis, calculate some descriptive statistics for the two groups. Click on **Stats**, **Summary Statistics** and **Summarise Statistics...** You may also want to save your output to a file so that you can look at your results later.

From looking at the means does there appear to be a difference between the two plantations?

Choose from the **Stats** pull down menu **Statistical Tests,** and then **One and two sample t-tests ...** You are now given the option of which test to do. Look to see what is available, but then choose **t-test (unpaired).** Note that you have the option of entering the data in two different ways. Check that your way (two columns of data) is the one which Genstat will act on. Enter *plant_a* and *plant_b* as your two data sets. Click on **OK**

Record your results below. The default analysis being used here gives you a two-sided test for the comparison of two independent groups, where the null hypothesis is one of no difference (zero difference).

Now take a look at the **Input Log** Window. Do you recognise the syntax which is presented here. Note that you could change the 95% confidence interval for the difference between the plantation means to whatever you want.

Record below your 95% confidence interval for the difference in tensile strength between plantations.

2. In the "Getting started" session, you entered some rainfall data and saved them into a Genstat spreadsheet called *rainfall.gsh.* The data were from two sites and different years. Go back and look at this example. You are now being asked to do a test to investigate whether or not there is any difference between the two sites with respect to their average rainfall.

Retrieve this spreadsheet. To do this you will need to choose **Data** from the menu bar, **Load,** and then **GSH Spreadsheet...** . Give the name of the spreadsheet. OK. The data are then loaded into Genstat.

Determine whether or not these data should be analysed as a paired or unpaired test. Then carry out whichever analysis you think is appropriate. Record your results below.

Estimate a 95% confidence interval for the difference between the two sites.

What do you conclude.

Chisquare Tests for Two-way Tables

1. Given below is a 5x2 table of observed frequencies showing the number of rubber trees that were not diseased with a particularly serious root rot disease in five different plantations in South East Asia. These five plantations were similar in terms of altitude, rainfall etc. Their management practices were also similar, except that they all used different sprays for the control of the disease.

Plantation	No. diseased	No. not diseased	Total
A	43	237	280
B	52	198	250
C	25	245	270
D	48	212	260
E	57	233	290
Total	225	1125	1350

You are required to carry out an analysis of these data to determine whether or not there are differences between the proportions of diseased trees in the five plantations.

First of all, you need to enter the table into Genstat. The way to do this is to first create a table. To do this you will need to choose **Spread** from the menu bar, **New,** and than **Create.** From the spreadsheets shown double click on **table**. Enter **5** for rows and **2** for columns. Enter the data.

Make sure you understand what has been done here.

If you are happy that your data are entered correctly, then you are now ready to do your chi-square test of association. Choose **Stats** from the menu bar, **Statistical tests**, and than **Contigency tables ….** In the available data window the table name is shown. Double click on it. The table name will show in the table window. Click on **Run** button.

What output do you get. Record it below.

Remember the null hypothesis here is that there is no association between disease and plantations. From your output, is there any evidence of association between disease and plantation. Is there therefore evidence to suggest that the plantations have different proportions of diseased trees.

Instead of just doing the chisquare test, you might also want to look at the fitted values that you get under the hypothesis of no association. Go back to the **Contigency Tables** dialogue window and click the **Options…** button. Click on the check box for **Expected Values**.

Record your fitted values below. How can you use these to further interpret your results?

2. In a cocoa plantation in region A of Java, 117 out of 1067 trees were found to be infected by phytophthora. In region B, 54 out of 402 trees were found to be infected. You are required to investigate whether or not there is evidence of a difference in infection rates between the two regions.

First of all, lay these data out in a 2x2 table of region (A/B)versus infection (y/n).

Now using the ideas of table data entry illustrated above, enter these data into Genstat table.

Carry out a chisquare test to compare the region. What do you conclude.

Note that had the 2x2 table been summarising a small dataset, with low expected values, you may have wanted to do a Fisher's exact test.

Simple Analyses of Variance using Genstat

Topics covered in this session are:

1. More on Genstat's spreadsheet
2. One-way analysis of variance for the completely randomised design
3. Two-way analysis of variance for the randomised complete block design
4. Model checking via residual plots
5. Genstat syntax for analysis of variance

Genstat has a fairly extensive menu for analysing balanced experiments. Here we are going to start looking at some simple designs. We will look at the default output produced by Genstat and further analysis which are available to you. Whilst these are introduced here for the simple designs the same defaults and additional outputs are available with all the different analyses of variance on the menu.

Section A: One-way anova for the Completely Randomised Design

Example: An experiment was conducted to compare the yields of five varieties of lentils under rainfed conditions. The experiment was planted in a completely randomized design with each variety replicated four times. During the growing season, however, sheep broke through the fence and heavily grazed four plots along the edge of the experiment before they were detected. The yields, in kilograms per hectare, from the remaining plots are given below:

Variety				
1	2	3	4	5
74	54	32	74	60
43	44	29	63	50
76	39		87	43
64				54

These data are not currently in a file, they need to be entered directly into Genstat. Do this using Genstat's spreadsheet facility. Remember, Genstat needs a column for each variable and for analysis of variance it needs separate variables for response, the blocking factors and the treatment factors. Here there is only the response and the treatment factor.

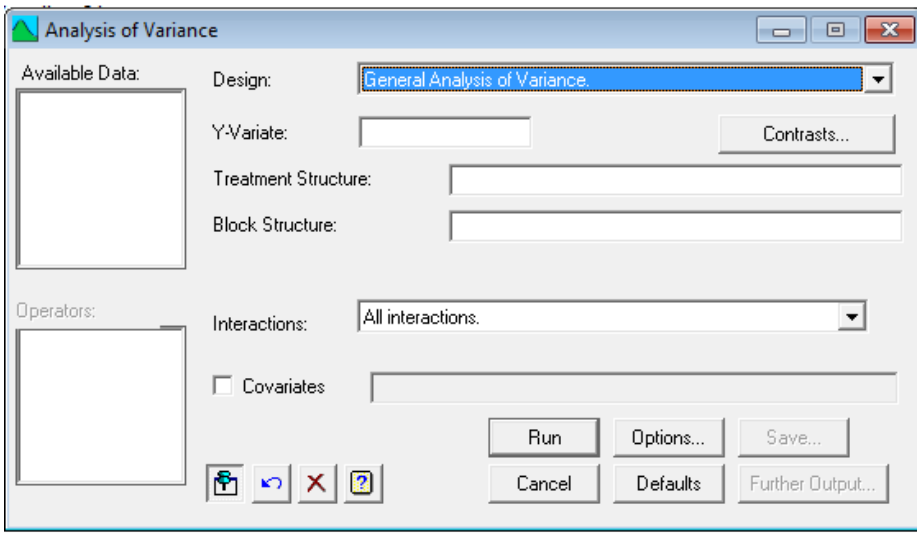
Choose **Spread** from the menu bar, **New** and then **Blank...** Set up the dimensions of the spreadsheet (number of columns and rows), name your columns and enter the data. Use the names *yield* and *variety*. Note though that because you have entered the data in a numeric form, Genstat automatically assumes that your data are *variates*. We will however wish to

treat *variety* as a factor. To change it from a variate to a factor, click on **Spread**, **Column** and **Convert**... Change the **Column type**: for *variety* from Variate to Factor.

Save these data into a Genstat spreadsheet for future work, using **Spread** and **Save As.** Record the name of your file.

You are now set up to do your analysis of the data.

Analysis of variance is one of the options available to you on the **Stats** menu. At the menu bar click on **Stats**, **Analysis of Variance, General**... You will then get the Analysis of Variance dialogue box which looks like:



At **Design**: it says General Analysis of Variance. There are a choice of analyses available to you; have a look at the list.

Since this is a completely randomised design you want an analysis of variance with just treatments and residual as the sources of variation. Sketch out below the analysis of variance table you expect: sources and degrees of freedom are sufficient here. Now to get the correct analysis choose **One-way ANOVA [no blocking].**

Source	Df
Total	

Note that the dialogue box now changes. You are asked to specify your y-variate and your treatment factor.
From the available data, select *yield* as your y-variate; and *variety* as your treatment factor.

For the time being, ignore the **Covariates** option. Below this there are seven buttons, but two of them are shaded out (just as in the dialogue box shown on the previous page). Click on **Run**.

Move to the **Output** window to see your results. You should find an analysis of variance table, a table of means and standard errors of the differences between means.

Is the format of the analysis of variance the same as you sketched out earlier? From the output answer the following questions:

Which variety mean is based on the fewest replicates? And which the most?
Which variety had the highest yield?

Now look at the standard errors output. Make sure you understand what is being presented here.

We will now look to see what else can be produced using this menu. Go back to the **Analysis of Variance** dialogue box. You will see that the **Save...** and **Further Output...** buttons are now no longer shaded out. Click on **Save...** What does this allow you to do?

Go back to the previous dialogue box, and click on **Options...** This shows you the default options used to produce the analysis of variance output. If you had wanted to change these options prior to running the analysis (e.g. perhaps requesting that the CV be printed) then you could have done so here.

Alternatively, if you had already run the analysis and then realised you wanted some further information, you could do so using **Further Output...** To see what this does cancel the Anova Options. Click on **Further Output...**

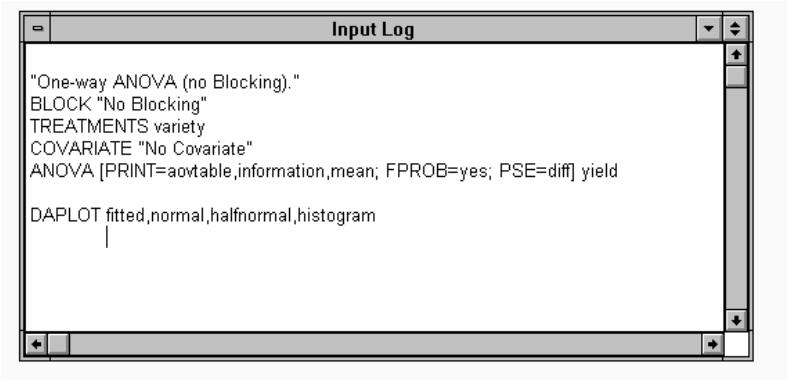
Here you have the option of asking for more output to be displayed. This gives you considerable flexibility when producing anova output. In addition to these facilities you can also produce some graphs. Two options are available to you: **Residual Plots...** and **Means Plots...** The **Residual Plots...** option will allow you to check the assumptions behind your analysis of variance. We will explore this further here (**Means Plots...** will be discussed later).

Click on **Residual Plots...** There are 5 different types of plot available to you; four of which are default plots. If you are happy to have all of these click on **Run**. Otherwise alter the selection to suit yourself. If you do not know what any of them are, and why you should use them then please ask.

From these plots can you decide whether the following assumptions underlying your analysis of variance are appropriate.

(i) constant variance? Yes/No
(ii) normality? Yes/No

Have a look at what Genstat commands you used to generate your analyses. Go to the **Input Log** Window to see these. This window should look like:



The parts in quotes (e.g. "No Blocking") are comments and do not add to the analysis. The same analysis could have been produced by using only the TREATMENTS, ANOVA and DAPLOTS statements. The TREATMENTS and ANOVA statements were responsible for the initial analysis, whilst DAPLOTS produced the residual plots. You can see this below. Edit the **Input Log** to remove the BLOCKS and COVARIATE lines. Highlight the TREATMENTS and ANOVA lines; click on **Run** at the menu bar, and **Submit Selection**. Check to see what is produced at the **Output** window.

Go back to the **Input Log**, place the cursor at the end of the DAPLOT line; then click on **Run** and **Submit Line** to produce your residual plots. If you feel that you now know how to do a 1-way analysis of variance in Genstat, you can stop now. Otherwise seek help on any parts you are unsure about.

Section B: Randomised Complete Block Design

Example: A bacteriologist wanted to determine the effect of four different sources of seed inoculum A, B, C and D, and a control, E, on the dry matter yield of irrigated alfalfa. The plots were to be furrow irrigated and the agronomist suspected that there might be a differential effect of irrigation from one end of the furrow to the other. He decided to use a randomised block design. The layout of the experiment and the data, i.e. total dry matter yields (kg/plot), appear below.

										Water Flow
Block	1	A	34.8	B	33.7	D	30.4	C	32.7	E	24.4	↓
	2	D	37.0	E	28.8	B	33.5	A	35.6	C	33.4	↓
	3	C	35.8	D	35.6	A	37.9	E	26.7	B	35.1	↓
	4	B	34.8	C	39.1	D	32.7	E	32.4	A	37.4	↓

The data are contained in an ASCII file named *alfalfa.txt* in columns representing yields, blocks and inoculum respectively.

Read these data into Genstat, using the **Data** menu and by choosing **Load** and then **Ascii file...** Enter the file name (using **Browse...** if you need help to find it).

[Note too that the inoculum are called A, B, C, D, E in the file]

From the first few lines you can see that the variable names are at the top of the file, so click on the option **Read Column names from file**.

Because these data are from a randomised block experiment, we have two factors in our file - block and inoculum. We therefore want to identify these as grouping variables. Click on the box **Automatically Group Data**. **OK**.

Genstat will now check whether you wish to define *block* as a variable or factor, and *inoculum* as text or factor. Define both as factors.

For a randomised complete block design, providing the assumptions hold, you will want to do an analysis of variance with effects for block and treatment (in this case inoculum). Sketch out a skeleton anova with sources and degrees of freedom so that you can check back on this later.

From the menu bar choose **Stats** and then **Analysis of Variance** and then **General...**

At the **Analysis of Variance** window select the appropriate **Design:** from the list available. What do you think if should be?

Either the **One-way ANOVA [in Randomised Blocks]** or **Two-way ANOVA [no Blocking]** will produce the correct analysis. Choose **One-way ANOVA [in Randomised Blocks].** At the **Y-variate:** specify that yield is your response, at **Treatment:** that your treatment factor is inoculum; and at **Blocks:** specify that your blocking factor is called block. **OK**.

Move to the **Output** window to observe your results. You should find an analysis of variance table with effects for block and inoculum. Is there a significant difference between the inoculums? You should also find a Table of means, and a summary of the standard errors of differences between means.

Which inoculum gives the lowest mean yield.
How many replicates are each of these means based on?

What is the standard error of the difference between the means for inoculums A and C, and how many degrees of freedom are associated with it.

Now take a look at the **Input Log** window to see what commands were generated by your choices. What do you see?

Go back to the **Analysis of Variance** window and run the analysis again, this time specifying **Two-way ANOVA [no Blocking]**. Now you will have to specify *block* as **Treatment1:** and *inoculum* as **Treatment2:** At **Interactions**: choose the option for only main effects. **Run**.

At your **Output** window your results should include an analysis of variance table, a table of means and standard errors of differences. How does this output compare with your previous analyses of these data. Which do you prefer?

Now take a look at the commands which were used to generate these analyses. How do they differ from previous analyses?

Another alternative would be to use the commands

TREATMENTS block + inoculum
ANOVA[PRINT = a, i, m; FPROB=y]yield

Try running these by typing them in at the **Input log** window. You can do this either by typing the data in directly or by copying and editing lines already in the **Input Log**. Make sure that the output is still the same.

If you are happy that you understand what you have done in this section then stop now. If you wish to explore the procedures further then try changing some of the options which are available to you. You might also like to try using Genstat's command language to produce different parts of the output, or to do additional analyses (e.g. residual plots).

And finally, if there is anything you are still unclear about then please ask.

Further Analyses of Variance

Topics covered in this session are:

1. Factorial experiments
2. Split-plot designs
3. More on Genstat syntax

Section A: Factorial Experiments

When the treatments being studied in an experiment have a factorial structure there are several different ways that the analysis of variance of the data can be constructed using the Genstat menu system. From the **Stats** pull-down menu, choose **Analysis of Variance, General ...** and spend a few minutes exploring the dialogue boxes for

(i) general treatment structure (no blocking)

(ii) completely randomised design

(iii) general treatment structure (in randomised blocks)

(iv) general analysis of variance

They all have a similar format.

Treatment structure can be specified as having one or more factors. Blocking, if there is any, can also be constructed from one or more factors in the dataset.

Let us now try out one example. The example is a relatively straightforward one, but the ideas introduced here with it apply also to more general situations.

Example: An experiment was conducted on maize in 1982/83 at Mangwende to test the effect of two times of application of basal fertilizer on maize yield:

: at planting
: after emergence

and the effect of type of basal fertilizer:

: Compound D
: Compound P

The treatments were as follows:

Treatment 1: Time of fertilizer application - after emergence (1)
Type of fertilizer - Compound D (1)

Treatment 2: Time of fertilizer application - after emergence (1)
Type of fertilizer - Compound P (2)

Treatment 3: Time of fertilizer application - at planting (2)
Type of fertilizer - Compound D (1)

Treatment 4: Time of fertilizer application - at planting (2)
Type of fertilizer - Compound P (2)

(A 2 x 2 factorial)

The treatments were laid out at random in six blocks, each containing 4 plots.

The data are stored in an ascii file called *maize.txt*. The first column of data are the grain yields; the second column, the block number (1-6); the third column, treatment number corresponding to the numbers given above (1-4); the fourth column gives the time of application and the fifth column the basal compound.

Read these data into Genstat, making sure that the block, time of fertiliser application and type of fertiliser are declared as factors.

Sketch out the analysis of variance table (sources and degrees of freedom) you would expect to produce for these data.

Source	Df
Total	

Go to the Analysis of Variance dialogue box. Which **Design:** do you think you should use?

Choose **General Treatment Structure [in Randomised Blocks]**. Specify that *yield* is your y-variate and that *block* is your blocking factor. At **Treatment Structure:** you have several choices depending on whatever model you wish to fit.

Terms		Model with:
(i)	applic	main effect for time application
(ii)	fert	main effect for type of fertiliser
(iii)	applic + fert	main effects for both time of application and type of fertiliser
(iv)	applic + fert + applic.fert	main effects and interaction

Enter your choice of model.

At **Interactions:** you can further specify how many of the interactions in your treatment structure you want produced. Choose the option you think is most appropriate.

Click on **Run**, and move to the **Output** window.

Are your results as you expected?

Is there a significant difference between compound D and compound P? And is that difference independent of the time of application?

Estimate the difference in yield between the two compounds. And what is the standard error of that difference?

Finally, do not forget to check your **Input Log** window to see what Genstat commands would produce this output. Is it what you would expect, from what you already know about the ANOVA, TREATMENTSTRUCTURE and BLOCKSTRUCTURE commands.

If you want a challenge: Produce some further output to go with this analysis. In particular, produce a **Means Plot...** (available from selecting **Further Output...** back at the Analysis of Variance dialogue box). Describe briefly below what your plot contains.

Section B: Split-Plot Experiments

Unlike a lot of other packages Genstat is very good at analysing data from multi-strata experiments such as split-plot designs. With relative ease it will produce the correct analysis of variance, with significance testing at the correct level and all the correct standard errors for the different comparisons you might want to make - provided of course you have specified your design correctly!

We will use the example below to illustrate the analysis of split-plot experiments using Genstat.

Example: Four strains of perennial ryegrass were grown as swards at each of two fertilizer levels. The four strains were S23, New Zealand, Kent and X (a "hypothetical" strain introduced to illustrate some points of statistical interest). The fertilizer levels were denoted by H, heavy, and A, average; the experiment was laid out as four blocks of four whole plots for the varieties each split in two for the application of fertilizer. The midsummer dry matter yields, in units of 10 lb/acre, were as follows:

	Manuring	Block			
		1	2	3	4
S23	H	299	318	284	279
	A	247	202	171	183
New Zealand	H	315	247	289	307
	A	257	175	188	174
X	H	403	439	355	324
	A	222	170	192	176
Kent	H	382	353	383	310
	A	233	216	200	143

The data are in a Genstat spreadsheet called *strains.gsh.* Load the spreadsheet into Genstat. You should see that there are six columns of data, five of which are factors; the remaining one contains the dry matter yields. As well as factors for block, fertiliser and strain, there are two others factors.

What are they and how are they coded? Make sure you understand how the data have been entered.

As usual, sketch out the analysis of variance (sources and d.f.) table you want to produce for these data.

Source	Df
Total	

At the Analysis of Variance dialogue box, choose **Split-Plot Design** and specify that the y-variate is *yield*. Enter the treatment structure you think is appropriate.

And specify which factors are **Blocks:**, **Whole Plots:** and **Sub-plots:**. Click on **OK,** and then move to your **Output** window.

Unless you have altered the default settings your output should include the anova table, a table of means and standard errors. Study this output and then answer the following questions.

1. Is there a significant fertiliser effect, and is it dependent on the strain of rye grass being grown.

2. Is there a significant difference between heavy and average fertiliser application with the known strain. Carry out a t-test to answer this [t = difference/s.e.d.]

3. Is there a significant difference between Kent and StrainX when average manuring is used? What s.e.d. do you use here for your t-tests?

Finally, take a look at the Genstat commands which were used to generate this analysis. What does *block/mainplot/subplot* on the BLOCK statement mean?

Note that the same analysis can be produced using the **General Analysis of Variance** option. Try it. The y-variate and treatment structure details should be straightforward; but what should you put in the **Block Structure:** section?

Additional Analysis of Variance Examples using Genstat

1. These data are from a nursery experiment looking at the effects of adding different amounts of N, P and K to the tube soil in which seedlings of *Pinus. patula* were grown (50 tubes per treatment per block). Two levels of each of N, P and K were studied, a low and a high level (2^3 factorial). The results are seedling height (in inches).

			Block			
K	P	N	1	2	3	4
Low	Low	Low	3.42	3.25	3.39	3.57
		High	4.09	4.23	4.67	4.15
	High	Low	4.22	3.71	4.44	3.88
		High	4.12	4.63	4.81	4.30
High	Low	Low	3.92	3.12	3.55	3.99
		High	4.37	4.20	4.67	4.50
	High	Low	3.63	3.89	4.48	4.22
		High	4.16	4.24	4.57	4.33

These data are in an ascii file called *patula.txt* where the data in the first column is the block number, in the second column is the amount of N, in the third is P, in the fourth is K(1 = low, 2 = high), and in the fifth is seedling height.

Read these data into Genstat and carry out an analysis of variance looking at main effects and interactions of the treatment factors. Estimate effects (with standard errors).

2. The following example involves 4 treatments laid out as a randomised complete block with 4 blocks, four replicates each for treatment 1 and treatment 2 (one replication each per block) and eight replications each for treatment 3 and 4 (two replications per block). Treatments 3 and 4 are more important treatments having two times as many replications as treatments 1 and 2.

Treatment	Blocks				
number	1	2	3	4	Mean
1	2.2	3.2	2.9	1.9	2.550
2	2.7	3.4	3.0	2.4	2.875
3	2.4	3.3	3.0	2.5	
	2.6	3.6	2.8	2.8	2.875
4	2.6	3.4	3.1	3.2	
	3.0	3.2	3.4	2.9	3.100

Enter these data into Genstat. Carry out an analysis to investigate whether or not there are significant differences between the treatments.

Regression Methods

Topics covered in this session are:

1. Simple linear regression
2. Multiple linear regression
3. Model selection
4. Genstat syntax for regression analysis

Genstat's FIT statement allows the user to fit to data all kinds of linear models ranging from the simple linear regression model for normally distributed data through to generalised linear models such as log-linear models for contingency table data and logistic regression for binomial data.

Here we are going to look at models which fall within the general linear model framework, and will concentrate on model fitting, selection and checking. To do this we will use the example below.

Example: Information was collected from a random sample of 40 trees which were part of a mangium progeny trial (planted in 1991) to examine the relationship between tree height at 36 months (y) and the following explanatory variables:

(i) diameter at breast height at 26 months
(ii) diameter at breast height at 35 months
(iii) height to first fork at 35 months

The data are available in an ascii file called *mangium.txt* as follows:

Column	Variable
1	Tree number (treeno.)
2	diameter at 26 months (dbh26)
3	tree height at 36 months (height)
4	diameter at 35 months (dbh35)
5	height to first fork (fork)

First of all, open the file and read the data into Genstat, using variable names *treeno, dbh26, height, dbh35* and *fork*.

We will consider a simple linear regression of tree height on diameter at 35 months (i.e. $y = \beta_0 + \beta_1 x$) .

Using the **Graphics** pull down menu and **Point Plot**, produce a scatterplot of your response (*height*) against your explanatory variable (*dbh35).* Does there appear to be a linear

relationship between the two? Out of interest you might also like to look at the relationship between tree height and the other two variables too.

Now go back to Genstat (remember, you need to choose **Options** from the Genstat Graphics menu bar, and **Goto Genstat**. At the **Stats** pull down menu choose **Regression Analysis** and then **Linear...**

There are several different options available to you: simple linear regression and multiple regression both with and without groups, and general linear regression. Spend a few minutes looking at the dialogue box for each of these. What differences do you see amongst them? Which one seems to be the most different?

Since we have no groupings here choose **Simple Linear Regression** and supply the names of your response and explanatory variates (here *height* and *dbh35).* The buttons at the bottom of the screen are similar to the ones seen for Analysis of Variance. Click on **Run**.

Before moving to the **Output** window check to see, using **Options...** what the default output is for this type of regression analysis. [As with Analysis of Variance the defaults are the same no matter which method you are using]. From this, what do you think you will see at the **Output** window.

Now move to this window and see if you are correct. You should find an analysis of variance partitioned into regression and residual components, and estimates of the regression coefficients.

Answer the following questions using the output.

1. Is there a significant regression relationship.

2. Record the fitted equation ($y = \beta_0 + \beta_1 x$) .

3. A common summary statistic with regression analysis is R^2, the coefficient of determination. [R^2 = regression SS/total SS]. Calculate R^2 here.
   How do you interpret this?

4. In the "Estimate of regression coefficients" section there are four columns of figures. Make sure you know what they all are, particular the t(38) and t pr.

As with Analysis of Variance, you have the option of producing further output. Of particular interest might be graphical output. Two different types are available to you: (i) a scatterplot of the data with the fitted line and (ii) model checking graphics.

Explore both of them, using **Further Output...**

In the **Model Checking...** section you have the opportunity of looking at both residuals and leverage and influence diagnostics. You also have the choice of standardised or ordinary or deletion residuals. And you can do index plots, plots versus fitted values and normal probability plots. It is unlikely that you would want to produce all the choices available to

you. List below 5 model checking plots which you would regularly like to do, and the reason why.

Plot	Reason
1.	
2.	
3.	
4.	
5.	

We will come back to these later when you look at multiple regression.

Before moving on, take a look at the Genstat syntax which has been used to produce the above analysis. There are three core statements which are used; record them below. What do you think each one does.

Statement	What it does?

If you are happy that you understand what Genstat is doing with simple linear regression, and how the syntax works then you can stop here. If there is anything you are unsure about then please ask.

Section B: Multiple Regression

Models with more than one explanatory variable

Since the simple linear regression of tree height on diameter at 35 months only explains around 60% of the variation in tree height, you may want to go on and fit a multiple regression. If you wished to fit a multiple regression of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

and know exactly which explanatory variables to include in the model then you could do so easily using **Multiple Linear Regression** as your regression method. The only difference between this and **Simple Linear Regression** is that you now enter more than one explanatory variable. Try this out. Go back to the **Regression** dialogue box and choose **Multiple Linear Regression**. For a model, for example, with both dbh26 and dbh35 you would enter both of these as explanatory variables, separating them by either a space, comma or a + sign.

However, it is more likely to be the case that the analyst is confronted with several explanatory variables and is faced with the problem of trying to find an appropriate multiple regression model to summarise the variation in his response. This usually means finding the minimal model from a selection of models and forward selection and backward elimination are two commonly used approaches for this. To do such an analysis you would choose the **General Linear Regression** option.

Model Selection

Having chosen **General Linear Regression** let us now consider trying to find an adequate model using *dbh26, dbh35* and *fork* as explanatory variables which explains the variation in tree height. Specify *height* as your explanatory variable. The **Maximal Model:** is equivalent to the TERMS statement in the Genstat syntax. Here you specify the most complicated model you are going to investigate. In this case our maximal model will be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where x_1, x_2 and x_3 are *dbh26*, *dbh35* and *fork* respectively. Enter this model here. For the purpose of illustrating how to do model selection in Genstat we will use a Forward Selection approach. The ideas for Backward Elimination and even Stepwise regression are similar. So start by fitting the model with only *dbh35*. (You can take it from me that of the three simple linear regression models, this one explained more variation than either of the other two on their own. However, check it out if you want.). Before clicking on **OK**, you should change the output options available to you. Click on **Options...** and then indicate that you want the **Accumulated** analysis of variance table. This will give you sequential sums of squares. Now click on **OK** to fit this first model.

Note that when the procedure has finished running the **Change Model...** button is no longer shaded out.

Your output should be similar to the one seen earlier, except that now you also have an analysis of variance table with sums of squares for *dbh35*.

Now we want to look at adding at least one of the other terms to the model.

Go back to the **Linear Regression** dialogue box and **Change Model...**

There are four options available to you:

ADD - adds a term to your current model,

DROP - drops a term for your current model.

Both of these effect a change in the current model.

TRY - looks to see the effect of adding a term to the current model, but does not change it.

SWITCH - looks at the effect of adding or dropping a term depending whether it is absent or present from the current model. Again the model is not changed.

Since we have no idea at this stage whether to add *dbh26* or *fork* we can try adding each of these in turn to the model with just dbh35. **Try** *dbh26* first. From the accumulated anova produced at the **Output** window, is the relationship between height and *dbh26* significant when you have allowed for the effect of *dbh35*. Go back and **Try** *fork*. What do the results of this analysis show.

If either *dbh26* or *fork* appears to substantially explain some more of the variation once the relationship with *dbh35* has been allowed for, then add the most significant term to the model, using **Add**. The current model will now contain two explanatory variables. Repeat the procedure until you have got to a model you are happy with. Summarise your results below.

Having found a satisfactory model you might like to do some model checking on your final fitted model. Go back to **Further Output...** to do this. You can now produce your five "most wanted" plots. Since the dialogue box only allows you to produce them one at a time you might want to enter the command (or commands) at the **Input Log** window and submit these instead.

Now answer the following questions

1. Are there any outliers?

2. Are there any influential observations, or points of high leverage?

3. If the answer is yes to either 1 or 2 which observations are they?

4. Are the assumptions of normality and constant variance underlying the analysis reasonable?

That is the end of this section. If you are unclear about any particular point we have covered then please ask.

Write down below the four most useful you have learnt in this whole session.

1.

2.

3.

4.