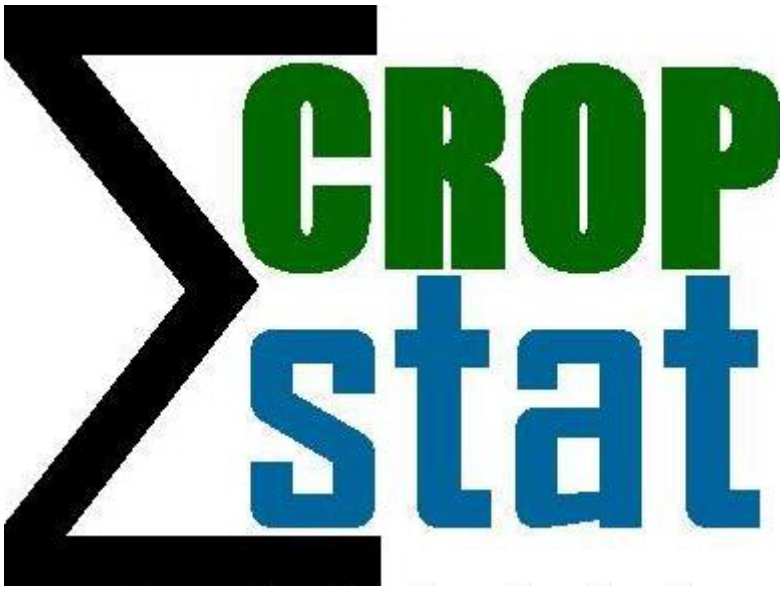


V E R S I O N 7 . 2

Tutorial Manual Part 2

Crop Research Informatics Laboratory
INTERNATIONAL RICE RESEARCH INSTITUTE

A NOTE TO THE READER:
An electronic copy of this tutorial manual comes with the CropStat installer. The CropStat Tutorial may be printed/copied and distributed to any number of users. CropStat is a freeware developed for non-profit use. Hence, selling of either the software or the tutorial is prohibited.

ISBN 978-971-22-0218-6
Revised July 2009

Tutorial Part 2: Contents

Tutorial Part 2: Contents

AMMI MODEL

At the end of the tutorial, the user should be able to

- perform genotype by environment (G×E) analysis
- generate cross-site and stability analysis and AMMI models

I. Sample Problem: Mean Yield Data Set generated by Single Site Analysis

The mean yield dataset *MYIELD94.SYS* generated by the single-site analysis example will be used to illustrate the use of CropStat in running an additive main effects and multiplicative interaction (AMMI) model analysis. This file contains the following variables:

Variable	Description
SET$	Contains the location identification
VTYNO	Contains the genotype levels
VARIETY$	Contains the genotype names
NOS	Contains the number of non-missing observations for each genotype in each environment
YIELD	Mean yield across replications

- Open the data file *MYIELD94.SYS* from the *CROPSTAT7.2\TUTORIAL\ TUTORIAL DATASETS* folder.

- Select **File ⇒ Save-as**. Click the **Save in** box and go go inside working folder *C:\MY CROPSTAT*. Create a subfolder AMMI then click **Save**

II. Genotype by Environment (G×E) Analysis

The input data file may be a raw data file with replicate observations for each genotype x environment (GxE) cell, or a file of means or adjusted means saved from a single site analysis. In the former case, arithmetic means are computed as the data are read into the GxE table. CropStat takes account of missing observations when forming means.

The advantages of using the means file outputted by single site-analysis are:

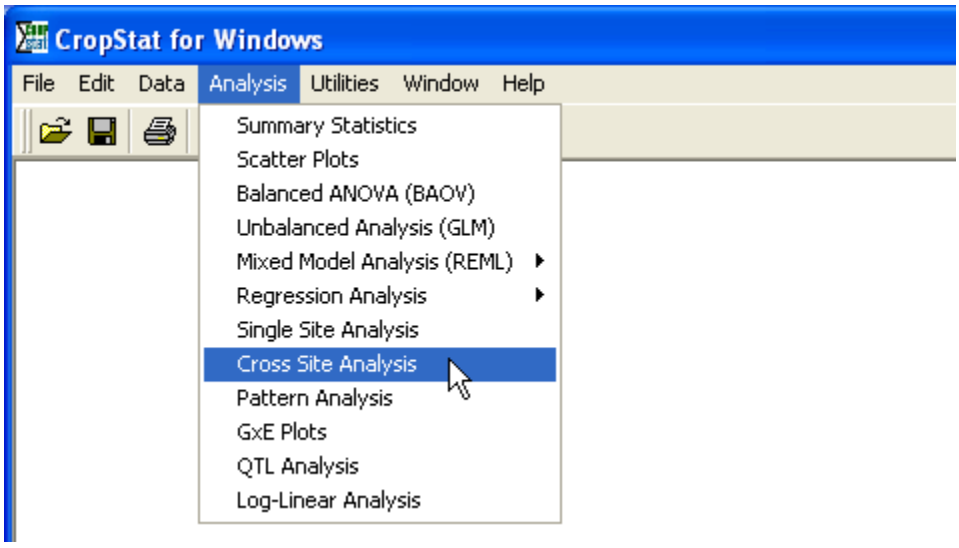
1. Means are adjusted especially if design is augmented or one of the lattices. If raw data are used, arithmetic means are computed. Thus, the outputs will be different.

2. Single-site analysis also saves the EMS of individual ANOVA and thus AMMI will provide LSD values for each site. Using the raw data will not give LSD values.

Treatment by variate tables must have less than 120 treatments (NROW) and 100 variates (NCOL) and (NROW+4) * (NCOL+1) must be less than 8200. Treatment × site tables are similarly restricted (where NCOL is the number of sites) but in addition NROW+NCOL must be less than 126.

The following are the steps in performing a G×E analysis in CropStat.

- Select **Analysis|Cross Site Analysis** from the Main Window.

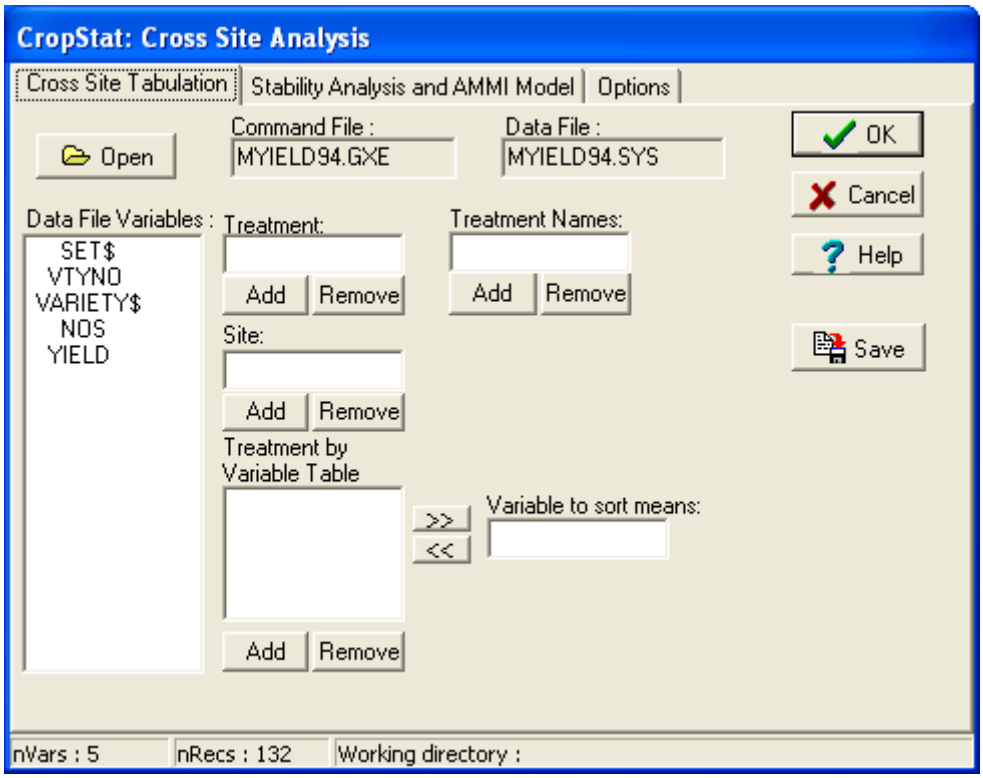


- Click the **Look In** box and select drive *C:\MY CROPSTAT\AMMI* folder.

- In the **File Name** box, type *MYIELD94* as the name of the command file. Click **Open**.

- Since no command file called *MYIELD94.GXE* exists, CropStat will display a message box confirming if you want to create one. Click **Yes**.

- Enter *MYIELD94* as the name of the data file to be used in this analysis. Click **Open**. The **Cross Site Analysis** dialog box will appear.



- Specify the treatment variate defining the row factor levels, usually genotypes. Select *VTYNO* from the **Data File Variables** list box. Click **Add** button under the **Treatment** edit box.

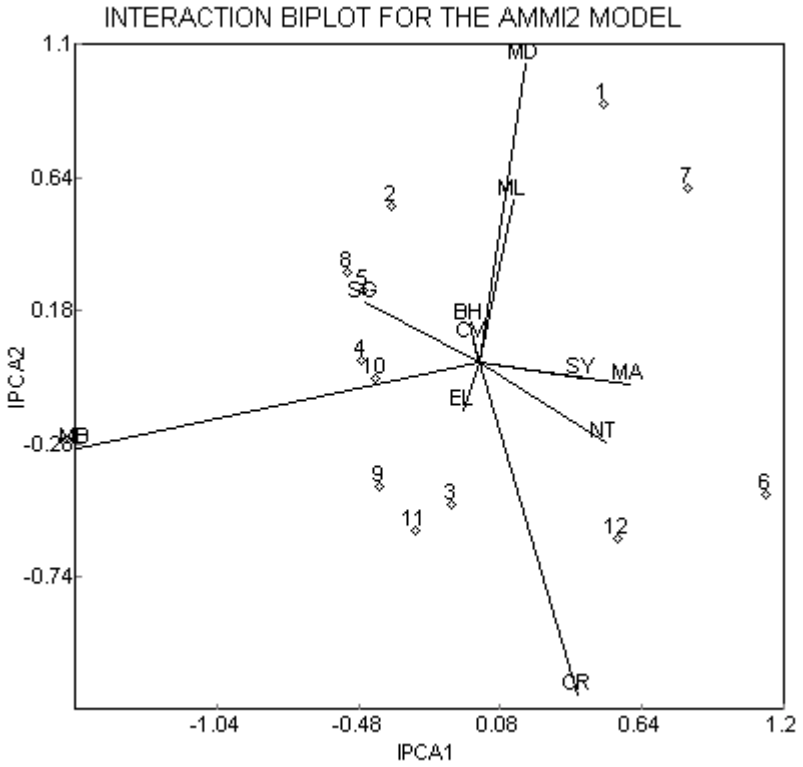
*(**Note**:* Treatment variate may be a character variate or a numeric variate, but it must specify unique row levels for the G×E matrix. Clarity of graphical output is enhanced if the first two characters of the treatment factor are unique.)

- To specify a variate containing treatment names, select *VARIETY$* from the **Data File Variables** list box. Click **Add** button under the **Treatment Names** edit box. These names are printed adjacent to the levels specified by the row factor in output tables. If the row factor is a numeric variate, this allows inclusion of names as well as numbers for treatments. If the row factor is a character variate, the names variate effectively extends the length of the treatment names by 12 characters although the first part of the name (contained in the row factor variate) must uniquely define all the row factor levels. Treatment names need not uniquely define all treatment levels.

```
TREATMENT MEANS AND COUNTS OVER SITES FOR EACH VARIATE.  FILE MYIELD94   12/ 2/ 4
13:21
 ------------------------------------------------------------------------------
:PAGE   2
 VARIETY\SITE              |YIELD       |
 --------------------------|------------
 11          W56-50        | 1.528    11|
 3           GUAR          | 1.484    11|
 7           UPL5          | 1.413    11|
 6           OS6           | 1.326    11|
 9           W181-18       | 1.299    11|
                           |              |
 12          W96-1-1       | 1.286    11|
 5           OL5           | 1.273    11|
 1           AZU           | 1.211    11|
 8           VAND          | 1.195    11|
 4           IT146         | 1.182    11|
                           |              |
 10          W56-125       | 1.140    11|
 2           BGORA         | 1.006    11|
 --------------------------|------------
 SITE MEANS                | 1.278   132|
```

It is useful to include variate containing treatment names in addition to the treatment variate containing levels so that output is well annotated. These names are printed adjacent to the levels specified by the treatment variate in output tables.

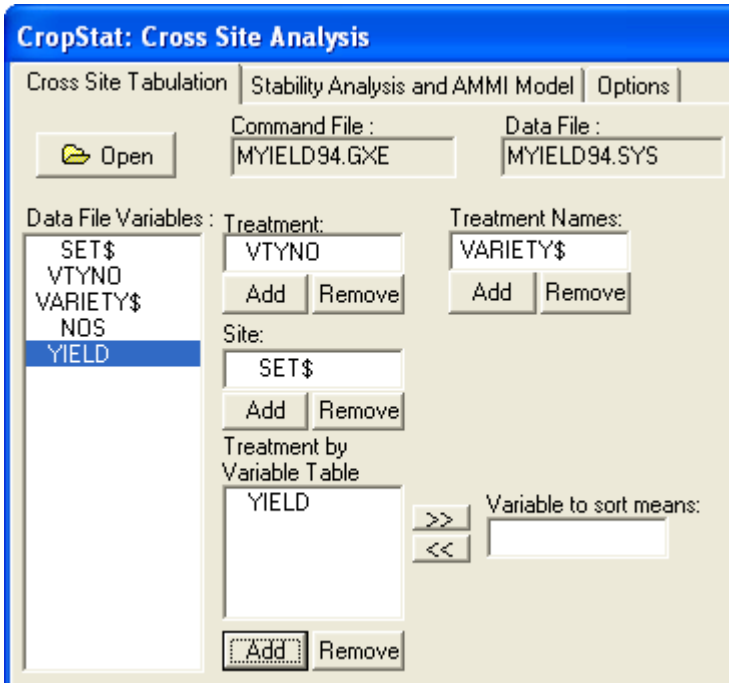
- To specify site variate, select *SET$* from the **Data File Variables** list box. Click **Add** button under the **Site** edit box. This variate may be a character variate or a numeric variate but it must specify unique column levels for the G×E matrix.



VARIATE: YIELD DATA FILE: MYIELD94 MODEL FIT: 66.0% OF GXE S

Clarity of graphical output is enhanced if the first two characters of the treatment and site factors are unique and the site codes are different from the codes for the treatment factor. If there are fewer than 100 rows or genotypes, digits 01 to 99 are useful first characters for the row factor, similarly the letters A…Z, AA…ZZ, etc. form suitable characters for the columns or environments. These abbreviations are useful to distinguish entities on the AMMI biplots.

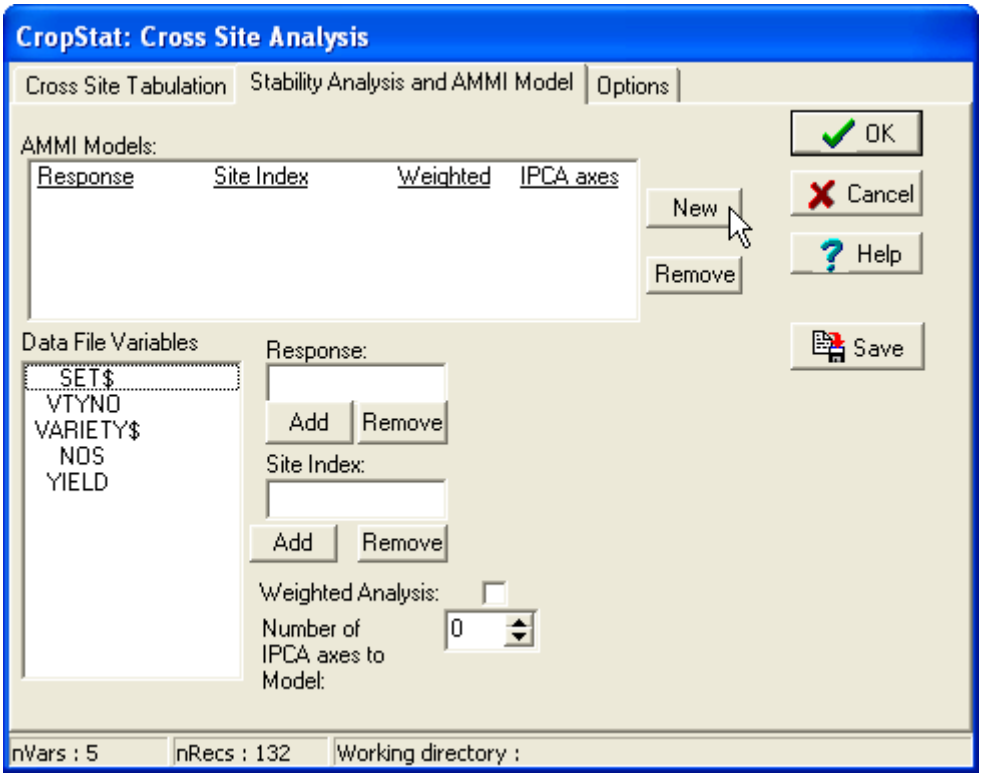
- To request a summary table of treatment means, averaged over sites, for variate yield, select *YIELD* from the **Data File Variables** list box. Click **Add** button under the **Treatment by Variable Table** list.



If the user requires a summary table of treatment means, averaged over sites, for several variates, the list of variates is specified in the **Treatment by Variable** list. The user may also specify one of the variates in the treatment by variate means table on which to sort the rows of all tables of means produced.

- To request for stability and AMMI model analysis, click the **Stability Analysis and AMMI Model** tab. This will bring you to **Stability Analysis and AMMI model** page.

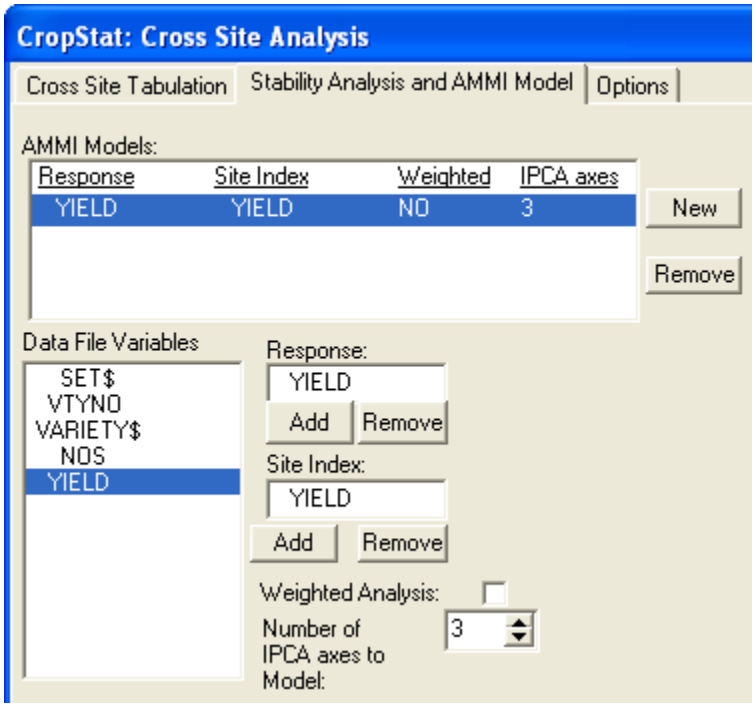
- To specify a model, click **New.**



- Select *YIELD* from **Data File Variables** list box. Click **Add** button under **Response** edit box.

- To specify the variate whose average per site you wish to use as site index in the stability regression, select *YIELD* from **Data File Variables** list box. Click **Add** under **Site Index** edit box. (***Note*:** Site index is usually the response variate but any numeric variate may be specified.)

- To specify the number of IPCA axes to be included in the AMMI model, click on **the Number of IPCA axes to Model** box. Increase the number to *3*.



- Click **OK** to run the analysis.

III. Sample Output

The following output will be displayed in the Text Editor. This is saved in *MYIELD.OUT*.

1. List of treatment and site levels

```
 PBGXE - CROSS SITE ANALYSIS  FILE MYIELD94   12/ 2/ 4 10:33
 ------------------------------------------------------------------- :PAGE   1

  12    VTYNO     CODES:
  1 1           AZU           2 2           BGORA         3 3           GUAR
  4 4           IT146         5 5           OL5           6 6           OS6
  7 7           UPL5          8 8           VAND          9 9           W181-18
 10 10          W56-125      11 11          W56-50       12 12          W96-1-1

  11      SET$    CODES:
  1 BH            2 CR            3 CV            4 EL            5 MA
  6 MB            7 MD            8 ML            9 NT           10 SG
 11 SY

 TREATMENT BY VARIATE MEANS HAVE BEEN REQUESTED FOR  1 VARIATES:
 YIELD
 ROWS OF MEANS TABLES TO BE SORTED ON VARIATE YIELD
 GXE ANALYSIS HAS BEEN REQUESTED FOR  1 VARIATES
```

2. Treatment × variate table of means

```
TREATMENT MEANS AND COUNTS OVER SITES FOR EACH VARIATE.  FILE MYIELD94   12/ 2/
4 10:33
 ------------------------------------------------------------------------------
:PAGE   2
 VARIETY\SITE              |YIELD       |
 --------------------------|------------
 11          W56-50        | 1.528    11|
 3           GUAR          | 1.484    11|
 7           UPL5          | 1.413    11|
 6           OS6           | 1.326    11|
 9           W181-18       | 1.299    11|
                           |            |
 12          W96-1-1       | 1.286    11|
 5           OL5           | 1.273    11|
 1           AZU           | 1.211    11|
 8           VAND          | 1.195    11|
 4           IT146         | 1.182    11|
                           |            |
 10          W56-125       | 1.140    11|
 2           BGORA         | 1.006    11|
 --------------------------|------------
 SITE MEANS                | 1.278   132|
```

3. Treatment × environment table of means

```
  12 X  11 MATRIX OF TREATMENT BY SITE MEANS FOR VARIATE    YIELD  FILE MYIELD94   12/ 2/ 4 10:33
------------------------------------------------------------------------------- :PAGE   3

SECTION  1

VARIETY\SITE             |BH          |CR          |CV          |EL          |MA          |MB          |MD          |ML          |
-------------------------|------------|------------|------------|------------|------------|------------|------------|------------|
11          W56-50       | 1.568      | 2.687      |0.5608      |0.9503      | 1.739      | 2.886      |0.6622      | 1.118      |
3           GUAR         | 1.296      | 2.999      |0.3392      | 1.402      | 1.364      | 2.315      |0.9897      | 1.270      |
7           UPL5         | 1.468      | 1.922      |0.2748      |0.9427      | 2.149      |0.6325      | 1.922      | 2.324      |
6           OS6          | 1.388      | 2.995      |0.3518      |0.7097      | 2.494      |0.1328      |0.6000      | 1.250      |
9           W181-18      | 1.576      | 2.281      |0.3715      | 1.674      |0.7015      | 2.252      |0.3625      |0.9295      |
                         |            |            |            |            |            |            |            |            |
12          W96-1-1      | 1.012      | 3.272      |0.3390      | 1.543      | 1.362      | 1.001      |0.7757      | 1.174      |
5           OL5          | 1.524      | 1.422      |0.1348      | 1.025      | 1.542      | 2.414      |0.7732      | 1.665      |
1           AZU          | 1.344      | 1.097      |0.4838      |0.7437      | 1.095      |0.7377      | 2.270      | 1.421      |
8           VAND         | 1.752      | 2.079      |0.2125      |0.1192E-06  | 1.038      | 2.408      | 1.393      | 1.470      |
4           IT146        |0.6880      | 2.079      |0.1535      |0.9647      |0.4965      | 2.782      | 1.501      | 1.366      |
                         |            |            |            |            |            |            |            |            |
10          W56-125      | 1.080      | 1.508      |0.5073      |0.5350      | 1.820      | 2.662      |0.5667      | 1.096      |
2           BGORA        |0.8480      |0.8477      |0.2168      | 1.520      |0.7500      | 1.578      | 1.058      | 1.244      |
-------------------------|------------|------------|------------|------------|------------|------------|------------|------------|
SITE MEANS               | 1.295      | 2.099      |0.3288      | 1.001      | 1.379      | 1.817      | 1.073      | 1.361      |
SITE INDEX               | 1.295      | 2.099      |0.3288      | 1.001      | 1.379      | 1.817      | 1.073      | 1.361      |
SE OF MEANS              |0.1336      |0.2363      |0.5772E-01  |0.1717      |0.1811      |0.2563      |0.2499      |0.2417      |
                         |            |            |            |            |            |            |            |            |
LSD(5%)                  |0.3844      |0.6931      |0.1661      |0.5036      |0.5211      |0.7375      |0.7190      |0.6953      |

SECTION  2

VARIETY\SITE             |NT          |SG          |SY          |TRT MEANS   |
-------------------------|------------|------------|------------|------------|
11          W56-50       | 1.791      | 1.347      | 1.495      | 1.528      |
3           GUAR         | 1.501      | 1.590      | 1.260      | 1.484      |
7           UPL5         | 1.326      |0.9150      | 1.667      | 1.413      |
6           OS6          | 1.812      | 1.188      | 1.670      | 1.326      |
9           W181-18      | 1.236      | 1.825      | 1.076      | 1.299      |
                         |            |            |            |            |
12          W96-1-1      | 1.271      |0.8950      | 1.497      | 1.286      |
5           OL5          |0.5510      | 1.860      | 1.087      | 1.273      |
1           AZU          | 1.565      | 1.140      | 1.418      | 1.211      |
8           VAND         |    0.0000  | 1.883      |0.9123      | 1.195      |
4           IT146        | 1.234      |0.7925      |0.9420      | 1.182      |
                         |            |            |            |            |
10          W56-125      |0.7120      | 1.027      | 1.024      | 1.140      |
2           BGORA        |0.6473      | 1.898      |0.4563      | 1.006      |
-------------------------|------------|------------|------------|------------|
SITE MEANS               | 1.137      | 1.363      | 1.209      | 1.278      |
SITE INDEX               | 1.137      | 1.363      | 1.209      | 1.278      |
SE OF MEANS              |0.2917      |0.3215      |0.2177      |    .       |
                         |            |            |            |            |
LSD(5%)                  |0.8554      |0.9249      |0.6386      |    .       |

ANALYSIS OF RESIDUAL VARIATION WITHIN SITES
POOLED ERROR MEAN SQUARES FOR 11 SITES WITH  319 D.F.= 0.18736
BARTLETT'S STATISTIC=   93.79 P-VALUE (CHI^2 WITH 10 D.F.) = 1.000
```

4. Main effects analysis
   a) Environment means

```
PREDICTED TREATMENT AND ENVIRONMENT MEANS  FILE MYIELD94   12/ 2/ 4 10:33
 ------------------------------------------------------------------- :PAGE   4

 PREDICTED MEANS, SES AND MULTIPLE COMPARISONS

 ENVIRONMENT        MEAN          SE    DUNCAN GROUPS  LSD TESTS
 CR               2.0991     0.16070     |   .
 MB               1.8167     0.16070     ||  ..
 MA               1.3792     0.16070      || 2..
 SG               1.3634     0.16070      || 21..
 ML               1.3608     0.16070      || 21...
 BH               1.2953     0.16070       | 31....
 SY               1.2087     0.16070       | 32.....
 NT               1.1372     0.16070       | 32......
 MD               1.0728     0.16070       | 32.......
 EL               1.0008     0.16070       | 33........
 CV              0.32879     0.16070         3333333322.
```

b) Treatment means

```
PREDICTED MEANS, SES AND MULTIPLE COMPARISONS

 TREATMENT                   MEAN          SE    DUNCAN GROUPS  LSD TESTS
11          W56-50         1.5277      0.16785     | .
3           GUAR           1.4842      0.16785     | ..
7           UPL5           1.4130      0.16785     | ...
6           OS6            1.3264      0.16785     | ....
9           W181-18        1.2986      0.16785     | .....
12          W96-1-1        1.2856      0.16785     | ......
5           OL5            1.2726      0.16785     | .......
1           AZU            1.2105      0.16785     | ........
8           VAND           1.1953      0.16785     | .........
4           IT146          1.1818      0.16785     | ..........
10          W56-125        1.1399      0.16785     | ...........
2           BGORA          1.0058      0.16785     | 11.........
```

5. Analysis of residuals from the main effects model
   a) Table of studentized residuals

```
RESIDUALS FROM THE ADDITIVE TREATMENT BY SITE MODEL

(ENTRIES ARE SIZE OF RESIDUAL IN STANDARD ERRORS,
 ROWS AND COLUMNS SORDED ACCORDING TO MARGINAL MEANS)
                                                                |
                                                                |
                                                                |
                                                                |
                                                                |
                                                                |   T
                                                                |   -
                                                                |   E
                                                                |   F
                                                                |   C
                         C   M   M   S   M   B   S   N   M   E   C |   T
                         R   B   A   G   L   H   Y   T   D   L   V |   S
                        ---------------------------------------------
11          W56-50       0   1   0   0   0   0   0   0  -1   0   0 |   1
3           GUAR         1   0   0   0   0   0   0   0   0   0   0 |   1
7           UPL5         0  -2   1  -1   1   0   0   0   1   0   0 |   0
6           OS6          1  -3   2   0   0   0   0   1  -1   0   0 |   0
9           W181-18      0   0  -1   0   0   0   0   0  -1   1   0 |   0
12          W96-1-1      2  -1   0   0   0   0   0   0   0   1   0 |   0
5           OL5         -1   1   0   0   0   0   0  -1   0   0   0 |   0
1           AZU         -1  -1   0   0   0   0   0   0   2   0   0 |   0
8           VAND         0   1   0   1   0   1   0  -2   0  -1   0 |   0
4           IT146        0   2  -1   0   0  -1   0   0   1   0   0 |   0
10          W56-125      0   1   1   0   0   0   0   0   0   0   0 |   0
2           BGORA       -1   0   0   1   0   0   0   0   0   1   0 |  -1
                        ---------------------------------------------
SITE EFFECTS             5   3   0   0   0   0   0   0  -1  -1  -6 |  26
```

b) Box plot of studentized residuals

```
 BOX PLOT OF  132 STUDENTIZED RESIDUALS FROM LPLT= -3.408    TO ULPT=  2.490
 NO.<LPLT
NO.>UPLT
        0 *       *     ----------------I     +     I----------------- *   0

 MEDIAN=  0.2354E-01 ANDERSON-DARLING STATISTIC=  0.474
```

c) Analysis of variance for the additive model

ANALYSIS OF VARIANCE FOR THE ADDITIVE MODEL

SOURCE	D.F.	S.S.	M.S.	F	FPROB
TREATMENTS	11	2.63722	0.239747		
LOCATIONS	10	24.4041	2.44041		
TREATMENT X SITES	110	34.0890	0.309900		
POOLED ERROR(PER MEAN)	319	16.4361	0.515238E-01		
TOTAL	131	61.1303			

ESTIMATED MAXIMUM STRUCTURAL CONTENT OF TREATMENT X SITE SS IS 83.37%

d) Table of raw residuals from the additive site × treatment additive model

RESIDUALS FROM THE ADDITIVE SITE X TREATMENT MODEL FILE MYIELD94 12/ 2/ 4 10:33 :PAGE 5

SECTION 1

VARIETY\SITE		BH	CR	CV	EL	MA	MB	MD	ML
11	W56-50	0.2340E-01	0.3386	-.1731E-01	-.2998	0.1103	0.8198	-.6599	-.4915
3	GUAR	-.2050	0.6938	-.1953	0.1955	-.2207	0.2926	-.2888	-.2960
7	UPL5	0.3815E-01	-.3120	-.1886	-.1927	0.6358	-1.319 *	0.7146	0.8282
6	OS6	0.4473E-01	0.8483	-.2498E-01	-.3391	1.067 *	-1.732 ***	-.5208	-.1590
9	W181-18	0.2605	0.1617	0.2255E-01	0.6530	-.6979	0.4154	-.7305	-.4514
12	W96-1-1	-.2905	1.166 *	0.3009E-02	0.5350	-.2493E-01	-.8229	-.3043	-.1935
5	OL5	0.2345	-.6712	-.1882	0.2969E-01	0.1689	0.6034	-.2937	0.3103
1	AZU	0.1166	-.9342	0.2229	-.1893	-.2166	-1.011 *	1.265 *	0.1286
8	VAND	0.5398	0.6338E-01	-.3313E-01	-.9177	-.2581	0.6742	0.4033	0.1926
4	IT146	-.5106	0.7691E-01	-.7860E-01	0.6052E-01	-.7860	1.062 *	0.5246	0.1017
10	W56-125	-.7684E-01	-.4523	0.3170	-.3273	0.5793	0.9838	-.3676	-.1260
2	BGORA	-.1747	-.9788	0.1606	0.7922	-.3565	0.3370E-01	0.2576	0.1559
SITE EFFECTS		0.1689E-01	0.8207 ***	-.9497 ***	-.2776	0.1008	0.5383 ***	-.2056	0.8233E-01

SECTION 2

VARIETY\SITE		NT	SG	SY	T-EFCTS
11	W56-50	0.4045	-.2651	0.3707E-01	0.2493
3	GUAR	0.1577	0.2091E-01	-.1547	0.2057
7	UPL5	0.5459E-01	-.5829	0.3235	0.1345
6	OS6	0.6268	-.2238	0.4131	0.4793E-01
9	W181-18	0.7829E-01	0.4415	-.1532	0.2015E-01
12	W96-1-1	0.1266	-.4756	0.2815	0.7199E-02
5	OL5	-.5804	0.5025	-.1158	-.5862E-02
1	AZU	0.4960	-.1555	0.2772	-.6790E-01
8	VAND	-1.054 *	0.6028	-.2132	-.8316E-01
4	IT146	0.1938	-.4742	-.1700	-.9669E-01
10	W56-125	-.2867	-.1974	-.4584E-01	-.1385
2	BGORA	-.2172	0.8068	-.4797	-.2727
SITE EFFECTS		-.1412	0.8493E-01	-.6978E-01	1.278 ***

6. Table of fitted values from the additive site × treatment model

FITTED VALUES FROM ADDITIVE SITE X TREATMENT MODEL FILE MYIELD94 12/ 2/ 4 10:33
:PAGE 6

SECTION 1

VARIETY\SITE		BH	CR	CV	EL	MA	MB	MD	ML
11	W56-50	1.545	2.348	0.5781	1.250	1.628	2.066	1.322	1.610
3	GUAR	1.501	2.305	0.5345	1.207	1.585	2.022	1.279	1.566
7	UPL5	1.430	2.234	0.4633	1.135	1.514	1.951	1.207	1.495
6	OS6	1.343	2.147	0.3767	1.049	1.427	1.865	1.121	1.409
9	W181-18	1.315	2.119	0.3489	1.021	1.399	1.837	1.093	1.381
12	W96-1-1	1.303	2.106	0.3360	1.008	1.386	1.824	1.080	1.368
5	OL5	1.289	2.093	0.3229	0.9950	1.373	1.811	1.067	1.355
1	AZU	1.227	2.031	0.2609	0.9329	1.311	1.749	1.005	1.293
8	VAND	1.212	2.016	0.2456	0.9177	1.296	1.734	0.9897	1.278
4	IT146	1.199	2.002	0.2321	0.9041	1.283	1.720	0.9761	1.264
10	W56-125	1.157	1.961	0.1903	0.8623	1.241	1.678	0.9343	1.222
2	BGORA	1.023	1.826	0.5611E-01	0.7282	1.107	1.544	0.8002	1.088
SITE ESTS.		1.295	2.099	0.3288	1.001	1.379	1.817	1.073	1.361

SECTION 2

VARIETY\SITE		NT	SG	SY	T-ESTS.
11	W56-50	1.386	1.613	1.458	1.528
3	GUAR	1.343	1.569	1.414	1.484
7	UPL5	1.272	1.498	1.343	1.413
6	OS6	1.185	1.411	1.257	1.326
9	W181-18	1.157	1.384	1.229	1.299
12	W96-1-1	1.144	1.371	1.216	1.286
5	OL5	1.131	1.358	1.203	1.273
1	AZU	1.069	1.295	1.141	1.211
8	VAND	1.054	1.280	1.126	1.195
4	IT146	1.041	1.267	1.112	1.182
10	W56-125	0.9987	1.225	1.070	1.140
2	BGORA	0.8645	1.091	0.9360	1.006
SITE ESTS.		1.137	1.363	1.209	1.278

7. Stability analysis

ANOVA AND STABILITY REGRESSIONS FILE MYIELD94 12/ 2/ 4 10:33
:PAGE 7

REGRESSIONS OF YIELD FOR EACH VARIETY ON MEANS OF YIELD AT EACH SITE

VARIETY		MEAN	SLOPE	SE	MS-TXL	MS-REG	MS-DEV	R**2(%)
1	AZU	1.21	0.094*	0.357	0.40	1.67	0.26	42.
2	BGORA	1.01	0.457	0.346	0.28	0.60	0.24	21.
3	GUAR	1.48	1.422*	0.180	0.10	0.36	0.07	38.
4	IT146	1.18	1.221	0.370	0.26	0.10	0.28	4.
5	OL5	1.27	1.090	0.305	0.17	0.02	0.19	1.
6	OS6	1.33	0.974	0.567	0.59	0.00	0.65	0.
7	UPL5	1.41	0.593	0.445	0.40	0.34	0.40	8.
8	VAND	1.20	1.409	0.407	0.34	0.34	0.34	10.
9	W181-18	1.30	1.117	0.340	0.21	0.03	0.23	1.
10	W56-125	1.14	1.048	0.330	0.20	0.00	0.22	0.
11	W56-50	1.53	1.415	0.282	0.18	0.35	0.16	19.
12	W96-1-1	1.29	1.159	0.392	0.29	0.05	0.31	2.

SLOPE - SLOPES OF REGRESSIONS OF VARIETY MEANS ON SITE INDEX.
* INDICATES SLOPES SIGNIFICANTLY DIFFERENT FROM THE
SLOPE FOR THE OVERALL REGRESSION WHICH IS 1.00
MS-TXL - CONTRIBUTION OF EACH VARIETY TO INTERACTION MS
MS-REG - CONTRIBUTION OF EACH VARIETY TO THE REGRESSION
COMPONENT OF THE TREATMENT BY LOCATION INTERACTION
MS-DEV - DEVIATIONS FROM REGRESSION COMPONENT OF INTERACTION
R**2 - SQUARED CORRELATION BETWEEN RESIDUALS FROM THE MAIN
EFFECTS MODEL AND THE SITE INDEX.

VARIATE YIELD WAS SITE INDEX WITH OVERALL MEAN 1.278
THE FOLLOWING SITE MEANS OF YIELD WERE USED AS X-VARIATES
1.295 2.099 0.3288 1.001 1.379 1.817 1.073 1.361 1.137 1.363
1.209

ANOVA FOR VARIABLE YIELD WITH SITE REGRESSIONS ON YIELD

SOURCE	D.F.	S.S.	M.S.	F	FPROB
TREATMENTS	11	2.63722	0.239747		
LOCATIONS	10	24.4041	2.44041		
TREATMENT X SITES	110	34.0890	0.309900		
TRT X SITE REG	11	3.85730	0.350664	1.148	0.333
DEVIATIONS	99	30.2317	0.305370		
TOTAL	131	61.1303			

8. AMMI models

a) Singular values and IPCA scores

```
AMMI ANALYSIS  FILE MYIELD94   12/ 2/ 4 10:33
 ------------------------------------------------------------------ :PAGE   8
 SINGULAR VALUES OF INTERACTION MATRIX (CONDITION=  0)
 3.7189    2.9465    2.0665    1.8115    1.4686    1.0853    .59530    .53873    .17090
.12911

 SCORES FOR FIRST 4 AMMI COMPONENTS FOR TREATMENTS

 1  1           AZU          0.484989E+00 0.894846E+00-0.341668E+00-0.260893E+00
 2  2           BGORA       -0.347714E+00 0.542918E+00-0.542124E+00 0.633119E+00
 3  3           GUAR        -0.112153E+00-0.485549E+00-0.203158E+00-0.906262E-01
 4  4           IT146       -0.472739E+00 0.101247E-01-0.427652E+00-0.855010E+00
 5  5           OL5         -0.457565E+00 0.248741E+00 0.335581E+00 0.440063E+00
 6  6           OS6          0.113068E+01-0.456813E+00 0.352428E+00 0.292901E+00
 7  7           UPL5         0.818176E+00 0.600134E+00 0.195874E+00-0.158178E+00
 8  8           VAND        -0.523923E+00 0.310660E+00 0.673755E+00-0.486120E-01
 9  9           W181-18     -0.399933E+00-0.425652E+00-0.461328E+00 0.460309E+00
 10 10          W56-125     -0.413494E+00-0.532487E-01 0.556074E+00-0.105940E+00
 11 11          W56-50      -0.253762E+00-0.579493E+00 0.223528E+00-0.268181E+00
 12 12          W96-1-1      0.547436E+00-0.606668E+00-0.361309E+00-0.389522E-01

 SCORES FOR FIRST 4 AMMI COMPONENTS FOR ENVIRONMENTS

 BH             BH          -0.441225E-01 0.131777E+00 0.352211E+00 0.293306E+00
 CR             CR           0.383300E+00-0.114499E+01 0.288672E-01-0.286554E+00
 CV             CV          -0.273747E-01 0.690729E-01-0.291638E-01 0.282094E-01
 EL             EL          -0.668252E-01-0.164046E+00-0.938567E+00 0.477560E+00
 MA             MA           0.590848E+00-0.728330E-01 0.827118E+00 0.226515E+00
 MB             MB          -0.159586E+01-0.293058E+00 0.211775E+00-0.450511E+00
 MD             MD           0.180824E+00 0.102976E+01-0.248242E+00-0.614023E+00
 ML             ML           0.134380E+00 0.560760E+00 0.158012E+00-0.554201E-01
 NT             NT           0.492593E+00-0.272652E+00-0.479483E+00-0.288764E+00
 SG             SG          -0.456540E+00 0.208176E+00-0.236278E-02 0.820922E+00
 SY             SY           0.408781E+00-0.519706E-01 0.119835E+00-0.151240E+00
```

b) Residuals and AMMI ANOVA for the specified model

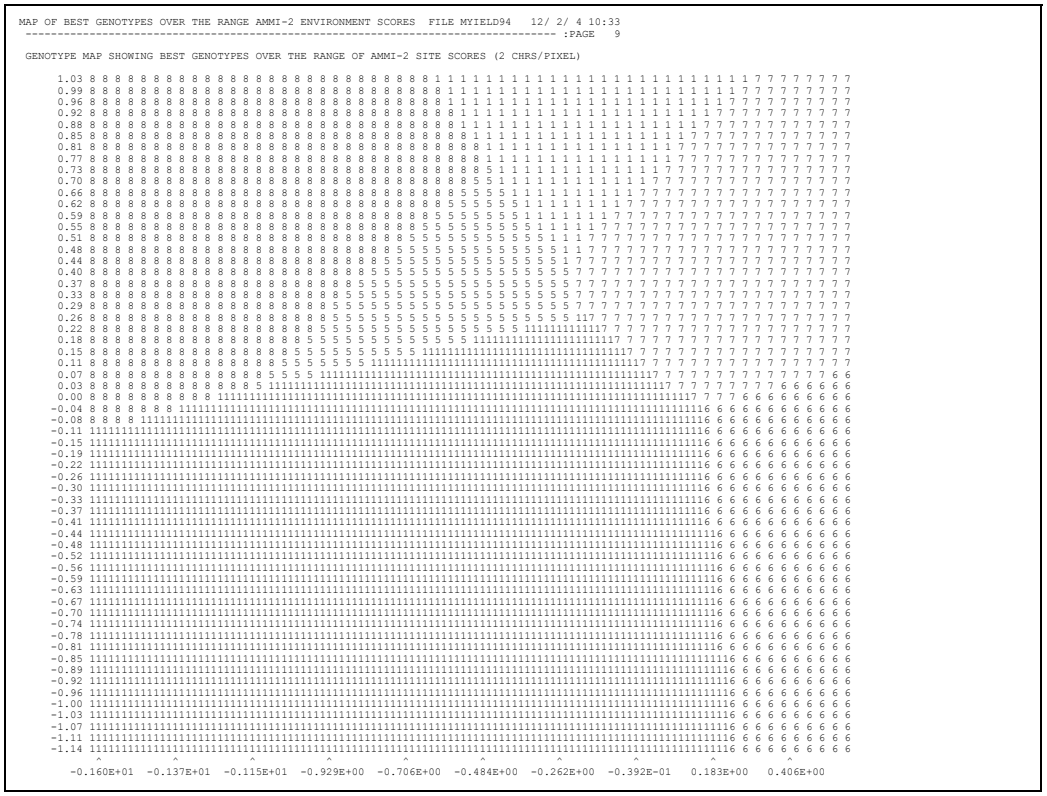
```
 RESIDUALS FROM THE AMMI-3 MODEL

 (ENTRIES ARE SIZE OF RESIDUAL IN UNITS OF ROOT (RESIDUAL GXE MS),  ROWS AND COLUMNS SORDED ACCORDING TO
MARGINAL MEANS)
                                                                 |
                                                                 |
                                                                 |
                                                                 |
                                                                 |
                                                                 |   T
                                                                 |   -
                                                                 |   E
                                                                 |   F
                                                                 |   C
                          C  M  M  S  M  B  S  N  M  E  C |   T
                          R  B  A  G  L  H  Y  T  D  L  V |   S
                         ----------------------------------------
 11          W56-50       0  0  0  0  0  0  0  1  0  0  0 |  26
 3           GUAR         0  0  0  0  0  0  0  0  0  0  0 |  26
 7           UPL5         0  0  0  0  0  0  0  0  0  0  0 |  26
 6           OS6          0  0  0  1  0  0  0  0  0  0  0 |  26
 9           W181-18      0  0  0  0  0  1  0  0  0  0  0 |  26
 12          W96-1-1      0  0  0  0  0  0  0 -1  0  0  0 |  26
 5           OL5          0  0  0  0  0  0  0  0 -1  0  0 |  26
 1           AZU          0  0  0  0 -1  0  0  0  0  0  0 |  26
 8           VAND         1  0 -1  0  0  0  0 -1  0  0  0 |  26
 4           IT146        0  1  0 -1  0 -1  0  0  1 -1  0 |  26
 10          W56-125     -1  0  0 -1  0  0  0  0  0  0  0 |  26
 2           BGORA        0  0  0  1  0  0  0  0 -1  0  0 |  26

 BOX PLOT OF  132 STANDERSIZED RESIDUALS FROM LPLT= -1.919    TO ULPT=  1.662
 NO.<LPLT                                                        NO.>UPLT
       0 *         ---------------I     +       I--------------------   0
```

c) Analysis of variance for the AMMI model

ANALYSIS OF VARIANCE FOR THE AMMI MODEL

SOURCE	D.F.	S.S.	M.S.	F	FPROB
TREATMENTS	11	2.63722	0.239747		
LOCATIONS	10	24.4041	2.44041		
TREATMENT X SITES	110	34.0890	0.309900		
AMMI COMPONENT 1	20	13.8302	0.691510	3.072	0.000
AMMI COMPONENT 2	18	8.68165	0.482314	3.000	0.001
AMMI COMPONENT 3	16	4.27033	0.266896	2.046	0.025
AMMI COMPONENT 4	14	3.28167	0.234405	2.446	0.013
GXE RESIDUAL	42	4.02510			
TOTAL	131	61.1303			

d) Map of best genotypes



e) AMMI residuals, additive effects, and multiplicative scores

AMMI RESIDUALS, ADDITIVE EFFECTS AND MULTIPLICATIVE SCORES FILE MYIELD94 12/ 2/ 4 10:33 :PAGE 10

SECTION 1

VARIETY\SITE		BH	CR	CV	EL	MA	MB	MD	ML
11	W56-50	0.9838E-02	-.2341	0.2229E-01	-.2020	0.3310E-01	0.1976	0.3826E-01	-.1678
3	GUAR	-.7446E-01	0.1867	-.1707	-.8237E-01	-.2176E-01	0.1431E-01	0.1810	0.2346E-01
7	UPL5	-.7382E-01	0.5592E-01	-.2019	0.1443	0.3403E-01	0.1213	-.2667E-02	0.3508
6	OS6	0.3069E-01	-.1183	0.4781E-01	-.7704E-02	0.7376E-01	-.1360	-.1673	-.1104
9	W181-18	0.4614	-.1590	0.2755E-01	0.1235	-.1110	-.2499	-.3344	-.8610E-01
12	W96-1-1	-.5918E-01	0.2717	0.4936E-01	0.1329	-.9372E-01	-.5057E-01	0.1318	0.1303
5	OL5	0.6337E-01	-.2207	-.2081	0.3549	0.1798	-.1250	-.3838	0.1793
1	AZU	0.1404	-.8566E-01	0.1644	-.3307	-.1554	0.9750E-01	0.1713	-.3843
8	VAND	0.2385	0.6005	-.4929E-01	-.2694	-.4832	-.2136	0.3454	-.1763E-01
4	IT146	-.3822	0.2820	-.1047	-.3708	-.1523	0.4011	0.4935	0.2271
10	W56-125	-.2839	-.3708	0.3255	0.1582	0.3598	0.1905	-.9994E-01	-.1285
2	BGORA	-.7060E-01	-.2082	0.9781E-01	0.3492	0.3368	-.2473	-.3732	-.1615E-01
SITE EFFECTS		0.1689E-01	0.8207 ***	-.9497 ***	-.2776	0.1008	0.5383	-.2056 ***	0.8233E-01
AMMI1 SITE		-.4412E-01	0.3833	-.2737E-01	-.6683E-01	0.5908	-1.596 ***	0.1808 ***	0.1344
AMMI2 SITE		0.1318 ***	-1.145	0.6907E-01	-.1640	-.7283E-01	-.2931	1.030	0.5608
AMMI3 SITE		0.3522 ***	0.2887E-01	-.2916E-01	-.9386	0.8271	0.2118 ***	-.2482 ***	0.1580 ***

SECTION 2

VARIETY\SITE		NT	SG	SY	T-EFCTS	AMMI1 TRT	AMMI2 TRT	AMMI3 TRT
11	W56-50	0.4787	-.2598	0.8390E-01	0.2493 ***	-.2538	-.5795 ***	0.2235 ***
3	GUAR	-.1682E-01	0.7031E-01	-.1098	0.2057	-.1122 ***	-.4855 ***	-.2032 ***
7	UPL5	-.9089E-01	-.3338	-.3255E-02	0.1345	0.8182 ***	0.6001 ***	0.1959 ***
6	OS6	0.1143	0.3883	-.1151	0.4793E-01	1.131 ***	-.4568	0.3524 ***
9	W181-18	-.6196E-01	0.3464	0.4349E-01	0.2015E-01	-.3999	-.4257 ***	-.4613 ***
12	W96-1-1	-.4817	-.1002	0.6945E-01	0.7199E-02	0.5474	-.6067	-.3613 ***
5	OL5	-.1262	0.2426	0.4395E-01	-.5862E-02	-.4576 ***	0.2487 ***	0.3356 ***
1	AZU	0.3373	-.1212	0.1664	-.6790E-01	0.4850	0.8948	-.3417 ***
8	VAND	-.3882	0.3005	-.6360E-01	-.8316E-01	-.5239 ***	0.3107	0.6738 ***
4	IT146	0.2244	-.6931	0.7504E-01	-.9669E-01	-.4727 ***	0.1012E-01	-.4277 ***
10	W56-125	0.1691	-.3738	0.5379E-01	-.1385 ***	-.4135 ***	-.5325E-01	0.5561 ***
2	BGORA	-.1578	0.5338	-.2443	-.2727	-.3477	0.5429 ***	-.5421 ***
SITE EFFECTS		-.1412	0.8493E-01	-.6978E-01	1.278 ***	.	.	.
AMMI1 SITE		0.4926 ***	-.4565	0.4088	.	.	.	.
AMMI2 SITE		-.2727	0.2082 ***	-.5197E-01	.	.	.	.
AMMI3 SITE		-.4795	-.2363E-02	0.1198 ***	.	.	.	.

f) Fitted values from the AMMI model

FITTED VALUES FROM THE AMMI MODEL FILE MYIELD94 12/ 2/ 4 10:33 :PAGE 11

SECTION 1

VARIETY\SITE		BH	CR	CV	EL	MA	MB	MD	ML
11	W56-50	1.558	2.921	0.5385	1.152	1.706	2.688	0.6240	1.286
3	GUAR	1.370	2.812	0.5100	1.484	1.386	2.301	0.8087	1.247
7	UPL5	1.542	1.866	0.4767	0.7984	2.115	0.5112	1.925	1.973
6	OS6	1.357	3.114	0.3039	0.7174	2.420	0.2688	0.7673	1.360
9	W181-18	1.115	2.440	0.3439	1.551	0.8125	2.502	0.6969	1.016
12	W96-1-1	1.071	3.000	0.2896	1.410	1.455	1.052	0.6440	1.044
5	OL5	1.461	1.643	0.3429	0.6698	1.362	2.539	1.157	1.486
1	AZU	1.204	1.183	0.3194	1.074	1.250	0.6403	2.099	1.806
8	VAND	1.514	1.479	0.2618	0.2694	1.521	2.621	1.048	1.488
4	IT146	1.070	1.797	0.2582	1.335	0.6488	2.381	1.007	1.139
10	W56-125	1.364	1.879	0.1817	0.3768	1.460	2.471	0.6667	1.225
2	BGORA	0.9186	1.056	0.1189	1.171	0.4132	1.825	1.431	1.260
SITE ESTS.		1.295	2.099	0.3288	1.001	1.379	1.817	1.073	1.361
AMMI1 SITE		-.4412E-01	0.3833	-.2737E-01	-.6683E-01	0.5908	-1.596	0.1808	0.1344
AMMI2 SITE		0.1318	-1.145	0.6907E-01	-.1640	-.7283E-01	-.2931	1.030	0.5608
AMMI3 SITE		0.3522	0.2887E-01	-.2916E-01	-.9386	0.8271	0.2118	-.2482	0.1580

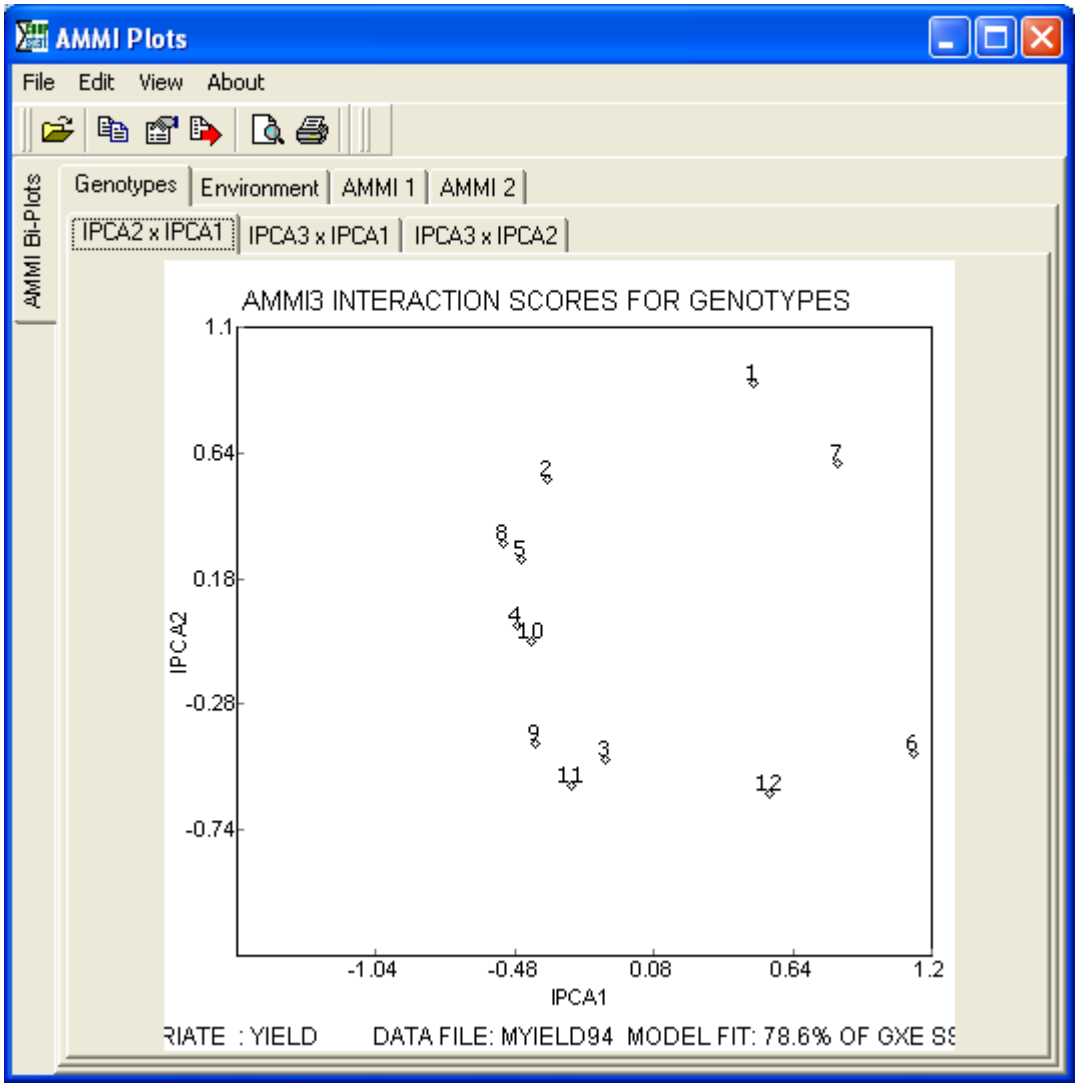
SECTION 2

VARIETY\SITE		NT	SG	SY	T-ESTS.	AMMI1 TRT	AMMI2 TRT	AMMI3 TRT
11	W56-50	1.312	1.607	1.411	1.528	-.2538	-.5795	0.2235
3	GUAR	1.517	1.520	1.369	1.484	-.1122	-.4855	-.2032
7	UPL5	1.417	1.249	1.670	1.413	0.8182	0.6001	0.1959
6	OS6	1.698	0.7992	1.785	1.326	1.131	-.4568	0.3524
9	W181-18	1.298	1.479	1.032	1.299	-.3999	-.4257	-.4613
12	W96-1-1	1.753	0.9952	1.428	1.286	0.5474	-.6067	-.3613
5	OL5	0.6772	1.617	1.043	1.273	-.4576	0.2487	0.3356
1	AZU	1.228	1.261	1.252	1.211	0.4850	0.8948	-.3417
8	VAND	0.3882	1.582	0.9759	1.195	-.5239	0.3107	0.6738
4	IT146	1.010	1.486	0.8670	1.182	-.4727	0.1012E-01	-.4277
10	W56-125	0.5429	1.401	0.9705	1.140	-.4135	-.5325E-01	0.5561
2	BGORA	0.8052	1.364	0.7007	1.006	-.3477	0.5429	-.5421
SITE ESTS.		1.137	1.363	1.209	1.278	.	.	.
AMMI1 SITE		0.4926	-.4565	0.4088	.	.	.	.
AMMI2 SITE		-.2727	0.2082	-.5197E-01	.	.	.	.
AMMI3 SITE		-.4795	-.2363E-02	0.1198	.	.	.	.

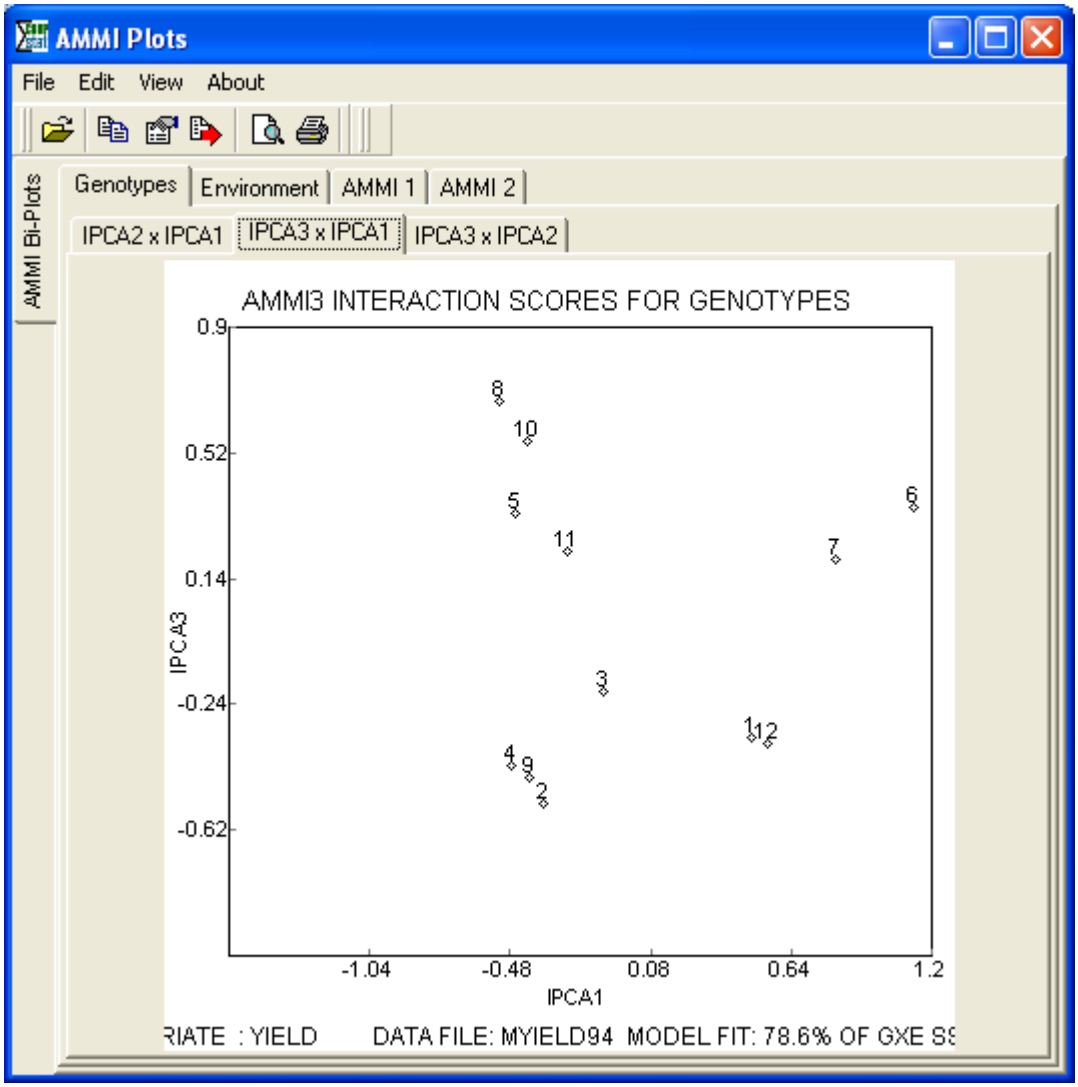
IV. Graphical Output

CropStat will also output the following AMMI plots.

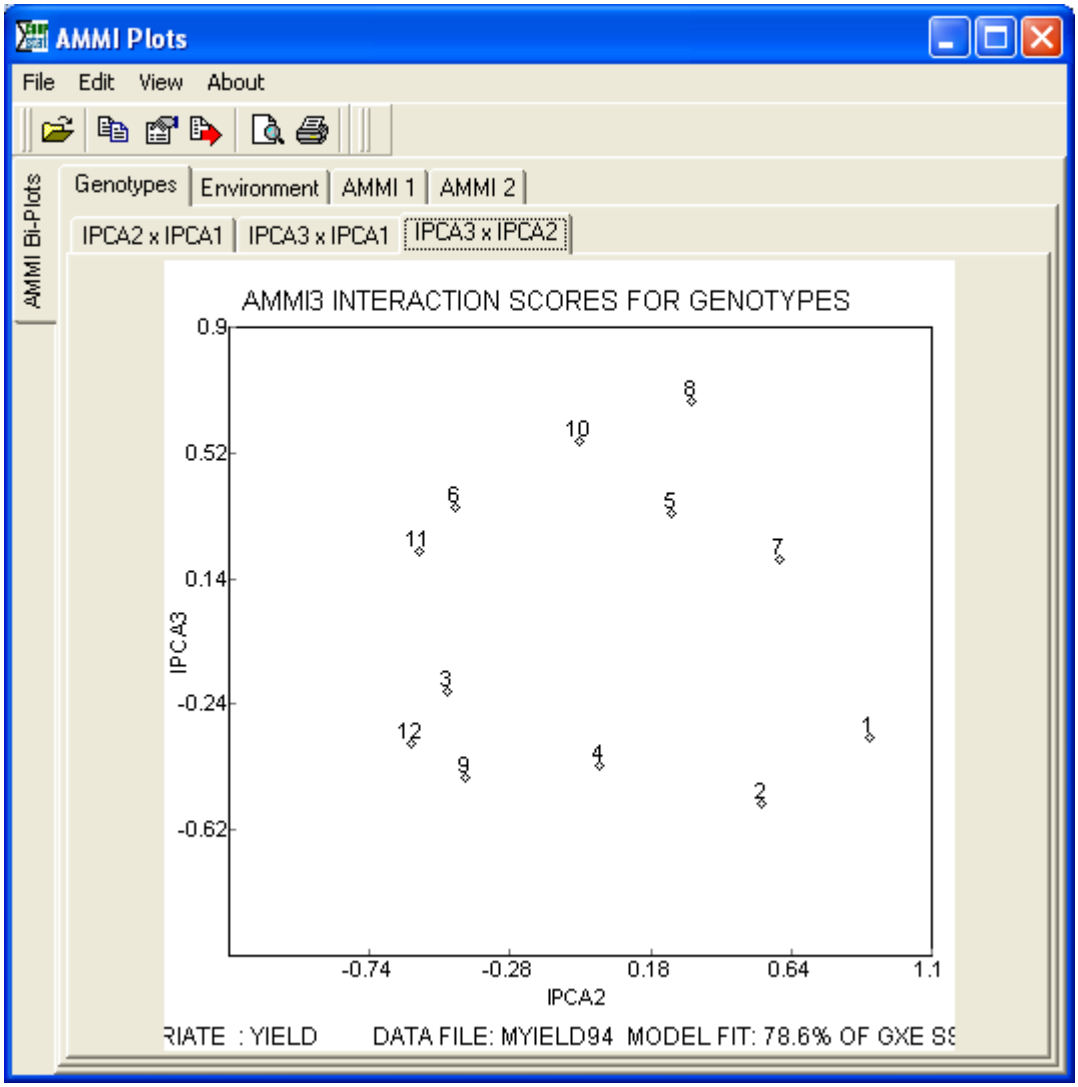
1. Genotypes
   a. IPCA2 × IPCA1



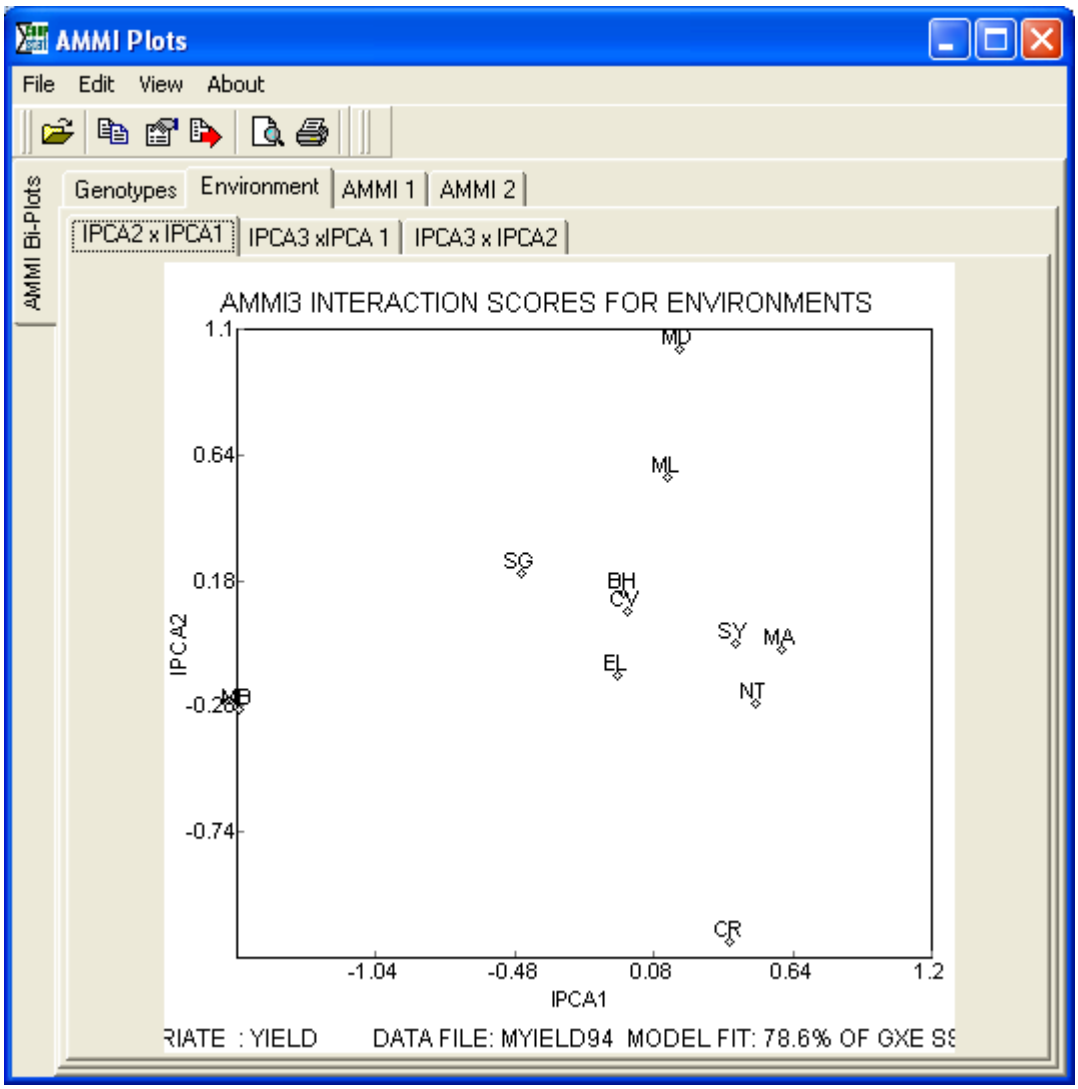
b. IPCA3 × IPCA1



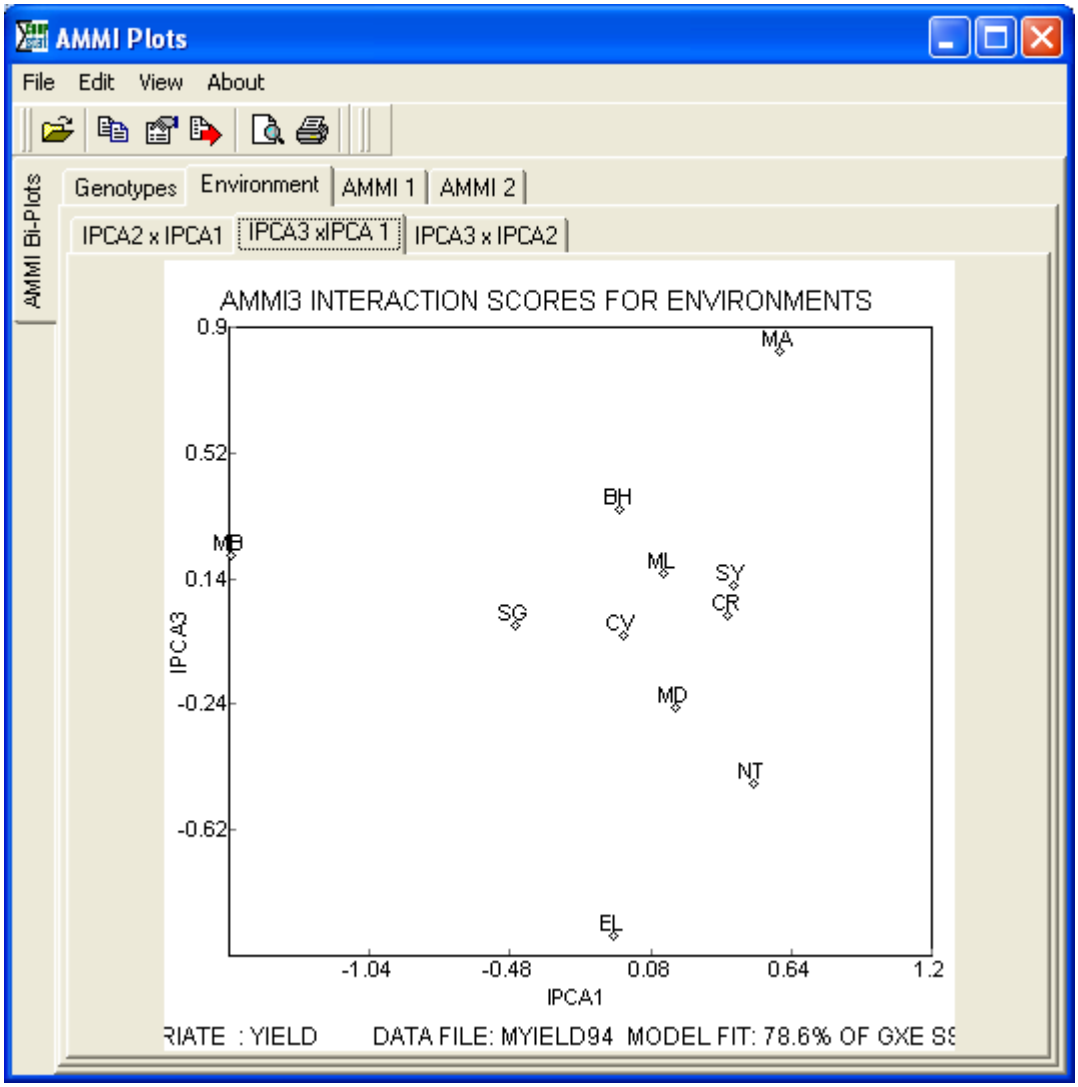
c. IPCA3 × IPCA2



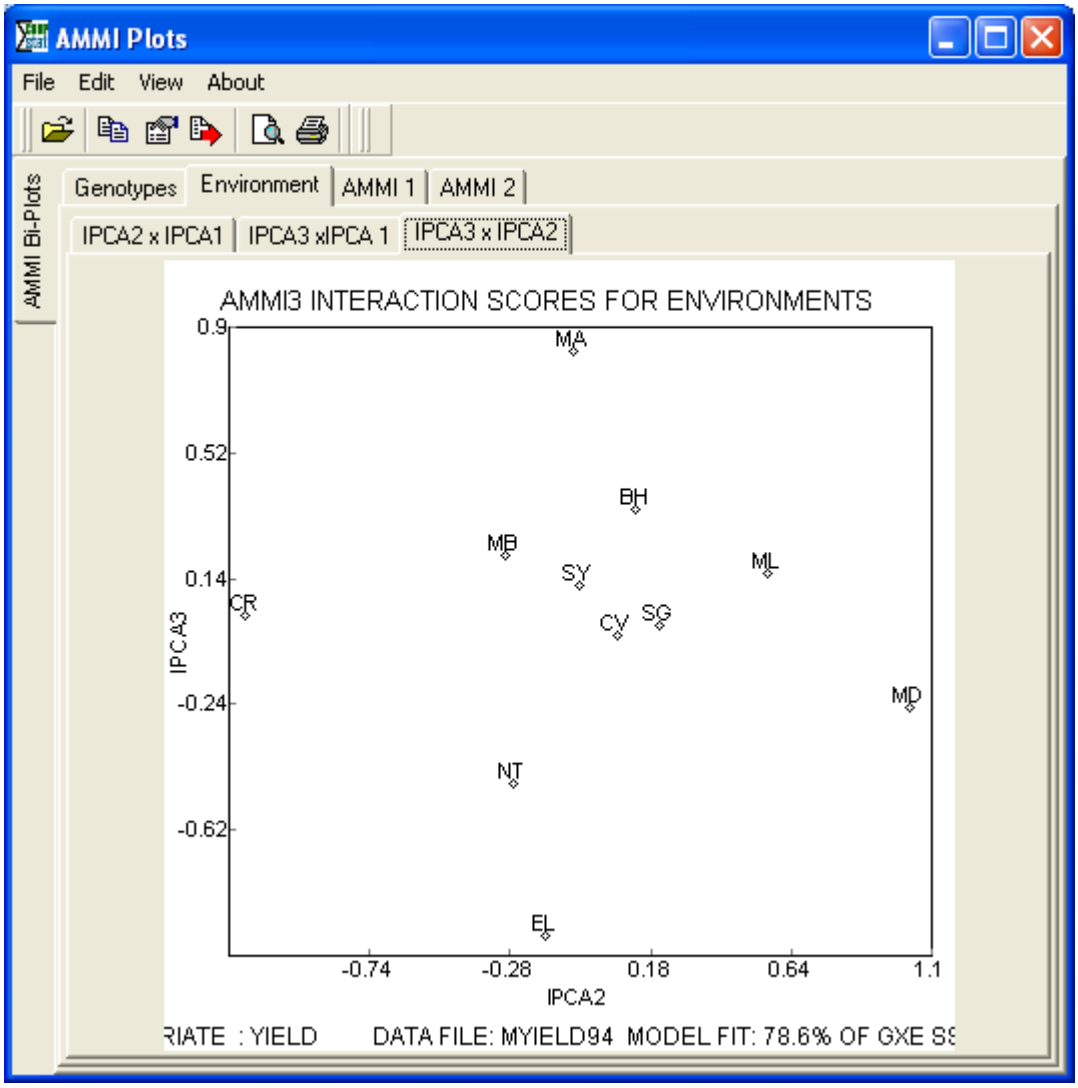
2. Environment
   a. IPCA2 × IPCA1



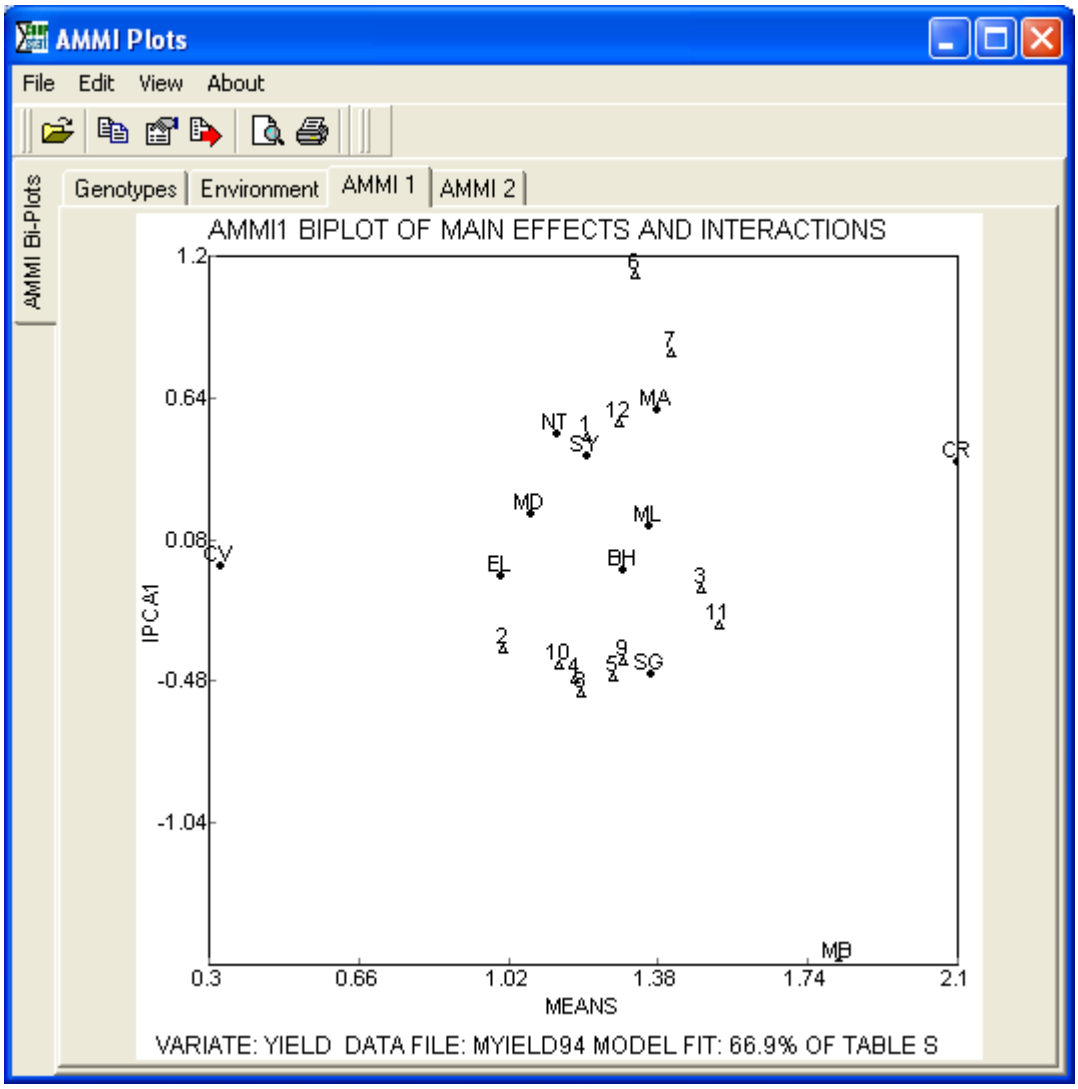
b. IPCA3 × IPCA1



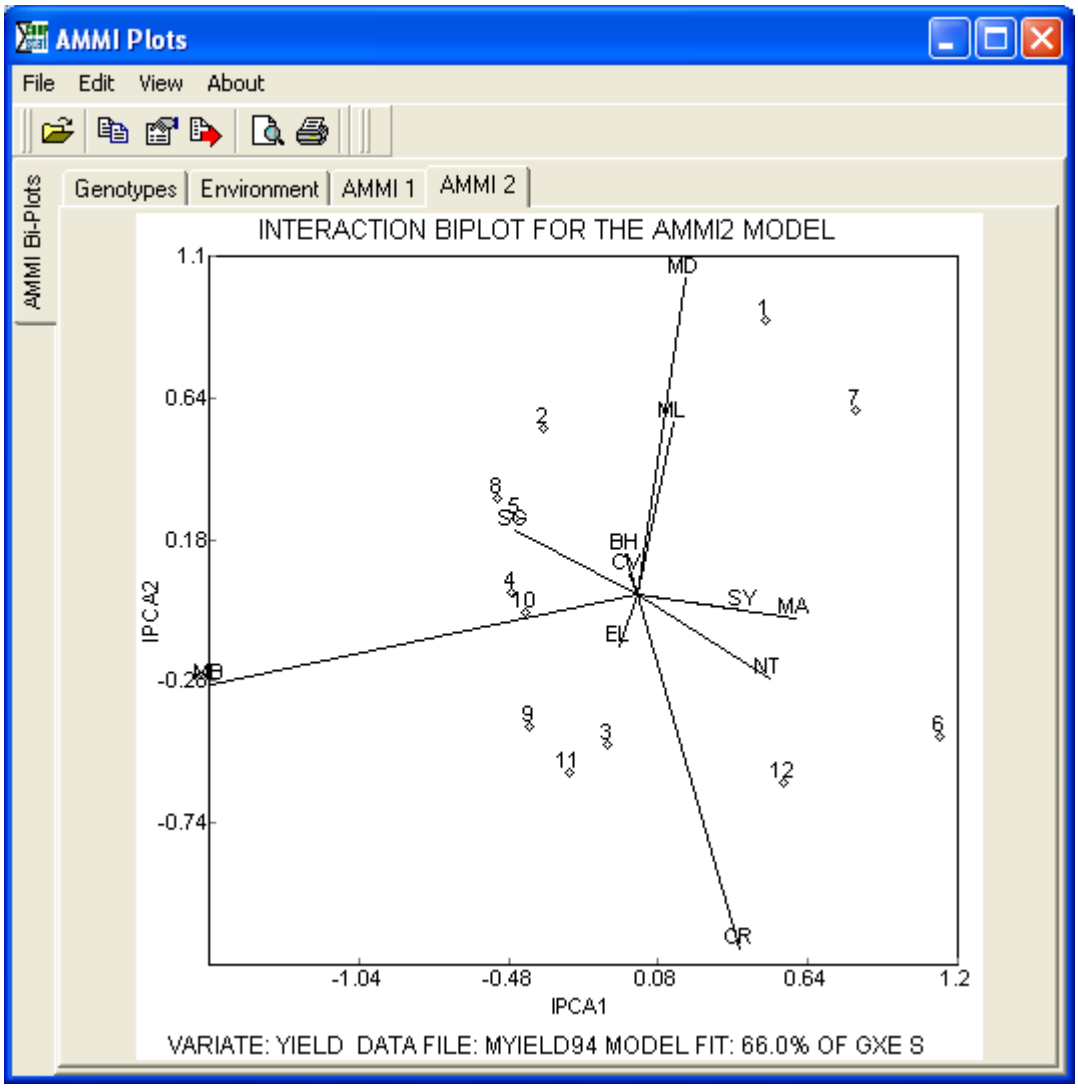
c. IPCA3 × IPCA2



3. AMMI1


AMMI Plots
File Edit View About
AMMI Bi-Plots
Genotypes Environment AMMI 1 AMMI 2
AMMI1 BIPLOT OF MAIN EFFECTS AND INTERACTIONS
1.2
0.64
0.08
-0.48
-1.04
IPCA1
0.3
0.66
1.02
1.38
1.74
2.1
MEANS
5
7
MA
12
NT
1
SY
CR
MD
ML
CV
BH
EL
3
11
2
10
4
9
SG
6
MB
VARIATE: YIELD DATA FILE: MYIELD94 MODEL FIT: 66.9% OF TABLE S

4. AMMI2



You can also produce the AMMI plots by running the PBGXEPLT.CMD, a command file outputted by CropStat when running the Genotype × Environment in the **Analysis-G×E plots.** AMMI also outputs a SYS file Pbstbplt.sys which graphs the result of the stability analysis.

PATTERN ANALYSIS

At the end of the tutorial, the user should be able to

- perform pattern analysis

I. Sample Problem

The mean yield data file *MYIELD94.SYS* produced by single-site analysis using the upland rice G×E data will be used.

- Open the data file *MYIELD94.SYS* from the *CROPSTAT7.2\TUTORIAL\ TUTORIAL DATASETS* folder.

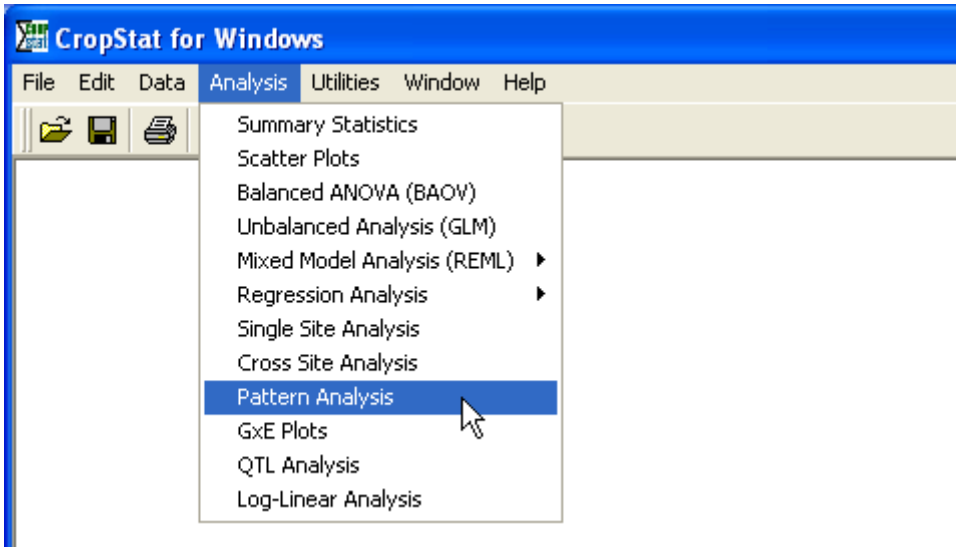
- Create a subfolder PATTERN ANALYSIS inside your working directory C:\MY CROPSTAT. Save *MYIELD94*.SYS inside this created folder by selecting **File ⇒ Save-as**.

II. Pattern Analysis

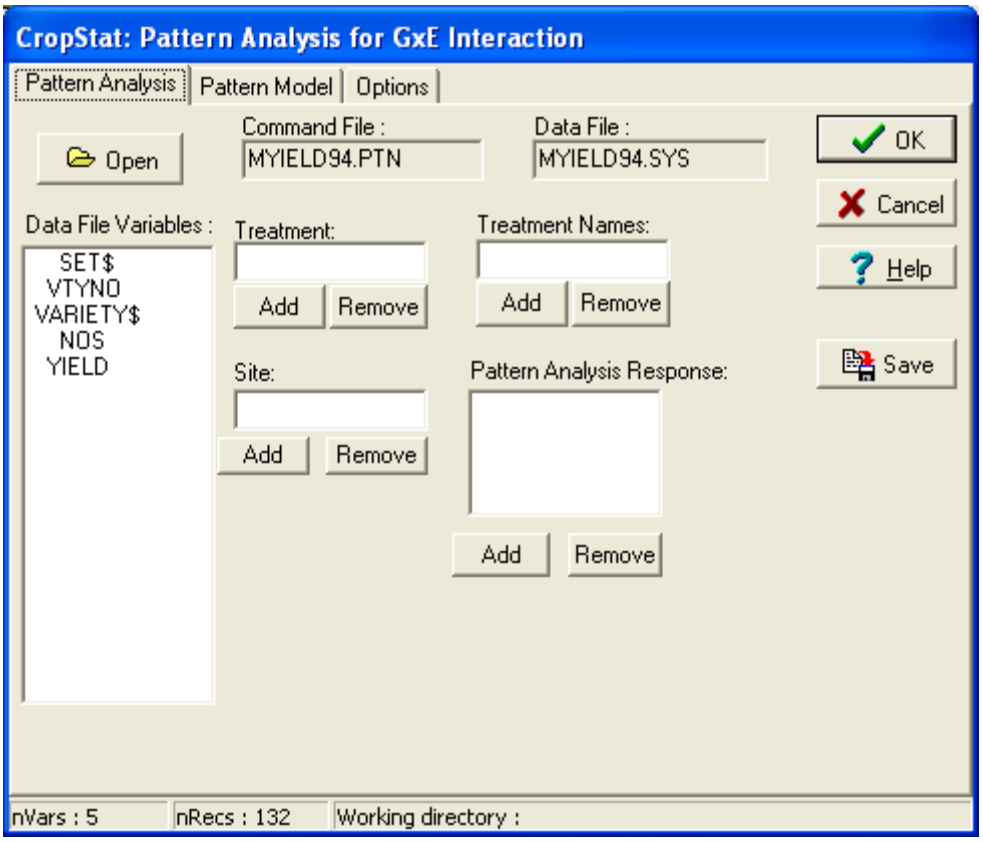
The input data file may be a raw data file with replicate observations for each G×E cell, or a file of means or adjusted means saved from a single-site analysis. In the former case, means are computed as the data are read into the G×E table. CropStat takes account of missing observations when forming means.

The G×E matrix must have less than 180 rows (usually genotypes) and less than 100 columns (usually environments), but in total, it must not have more than about 8000 entries (NROWS*(NCOLS+2)<8200). Also, if the G×E matrix has more than 4000 entries, the ordination cannot be performed.

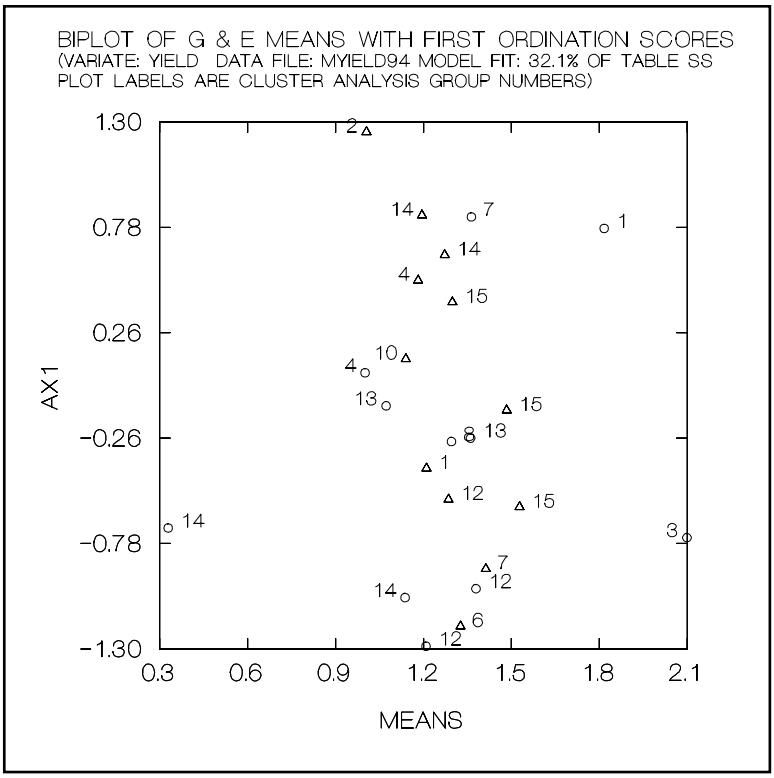
- Select **Analysis|Pattern Analysis** from the Main Window.



- In the **Open** Command file dialog box, click the **Look In** box and select drive *C:\MY CROPSTAT\PATTERN ANALYSIS*.
- In the **File name** box, enter *MYIELD94*. Click **Open**. Since *MYIELD94.PTN* does not exist, a message box will ask if you want to create this command file. Click **Yes**.
- Specify the data file to be used by entering *MYIELD94* as **File name** in the **Open Data File** dialog box. Click **Open**.
- The **Pattern Analysis** dialog box will appear.



- Specify the treatment variate (row factor) by selecting *VTYNO* from **Data File Variables** list, then click the **Add** button under the **Treatment** edit box. Treatment variate may be a character variate or a numeric variate but it must specify unique row levels for the G×E matrix. Clarity of graphical output is enhanced if the first two characters of the treatment factor are unique.



Clarity of graphical output is enhanced if the first two characters of the treatment factor are unique.

- Specify the variate containing row factor or treatment names in addition to the factor levels specified by treatment variate. Select VARIETY$ from **Data File Variables** list box, then click the **Add** button under **Treatment names** edit box. These names are printed adjacent to the levels specified by the treatment factor in the output tables.

GENOTYPE X ENVIRONMENT MEANS FOR VARIATE YIELD FILE MYIELD94 4/10/ 4 16:52
:PAGE 2

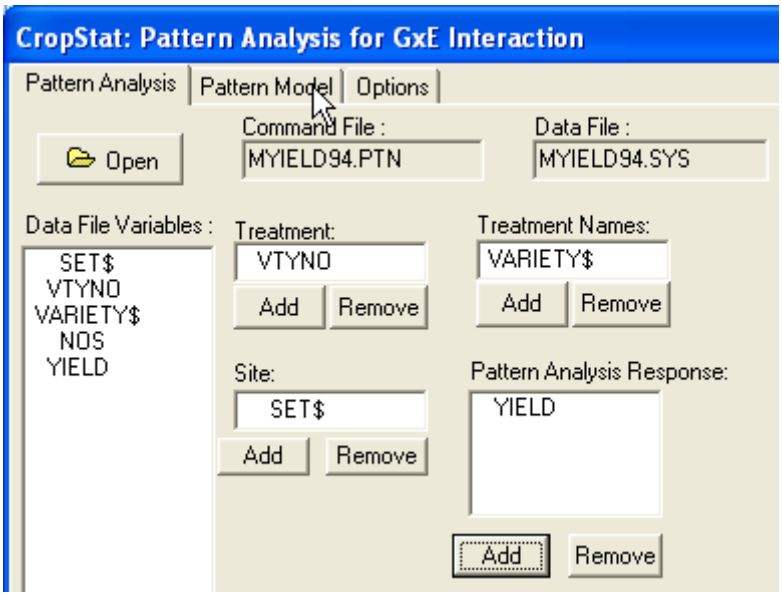
SECTION 1

VARIETY\SITE		BH	CR	CV	EL	MA	MB	MD
1	AZU	1.344	1.097	0.4838	0.7437	1.095	0.7377	2.270
2	BGORA	0.8480	0.8477	0.2168	1.520	0.7500	1.578	1.058
3	GUAR	1.296	2.999	0.3392	1.402	1.364	2.315	0.9897
4	IT146	0.6880	2.079	0.1535	0.9647	0.4965	2.782	1.501
5	OL5	1.524	1.422	0.1348	1.025	1.542	2.414	0.7732
6	OS6	1.388	2.995	0.3518	0.7097	2.494	0.1328	0.6000
7	UPL5	1.468	1.922	0.2748	0.9427	2.149	0.6325	1.922
8	VAND	1.752	2.079	0.2125	0.1192E-06	1.038	2.408	1.393
9	W181-18	1.576	2.281	0.3715	1.674	0.7015	2.252	0.3625
10	W56-125	1.080	1.508	0.5073	0.5350	1.820	2.662	0.5667
11	W56-50	1.568	2.687	0.5608	0.9503	1.739	2.886	0.6622
12	W96-1-1	1.012	3.272	0.3390	1.543	1.362	1.001	0.7757
SITE MEANS		1.295	2.099	0.3288	1.001	1.379	1.817	1.073

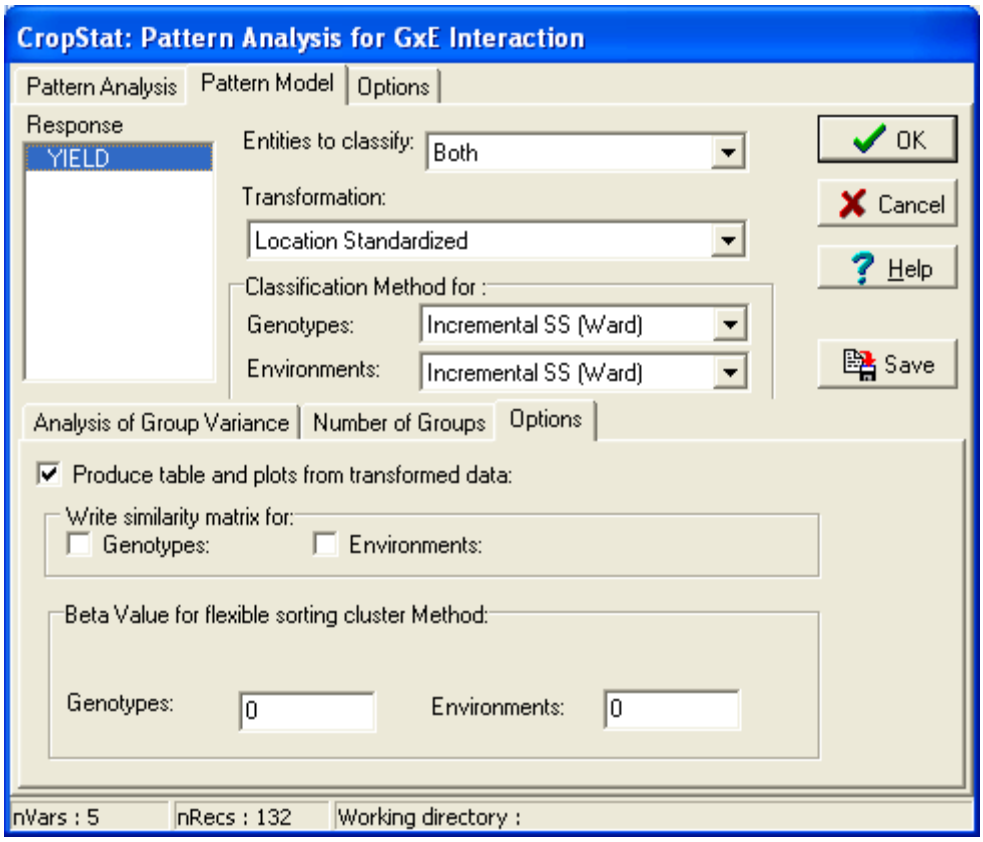
It is useful to include variate containing treatment names in addition to the treatment variate containing levels so that output is well annotated. These names are printed adjacent to the levels specified by the treatment factor in the output tables. If the treatment factor is a numeric variate, this allows inclusion of names as well as numbers for treatments. If the treatment factor is character variate, the names variate effectively extends the length of the treatment names by 12 characters although the first part of the name (contained in the row factor variate) must uniquely define all the row factor levels. Treatment names need not uniquely define all treatment levels.

- To specify the variate defining sites or environment (column factor), select *SET$* from **Data File Variables** list box, then click the **Add** button under **Site** edit box. The site variate may be a character or a numeric variate but it must specify unique column levels for the G×E matrix. Clarity of graphical output is enhanced if the first two characters of the site factor are unique and different from the codes for the treatment factor.

- Specify the variate to be analyzed. Select *YIELD* from **Data File Variables** list box, then click the **Add** button under **Pattern Analysis Response** edit box.



- To specify the model, click **Pattern Model** tab. In the **Pattern Model** page, click **Entities to classify** box and select *BOTH* from the drop-down list. Both genotype and environment will be classified in the analysis.



- To select the transformation, click the **Transformation** box from the list. *LOCATION STANDARDIZED* is the default transformation. This is the standardization technique that will be applied to the G×E matrix before pattern analysis.

The following transformation techniques are available in CropStat:

1. Raw data -- no standardization
2. Location centered -- column means subtracted for each entry
3. Location standardized -- subtraction of column means and division by column standard deviations (default)
4. Mean polish -- residuals after fitting row and column main effects by least squares (same ordination as AMMI analysis)

Column standardized mean polish-- mean polish residuals divided by standard deviations of the residuals in each column

- To specify the classification method for genotype and environment classification, click **Classification Method for Genotype** box and select *INCREMENTAL SS (WARD)* from the drop-down list. Click also the **Environment** box and select

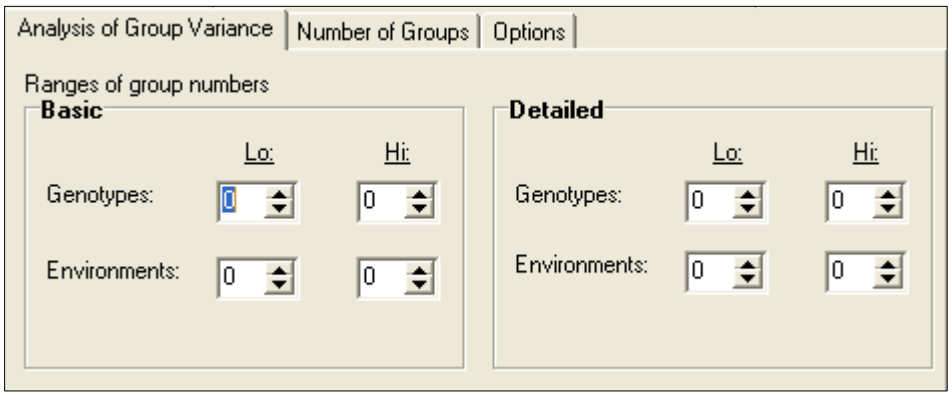
INCREMENTAL SS (WARD) from the drop-down list. (***Note***: *INCREMENTAL SS (WARD*) is the default classification method.)

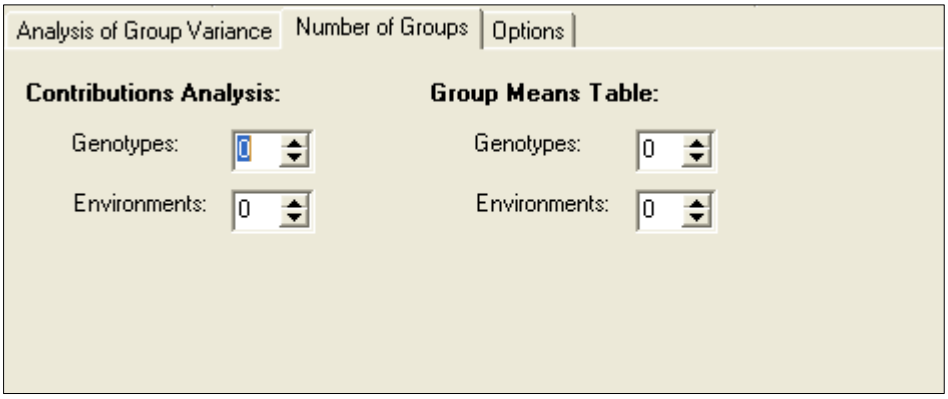
> Several methods of hierarchical, agglomerative cluster analysis are available in pattern analysis. The choice of method depends on the purpose of the analysis and the type of data being classified. Each choice corresponds to different measures of distance between entities being classified (genotypes or environment levels). The methods available are as follows:
>
> 1. None -- omit classification
> 2. Nearest neighbor
> 3. Farthest neighbor
> 4. Group average
> 5. Median
> 6. Centroid
> 7. Flexible sorting (WPGMA)
> 8. Flexible sorting (UPGMA)
> 9. Incremental SS (Ward) – default

- To produce tables and plots from transformed data, click the **Produce Table and Plots from Transformed Data** option. This instructs CropStat to generate tables of genotype group by environment group means from raw or transformed data. Fusion plots and response plots are also produced from the same data.

- To request for printing of similarity matrix for both genotypes and environments, click **Genotype** and **Environment** boxes in **Write Similarity Matrix** group. The similarity matrix between all pairs of genotypes or/and between all pairs of environments will be printed. Specify the beta value if flexible sorting methods (WPGMA & UPGMA) are chosen as classification method for genotype and/or environment. In our example there is no need to change the default beta values.

- Leave the remaining components of the model page unchanged.

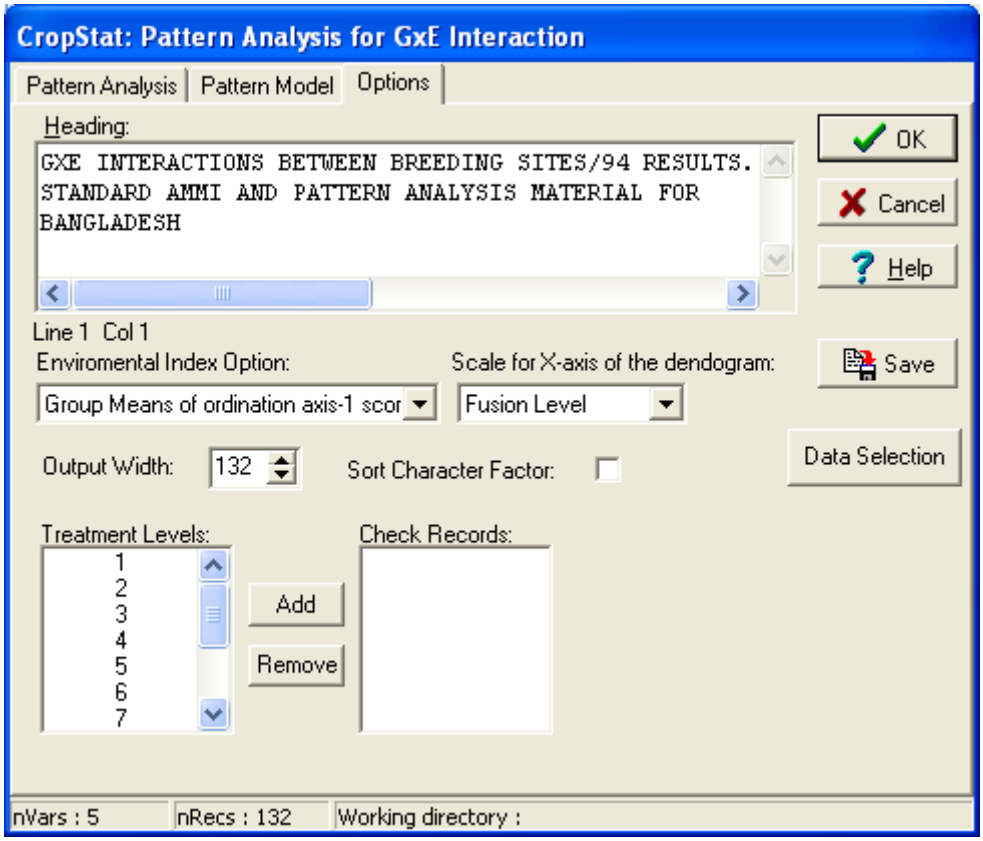




In the Analysis of Group Variance and Number of Groups tabs, the user can specify the final number of genotype and environment groups to be used in the presentation of group means of standardized and raw data. Zeroes are default values which are the minimum limits for group ANOVAs.

- Click **Options** tab. This will bring you to the **Options** page.
- In the heading box, type *GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS. STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.*
- Click **Environmental Index Option** box. Select *GROUP MEANS OF ORDINATION AXIS-1 SCORES* from the drop-down list. This is the index that will be used as x-axis in response and fusion plots.
- Click **Scale for X-axis of the dendogram** box to specify the values we wish to put on the x-axis of the dendogram. Select *FUSION LEVEL* from the drop-down list. Fusion number results in dendograms with equal steps which sometimes corrects for crowding at the low fusion levels. Square root fusion level should represent error differences at the lower fusion levels.

- The **Output width** box allows the user to specify the width of the output device in number of print characters. The default is 132 but any value between 80 and 225 is acceptable. Wider output lines (which maybe printed by using compressed printing) allow large G×E matrices to be printed in fewer sections. For this example, the default value 132 will be used.
- Leave the **Check Records** box blank since no special treatment check varieties is implemented in pattern analysis.
- Click **OK** to run the analysis.



III. Sample Output

The following output will appear in the Text Editor. This is saved in *MYIELD.OUT*.

1. Analysis specifications

```
GEBEI PATTERN ANALYSIS FOR VARIATE YIELD  FILE MYIELD94   4/10/ 4 16:52
------------------------------------------------------------ :PAGE   1
                        GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
                        STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.

THE GEBEI PROGRAM HAS BEEN ADAPTED FROM RESEARCH PROGRAMS OF DR. IAN DELACY, UNIVERSITY OF QUEENSLAND, AUSTRALIA

GXE MATRIX FOR YIELD:  12 GENOTYPES AND  11 ENVIRONMENTS

 12    VTYNO     CODES:
1           AZU          2           BGORA        3           GUAR
4           IT146        5           OL5          6           OS6
7           UPL5         8           VAND         9           W181-18
10          W56-125      11          W56-50       12          W96-1-1

 11     SET$    CODES:
BH           CR           CV           EL           MA
MB           MD           ML           NT           SG
SY

DATA TRANSFORMATION: Location standardised
DISTANCE MEASURE: SED
CLASSIFICATION METHOD FOR GENOTYPES: Incremental SS (Ward)       0.0000
WRITE GENOTYPE SIMILARITY MATRIX: Yes
CLASSIFICATION METHOD FOR ENVIRONMENTS: Incremental SS (Ward)       0.0000
WRITE ENVIRONMENT SIMILARITY MATRIX: Yes
ANALYSE GROUP VARIANCE FOR  4 TO 12 GENOTYPE AND  4 TO 11 ENVIRONMENT GROUPS
DETAILED GROUP ANOVA FOR  9 TO  9 GENOTYPE AND  8 TO  8 ENVIRONMENT GROUPS
CONTRIBUTIONS ANALYSIS DOWN TO  0 GENOTYPE AND  0 ENVIRONMENT GROUPS
GROUP MEANS TABLES FOR  9 GENOTYPE AND  8 ENVIRONMENT GROUPS
```

2. Table of means

```
 GENOTYPE X ENVIRONMENT MEANS FOR VARIATE YIELD  FILE MYIELD94   4/10/ 4 16:52
------------------------------------------------------------ :PAGE   2
                        GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
                        STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.


SECTION  1

VARIETY\SITE             |BH          |CR          |CV          |EL          |MA          |MB          |MD          |ML          |
-------------------------|------------|------------|------------|------------|------------|------------|------------|------------|
1           AZU          | 1.344      | 1.097      |0.4838      |0.7437      | 1.095      |0.7377      | 2.270      | 1.421      |
2           BGORA        |0.8480      |0.8477      |0.2168      | 1.520      |0.7500      | 1.578      | 1.058      | 1.244      |
3           GUAR         | 1.296      | 2.999      |0.3392      | 1.402      | 1.364      | 2.315      |0.9897      | 1.270      |
4           IT146        |0.6880      | 2.079      |0.1535      |0.9647      |0.4965      | 2.782      | 1.501      | 1.366      |
5           OL5          | 1.524      | 1.422      |0.1348      | 1.025      | 1.542      | 2.414      |0.7732      | 1.665      |
                         |            |            |            |            |            |            |            |            |
6           OS6          | 1.388      | 2.995      |0.3518      |0.7097      | 2.494      |0.1328      |0.6000      | 1.250      |
7           UPL5         | 1.468      | 1.922      |0.2748      |0.9427      | 2.149      |0.6325      | 1.922      | 2.324      |
8           VAND         | 1.752      | 2.079      |0.2125      |0.1192E-06  | 1.038      | 2.408      | 1.393      | 1.470      |
9           W181-18      | 1.576      | 2.281      |0.3715      | 1.674      |0.7015      | 2.252      |0.3625      |0.9295      |
10          W56-125      | 1.080      | 1.508      |0.5073      |0.5350      | 1.820      | 2.662      |0.5667      | 1.096      |
                         |            |            |            |            |            |            |            |            |
11          W56-50       | 1.568      | 2.687      |0.5608      |0.9503      | 1.739      | 2.886      |0.6622      | 1.118      |
12          W96-1-1      | 1.012      | 3.272      |0.3390      | 1.543      | 1.362      | 1.001      |0.7757      | 1.174      |
-------------------------|------------|------------|------------|------------|------------|------------|------------|------------|
SITE MEANS               | 1.295      | 2.099      |0.3288      | 1.001      | 1.379      | 1.817      | 1.073      | 1.361      |

SECTION  2

VARIETY\SITE             |NT          |SG          |SY          |TRT MEANS   |
-------------------------|------------|------------|------------|------------|
1           AZU          | 1.565      | 1.140      | 1.418      | 1.211      |
2           BGORA        |0.6473      | 1.898      |0.4563      | 1.006      |
3           GUAR         | 1.501      | 1.590      | 1.260      | 1.484      |
4           IT146        | 1.234      |0.7925      |0.9420      | 1.182      |
5           OL5          |0.5510      | 1.860      | 1.087      | 1.273      |
                         |            |            |            |            |
6           OS6          | 1.812      | 1.188      | 1.670      | 1.326      |
7           UPL5         | 1.326      |0.9150      | 1.667      | 1.413      |
8           VAND         |    0.0000  | 1.883      |0.9123      | 1.195      |
9           W181-18      | 1.236      | 1.825      | 1.076      | 1.299      |
10          W56-125      |0.7120      | 1.027      | 1.024      | 1.140      |
                         |            |            |            |            |
11          W56-50       | 1.791      | 1.347      | 1.495      | 1.528      |
12          W96-1-1      | 1.271      |0.8950      | 1.497      | 1.286      |
-------------------------|------------|------------|------------|------------
SITE MEANS               | 1.137      | 1.363      | 1.209      | 1.278      |
```

3. Matrix of transformed means

```
TRANSFORMED GXE MATRIX FOR VARIATE YIELD  FILE MYIELD94   4/10/ 4 16:52
------------------------------------------------------------ :PAGE   3
                        GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
                        STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.


SECTION  1

VARIETY\SITE            |BH          |CR          |CV          |EL          |MA          |MB          |MD          |ML          |
------------------------|------------|------------|------------|------------|------------|------------|------------|------------|
1           AZU         |0.1505      |-1.277      | 1.123      |-.5336      |-.4707      |-1.128      | 2.043      |0.1691      |
2           BGORA       |-1.383      |-1.594      |-.8117      | 1.078      |-1.041      |-.2499      |-.2574E-01  |-.3251      |
3           GUAR        |0.2061E-02  | 1.146      |0.7576E-01  |0.8323      |-.2478E-01  |0.5211      |-.1418      |-.2514      |
4           IT146       |-1.878      |-.2520E-01  |-1.270      |-.7504E-01  |-1.461      | 1.009      |0.7301      |0.1386E-01  |
5           OL5         |0.7071      |-.8627      |-1.406      |0.4945E-01  |0.2697      |0.6249      |-.5112      |0.8478      |
                        |            |            |            |            |            |            |            |            |
6           OS6         |0.2866      | 1.142      |0.1663      |-.6041      | 1.844      |-1.761      |-.8068      |-.3091      |
7           UPL5        |0.5339      |-.2261      |-.3915      |-.1207      | 1.274      |-1.238      | 1.449      | 2.681      |
8           VAND        | 1.412      |-.2520E-01  |-.8424      |-2.077      |-.5646      |0.6181      |0.5463      |0.3048      |
9           W181-18     |0.8679      |0.2317      |0.3094      | 1.397      |-1.121      |0.4555      |-1.212      |-1.201      |
10          W56-125     |-.6659      |-.7527      | 1.293      |-.9665      |0.7293      |0.8840      |-.8635      |-.7366      |
                        |            |            |            |            |            |            |            |            |
11          W56-50      |0.8432      |0.7490      | 1.680      |-.1048      |0.5949      | 1.118      |-.7005      |-.6746      |
12          W96-1-1     |-.8762      | 1.494      |0.7395E-01  | 1.125      |-.2933E-01  |-.8531      |-.5069      |-.5187      |
------------------------|------------|------------|------------|------------|------------|------------|------------|------------|

SECTION  2

VARIETY\SITE            |NT          |SG          |SY          |
------------------------|------------|------------|------------|
1           AZU         |0.7783      |-.5230      |0.5815      |
2           BGORA       |-.8906      | 1.251      |-2.090      |
3           GUAR        |0.6607      |0.5306      |0.1417      |
4           IT146       |0.1765      |-1.337      |-.7408      |
5           OL5         |-1.066      | 1.163      |-.3380      |
                        |            |            |            |
6           OS6         | 1.227      |-.4118      | 1.281      |
7           UPL5        |0.3438      |-1.050      | 1.272      |
8           VAND        |-2.067      | 1.217      |-.8232      |
9           W181-18     |0.1790      | 1.081      |-.3695      |
10          W56-125     |-.7730      |-.7864      |-.5121      |
                        |            |            |            |
11          W56-50      | 1.188      |-.3717E-01  |0.7954      |
12          W96-1-1     |0.2432      |-1.097      |0.8019      |
------------------------|------------|------------|------------
```

4. Proximity matrix and fusion table for genotypes

```
CLUSTER ANALYSIS FOR VARIATE YIELD  FILE MYIELD94   4/10/ 4 16:52
------------------------------------------------------------ :PAGE   4
                        GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
                        STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.


  SECTION  1  DISSIMILARITY MATRIX FOR GENOTYPES
           1             2             3             4             5             6
1
2           0.2498162E+01
3           0.1640048E+01  0.1802566E+01
4           0.1971440E+01  0.1486083E+01  0.1379180E+01
5           0.2264308E+01  0.1228961E+01  0.1126540E+01  0.1886261E+01
6           0.1963476E+01  0.3989616E+01  0.1254960E+01  0.3227008E+01  0.2458570E+01
7           0.1306940E+01  0.3893648E+01  0.2098171E+01  0.2814254E+01  0.2059137E+01  0.1662592E+01
8           0.2528595E+01  0.2265742E+01  0.2053825E+01  0.2513581E+01  0.8532936E+00  0.3401126E+01
9           0.2401554E+01  0.1506533E+01  0.5462394E+00  0.2174337E+01  0.1289922E+01  0.2354961E+01
10          0.1780632E+01  0.1979573E+01  0.1312945E+01  0.1684661E+01  0.1549019E+01  0.1970366E+01
11          0.1813547E+01  0.3408041E+01  0.5980002E+00  0.2594474E+01  0.2061062E+01  0.1216101E+01
12          0.1862951E+01  0.2490697E+01  0.5753903E+00  0.1484679E+01  0.2151486E+01  0.9643190E+00

  SECTION  2  DISSIMILARITY MATRIX FOR GENOTYPES
           7             8             9            10            11            12
7
8           0.3042713E+01
9           0.3739393E+01  0.2247032E+01
10          0.2872392E+01  0.1930903E+01  0.1657462E+01
11          0.2654386E+01  0.2734167E+01  0.1086459E+01  0.1058775E+01
12          0.2078122E+01  0.3291159E+01  0.1340913E+01  0.1597767E+01  0.1268786E+01

Classification of Genotypes

GpI + GpJ = GpIJ at Fusion level  NO ELEMENTS  NAMES OF FUSING ELEMENTS

  3 +   9 =  13      .54624        2   3           GUAR          + 9           W181-18
  5 +   8 =  14      .85329        2   5           OL5           + 8           VAND
 13 +  11 =  15      .94089        3   GGp 13                    + 11          W56-50
  6 +  12 =  16      .96432        2   6           OS6           + 12          W96-1-1
  1 +   7 =  17     1.3069         2   1           AZU           + 7           UPL5
  2 +   4 =  18     1.4861         2   2           BGORA         + 4           IT146
 15 +  10 =  19     1.6428         4   GGp 15                    + 10          W56-125
 19 +  16 =  20     2.1736         6   GGp 19                    + GGp 16
 18 +  14 =  21     2.7776         4   GGp 18                    + GGp 14
 17 +  20 =  22     4.0947         8   GGp 17                    + GGp 20
 22 +  21 =  23     5.2136        12   GGp 22                    + GGp 21
```

5. Proximity matrix and fusion table for environments

```
  SECTION  1  DISSIMILARITY MATRIX FOR ENVIRONMENTS
            BH             CR             CV             EL             MA             MB
BH
CR            0.1558574E+01
CV            0.1484759E+01  0.1506712E+01
EL            0.2463628E+01  0.1513462E+01  0.1933183E+01
MA            0.1202815E+01  0.1256456E+01  0.1222160E+01  0.2371763E+01
MB            0.1845556E+01  0.1978637E+01  0.1902356E+01  0.1940587E+01  0.2646384E+01
MD            0.1949919E+01  0.2580480E+01  0.2128765E+01  0.2415456E+01  0.2126291E+01  0.2442856E+01
ML            0.1474122E+01  0.2281367E+01  0.2604685E+01  0.2309487E+01  0.1238110E+01  0.2465593E+01
NT            0.1943328E+01  0.9416354E+00  0.8662494E+00  0.1205791E+01  0.1199660E+01  0.2523885E+01
SG            0.1007389E+01  0.2295856E+01  0.2334562E+01  0.1647418E+01  0.2393805E+01  0.1266714E+01
SY            0.1166117E+01  0.7747416E+00  0.1032728E+01  0.1971058E+01  0.5286838E+00  0.2795747E+01

  SECTION  2  DISSIMILARITY MATRIX FOR ENVIRONMENTS
            MD             ML             NT             SG             SY
MD
ML            0.6511404E+00
NT            0.1822207E+01  0.2051208E+01
SG            0.2386561E+01  0.2183672E+01  0.2755997E+01
SY            0.1643255E+01  0.1320395E+01  0.5305052E+00  0.2835927E+01

Classification of Environments

GpI + GpJ = GpIJ at Fusion level  NO ELEMENTS  NAMES OF FUSING ELEMENTS

  5 +  11 =  12      .52868        2   MA                      + SY
  7 +   8 =  13      .65114        2   MD                      + ML
  3 +   9 =  14      .86625        2   CV                      + NT
  1 +  10 =  15     1.0074         2   BH                      + SG
  2 +  12 =  16     1.1779         3   CR                      + LGp 12
 16 +  14 =  17     1.3710         5   LGp 16                  + LGp 14
 15 +   6 =  18     1.7391         3   LGp 15                  + MB
 18 +   4 =  19     2.3392         4   LGp 18                  + EL
 19 +  13 =  20     3.7466         6   LGp 19                  + LGp 13
 20 +  17 =  21     4.5583        11   LGp 20                  + LGp 17
```

6. Basic group ANOVA

```
COMBINED ANOVA FOR RAW DATA VALUES OF YIELD  FILE MYIELD94   4/10/ 4 16:52
------------------------------------------------------------------------ :PAGE   5
                        GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
                        STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.          1


SOURCE       DF  SUM OF SQUARES    MEAN SQUARES  % OF TOT.

ENVS.        10       24.404085       2.440408   39.9% OF TOTSS
GENS.        11        2.637222       0.239747    4.3% OF TOTSS
GEN X ENV   110       34.088966       0.309900   55.8% OF TOTSS
HET          11        3.857300       0.350664          11.3% OF GESS
DEV          99       30.231667       0.305370          88.7% OF GESS
TOTAL       131       61.130272
TOT GSS               36.726189                  60.1% OF TOTSS
REORDER SS            12.889073                         35.1% OF TGSS
                                                        37.8% OF GESS
NON REO SS            23.837103                         64.9% OF TGSS
TOT ESS               58.493050                 95.7% OF TOTSS
BASIC GROUP ANOVA FOR TRANSFORMED VALUES OF VARIATE YIELD  FILE MYIELD94   4/10/ 4 16:52
------------------------------------------------------------------------ :PAGE   6
                        GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
                        STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.


BETWEEN GROUP SS AS A PERCENTAGE OF TOTAL SS (BSS%), WITHIN GROUP DF (WDF) AND MS (WMS)
FOR A RANGE OF DATA ARRAYS REDUCED FROM THE ORIGINAL 12 GENOTYPE BY 11
ENVIRONMENT ARRAY ( 132 CELLS) BY CLUSTER ANALYSIS (TOTAL SS =  121.00    )

NO. OF  |     NO. OF ENVIRONMENT GROUPS
GEN GPS |     4          5          6          7          8          9          10         11
        |---------|----------|----------|----------|----------|----------|----------|----------|
    BSS%|   41.85 |    44.80 |    47.21 |    50.07 |    53.32 |    53.95 |    54.53 |    54.94 |
  4 WDF |     116 |      112 |      108 |      104 |      100 |       96 |       92 |       88 |
    WMS | 0.6065  | 0.5963   | 0.5914   | 0.5809   | 0.5648   | 0.5804   | 0.5980   | 0.6196   |
        |---------|----------|----------|----------|----------|----------|----------|----------|
    BSS%|   48.04 |    52.23 |    56.24 |    59.11 |    62.44 |    63.79 |    64.39 |    64.82 |
  5 WDF |     112 |      107 |      102 |       97 |       92 |       87 |       82 |       77 |
    WMS | 0.5614  | 0.5403   | 0.5191   | 0.5101   | 0.4940   | 0.5036   | 0.5254   | 0.5529   |
        |---------|----------|----------|----------|----------|----------|----------|----------|
    BSS%|   51.49 |    56.55 |    60.57 |    64.45 |    67.79 |    70.44 |    71.05 |    72.28 |
  6 WDF |     108 |      102 |       96 |       90 |       84 |       78 |       72 |       66 |
    WMS | 0.5435  | 0.5154   | 0.4970   | 0.4779   | 0.4640   | 0.4585   | 0.4865   | 0.5081   |
        |---------|----------|----------|----------|----------|----------|----------|----------|
    BSS%|   53.54 |    60.77 |    64.92 |    69.14 |    73.38 |    76.51 |    77.16 |    79.04 |
  7 WDF |     104 |       97 |       90 |       83 |       76 |       69 |       62 |       55 |
    WMS | 0.5406  | 0.4894   | 0.4716   | 0.4499   | 0.4239   | 0.4119   | 0.4457   | 0.4612   |
        |---------|----------|----------|----------|----------|----------|----------|----------|
    BSS%|   54.57 |    61.80 |    68.22 |    72.45 |    76.85 |    80.23 |    82.87 |    84.98 |
  8 WDF |     100 |       92 |       84 |       76 |       68 |       60 |       52 |       44 |
    WMS | 0.5497  | 0.5024   | 0.4578   | 0.4387   | 0.4119   | 0.3987   | 0.3985   | 0.4131   |
        |---------|----------|----------|----------|----------|----------|----------|----------|
    BSS%|   56.71 |    64.87 |    71.29 |    76.16 |    80.62 |    84.16 |    86.85 |    89.36 |
  9 WDF |      96 |       87 |       78 |       69 |       60 |       51 |       42 |       33 |
    WMS | 0.5456  | 0.4886   | 0.4453   | 0.4180   | 0.3909   | 0.3758   | 0.3787   | 0.3901   |
        |---------|----------|----------|----------|----------|----------|----------|----------|
    BSS%|   59.67 |    68.08 |    74.63 |    79.85 |    84.73 |    88.42 |    91.11 |    93.64 |
 10 WDF |      92 |       82 |       72 |       62 |       52 |       42 |       32 |       22 |
    WMS | 0.5305  | 0.4710   | 0.4264   | 0.3933   | 0.3553   | 0.3337   | 0.3360   | 0.3499   |
        |---------|----------|----------|----------|----------|----------|----------|----------|
    BSS%|   61.74 |    70.20 |    76.74 |    82.58 |    87.55 |    91.74 |    94.97 |    97.52 |
 11 WDF |      88 |       77 |       66 |       55 |       44 |       33 |       22 |       11 |
    WMS | 0.5261  | 0.4684   | 0.4265   | 0.3833   | 0.3424   | 0.3028   | 0.2769   | 0.2731   |
        |---------|----------|----------|----------|----------|----------|----------|----------|
    BSS%|   63.60 |    72.22 |    79.02 |    84.86 |    89.85 |    94.15 |    97.38 |   100.00 |
 12 WDF |      84 |       72 |       60 |       48 |       36 |       24 |       12 |        0 |
    WMS | 0.5244  | 0.4669   | 0.4231   | 0.3817   | 0.3410   | 0.2950   | 0.2643   |          |
        |---------|----------|----------|----------|----------|----------|----------|----------|
```

(continuation of Basic Group ANOVA)

PERCENTAGES OF ENVIRONMENT S.S. RETAINED AMONG ENVIRONMENT GROUPS (%ESS),
GENOTYPE S.S. RETAINED AMONG GENOTYPE GROUPS (%GSS), AND INTERACTION S.S.
RETAINED IN AMONG X AMONG GROUPS (%GXESS) FOR A RANGE OF REDUCED DATA ARRAYS
PRODUCED BY CLUSTER ANALYSIS

%ESS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

%GXESS

%GSS	NO. ENVIRONMENT GROUPS 4	5	6	7	8	9	10	11
62.63	4	39.33	42.64	45.34	48.55	52.19	52.90	53.55	54.00
63.02	5	46.22	50.92	55.42	58.63	62.37	63.88	64.56	65.03
83.95	6	47.55	53.23	57.74	62.09	65.83	68.80	69.49	70.87
84.47	NO. 7	49.79	57.89	62.55	67.28	72.03	75.55	76.28	78.38
89.01	GEN 8	50.40	58.50	65.70	70.44	75.38	79.17	82.13	84.49
90.68	GPS 9	52.59	61.74	68.94	74.40	79.40	83.37	86.39	89.20
96.03	10	55.26	64.69	72.03	77.88	83.36	87.49	90.52	93.35
97.13	11	57.44	66.93	74.27	80.81	86.39	91.09	94.70	97.56
100.00	12	59.18	68.85	76.47	83.02	88.62	93.44	97.06	100.00

7. Detailed group ANOVA

DETAILED ANOVA OF TRANSFORMED DATA FOR 9 GENOTYPE GROUPS (VTYNO) AND 8 ENVIRONMENT GROUPS (SET$)

SOURCE	SS	DF	MS	SS	DF	MS	SS	DF	MS
GEN	0.1309E+02	11	0.1190E+01	VARIABILITY BEFORE CLUSTERING					
AMG GEN GPS	0.1187E+02	8	0.1483E+01	VARIABILITY RETAINED BY CLUSTERING					
W/I GEN GPS	0.1219E+01	3	0.4064E+00	VARIABILITY LOST WITHIN CLUSTERS					
W/I GEN GP I	0.1075E+01	2	0.5376E+00	0.1440E+00	1	0.1440E+00			
GEN x ENV	0.1079E+03	110	0.9810E+00	VARIABILITY BEFORE CLUSTERING					
AMG GEN GPSxAMG ENV GPS	0.8568E+02	56	0.1530E+01	VARIABILITY RETAINED BY CLUSTERING					
AMG GEN GPSxW/I ENV GPS	0.1058E+02	24	0.4408E+00	VARIABILITY LOST WITHIN CLUSTERS					
AMG GEN GPSxW/I ENV GPJ	0.4285E+01	8	0.5356E+00	0.3261E+01	8	0.4076E+00	0.3034E+01	8	0.3792E+00
AMG ENV GPSxW/I GEN GPS	0.9956E+01	21	0.4741E+00	VARIABILITY LOST WITHIN CLUSTERS					
AMG ENV GPSxW/I GEN GPI	0.6690E+01	14	0.4779E+00	0.3266E+01	7	0.4666E+00			
W/I GEN x W/I ENV GPS	0.1697E+01	9	0.1885E+00	VARIABILITY LOST WITHIN CLUSTERS					
W/I G GP 15 xW/I E GP J	0.3003E+00	2	0.1502E+00	0.5527E-02	2	0.2764E-02	0.1080E+00	2	0.5399E-01
W/I G GP 14 xW/I E GP J	0.6122E+00	1	0.6122E+00	0.6403E+00	1	0.6403E+00	0.3047E-01	1	0.3047E-01
TOTAL(TRANSFORMED DATA)	0.1210E+03	131	0.9237E+00						

ORDER OF GEN GROUPS (I) IN OUTPUT ABOVE: 15 14 12 10 7 6 4 2 1
ORDER OF ENV GROUPS (J) IN OUTPUT ABOVE: 14 13 12 10 6 4 2 1
*, ** INDICATE COMPONENT MSs SIGNIFICANTLY LARGER THAN CORRESPONDING POOLED MS AT THE 5% AND 1% LEVELS

% TOTAL SS IN AMONG,AMONG*AMONG GPS	80.62
% GEN SS IN AMONG GEN GROUPS	90.68
% G*E IN AMONG GEN GPS * AMONG ENV GPS	79.40

(continuation of detailed group ANOVA)

```
PARTITION OF VARIATION FOR YIELD
INTO AMONG- AND WITHIN- GROUP COMPONENTS FOR (1) VTYNO      (2) SET$
 AND (3) TWO-WAY VTYNO      BY SET$        CLASSIFICATION

                                                                        PARTITION OF
                                                                        SS AMG & W/I
***********SOURCE***********    *DF*      **SSQ**      **MSQ**    GROUPS (%)

-------------------------------------------------------------------------------

VTYNO      (V)                      11        13.09         1.19        -
 AMONG V GROUPS                      8        11.87         1.48      90.68
 WITHIN V GROUPS                     3         1.22         0.41       9.32
(1)  SET$       (S)                 10         0.00         0.00        -
VTYNO      * SET$                  110       107.91         0.98        -
 AMONG V GROUPS * S                 80        96.26         1.20      89.20
 REMAINDER                          30        11.65         0.39      10.80


VTYNO      (V)                      11        13.09         1.19        -
 SET$        (S)                    10         0.00         0.00        -
 AMONG S GROUPS                      7         0.00         0.00      95.25
(2)  WITHIN S GROUPS                 3         0.00         0.00       4.75
VTYNO      * SET$                  110       107.91         0.98        -
 V * AMONG S GROUPS                 77        95.64         1.24      88.62
 REMAINDER                          33        12.28         0.37      11.38


VTYNO      (V)                      11        13.09         1.19        -
 AMONG V GROUPS                      8        11.87         1.48      90.68
 WITHIN V GROUPS                     3         1.22         0.41       9.32
 SET$        (S)                    10         0.00         0.00        -
 AMONG S GROUPS                      7         0.00         0.00      95.25
(3)  WITHIN S GROUPS                 3         0.00         0.00       4.75
VTYNO      * SET$                  110       107.91         0.98        -
 AMONG V GRPS * AMONG S GRPS        56        85.68         1.53      79.40
 REMAINDER                          54        22.23         0.41      20.60
 AMONG V GRPS * WITHIN S GRPS       24        10.58         0.44       9.80
 WITHIN V GRPS * AMONG S GRPS       21         9.96         0.47       9.23
 WITHIN V GRPS * WITHIN S GRPS       9         1.70         0.19       1.57

-------------------------------------------------------------------------------

TOTAL SUM OF SQUARES                        121.00
TOTAL SUM OF SQUARES AMONG GROUPS            97.55
PERCENTAGE OF TOTAL SUM OF SQUARES RETAINED AMONG GROUPS              80.62

          PARTITION OF VARIATION IN SEVERAL DATA ARRAYS AFTER REDUCTION OF THE
          ORIGINAL  12 VTYNO      X  11 SET$         ARRAY ( 132 CELLS) BY
          CLUSTER ANALYSIS

                                              PORTION OF S.S. RETAINED AMONG GROUPS (%)
                                              ----------------------------------------

N0. VTY    N0.  SE    ARRAY         %       TOTAL  VTYNO     SET$          V X T
 GRPS.      GRPS.      SIZE     REDUCTION    S.S.      S.S.       S.S.       S.S.

   9          8         72        45.45     80.62     90.68      0.00      79.40

-------------------------------------------------------------------------------
```

8. Ordination analysis

```
ORDINATION ANALYSIS OF TRANSFORMED GXE MATRIX  FILE MYIELD94   4/10/ 4 16:52
------------------------------------------------------------- :PAGE   7
                      GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
                      STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.

SINGULAR VALUES OF TRANSFORMED GXE MATRIX (CONDITION=  0)
6.2285    5.1433    4.4085    3.4863    3.0238    2.6726    2.1749    1.3769    1.0137    .46688
.73792E-01

SCORES FOR FIRST FOUR ORDINATION AXES FOR GENOTYPE

1  1           AZU         -0.409547E+00-0.705356E+00-0.408650E+00-0.710545E+00
2  2           BGORA        0.125028E+01 0.464686E-03-0.632042E+00 0.313403E+00
3  3           GUAR        -0.123719E+00 0.523727E+00-0.673063E-02 0.372324E+00
4  4           IT146        0.518712E+00-0.130193E+00-0.116774E+01-0.321888E+00
5  5           OL5          0.642778E+00-0.387686E+00 0.590299E+00 0.613573E+00
6  6           OS6         -0.118927E+01 0.160810E+00 0.289844E+00 0.336052E+00
7  7           UPL5        -0.905027E+00-0.132677E+01-0.116595E+00 0.449869E+00
8  8           VAND         0.838857E+00-0.702505E+00 0.110945E+01-0.198939E+00
9  9           W181-18      0.410367E+00 0.982292E+00 0.251954E+00 0.408401E+00
10 10          W56-125      0.130100E+00 0.274233E+00 0.180312E+00-0.113816E+01
11 11          W56-50      -0.600708E+00 0.767450E+00 0.547153E+00-0.530933E+00
12 12          W96-1-1     -0.562822E+00 0.543533E+00-0.637254E+00 0.406838E+00

SCORES FOR FIRST FOUR ORDINATION AXES FOR ENVIRONMENTS

BH             BH          -0.272028E+00-0.184497E+00 0.139028E+01 0.247390E+00
CR             CR          -0.746269E+00 0.728574E+00 0.170469E+00 0.505290E+00
CV             CV          -0.698756E+00 0.599622E+00 0.238038E+00-0.104332E+01
EL             EL           0.672760E-01 0.770525E+00-0.774261E+00 0.981226E+00
MA             MA          -0.998056E+00-0.205041E+00 0.605187E+00 0.929833E-01
MB             MB           0.779411E+00 0.525896E+00 0.352672E+00-0.589932E+00
MD             MD          -0.960675E-01-0.118379E+01-0.519460E+00-0.325744E+00
ML             ML          -0.254757E+00-0.128000E+01 0.191592E-01 0.498115E+00
NT             NT          -0.104189E+01 0.464368E+00-0.492668E+00 0.293082E-01
SG             SG           0.836051E+00 0.206871E+00 0.868494E+00 0.620184E+00
SY             SY          -0.128163E+01-0.914453E-01 0.180604E+00 0.148577E+00

ANOVA FOR THE PCA ORDINATION OF THE TRANSFORMED DATA

-------------------------------------------------------------------
SOURCE                 D.F.      S.S.        % TABLE SS
-------------------------------------------------------------------
  PCA COMPONENT 1        20    38.7943        32.1
  PCA COMPONENT 2        18    26.4532        21.9
  PCA COMPONENT 3        16    19.4352        16.1
  PCA COMPONENT 4        14    12.1542        10.0
 RESIDUAL AFTER 2        94    55.7526
-------------------------------------------------------------------
TOTAL MATRIX            132    121.000
```

9. Residual matrix

RESIDUALS FROM THE PCA-2 MODEL

(ENTRIES ARE SIZE OF RESIDUAL IN UNITS OF ROOT (RESIDUAL GXE MS)

		BH	CR	CV	EL	MA	MB	MD	ML	NT	SG	SY
1	AZU	0	-1	1	0	-1	0	1	-1	0	0	0
2	BGORA	-1	0	0	1	0	-1	0	0	0	0	0
3	GUAR	0	0	0	0	0	0	0	0	0	0	0
4	IT146	-2	0	-1	0	-1	0	0	0	1	-2	0
5	OL5	1	0	0	0	1	0	-1	0	0	0	0
6	OS6	0	0	0	0	0	-1	0	0	0	0	0
7	UPL5	0	0	0	1	0	0	0	0	0	0	0
8	VAND	1	1	0	-2	0	0	0	0	-1	0	0
9	W181-18	1	0	0	0	0	0	0	0	0	0	0
10	W56-125	0	-1	1	-1	1	0	0	0	0	-1	0
11	W56-50	1	0	1	0	0	1	0	0	0	0	0
12	W96-1-1	-1	0	0	0	0	0	0	0	0	0	0

BOX PLOT OF 132 STANDERSIZED RESIDUALS FROM LPLT= -2.287 TO ULPT= 1.962

NO.<LPLT NO.>UPLT

0 ----------I + I---------- 0

10. Group membership and group means

GROUP MEMBERSHIP FOR Location standardised VALUES OF YIELD FILE MYIELD94 4/10/ 4 16:52
:PAGE 8
GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.

Membership at the 9 group level

Group name	Number in Gp.	Members of group NO.	NAME					
GGp__1	1	1	1	AZU				
GGp__2	1	2	2	BGORA				
GGp__4	1	4	4	IT146				
GGp__6	1	6	6	OS6				
GGp__7	1	7	7	UPL5				
GGp_10	1	10	10	W56-125				
GGp_12	1	12	12	W96-1-1				
GGp_14	2	5	5	OL5	8	8	VAND	
GGp_15	3	3	3	GUAR	9	9	W181-18	11 11 W56-50

Membership at the 8 group level

Group name	Number in Gp.	Members of group NO.	NAME		
LGp__1	1	1	BH		
LGp__2	1	2	CR		
LGp__4	1	4	EL		
LGp__6	1	6	MB		
LGp_10	1	10	SG		
LGp_12	2	5	MA	11	SY
LGp_13	2	7	MD	8	ML
LGp_14	2	3	CV	9	NT

GENOTYPE GROUP MEANS FOR Location standardised VALUES OF YIELD FILE MYIELD94 4/10/ 4 16:52
:PAGE 9
GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.

Genotype group means - Section 1

Group Name	Group Mean	Location BH	CR	CV	EL	MA	MB	MD	ML
GGp__1	0.083	0.150	-1.277	1.123	-0.534	-0.471	-1.128	2.043	0.169
GGp__2	-0.553	-1.383	-1.594	-0.812	1.078	-1.041	-0.250	-0.026	-0.325
GGp__4	-0.441	-1.878	-0.025	-1.270	-0.075	-1.461	1.009	0.730	0.014
GGp__6	0.187	0.287	1.142	0.166	-0.604	1.844	-1.761	-0.807	-0.309
GGp__7	0.412	0.534	-0.226	-0.391	-0.121	1.274	-1.238	1.449	2.681
GGp_10	-0.286	-0.666	-0.753	1.293	-0.967	0.729	0.884	-0.863	-0.737
GGp_12	-0.013	-0.876	1.494	0.074	1.125	-0.029	-0.853	-0.507	-0.519
GGp_14	-0.128	1.060	-0.444	-1.124	-1.014	-0.147	0.621	0.018	0.576
GGp_15	0.290	0.571	0.709	0.689	0.708	-0.184	0.698	-0.685	-0.709
Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Genotype group means - Section 2

Group Name	Group Mean	Location NT	SG	SY
GGp__1	0.083	0.778	-0.523	0.582
GGp__2	-0.553	-0.891	1.251	-2.090
GGp__4	-0.441	0.177	-1.337	-0.741
GGp__6	0.187	1.227	-0.412	1.281
GGp__7	0.412	0.344	-1.050	1.272
GGp_10	-0.286	-0.773	-0.786	-0.512
GGp_12	-0.013	0.243	-1.097	0.802
GGp_14	-0.128	-1.566	1.190	-0.581
GGp_15	0.290	0.676	0.525	0.189
Mean	0.000	0.000	0.000	0.000

LOCATION GROUP MEANS FOR Location standardised VALUES OF YIELD FILE MYIELD94 4/10/ 4 16:52
:PAGE 10
GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.

Location group means - Section 1

Group Name	Group Mean	Genotype 1 AZU	2 BGORA	3 GUAR	4 IT146	5 OL5	6 OS6	7 UPL5	8 VAND
LGp__1	0.000	0.150	-1.383	0.002	-1.878	0.707	0.287	0.534	1.412
LGp__2	0.000	-1.277	-1.594	1.146	-0.025	-0.863	1.142	-0.226	-0.025
LGp__4	0.000	-0.534	1.078	0.832	-0.075	0.049	-0.604	-0.121	-2.077
LGp__6	0.000	-1.128	-0.250	0.521	1.009	0.625	-1.761	-1.238	0.618
LGp_10	0.000	-0.523	1.251	0.531	-1.337	1.163	-0.412	-1.050	1.217
LGp_12	0.000	0.055	-1.566	0.058	-1.101	-0.034	1.562	1.273	-0.694
LGp_13	0.000	1.106	-0.175	-0.197	0.372	0.168	-0.558	2.065	0.426
LGp_14	0.000	0.950	-0.851	0.368	-0.547	-1.236	0.696	-0.024	-1.455
Mean	0.000	0.083	-0.553	0.317	-0.441	-0.047	0.187	0.412	-0.209

(continuation of group membership and group means)

Location group means - Section 2

Group Name	Group Mean	Genotype 9 W181-18	10 W56-125	11 W56-50	12 W96-1-1
LGp__1	0.000	0.868	-0.666	0.843	-0.876
LGp__2	0.000	0.232	-0.753	0.749	1.494
LGp__4	0.000	1.397	-0.967	-0.105	1.125
LGp__6	0.000	0.455	0.884	1.118	-0.853
LGp_10	0.000	1.081	-0.786	-0.037	-1.097
LGp_12	0.000	-0.745	0.109	0.695	0.386
LGp_13	0.000	-1.206	-0.800	-0.688	-0.513
LGp_14	0.000	0.244	0.260	1.434	0.159
Mean	0.000	0.056	-0.286	0.496	-0.013

GXE GROUP MEANS FOR Location standardised VALUES OF YIELD FILE MYIELD94 4/10/ 4 16:52 :PAGE 11
GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.

Genotype group by location group means

Group Name	Group Mean	Location group LGp__1	LGp__2	LGp__4	LGp__6	LGp_10	LGp_12	LGp_13	LGp_14
GGp__1	0.083	0.150	-1.277	-0.534	-1.128	-0.523	0.055	1.106	0.950
GGp__2	-0.553	-1.383	-1.594	1.078	-0.250	1.251	-1.566	-0.175	-0.851
GGp__4	-0.441	-1.878	-0.025	-0.075	1.009	-1.337	-1.101	0.372	-0.547
GGp__6	0.187	0.287	1.142	-0.604	-1.761	-0.412	1.562	-0.558	0.696
GGp__7	0.412	0.534	-0.226	-0.121	-1.238	-1.050	1.273	2.065	-0.024
GGp_10	-0.286	-0.666	-0.753	-0.967	0.884	-0.786	0.109	-0.800	0.260
GGp_12	-0.013	-0.876	1.494	1.125	-0.853	-1.097	0.386	-0.513	0.159
GGp_14	-0.128	1.060	-0.444	-1.014	0.621	1.190	-0.364	0.297	-1.345
GGp_15	0.290	0.571	0.709	0.708	0.698	0.525	0.003	-0.697	0.682
Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Location index		-0.272	-0.746	0.067	0.779	0.836	-1.140	-0.175	-0.870

GXE GROUP AND MEMBER MEANS FOR Location standardised VALUES OF YIELD FILE MYIELD94 4/10/ 4 16:52 :PAGE 12
GXE INTERACTIONS BETWEEN BREEDING SITES/94 RESULTS.
STANDARD AMMI AND PATTERN ANALYSIS MATERIAL FOR BANGLADESH.

GXE GROUP AND MEMBER MEANS - SECTION 1

Genotype NAME		Location or location group BH	LGp__1	CR	LGp__2	EL	LGp__4	MB	LGp__6
1	AZU	0.150	0.150	-1.277	-1.277	-0.534	-0.534	-1.128	-1.128
GGp__1		0.150	0.150	-1.277	-1.277	-0.534	-0.534	-1.128	-1.128
2	BGORA	-1.383	-1.383	-1.594	-1.594	1.078	1.078	-0.250	-0.250
GGp__2		-1.383	-1.383	-1.594	-1.594	1.078	1.078	-0.250	-0.250
4	IT146	-1.878	-1.878	-0.025	-0.025	-0.075	-0.075	1.009	1.009
GGp__4		-1.878	-1.878	-0.025	-0.025	-0.075	-0.075	1.009	1.009
6	OS6	0.287	0.287	1.142	1.142	-0.604	-0.604	-1.761	-1.761
GGp__6		0.287	0.287	1.142	1.142	-0.604	-0.604	-1.761	-1.761
7	UPL5	0.534	0.534	-0.226	-0.226	-0.121	-0.121	-1.238	-1.238
GGp__7		0.534	0.534	-0.226	-0.226	-0.121	-0.121	-1.238	-1.238
10	W56-125	-0.666	-0.666	-0.753	-0.753	-0.967	-0.967	0.884	0.884
GGp_10		-0.666	-0.666	-0.753	-0.753	-0.967	-0.967	0.884	0.884
12	W96-1-1	-0.876	-0.876	1.494	1.494	1.125	1.125	-0.853	-0.853
GGp_12		-0.876	-0.876	1.494	1.494	1.125	1.125	-0.853	-0.853
5	OL5	0.707	0.707	-0.863	-0.863	0.049	0.049	0.625	0.625
8	VAND	1.412	1.412	-0.025	-0.025	-2.077	-2.077	0.618	0.618
GGp_14		1.060	1.060	-0.444	-0.444	-1.014	-1.014	0.621	0.621
3	GUAR	0.002	0.002	1.146	1.146	0.832	0.832	0.521	0.521
9	W181-18	0.868	0.868	0.232	0.232	1.397	1.397	0.455	0.455
11	W56-50	0.843	0.843	0.749	0.749	-0.105	-0.105	1.118	1.118
GGp_15		0.571	0.571	0.709	0.709	0.708	0.708	0.698	0.698
MEAN		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

(continuation of group membership and group means)

GXE GROUP AND MEMBER MEANS - SECTION 2

Genotype NAME		Location or location group SG	LGp_10	MA	SY	LGp_12	MD	ML	LGp_13
1	AZU	-0.523	-0.523	-0.471	0.582	0.055	2.043	0.169	1.106
GGp__1		-0.523	-0.523	-0.471	0.582	0.055	2.043	0.169	1.106
2	BGORA	1.251	1.251	-1.041	-2.090	-1.566	-0.026	-0.325	-0.175
GGp__2		1.251	1.251	-1.041	-2.090	-1.566	-0.026	-0.325	-0.175
4	IT146	-1.337	-1.337	-1.461	-0.741	-1.101	0.730	0.014	0.372
GGp__4		-1.337	-1.337	-1.461	-0.741	-1.101	0.730	0.014	0.372
6	OS6	-0.412	-0.412	1.844	1.281	1.562	-0.807	-0.309	-0.558
GGp__6		-0.412	-0.412	1.844	1.281	1.562	-0.807	-0.309	-0.558
7	UPL5	-1.050	-1.050	1.274	1.272	1.273	1.449	2.681	2.065
GGp__7		-1.050	-1.050	1.274	1.272	1.273	1.449	2.681	2.065
10	W56-125	-0.786	-0.786	0.729	-0.512	0.109	-0.863	-0.737	-0.800
GGp_10		-0.786	-0.786	0.729	-0.512	0.109	-0.863	-0.737	-0.800
12	W96-1-1	-1.097	-1.097	-0.029	0.802	0.386	-0.507	-0.519	-0.513
GGp_12		-1.097	-1.097	-0.029	0.802	0.386	-0.507	-0.519	-0.513
5	OL5	1.163	1.163	0.270	-0.338	-0.034	-0.511	0.848	0.168
8	VAND	1.217	1.217	-0.565	-0.823	-0.694	0.546	0.305	0.426
GGp_14		1.190	1.190	-0.147	-0.581	-0.364	0.018	0.576	0.297
3	GUAR	0.531	0.531	-0.025	0.142	0.058	-0.142	-0.251	-0.197
9	W181-18	1.081	1.081	-1.121	-0.369	-0.745	-1.212	-1.201	-1.206
11	W56-50	-0.037	-0.037	0.595	0.795	0.695	-0.701	-0.675	-0.688
GGp_15		0.525	0.525	-0.184	0.189	0.003	-0.685	-0.709	-0.697
MEAN		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

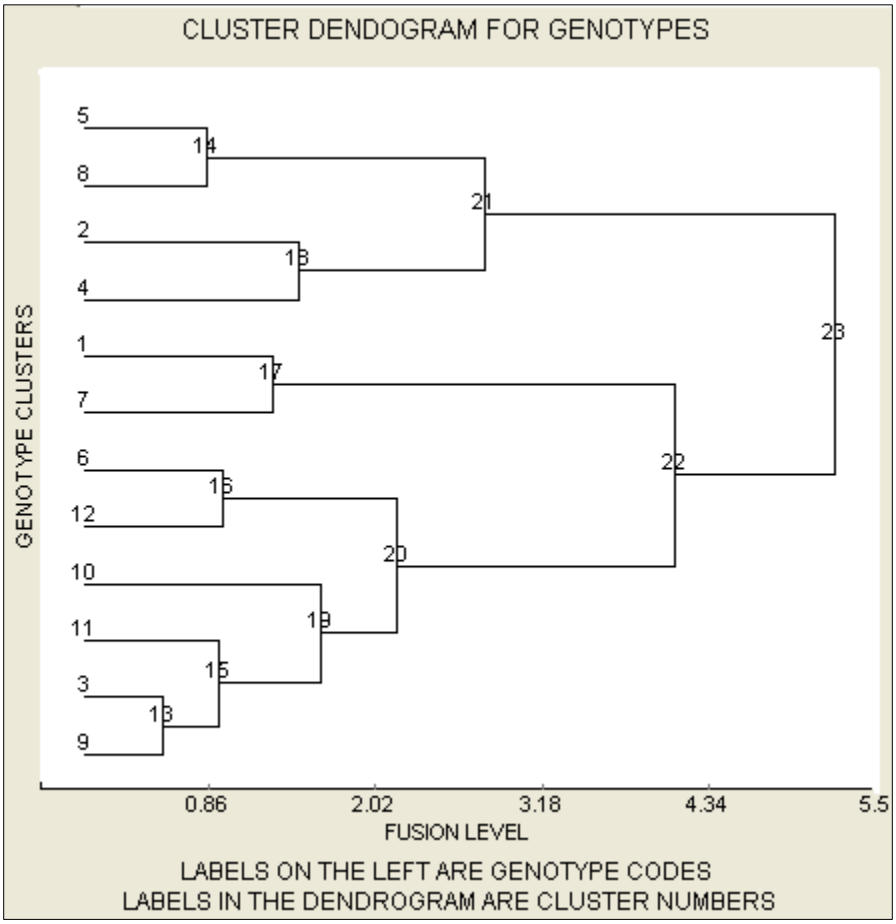
GXE GROUP AND MEMBER MEANS - SECTION 3

Genotype NAME		Location or location group CV	NT	LGp_14	MEAN
1	AZU	1.123	0.778	0.950	0.083
GGp__1		1.123	0.778	0.950	0.083
2	BGORA	-0.812	-0.891	-0.851	-0.553
GGp__2		-0.812	-0.891	-0.851	-0.553
4	IT146	-1.270	0.177	-0.547	-0.441
GGp__4		-1.270	0.177	-0.547	-0.441
6	OS6	0.166	1.227	0.696	0.187
GGp__6		0.166	1.227	0.696	0.187
7	UPL5	-0.391	0.344	-0.024	0.412
GGp__7		-0.391	0.344	-0.024	0.412
10	W56-125	1.293	-0.773	0.260	-0.286
GGp_10		1.293	-0.773	0.260	-0.286
12	W96-1-1	0.074	0.243	0.159	-0.013
GGp_12		0.074	0.243	0.159	-0.013
5	OL5	-1.406	-1.066	-1.236	-0.047
8	VAND	-0.842	-2.067	-1.455	-0.209
GGp_14		-1.124	-1.566	-1.345	-0.128
3	GUAR	0.076	0.661	0.368	0.317
9	W181-18	0.309	0.179	0.244	0.056
11	W56-50	1.680	1.188	1.434	0.496
GGp_15		0.689	0.676	0.682	0.290
MEAN		0.000	0.000	0.000	0.000

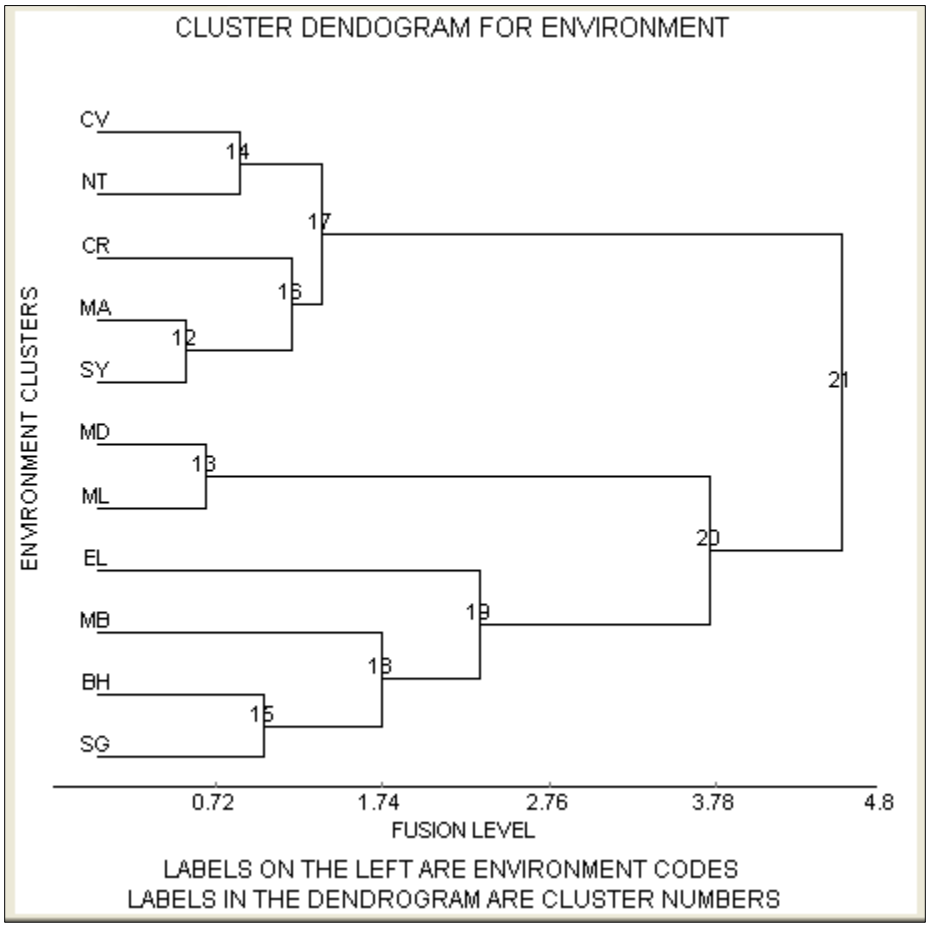
IV. Graphical Output

CropStat will also output the following plots.

1. Cluster
    a. Cluster dendogram for genotypes



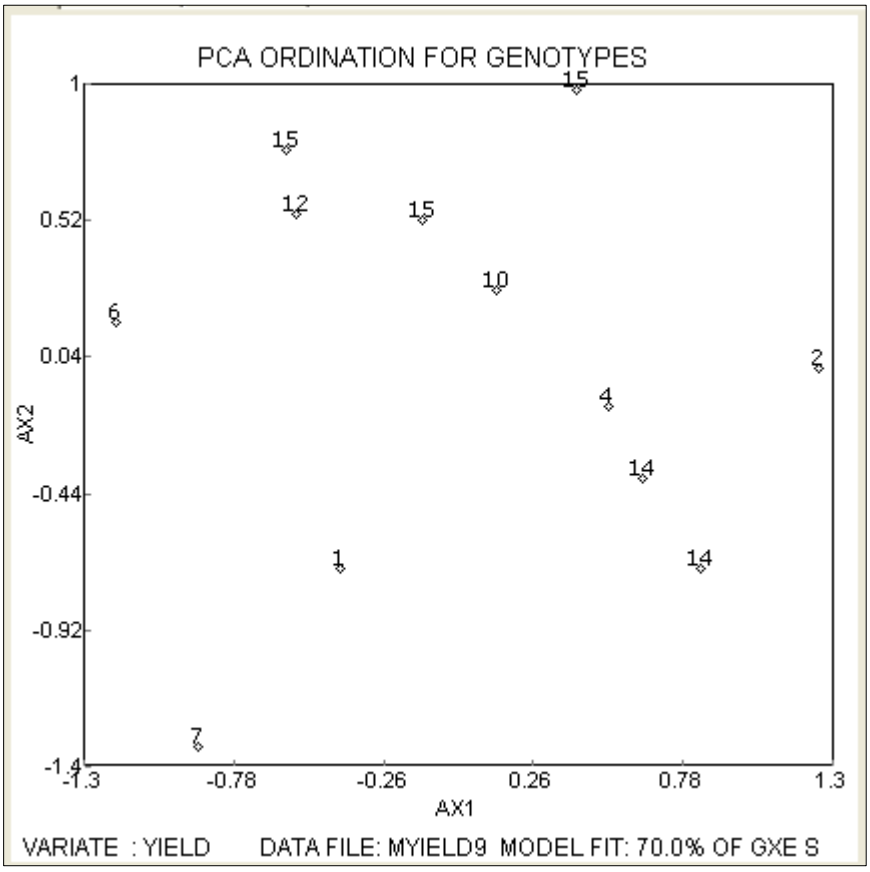
b. Cluster dendogram for environments



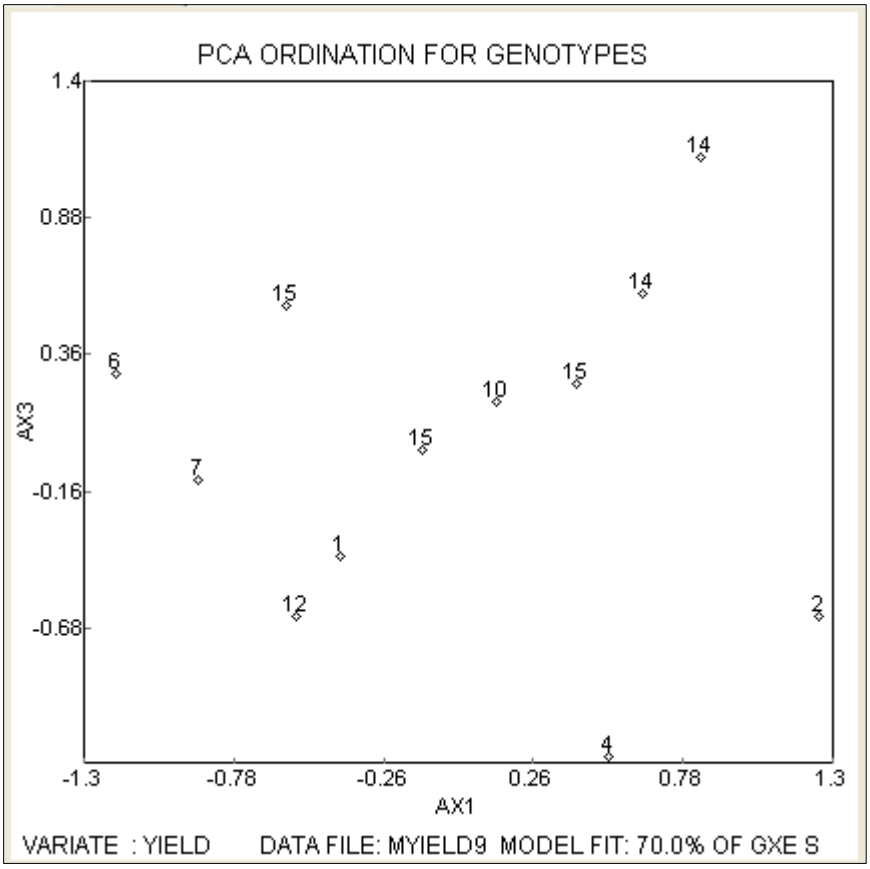
2. Ordination

a. Genotypes

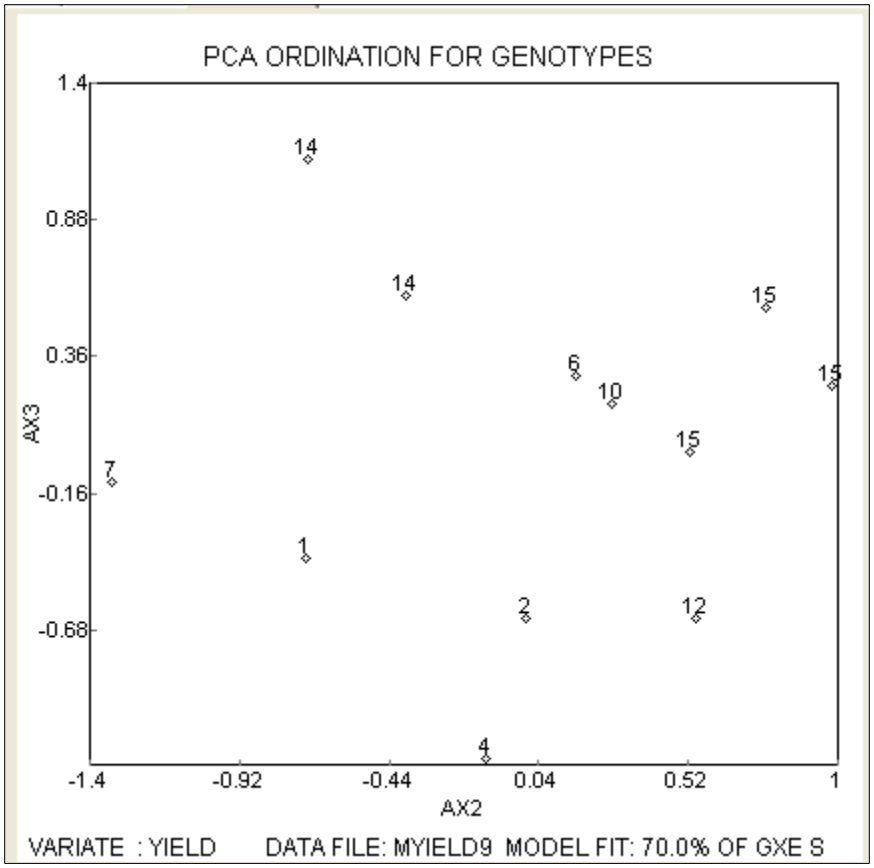
i. AX2 × AX1



ii. AX3 × AX1

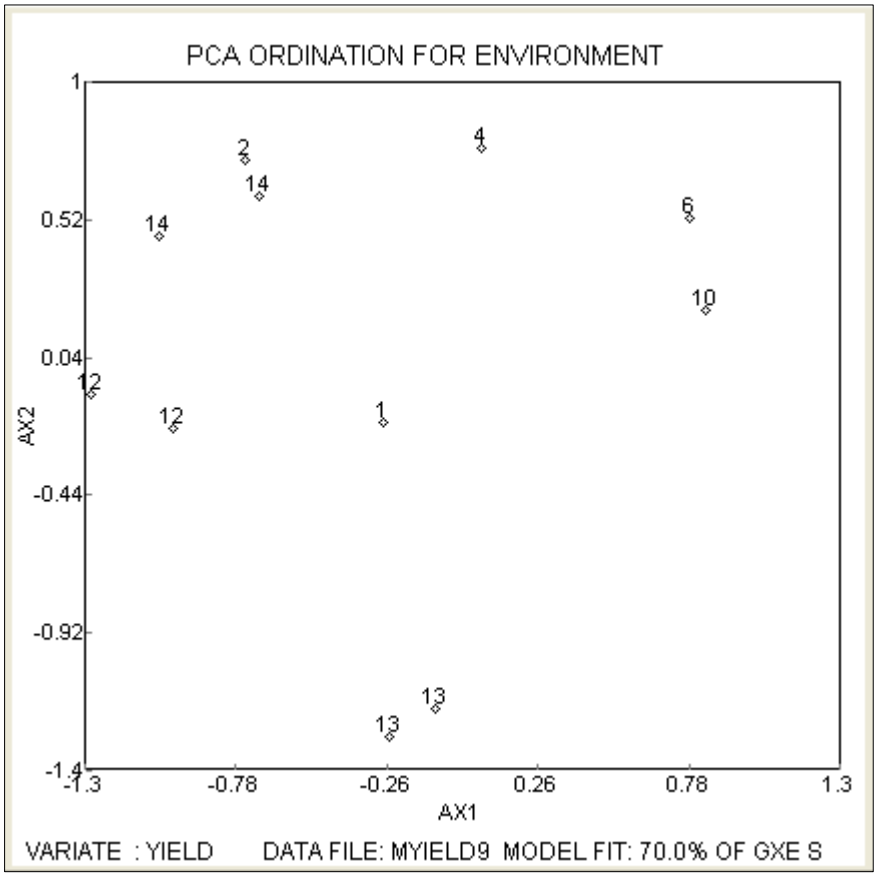

PCA ORDINATION FOR GENOTYPES
1.4
0.88
0.36
-0.16
-0.68
AX3
-1.3
-0.78
-0.26
0.26
0.78
1.3
AX1
14
14
15
6
15
10
15
7
1
12
2
4
VARIATE : YIELD
DATA FILE: MYIELD9 MODEL FIT: 70.0% OF GXE S

iii. AX3 × AX2

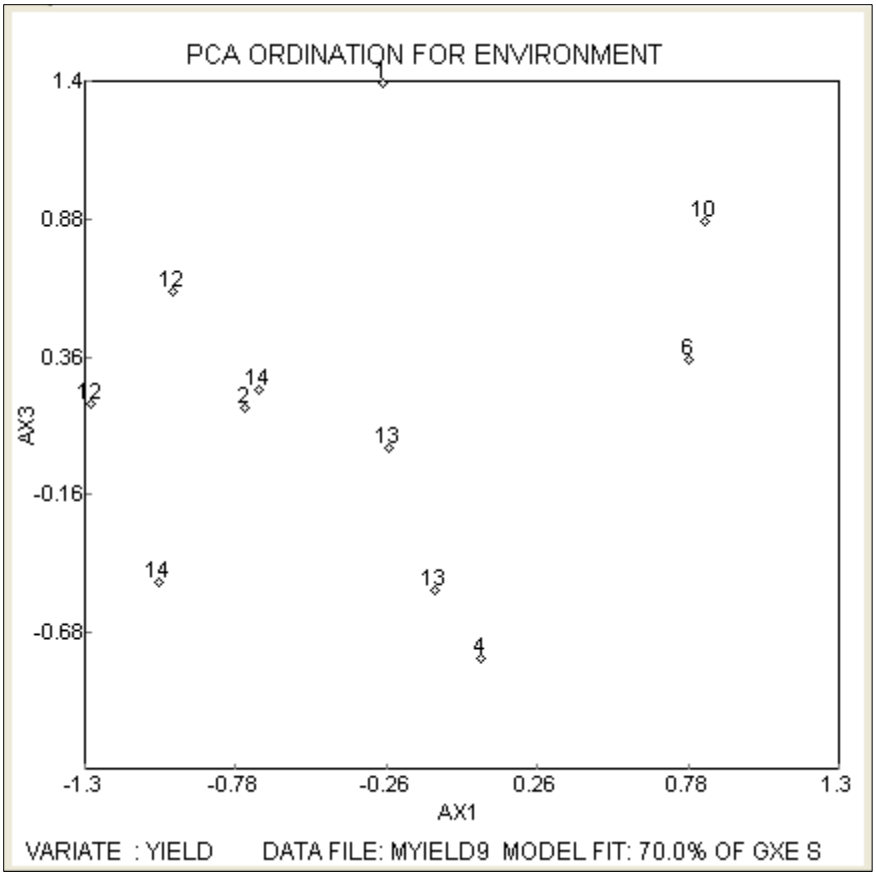

PCA ORDINATION FOR GENOTYPES
1.4
0.88
0.36
-0.16
-0.68
AX3
14
14
15
6
10
15
15
7
1
2
12
4
-1.4
-0.92
-0.44
0.04
0.52
1
AX2
VARIATE : YIELD
DATA FILE: MYIELD9 MODEL FIT: 70.0% OF GXE S

b. Environment

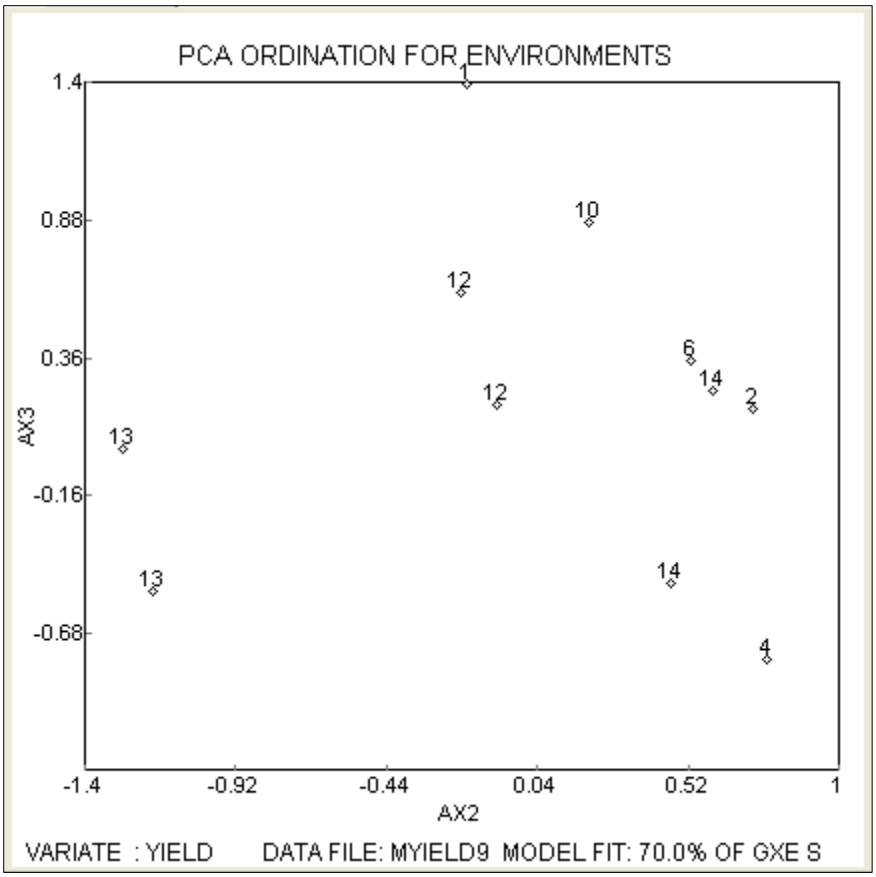
i. AX2 × AX1



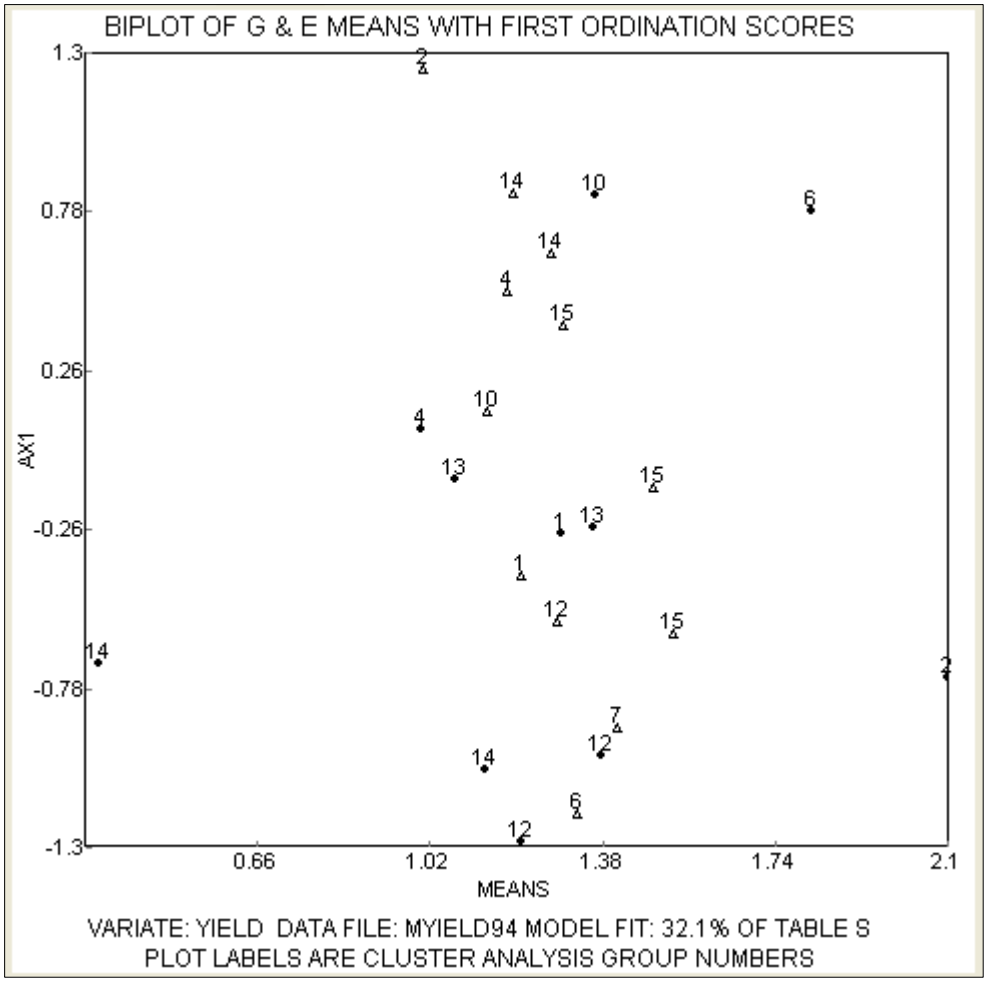
ii. AX3 × AX1


PCA ORDINATION FOR ENVIRONMENT
1.4
0.88
0.36
-0.16
-0.68
AX3
-1.3
-0.78
-0.26
0.26
0.78
1.3
AX1
VARIATE : YIELD
DATA FILE: MYIELD9 MODEL FIT: 70.0% OF GXE S

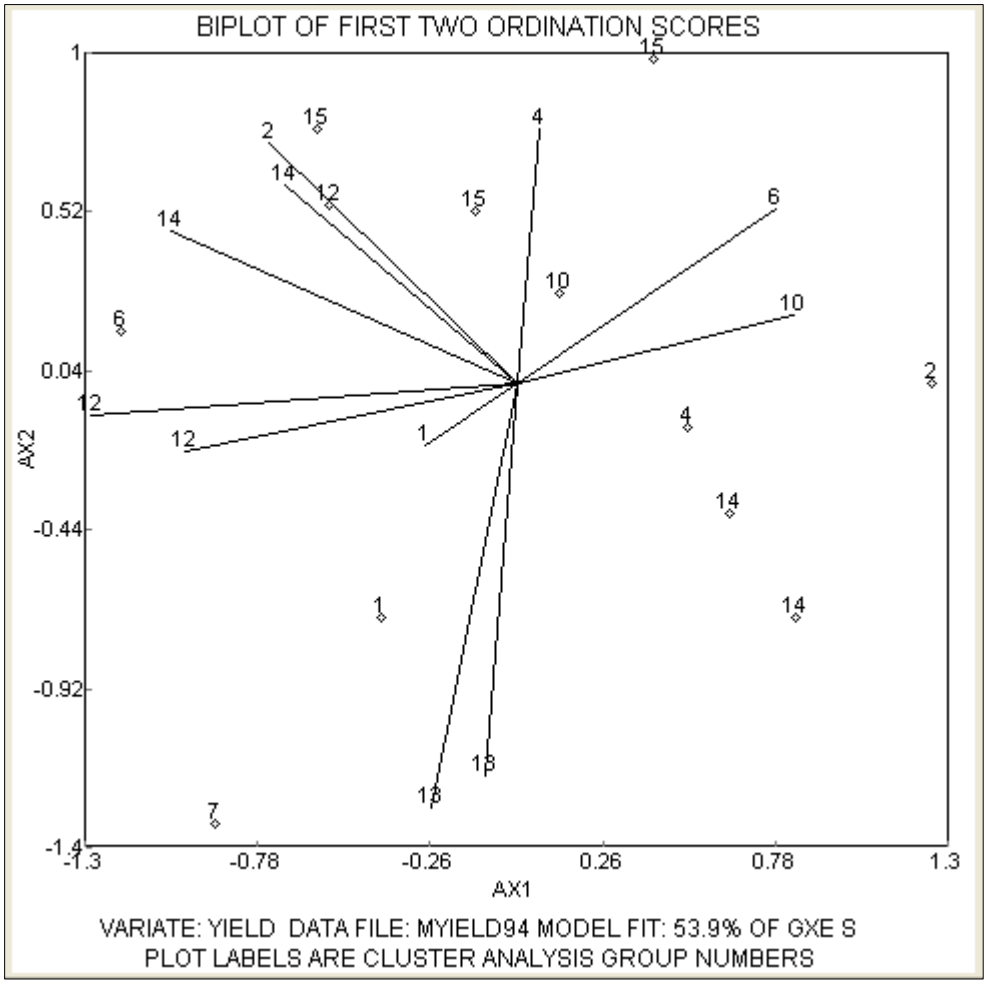
iii. AX3 × AX2


PCA ORDINATION FOR ENVIRONMENTS
1.4
0.88
0.36
-0.16
-0.68
AX3
-1.4
-0.92
-0.44
0.04
0.52
1
AX2
VARIATE : YIELD
DATA FILE: MYIELD9 MODEL FIT: 70.0% OF GXE S

c. Means


BIPLOT OF G & E MEANS WITH FIRST ORDINATION SCORES
1.3
0.78
0.26
-0.26
-0.78
-1.3
AX1
0.66
1.02
1.38
1.74
2.1
MEANS
VARIATE: YIELD DATA FILE: MYIELD94 MODEL FIT: 32.1% OF TABLE S
PLOT LABELS ARE CLUSTER ANALYSIS GROUP NUMBERS

d. Ordination Scores



These plots can also be produced if you run the GEBEIPLT.CMD, a command file outputted by CropStat when running pattern analysis in the **Analysis –G×E Plots.**

ANALYSIS OF QUANTITATIVE TRAIT LOCI

At the end of the tutorial, the user should be able to

- perform analysis of quantitative trait loci

I. Introduction

The QTL Analysis Item on the Analysis Menu of CropStat performs analysis of phenotypic and molecular marker data to provide information on the location and magnitude of QTLs on a genetic map and on the interaction of QTLs at different locations.

The basic methodology for mapping QTLs involves arranging a cross between two inbred strains differing substantially in a quantitative trait, then segregating progeny are scored for the trait and for a number of genetic markers. This leads to three types of data, a marker map which gives numbers, names and positions of molecular markers on chromosomes, marker data for a set of progeny from the cross and measurement data on phenotypic traits for the same progeny.

For this tutorial the segregating progeny was produced by doubling gamete chromosomes from the F_1 plants. Such plants have been referred to as DHL progeny. Occasionally in the tutorial a different origin will be assumed for the sake of illustration of the methods.

II. Data preparation

- Create a subfolder QTL ANALYSIS inside your working folder C:\MY CROPSTAT
- Import the marker map data *QTLMAP.ASC* stored in *CROPSTAT7.2\TUTORIAL\TUTORIAL DATASETS* folder. Save this data as *QTLMAP.SYS* inside your working folder *C:\MY CROPSTAT\QTL ANALYSIS*. To read an ASCII file into CropStat follow the procedure given in the Data and File management module.

- The next step is to input the marker data. Import the *QTLMKR.ASC* file and save as *QTLMKR.SYS* inside your working folder *C:\MY CROPSTAT\QTL ANALYSIS*.

The data for markers on chromosomes 1 and 9 of the rice genome is shown in Figure 2. The data are for DHL progeny so in principle only two genotypes are possible for each marker. These are represented as 1 and 3 in the data set. In fact, an indeterminate class is also apparent and these have been scored as 0 in the data set.

```
H RFLP MAP OF CHROMOSOMES 1 AND 9 OF THE RICE GENOME
 V001 MKRNO MARKER NUMBER
 V002 MARKER$ MARKER NAME
 V003 CHRMSM CHROMOSOME NUMBER
 V004 ORDER MARKER ORDER WITHIN CHROMOSOME
 V005 DIST DISTANCE ON CHROMOSOME FROM PREVIOUS MARKER
      (0 INDICATES THE FIRST KNOWN MARKER ON A CHROMOSOME)
 ////
 049  RG331   1  1   0.0
 082  RG810   1  2   9.2
 133  RZ801   1  3   9.5
 131  RZ730   1  4  35.4
 076  RG690   1  5  12.3
 093  RZ19    1  6   9.1
 054  RG381   1  7  22.1
 051  RG345   1  8   3.4
 100  RZ276   1  9  36.6
 028  RG146X  1 10   1.6
 035  RG173   1 11  14.4
 067  RG532   1 12  31.5
 044  RG246   1 13  16.7
 063  RG472   1 14  20.0
 078  RG757   9  1   0.0
 010  CDO590  9  2   3.5
 094  RZ206   9  3  21.4
 112  RZ422   9  4  38.6
 097  RZ228   9  5  16.0
 090  RZ12    9  6   4.9
 075  RG667   9  7   7.8
 060  RG451   9  8  18.3
 132  RZ792   9  9   6.0
 111  RZ404   9 10   0.9
```

Figure 1. Marker Map Data for QTL Analysis

- The next step is to input the phenotypic data. Import the *QTLYLD.ASC* file and save as *QTLYLD.SYS* inside your working folder C:\MY CROPSTAT\QTL ANALYSIS.

  The phenotypic data stored in QTLYLD.ASC is data from a RCB design with two reps are shown in Figure 3. In this case plant number, PN, is NOT nested within block or REP since replicate plant genotypes have been produced by collecting gametes from F_1 plants, doubling the chromosomes and generating homozygote plants which are selfed to produce replicate seeds. Hence plant 1 in rep 1 has the same genotype as plant 1 in rep 2. Convert this data to the CropStat file QTLYLD.SYS with the Import function of the Data Editor.

Figure 2. RFLP marker data for 94 plants

```
V001 PN
V002 RG331
V003 RG810
V004 RZ801
V005 RZ730
V006 RG690
V007 RZ19
V008 RG381
V009 RG345
V010 RZ276
V011 RG146X
V012 RG173
V013 RG532
V014 RG246
V015 RG472
V016 RG757
V017 CDO590
V018 RZ206
V019 RZ422
V020 RZ228
V021 RZ12
V022 RG667
V023 RG451
V024 RZ792
V025 RZ404
////TRIAL QTLTUT
 5 1 1 1 3 3 3 3 3 3 3 1 1 1 3 3 3 1 1 1 1 1 3 3 3
 7 1 1 1 1 1 1 3 3 3 3 1 1 1 1 1 1 1 3 1 1 1 1 1 1
10 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 3 3 1 1 3 3 3
12 1 1 1 1 1 1 1 1 3 3 3 3 3 1 1 3 3 3 1 1 1 1 0 3 3
13 1 1 1 1 1 1 1 1 3 3 0 0 1 1 1 1 1 1 0 1 1 1 0 3
20 3 3 3 1 1 1 1 1 3 3 3 3 3 3 3 3 1 1 1 3 3 1 0 1 1
21 3 3 0 0 3 0 3 1 3 3 0 1 0 1 0 0 0 3 3 1 1 0 1 1
22 1 1 1 0 3 3 3 3 3 3 3 1 1 3 3 0 3 0 3 3 3 3 3 3 3
25 3 3 3 1 1 1 1 1 3 3 3 3 3 1 1 1 1 1 1 1 1 3 3 3 3
27 1 1 1 1 1 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1
31 1 1 1 1 1 1 1 1 3 3 3 3 3 3 1 3 3 1 1 1 1 1 1 1 1
33 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 1 1 3 3 3 3 3 3
34 3 1 3 1 3 3 3 3 3 3 3 0 0 0 3 3 3 1 1 0 1 0 3 3 3
35 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 1 1 1 3 1 1 1 1 1 1
39 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 1 1 1 1 1 1
41 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 3 1 1 1 1 1 1 1 1
47 3 3 3 3 1 1 1 1 3 3 3 3 3 3 3 1 1 3 1 1 1 1 0 1 1
48 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 1 1 3 1 1 1 1 0 1 1
55 1 1 1 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
57 3 1 1 1 1 1 3 3 3 3 3 3 3 3 3 1 1 3 1 3 3 3 3 3 3
62 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 0 1 1
67 3 3 3 3 3 3 3 3 3 3 3 3 1 1 3 3 3 1 3 3 3 3 3 3 3
69 3 3 3 3 3 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1
74 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1
78 3 3 3 3 0 3 1 1 3 3 3 0 1 1 1 1 1 1 1 1 1 1 1 1 1
82 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1
84 3 1 1 1 1 1 1 1 3 3 3 3 3 1 1 3 3 3 1 1 1 1 1 1 1
87 1 1 1 3 3 0 1 1 3 3 3 1 1 1 1 0 3 1 1 3 3 3 3 3 3
88 3 3 3 1 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 3 3 3
89 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 3 3 3
90 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 1 1 1 1 1 1 1 1 1 1
94 3 3 3 1 1 1 1 1 1 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1
97 3 1 3 3 1 1 1 1 1 3 3 0 3 3 3 3 3 3 3 3 3 3 3 3 0
! 100 1 3 1 1 3 1 0 0 3 3 3 0 3 0 1 3 3 3 3 3 3 3 3 0 3
! 105 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
! 106 3 3 3 1 1 1 1 1 3 3 3 3 1 1 1 3 3 3 1 1 1 1 1 1 1
! 107 3 3 3 0 3 3 3 3 3 3 3 3 3 3 3 1 1 1 3 3 3 3 3 1 1
! 110 3 3 3 0 3 3 3 3 3 3 3 3 3 1 1 3 3 3 1 1 1 1 1 1 1
! 116 1 1 1 0 1 1 1 1 1 1 1 1 3 1 3 3 3 3 3 1 1 1 1 1 1
! 124 1 1 1 0 3 0 0 3 3 3 3 0 1 0 1 0 0 0 0 1 1 1 0 3 3
! 130 3 3 3 3 3 3 3 1 1 3 3 3 3 3 1 1 1 1 1 1 1 1 3 3 3
! 144 3 3 3 3 3 3 3 1 1 1 1 1 3 1 1 1 1 1 3 3 3 3 3 1 1
! 146 3 1 1 1 1 1 1 1 1 1 1 3 1 3 0 1 1 1 3 3 3 3 1 3 3
! 147 1 1 1 1 1 1 1 1 1 1 1 3 3 3 1 1 1 1 1 1 1 1 0 1 1
! 159 0 3 0 0 3 3 0 0 3 3 3 1 1 1 0 3 0 3 0 1 1 0 1 1
! 163 3 3 0 3 0 1 1 3 3 3 0 3 0 0 0 0 0 0 3 3 3 0 0 3
! 169 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 0 1 1
! 171 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 1 1 1 1 1 1 1 1 1 1
! 178 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 3 3 3 3 3 3 3 3
! 180 3 3 3 3 1 1 1 1 3 3 3 3 3 3 3 1 1 1 3 3 3 1 1 1 1
! 183 1 1 1 3 3 3 3 3 3 3 3 3 3 3 0 3 3 1 1 1 1 1 1 1 1
! 184 3 1 1 0 1 1 1 1 3 3 3 3 3 3 0 1 1 1 3 3 1 0 1 1 1
! 188 3 3 3 1 3 3 0 3 1 1 1 1 1 1 1 1 1 0 3 3 3 0 3 3 3
! 192 0 1 0 0 0 3 0 3 3 0 3 0 3 0 0 3 0 3 0 0 3 0 3 3 3
! 196 1 1 1 3 0 3 3 3 3 3 0 0 3 3 3 3 3 3 3 3 3 3 3 3 3
! 200 1 1 1 3 3 3 1 1 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1
! 210 3 3 3 1 1 1 1 1 3 3 3 1 1 1 1 1 1 1 1 3 3 3 3 3 3
! 226 1 3 3 0 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 0 3 1
! 228 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 3 3 3 3 3 3 3
! 232 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 3 3 3 3 3 3 3
! 236 1 1 1 3 3 3 1 1 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1
! 238 3 3 3 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
! 240 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1
! 241 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
! 251 1 1 1 1 1 1 1 3 3 3 3 3 3 1 1 1 1 1 3 3 3 3 3 3 3
! 252 1 1 1 1 1 1 1 1 1 1 3 3 3 3 1 1 1 3 3 3 3 3 3 3 3
! 253 3 3 0 0 0 0 1 0 3 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 3
! 262 3 3 3 3 3 3 3 3 0 1 3 3 0 1 1 3 1 3 3 1 1 1 0 1 1
! 268 3 3 3 3 3 3 1 1 1 1 1 3 3 3 3 0 1 1 1 1 1 1 1 1 1
! 269 1 1 1 3 3 3 3 3 3 3 3 3 3 3 1 0 3 3 1 1 1 1 1 1 1
! 272 1 1 1 1 1 1 1 1 3 3 3 3 3 3 1 3 3 3 1 1 1 1 1 1 1
! 281 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 1 1 1 3 1 3 3 3 3 3
! 284 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 3 3 1 1 1 1 1 1 1
! 288 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
! 290 1 3 3 1 3 3 3 3 3 3 3 3 1 1 1 1 1 3 3 3 3 3 3 3 3
! 291 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 0 1 1
! 295 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1
! 303 1 1 0 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1
! 313 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 3 3 3 3
! 326 3 3 3 3 1 1 1 1 1 3 3 3 3 3 3 3 3 3 1 1 1 1 1 0 1 1
! 329 3 3 3 3 1 1 1 1 1 1 1 1 1 3 3 3 3 3 1 1 3 3 3 3 3
! 331 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 1 1 3 3 3 3 3 3 1 1 1
! 332 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
! 333 1 1 1 1 1 1 3 3 3 3 3 3 1 1 1 3 3 3 3 3 3 3 3 3 3
! 335 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 0 1 1
! 336 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 3 3 3 3 3 3 3
! 350 3 3 0 1 1 1 0 0 3 3 3 3 3 3 3 1 1 0 3 0 1 0 0 0 0
! 351 1 1 1 1 1 1 1 3 3 3 3 3 3 1 1 1 1 1 3 3 3 3 3 3 3
! 355 3 3 3 3 3 3 3 3 3 3 3 3 0 3 3 1 1 0 3 3 3 0 3 3 0
! 356 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1
! 359 1 1 1 1 3 3 3 3 3 3 3 3 1 1 1 1 1 1 3 3 3 3 3 3 3
! 361 3 3 3 3 3 3 3 0 3 3 3 3 3 3 1 1 0 3 0 1 1 1 1 0 1 1
! 372 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3
```

Figure 3 Phenotypic data for QTL analysis

```
H PHENOTYPIC DATA FOR QTL MAPPING OF RICE YIELD COMPONENTS
V001 PN     PLANT GENOTYPE NUMBER
V002 REP    BLOCK (1-2)
V003 DUR    DURATION (PLANTING TO MATURITY)
V004 TIL45  TILLER NUMBER AT 45 DAT
V005 HGT    PLANT HEIGHT
V006 EXS    EXERSION
V007 PAN    PANICLE LENGTH
V008 NBPAN  NUMBER OF PANICLES
V009 TILMAT NUMBER OF TILLERS AT MATURITY
V010 PANWGT PANICLE WEIGHT
V011 NBG    NUMBER OF GRAINS
V012 STR    PERCENT STERILE GRAINS
V013 WGT10  GRAIN WEIGHT OF TEN PLANTS
V014 TGW    THOUSAND GRAIN WEIGHT
V015 WGTCOR GRAIN WEIGHT AT 14% MOISTURE
////TRIAL QTLTUT
  5 1 137  7.1  88.0 -1.4 26.1 17.6 22.4 1.86  95 36.6 125.3 24.81 141.5
  7 1 116  6.3  61.5 -8.6 21.4 13.1 15.0 2.50 128 18.2 152.2 22.23 185.3
 10 1 139  7.5 138.0  6.4 33.5 10.2 10.2 4.44 150 18.2 178.5 31.73 192.2
 12 1 118  9.3 109.0  5.1 24.2 11.9 12.0 5.00 191 16.3 237.0 24.01 249.5
 13 1 123  7.9  69.0 -8.8 25.6 14.3 14.7 3.10 152 32.2 218.2 25.48 258.9
 20 1 141 11.7 101.0  1.1 25.1 14.5 14.6 3.03 135 23.5 162.0 25.65 212.0
 22 1 134  8.9  72.5 -1.3 22.9 18.2 18.6 1.93 104 27.1 153.4 22.12 174.7
 25 1 124  9.2 116.0 -0.7 27.4 12.5 12.9 3.08 120 26.1 276.5 32.80 293.3
 27 1 127  9.1  60.0 -2.4 17.1 17.2 24.6 1.55  74 21.8 112.1 22.58 125.6
 31 1 125  9.6  90.0 -2.6 24.0 13.5 13.6 3.77 155 10.8 250.1 26.62 299.3
 33 1 133  3.9  79.0 -4.7 24.0  9.8 10.7 3.49 186 23.2  52.5 22.10  60.7
 34 1 127  8.4  84.0 -1.7 24.2 12.5 13.3 3.01 131 31.0 158.7 28.98 185.2
 35 1 120  8.5  58.5 -8.7 22.0 13.3 14.1 2.05 119 30.3 127.6 20.73 142.9
 39 1 140 10.9 118.5  4.6 28.0  9.9  9.9 2.84 143 23.5  34.1 26.61  43.8
 41 1 126  9.0 106.5  2.1 26.4  9.1  9.5 4.30 157 23.9 173.3 30.18 200.6
 47 1 125 10.9 138.0  3.6 27.5 13.5 13.6 3.70 181 28.0 211.9 26.90 245.3
 48 1 136  8.3  88.5 -2.8 23.0  9.9 10.0 5.07 237 13.3 208.1 20.74 224.1
 55 1 129 14.2 120.0  3.1 28.8 19.5 19.6 2.99 114 11.2 275.0 27.78 310.5
 57 1 123 11.5  84.5 -0.6 25.5 12.7 12.8 4.70 180 12.9 323.4 26.26 368.1
 62 1 118  8.5 108.0 -3.9 26.9  8.3  8.7 4.70 240 39.5 157.3 25.79 168.2
 67 1 140  6.3 147.5  0.3 32.0 10.0 10.2 4.19 211 30.9 152.9 31.15 171.3
 69 1 143  8.5 135.0  1.8 27.4  8.5  8.5 3.46 123 14.9  14.3 30.00  17.1
 74 1 125  8.4 108.0  0.6 27.9 12.0 12.7 4.43 162 18.4 108.3 31.17 133.0
 82 1 126  7.6 108.0  1.1 28.4 11.5 12.0 4.53 152 16.7 196.0 30.80 221.3
 84 1 125 11.0  98.0  2.1 29.4 12.9 12.9 3.51 127 17.7 217.9 30.75 246.1
 87 1 136  8.9  89.5  2.2 22.7 11.0 11.3 3.82 143  7.4 220.4 32.65 273.2
 88 1 122  7.8  88.5  0.7 21.7 11.6 11.7 3.26 138 17.6 162.4 27.62 186.4
 89 1 130 12.7  90.5  0.4 23.2 16.6 16.6 3.55 143  9.6 293.5 25.15 336.9
 90 1 120  8.2  85.5 -1.2 25.2 11.6 11.6 3.37 137  8.9 208.6 25.95 224.7
 94 1 132  8.9  93.5 -2.6 24.7 12.7 13.0 2.59 128 35.9 152.3 26.13 166.6
 97 1 140 10.3 146.0  1.9 31.6 12.9 13.2 3.78 129 23.0 115.8 33.17 124.7
100 1 124  8.8 114.5  0.3 31.5 10.8 10.9 5.37 258 17.8 213.0 26.62 248.5
105 1 127  8.1  94.5 -0.2 25.1  8.6  8.6 4.12 193 19.4  59.2 23.50  67.4
106 1 146  9.5  81.5  1.2 20.5  9.5  9.6 2.23 122 21.2  35.9 25.70  42.2
107 1 136  7.4 128.0  3.7 31.8  9.4  9.4 4.11 240 47.8 157.3 26.83 174.9
116 1 129 12.2 111.0  1.9 25.6 14.4 14.6 4.47 143 22.1 329.3 25.34 371.8
124 1 130  9.9  90.5  0.1 24.4 12.4 12.8 3.05 134 10.4 157.5 24.93 180.7
130 1 122  8.0 120.5  0.8 32.2  9.8  9.9 5.93 207 20.4 219.6 34.81 224.4
144 1 127 11.2 121.0  5.6 27.4  9.9  9.9 4.47 162 11.1 170.7 28.72 191.2
146 1 118  7.0  97.0 -1.6 25.3 11.5 11.8 4.04 172 24.1 151.3 22.64 169.5
147 1 118  5.7  88.0 -4.1 27.7  9.0  9.1 4.02 191 17.7 218.2 22.39 240.6
159 1 133  8.9 123.5  6.7 26.3 11.5 11.6 3.06 129 22.4 153.2 24.40 167.6
163 1 143  6.7 138.5  5.3 24.5  7.7  7.7 5.29 186  6.1 192.0 30.95 200.6
169 1 127  9.2 121.5  6.1 25.5  9.8  9.8 3.72 158 18.5 181.9 29.85 214.1
171 1 130 11.2 120.5  6.7 27.3 11.4 11.4 3.39 116 11.0 167.5 30.63 197.1
178 1 130 11.2 131.0  2.5 31.7 14.0 14.3 3.26 108 28.3  56.9 39.44  64.3
180 1 127  8.3 119.5  3.7 24.1 10.8 10.8 2.94 150 19.2 186.2 22.59 213.7
183 1 122  9.4 118.0  9.6 27.1  9.7  9.9 5.13 234 19.3 212.0 23.68 245.3
184 1 118  6.8  89.5  0.4 26.1 13.1 13.3 2.43  98 19.8 182.2 25.35 196.2
188 1 132 13.4  95.0  3.0 25.5 14.1 14.3 4.89 225 29.3 245.4 30.30 249.0
192 1 124  8.9 110.0  1.5 27.7 13.6 13.9 3.02 158 39.0 164.0 24.32 192.9
200 1 133  7.4 103.5  0.9 29.7  9.2  9.7 5.30 219 13.9 132.2 29.44 144.7
210 1 122 10.8  77.0 -6.8 23.8 13.6 13.8 3.33 129  3.4 189.0 24.85 211.7
226 1 129  6.1 102.0 -0.7 26.8  7.9  8.2 2.71 114 32.3 122.2 28.29 139.1
228 1 126 11.4 154.5  4.6 31.7 11.2 11.2 3.62 142 14.2 290.8 25.84 313.3
232 1 127  6.0 144.5  3.1 29.9  8.5  8.5 3.33 139 24.6 169.5 27.14 184.0
236 1 125  5.4 109.0  2.6 26.7  8.6  9.9 3.53 128 12.9 150.7 28.00 164.9
238 1 134  8.6 135.0  2.2 30.9 13.6 14.1 4.20 185 25.0 281.9 29.74 306.0
240 1 126  8.4  74.0 -2.6 21.8 12.9 16.1 2.69 110 10.2 103.4 26.83 116.7
241 1 127 10.7  96.5 -1.0 24.9 16.5 19.5 2.90 144 35.2 153.1 27.44 175.8
```

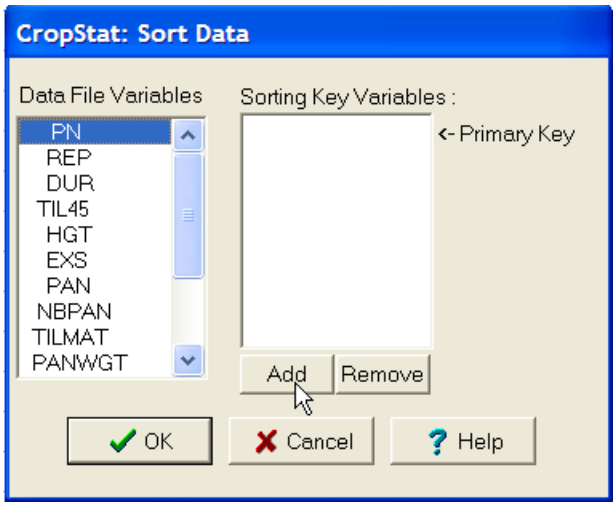
Figure 3 (CONTINUED) Phenotypic data for QTL analysis

```
 251 1 153  8.7 101.0  7.0 22.9 14.7 15.1 3.25 124  8.1 103.1 22.20 127.7  ! 97 2 146 10.6 127.0  1.3 29.7 11.5 11.5 3.27 124 18.1  26.0 32.79  30.6
 252 1 129 15.4  79.5 -1.4 25.2 18.7 18.8 2.34 103 28.1 223.9 28.74 265.8  !100 2 123  6.9 105.0 -3.3 32.0  9.4  9.5 5.00 227 20.1 149.0 23.17 163.0
 253 1 137  9.2 116.0  1.7 30.5  7.7  7.7 3.72 135 10.1  98.2 28.99 116.5  !105 2 131  6.7  94.0 -1.0 24.4 11.2 11.3 4.35 198 16.1 128.8 26.75 145.4
 262 1 131 10.3 114.0  0.0 26.5  9.5  9.5 2.78 124 11.9 131.2 25.44 143.5  !106 2 152 10.6  83.0  1.4 20.9 14.5 14.5 2.10 126 41.1  76.3 22.20  80.3
 268 1 130 10.7 152.0  8.5 29.6 11.5 11.8 3.73 121 25.2 110.5 37.19 125.8  !107 2 136  6.2 135.0  3.4 31.5  8.3  8.5 3.79 211 52.6 109.0 26.74 119.2
 269 1 130 10.2 126.0  2.0 26.1 13.2 13.9 3.18 140 15.1 268.2 22.29 300.5  !116 2 127 10.0  93.0  2.2 24.2 10.5 10.5 2.95 113 16.5 250.1 25.65 265.3
 272 1 139  9.9 104.0  0.1 24.9 13.7 14.1 4.01 245 20.0 246.1 19.25 289.5  !124 2 134 12.4  90.5  1.3 24.9 14.7 15.3 3.22 128  7.1 204.9 23.61 229.5
 281 1 137 10.1  86.5  2.4 25.6 11.4 11.4 3.77 138  9.8 246.4 28.51 278.2  !130 2 122  8.4 115.5 -0.8 32.0  8.9  9.1 5.23 192 21.9 292.2 32.57 327.3
 284 1 137 10.4 112.5  2.0 24.2 14.4 14.7 2.62 119 23.0 130.7 26.10 140.8  !144 2 129 10.1 128.0  5.8 27.7 15.7 15.7 4.90 180  8.3 370.8 27.94 396.3
 288 1 123 10.3 105.5 -2.6 25.6 12.6 12.9 2.78 148 39.6 149.2 24.39 175.5  !146 2 117  7.5  98.0 -0.2 25.3  8.1  8.1 3.31 162 29.8 180.9 23.99 194.8
 290 1 131  7.7 107.5  4.7 23.2 10.0 10.1 2.78 123 20.3 126.2 26.21 137.0  !147 2 114  4.5  85.5 -4.3 27.8  9.1  9.3 4.25 189  7.0 241.8 23.57 262.4
 295 1 128  8.8  99.5  4.5 22.5 11.4 11.7 2.61 113 24.6 141.0 26.43 158.0  !159 2 137  6.4 129.0  7.5 26.9 16.2 16.8 2.86 119 12.0 223.8 24.78 252.7
 303 1 137 11.2 120.5  4.2 23.6 11.7 12.3 2.98 125 10.2 179.2 24.87 190.1  !163 2 143  5.4 147.0  5.2 26.0  7.6  7.6 5.03 166  4.6 137.6 28.97 173.7
 313 1 137  8.2  99.0 -0.1 24.1  8.0  8.3 3.49 146 16.5 120.6 28.68 131.9  !169 2 127 12.4 134.0  5.9 26.3 11.1 11.1 5.09 190 25.2 224.1 28.26 265.9
 326 1 125  9.9 108.5  3.0 25.6 16.0 16.0 3.79 180  5.8 283.2 24.42 325.1  !171 2 131 11.2 109.0  6.9 27.9 10.0 10.9 3.07 105 11.9  80.7 30.48  88.3
 329 1 133  8.6 101.0  3.9 24.0  8.9  9.0 5.68 322 13.1 165.9 21.65 183.0  !178 2 132 10.1 132.5  2.8 31.8 13.8 14.0 3.25 106 30.7  38.2 39.84  43.1
 331 1 131 11.7 104.0  5.9 24.7 10.9 11.2 4.63 263 21.7 181.3 19.67 188.1  !180 2 126 11.1 122.0  2.9 24.3 11.3 11.3 3.22 144 17.7 212.4 22.46 247.9
 332 1 132  8.9 104.5  2.6 28.3 14.0 15.2 4.72 238 27.5 127.7 25.54 139.7  !183 2 118  7.1 114.5  6.9 26.9  8.7  9.1 5.00 258 29.1 205.0 23.16 220.8
 333 1 114  7.7  75.0 -4.2 23.6 12.3 14.2 3.50 159 38.7 104.3 32.08 130.4  !184 2 114  6.7  84.0 -0.6 26.5 12.1 12.3 2.23 101 25.8 146.3 25.45 160.0
 335 1 125 11.5 119.5  5.0 26.8 10.5 11.0 3.97 151 24.1 213.2 26.68 238.8  !188 2 132 13.9  90.0  2.2 25.1 11.3 11.6 5.34 225 24.8 295.3 28.06 318.1
 336 1 130  9.8 136.0  2.6 31.3  8.9  9.0 3.08 163 41.0  98.9 25.36 110.8  !192 2 125  8.2 107.5  1.0 27.7 11.9 12.4 2.94 156 36.7 166.2 25.54 197.2
 351 1 145 10.4  93.0  3.5 23.8 12.9 13.1 3.72 177 12.7 157.1 22.77 196.4  !200 2 131  5.2  97.0 -2.2 27.8  7.7  8.0 5.21 201 28.9  70.1 28.17  72.7
 355 1 134  9.0 122.0  3.3 24.1 13.0 13.0 3.33 142 23.1 221.3 29.12 242.1  !210 2 123 10.3  73.5 -4.5 23.0 14.9 15.1 3.08 121  7.1 213.9 24.49 247.5
 356 1 126  9.1  90.0 -1.8 23.8  9.8 10.2 4.12 161 11.0 142.3 27.19 164.6  !226 2 128  5.5 104.0 -1.1 26.2  9.0  9.5 3.46 150 32.5  87.0 29.38 101.5
 372 1 133  7.4 107.5  7.0 21.6  9.8 11.0 2.16 124 44.2  65.9 24.82  74.5  !228 2 122  7.2 159.5  3.5 29.8 11.6 11.7 3.43 161 20.9 197.3 26.85 219.2
   5 2 135  8.9  89.0 -0.4 25.7 14.1 17.1 1.71  91 36.3  69.9 24.91  77.1  !232 2 122  8.4 161.0  4.0 31.3 11.1 11.1 3.24 144 22.6 193.1 26.44 211.3
   7 2 116  5.6  63.5 -8.2 21.5 12.7 13.6 2.76 117  7.5 144.7 23.69 154.6  !236 2 123  8.3 102.5  1.7 27.3 10.7 12.6 2.65 215 68.8  82.7 27.15  93.4
  10 2 138  6.3 132.5  2.6 31.3 10.2 10.2 3.35 126 24.8 125.4 30.42 142.7  !238 2 138  8.2 134.0  3.4 30.6 12.9 13.3 4.61 159 13.5 272.1 26.09 302.3
  12 2 118  7.8 107.0  5.6 24.8 13.2 13.2 3.81 203 19.9 224.6 26.05 241.9  !240 2 126 10.4  76.0 -1.6 21.8 13.7 18.2 2.51 104 24.4 158.0 25.46 179.8
  13 2 123  8.0  66.5 -9.4 25.4 13.8 14.3 3.13 155 22.0 156.3 25.42 179.4  !241 2 122  8.0  92.0 -1.0 25.8 12.7 13.6 1.95 125 55.1  77.7 30.96  92.2
  20 2 141 11.7  97.5  0.3 24.9 16.1 16.1 3.22 159 21.3 100.5 24.32 131.5  !251 2 147  7.7  96.0  7.3 22.3 16.0 16.0 3.17 173 18.9  96.5 22.57 127.5
  22 2 130  5.9  72.5 -1.0 23.6 18.3 19.4 1.68  92 27.8 153.3 21.72 177.4  !252 2 127 11.4  77.0 -0.4 24.9 16.8 17.3 2.11 100 39.9 146.6 28.42 186.6
  25 2 120  8.8 107.5 -2.2 28.4 13.7 13.7 2.85 120 27.5 262.8 29.43 283.0  !253 2 138  7.7 116.0  0.6 31.2  8.4  8.5 3.83 151 17.5 149.3 27.94 163.4
  27 2 123  9.9  61.0 -3.0 18.7 18.4 23.1 1.58  78 15.9  40.1 24.48  46.4  !262 2 131  9.1 106.5  1.4 28.4  9.2  9.2 2.71 108  7.6 147.3 24.73 165.0
  31 2 125  8.0  89.5 -3.5 25.0 12.7 13.0 3.95 159  8.4 260.0 26.21 308.5  !268 2 127  9.1 138.5  8.1 28.5  8.5  8.9 3.11 117 26.3  66.8 33.76  78.6
  33 2 133  3.6  74.0 -7.1 23.8  9.8 14.1 3.00 164 29.6 100.7 24.63 109.3  !269 2 130 10.4 120.0  2.2 26.5 12.4 13.0 3.24 163 14.3 228.0 21.94 245.5
  34 2 126 11.0  93.0 -2.3 26.8 12.1 14.4 3.60 178 44.1 182.5 28.38 212.9  !272 2 138  9.9  98.5  0.2 24.4 13.2 13.4 3.06 199 21.1 150.7 20.25 155.1
  35 2 122  6.6  59.0 -7.4 22.0 13.9 14.2 2.79 160 17.5 160.8 20.45 189.2  !281 2 137  9.8  88.5  3.7 26.1  9.9 10.0 3.95 150 19.9 174.8 27.79 184.0
  39 2 139 12.1 131.5  4.8 27.6 13.0 13.2 3.34 156 18.3 122.2 24.57 136.9  !284 2 134 11.8 117.5  5.6 24.7 12.5 12.5 2.43 110 19.9 148.5 27.38 170.4
  41 2 125  8.5 112.0  2.7 26.0 10.1 10.2 3.82 148 16.7 124.7 30.37 143.1  !288 2 124  9.7 113.0 -2.9 27.8 12.1 13.6 3.98 197 30.4 111.1 25.04 132.9
  47 2 127  5.9 121.5  1.2 26.3 10.6 11.4 3.37 159 19.9 187.5 24.09 203.5  !290 2 132  4.9 104.0  4.7 25.1  9.2  9.9 2.53 121 34.6 134.1 25.62 139.1
  48 2 137  7.1  86.5 -2.0 23.3  9.6  9.8 3.48 181 13.1 130.0 22.40 137.9  !295 2 129  6.9 101.5  4.9 23.9 12.9 13.0 2.41 101 25.8 138.3 28.73 154.9
  55 2 127 11.2 111.0  1.7 28.3 14.7 15.1 3.06 101 11.6 103.4 26.58 117.7  !303 2 137  7.8 108.5  4.4 24.3  9.9 10.6 2.36 106 11.5 141.8 24.53 149.3
  57 2 127  9.3  81.0  0.9 23.8  9.6  9.6 4.16 180 12.1 207.2 23.59 221.4  !313 2 137  7.7 100.0 -0.6 24.3  8.8  9.7 3.04 124 23.1 139.3 28.45 151.2
  62 2 120  9.5 110.5 -5.0 26.7 10.1 11.4 3.14 200 58.4  83.8 24.84  92.4  !326 2 122  8.6  89.0  1.0 23.8 11.2 11.2 2.64 118  7.2 201.4 22.19 220.3
  67 2 139  9.3 139.0 -0.9 32.0  9.9 10.0 3.06 170 44.7 171.5 28.10 180.6  !329 2 133  8.8  97.0  3.4 24.5  9.8  9.8 4.79 261 13.9 217.6 21.28 229.1
  69 2 146  9.0 134.5  1.5 28.3  9.7  9.9 3.06 126 25.6  52.2 28.92  61.9  !331 2 137  8.7 106.0  6.4 24.8 11.1 11.4 5.55 284 17.7  29.0 21.80  31.0
  74 2 126  7.5 111.5  0.2 27.2 12.8 14.3 3.55 131 21.3 165.9 29.06 195.2  !332 2 131  7.6 101.5  3.1 29.0 13.8 14.0 4.21 217 29.1  49.4 24.59  54.0
  82 2 129 10.7 113.0  2.6 27.0 16.5 18.1 3.56 151 31.4 197.3 29.74 221.0  !333 2 114  7.1  66.0 -4.0 22.2 13.5 14.6 2.48  78 40.6  83.8 31.93  97.8
  84 2 125 10.7  97.0  2.2 29.6 12.6 13.8 3.52 127 20.2 198.3 32.91 223.9  !335 2 129 12.8 118.5  6.2 25.5 10.8 10.8 3.82 158 30.3 168.0 25.34 194.4
  87 2 137  8.7  92.0  2.4 23.9 10.4 11.1 3.67 141  9.4 210.1 28.07 213.1  !336 2 129  8.7 144.5  6.2 33.2  9.2  9.5 3.23 162 34.6 183.6 28.04 205.6
  88 2 122 12.6  85.0  3.0 22.0 17.4 17.6 2.98 123 18.5 296.7 27.75 343.3  !351 2 147  8.5  99.0  6.7 23.2 13.2 13.2 3.04 181 27.3  54.2 21.72  63.8
  89 2 132  9.3  84.5 -1.6 24.2 12.1 12.1 3.53 148 16.9 165.3 27.93 180.8  !355 2 133  8.7 115.5 -0.6 24.6 10.2 10.4 3.24 155 29.8 134.7 25.57 150.9
  90 2 125  8.8  55.0 -6.9 22.5 11.9 12.3 2.61 111 11.9 147.7 27.27 165.5  !356 2 126  6.9  87.0  0.0 23.5  9.7 10.8 3.70 148  6.8  73.7 27.11  85.3
  94 2 131  7.1  94.0 -2.8 26.7 11.7 12.3 2.74 142 39.2  88.7 25.55  98.6  !372 2 130  6.3 104.5  5.6 22.6  9.4 11.5 1.60  87 46.6  66.0 24.51  75.7
```

The next step in data preparation is to merge the marker and phenotypic data. The Merge Data Files Item on the CropStat Data menu requires that both files contain a common key variate and that they are sorted in ascending order of values of the key variate. From Figures 2 and 3 it is clear that the QTLMKR and QTLYLD data sets have a common variate PN which is suitable for a key variate, and QTLMKR is sorted on this variate, but QTLYLD is not.

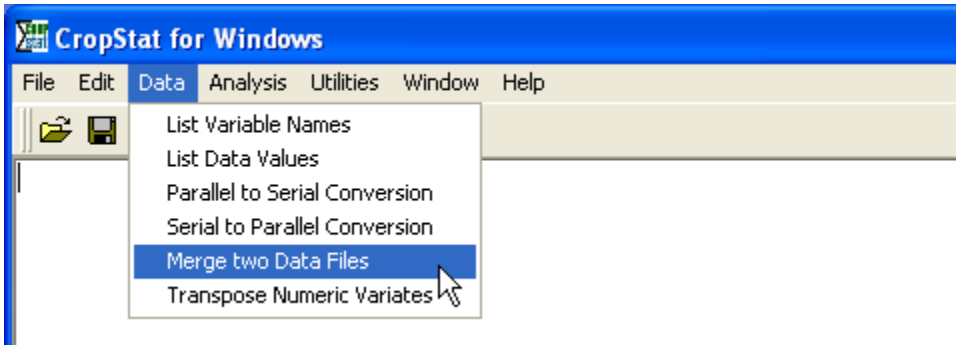
- To sort QTLYLD, first open the file in the Data Editor, and then select the **Sort** from the **Options Menu.**

- Under **Data File Variables** select PN. Then click **Add** under the Sorting Key Variables box.

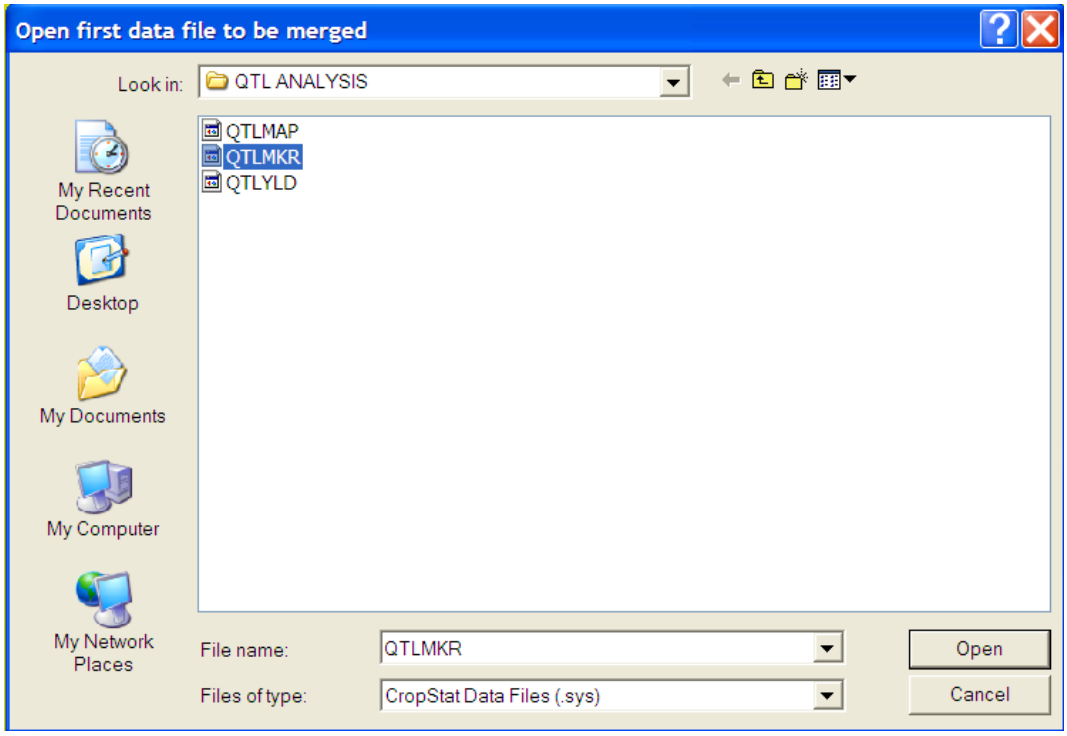


- Click **Ok** then save the changes to the original file and close the Data Editor.

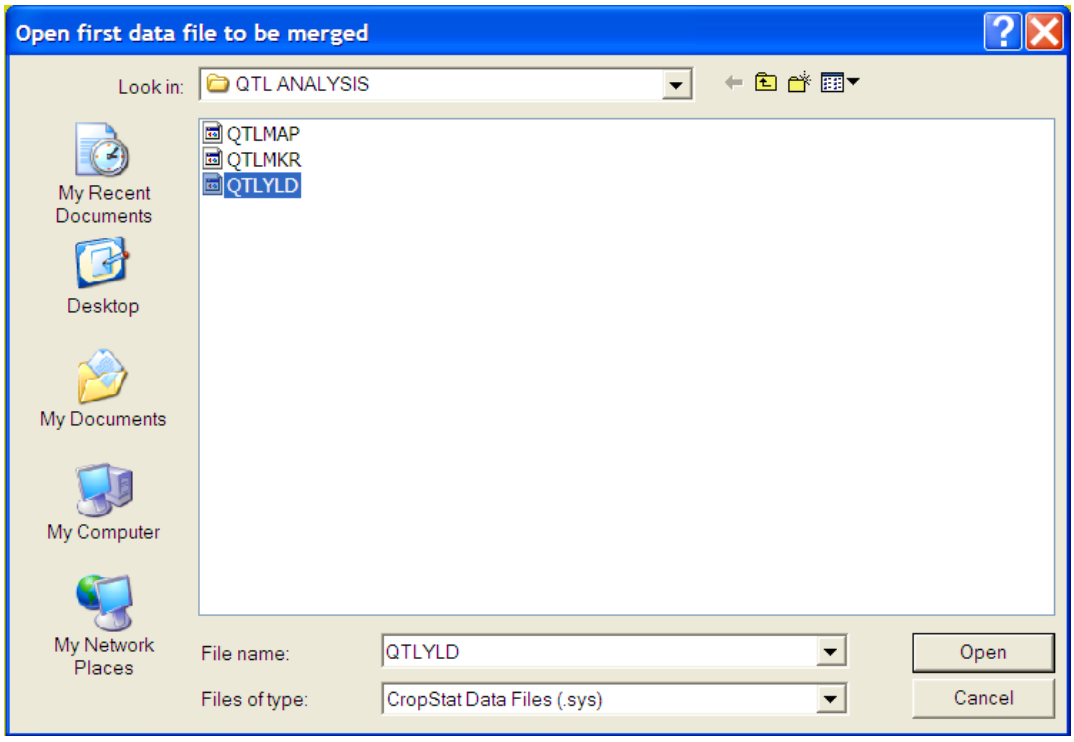
- To merge QTLMKR and QTLYLD select the **Merge two Data Files** item from the **Data menu** using the CropStat Main Window.



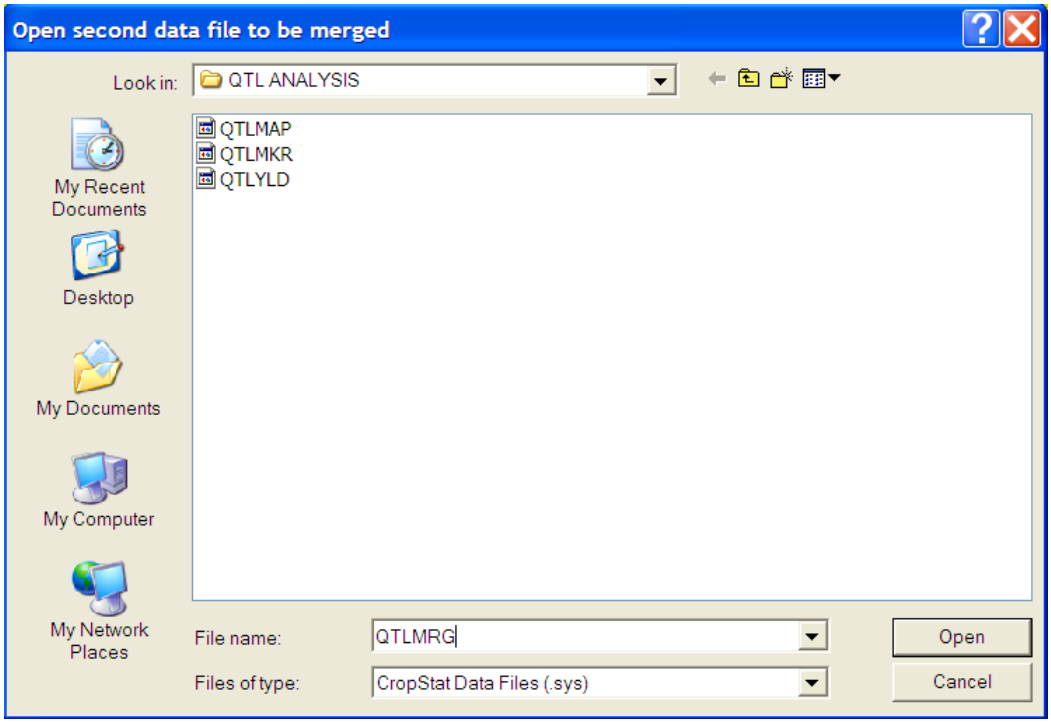
- Select *QTLMKR.SYS* from the list of data files. Click **Open**.



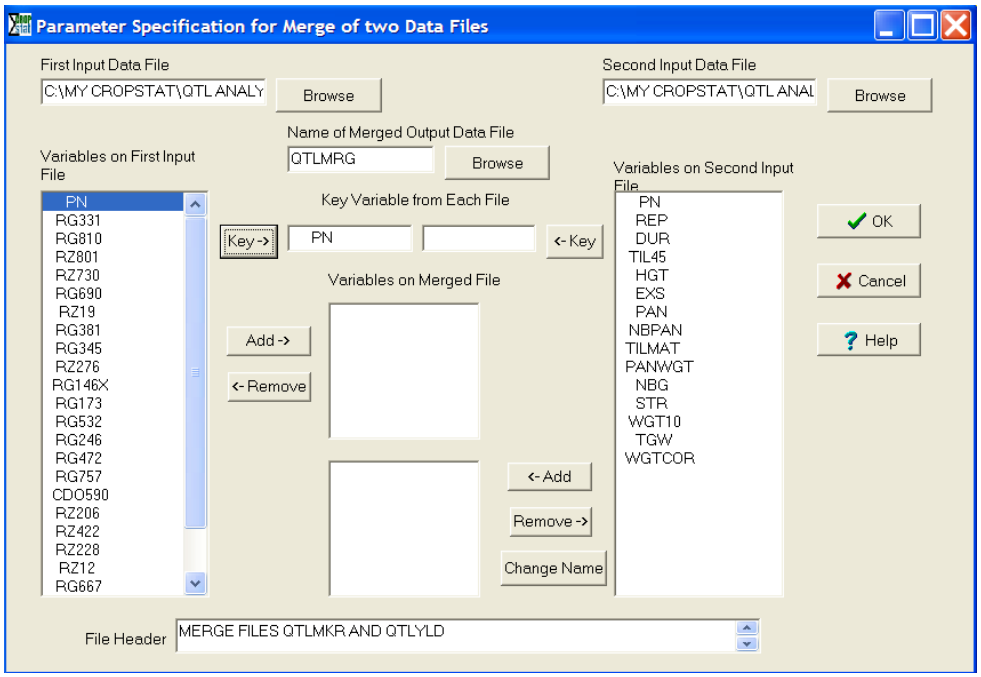
- Select QTLYLD.SYS as the second data file to be merged. Click **Open**.



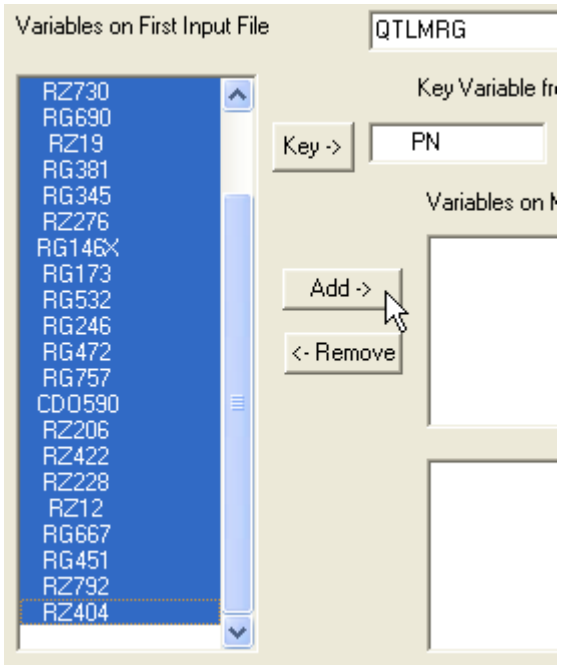
- Enter QTLMRG in the File name textbox as the name for the output file. Click **Save**.



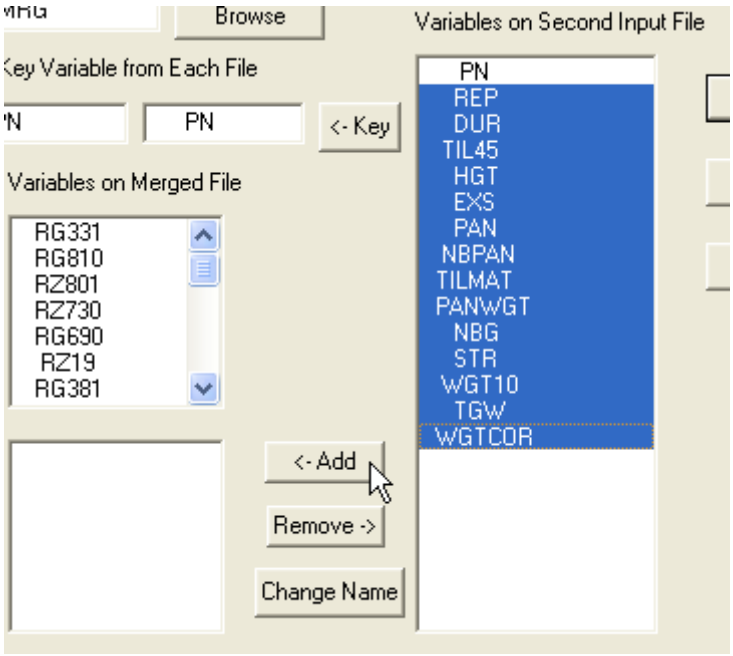
- Under the **Variables on First Input File** select PN and click **Key->**. Do the same for the Second Input Data File.



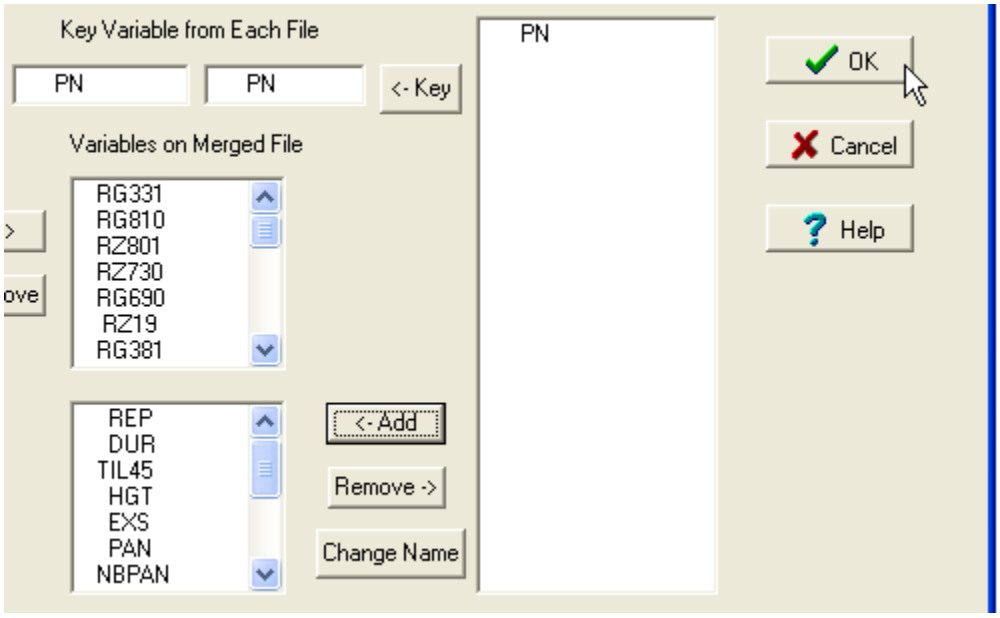
- Under the **Variables on First Input File** select all variates then click **Add->**.



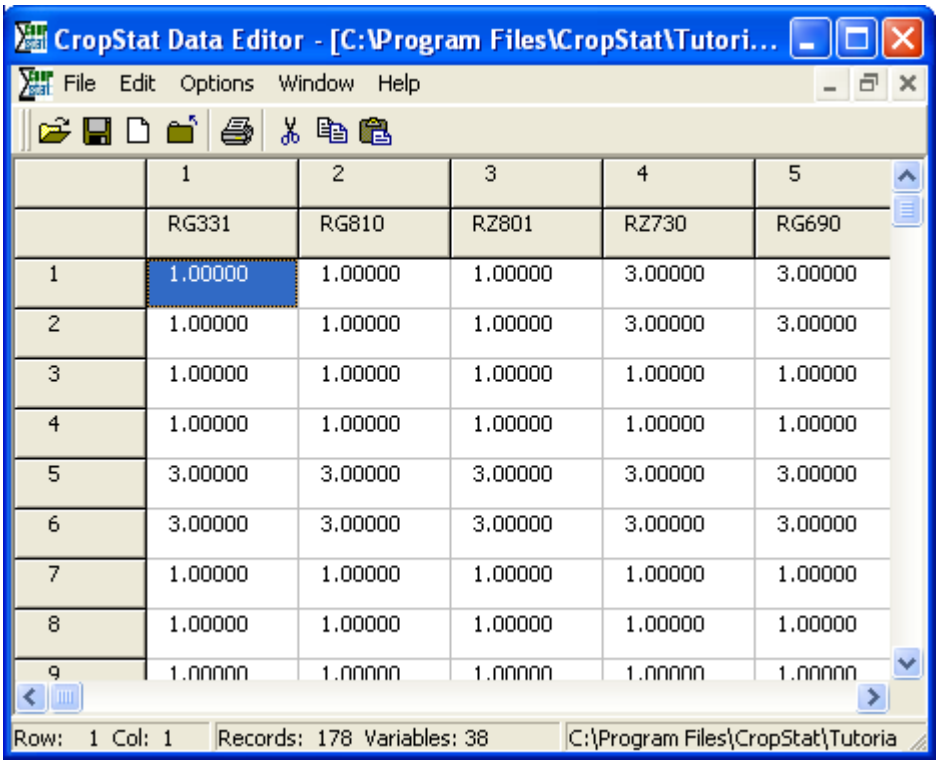
- For the Second Input Data File, select variates 2 to 15 (excluding PN) than click **<- Add**. PN was excluded since it is already selected from the First Input Data File.



- Click **OK**.


Key Variable from Each File
PN
PN
<- Key
Variables on Merged File
RG331
RG810
RZ801
RZ730
RG690
RZ19
RG381
REP
DUR
TIL45
HGT
EXS
PAN
NBPAN
<- Add
Remove ->
Change Name
PN
OK
Cancel
Help

- The merged file will open in the Data Editor. Note that each line of the QTLMKR data has been duplicated for the second replication of the QTLYLD data. Note also that where marker data is available for plants without phenotypic data, missing values have been entered for the phenotypic data.



	1	2	3	4	5
	RG331	RG810	RZ801	RZ730	RG690
1	1.00000	1.00000	1.00000	3.00000	3.00000
2	1.00000	1.00000	1.00000	3.00000	3.00000
3	1.00000	1.00000	1.00000	1.00000	1.00000
4	1.00000	1.00000	1.00000	1.00000	1.00000
5	3.00000	3.00000	3.00000	3.00000	3.00000
6	3.00000	3.00000	3.00000	3.00000	3.00000
7	1.00000	1.00000	1.00000	1.00000	1.00000
8	1.00000	1.00000	1.00000	1.00000	1.00000
9	1.00000	1.00000	1.00000	1.00000	1.00000

III. Model Selection

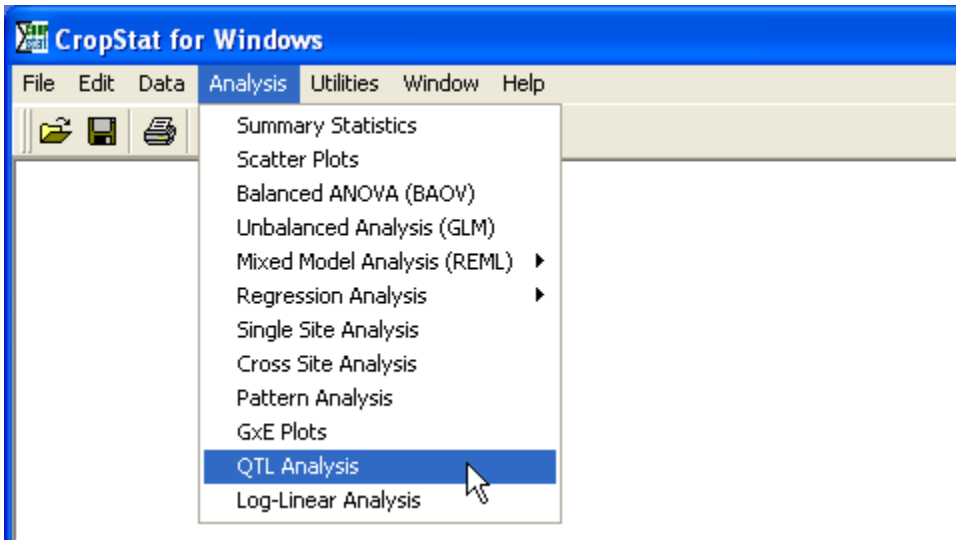
The factors classifying the phenotypic data for each marker or marker pair are: block or replication (BLK), marker (MKR) and genotype within marker (PN). We can consider blocks as a fixed effect since we are not interested in estimating responses over a population of possible blocks, the markers are fixed since there is no question of their being a random sample of possible markers in this case, and finally the plant genotypes are random since they represent a sample of the possible gamete genotypes derived from the F_1 through recombination and mutation and we are interested in the effects of any QTL over the population of such genotypes.

Possible effects in any linear model of marker on phenotypic response are BLK, MKR, BLK.MKR (block by marker interaction), PN/MKR (genotype in marker) and BLK.PN/MKR. The last term represents interactions between plant genotypes and blocks and is assumed to be due to measurement and experimental error. It does not therefore need to be explicitly stated in a model and is usually computed as a RESIDUAL effect. The BLK.MKR effect could be similarly omitted.

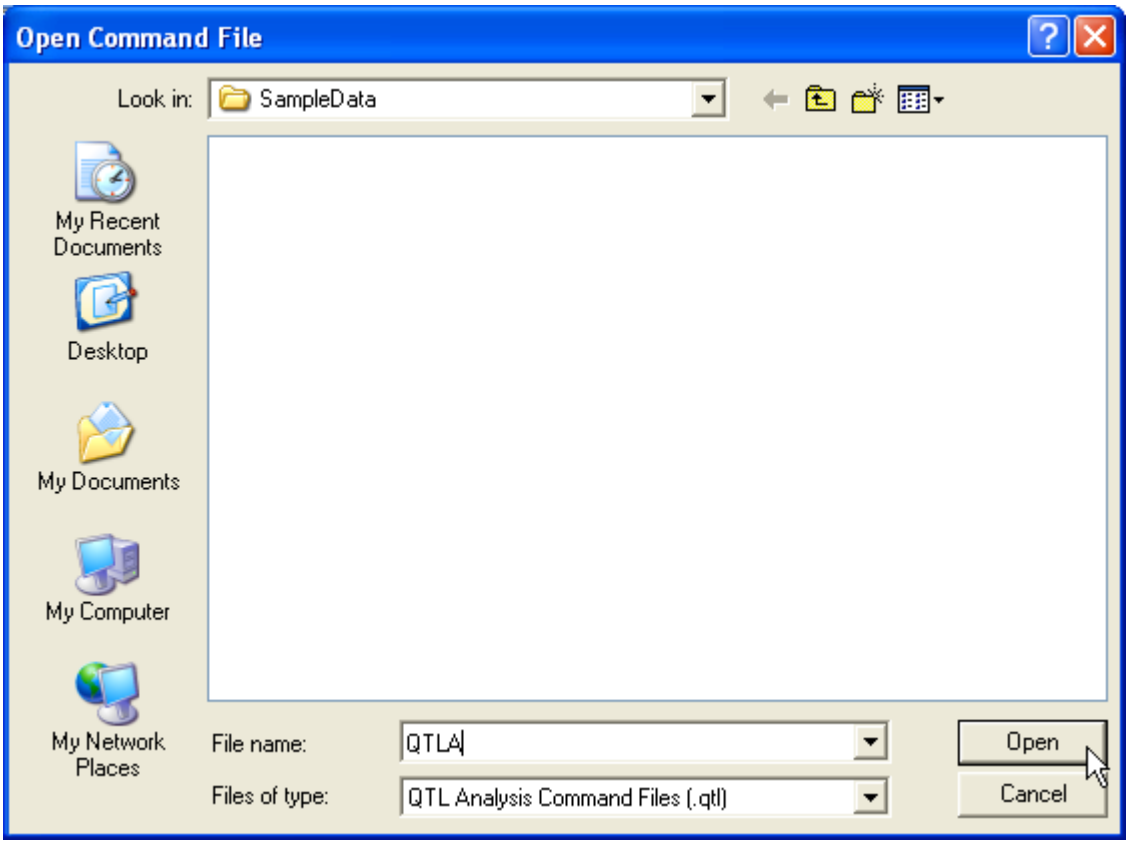
The interest in QTL analysis is in testing the effect MKR, or sub-effects of it. We therefore need to decide on the appropriate error term for comparisons between marker means. From a logical point of view it is clear that even if there is no MKR effect, differences between plant genotypes will be included in differences between marker class means (since one plant genotype cannot be represented across marker classes except perhaps by construction of isogenic lines). The BLK.PN/MKR or RESIDUAL effect does not contain differences between plant genotypes and is hence not suitable as an error term.

IV. QTL location by single marker ANOVA

- To detect QTLs by single marker ANOVA select **QTL Analysis** from the **Analysis** menu.



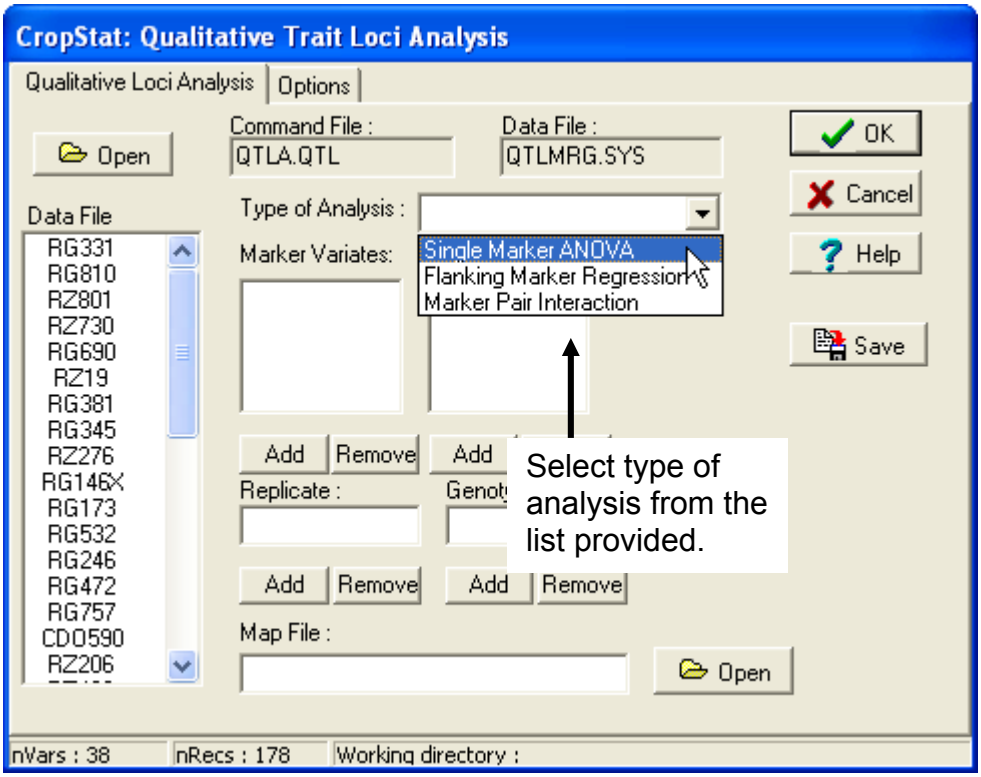
- Clik the **Look in** box and go inside you working folder *C:\MY CROPSTAT QTL ANALYSIS*, specify *QTLA.QTL* as command file name and click **Open**.



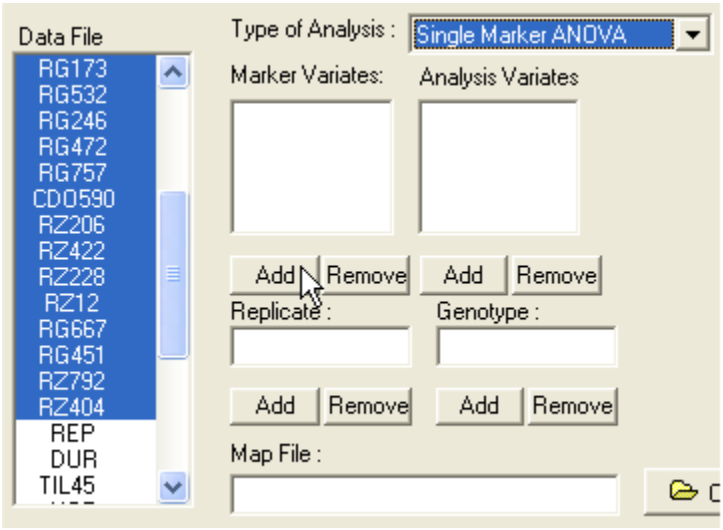
- Since the file *QTLA.QTL* does not exist CropStat will ask if you wish to create the file. Click **Yes.**

- In the **Open file with marker genotypes and phenotype values** dialog box, select *QTLMRG.SYS* as data file. Click **Open**.

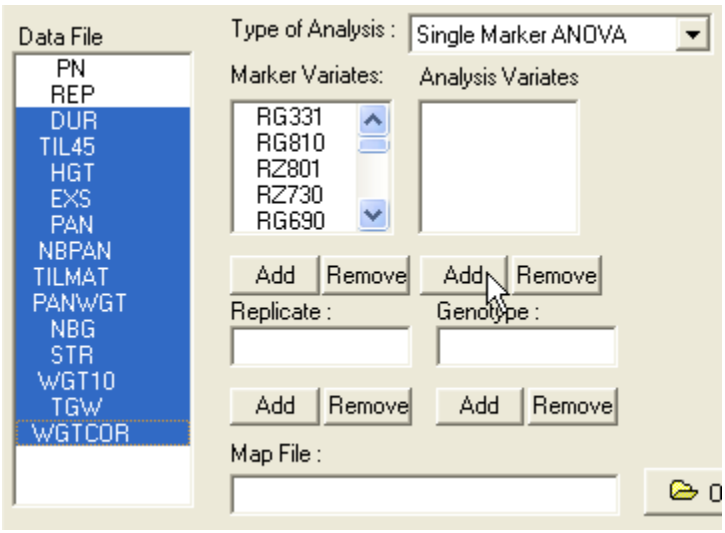
- To specify the **Type of Analysis** to perform, select *Single Marker ANOVA* from the list.



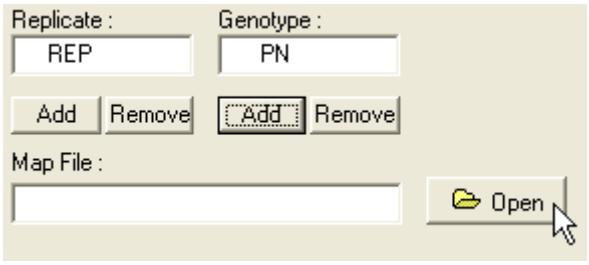
- Select the marker variates RG331 to RZ404 from the Data File and add them to the *Marker Variates* List.



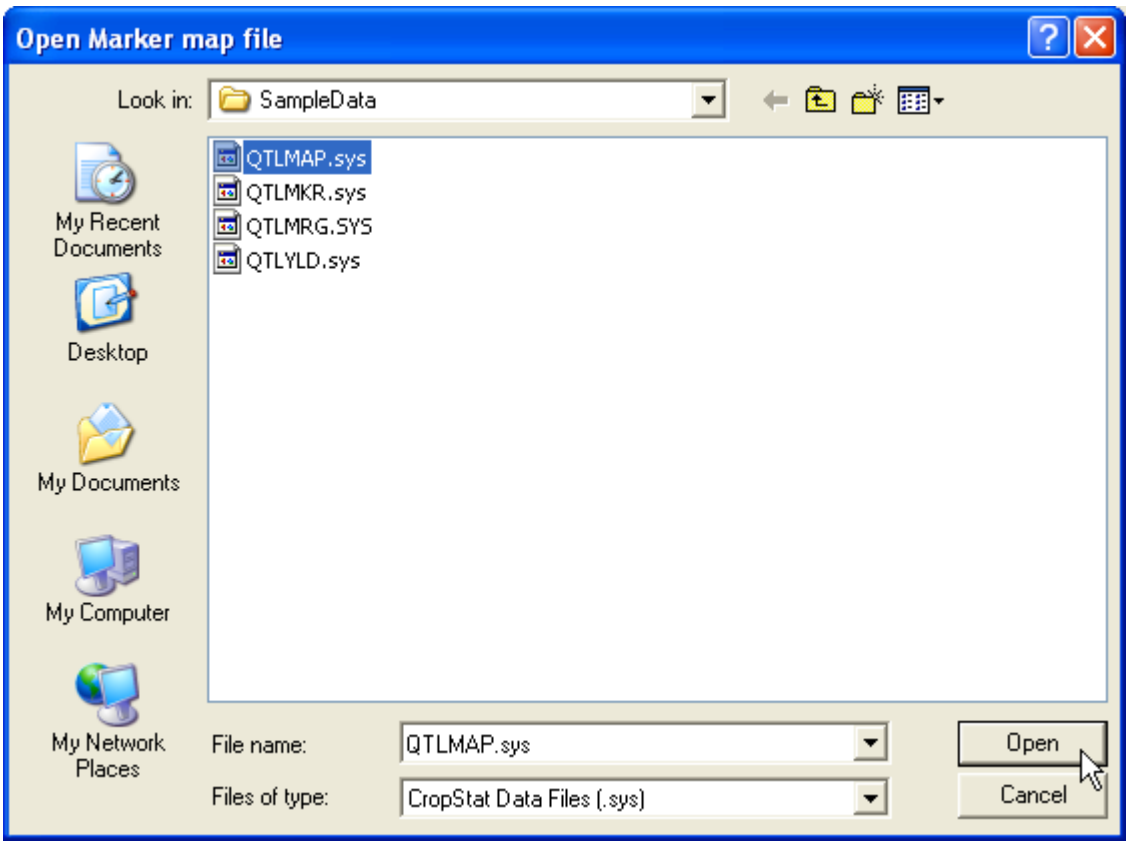
- Select the phenotypic variates, DUR to WGTCOR and add them to the *Phenotypic Analysis Variates* List.



- Add REP to the *Replicate Box* and PN to the *Genotype* Box to specify the variate which identifies genotypes in markers since the data is from replicated double haploid plants.

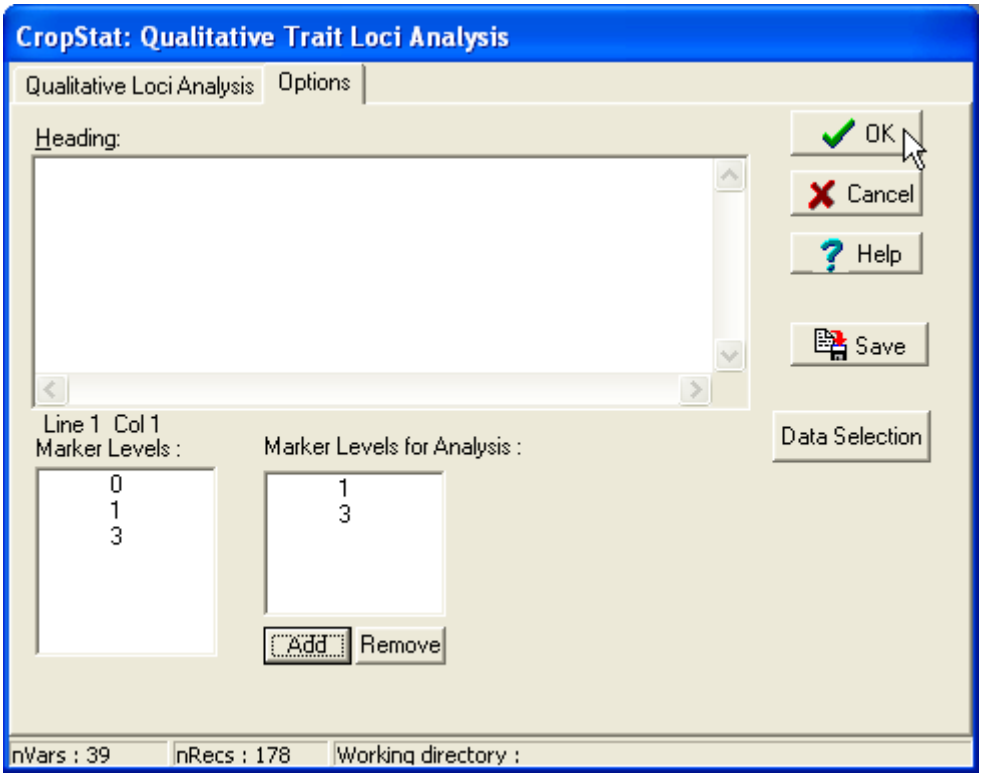


- Click the **Open** button adjacent to the *Map File* box to specify the Data file containing the marker map. Select *QTLMAP.SYS* from the list of files then click **Open**.



- Then on the **Options** page enter QTL Analysis Tutorial in the Title Box and **Add** 1 and 3 to the list of **Marker levels for Analysis**.

- Click **OK** to execute the analysis. A sample output is shown in the next section.



V. Sample Output

Partial output from single marker ANOVA

```
QTLA - SINGLE MARKER QTL ANOVA  FILE QTLMRG   5/10/ 4 16: 5
------------------------------------------------------------------ :PAGE   1
QTL LOCATION BY ANOVA FOR MARKER RG331         166 OBSERVATIONS
FACTOR RG331 HAS  2 LEVELS
FACTOR REP HAS  2 LEVELS
FACTOR PN HAS 38 LEVELS WITHIN LEVEL 1 OF FACTOR RG331
FACTOR PN HAS 45 LEVELS WITHIN LEVEL 3 OF FACTOR RG331

ANOVA RESULTS FOR 13 VARIATES WITH MARKER GENOTYPE DEFINED BY RG331
SOURCE:      REPSS      MKRSS        REP*MKR  PINMKR1   PINMKR2    RESIDUAL      MEAN      QTLEFF     SE      FPROB  MARKER
    DF:          1          1          1        37         44         81
DUR          0.6024      92.64      2.154      4520.      5025.      267.2      129.6      0.7497    0.8456    0.382  RG331
TIL45         11.61      18.53      11.12      291.0      234.5      128.9      8.931      0.3353    0.1984    0.091  RG331
HGT           170.0     0.2130E+05  68.39     0.2060E+05 0.3631E+05  2221.      104.3       11.37     2.065    0.000  RG331
EXS           1.320      221.9     0.1546      1280.      735.5      101.5      1.070      1.160     0.3886    0.004  RG331
PAN          0.2169      202.4     0.3024E-02  443.6      840.6      45.81      26.05      1.108     0.3102    0.001  RG331
NBPAN        0.6897      58.44      5.245      531.5      320.5      179.1      11.82     -0.5955    0.2526    0.020  RG331
TILMAT       0.5789E-01  129.6      7.934      822.4      343.8      213.7      12.38     -0.8869    0.2955    0.004  RG331
PANWGT        2.352      1.107     0.9231E-02  64.78      57.38      13.47      3.481      0.8194E-01 0.9566E-010.399  RG331
NBG           892.9      1077.      305.9     0.1628E+06 0.1248E+06 0.2767E+05  154.0      -2.556     4.642    0.590  RG331
STR           345.5      182.1      5.424      7231.      9608.      3644.      22.28      1.051     1.123    0.355  RG331
WGT10        0.1550E+05  2596.     0.1591E+05 0.2369E+06 0.3597E+06 0.1482E+06  161.3      3.969     6.685    0.562  RG331
TGW           7.524      302.4      13.07      594.6      1084.      106.6      26.64      1.355     0.3546    0.000  RG331
WGTCOR       0.2293E+05  2161.     0.2509E+05 0.2921E+06 0.4247E+06 0.1877E+06  181.1      3.621     7.327    0.628  RG331

QTLA - SINGLE MARKER QTL ANOVA  FILE QTLMRG   17/ 2/ 4  8:59
------------------------------------------------------------------ :PAGE   2
QTL LOCATION BY ANOVA FOR MARKER RG810         170 OBSERVATIONS
FACTOR RG810 HAS  2 LEVELS
FACTOR REP HAS  2 LEVELS
FACTOR PN HAS 42 LEVELS WITHIN LEVEL 1 OF FACTOR RG810
FACTOR PN HAS 43 LEVELS WITHIN LEVEL 3 OF FACTOR RG810

ANOVA RESULTS FOR 13 VARIATES WITH MARKER GENOTYPE DEFINED BY RG810
SOURCE:      REPSS      MKRSS        REP*MKR  PINMKR1   PINMKR2    RESIDUAL      MEAN      QTLEFF     SE      FPROB  MARKER
    DF:          1          1          1        41         42         83
DUR          0.1471      289.4      0.7195      5518.      3941.      277.6      129.6      1.305     0.8188    0.111  RG810
TIL45         13.05      2.097      3.872      296.7      248.7      138.1      8.911      0.1111    0.1966    0.581  RG810
HGT           160.1     0.2535E+05  81.84     0.2424E+05 0.2961E+05  2236.      104.6       12.21     1.954    0.000  RG810
EXS           1.237      287.9      0.1103      1257.      764.3      102.1      1.143      1.301     0.3786    0.001  RG810
PAN          0.2562      180.3      0.1312      471.3      841.0      45.82      26.08      1.030     0.3050    0.001  RG810
NBPAN        0.3488      91.64      15.90      488.4      340.2      181.3      11.85     -0.7342    0.2423    0.003  RG810
TILMAT       0.2118E-02  176.9      19.24      774.5      352.1      217.1      12.41      -1.020    0.2826    0.001  RG810
PANWGT        2.362      1.476     0.3270E-03  62.04      60.77      13.49      3.469      0.9320E-01 0.9330E-010.322  RG810
NBG           927.1      29.16      618.8     0.1554E+06 0.1351E+06 0.2737E+05  153.7     -0.4142    4.538    0.925  RG810
STR           302.6      145.0      5.319      8209.      9201.      3744.      22.40      0.9237    1.111    0.413  RG810
WGT10        0.1380E+05  82.44     0.1915E+05 0.2717E+06 0.3288E+06 0.1492E+06  161.7      0.6964    6.524    0.912  RG810
TGW           6.696      217.9      18.23      802.3      974.8      103.1      26.60      1.132     0.3549    0.002  RG810
WGTCOR       0.2039E+05 0.7738     0.2905E+05 0.3350E+06 0.3859E+06 0.1900E+06  181.6     -0.6747E-01  7.149    0.989  RG810
```

Since there are more than twenty pages of output in QTLA.OUT it is convenient to zero in on the significant tests. To do this list the most significant results from the SYS format output in file QTLA.SYS. Select **List Data Values** from the CropStat Data Menu, enter QTLA for file name, add MARKER$, TRAIT$, FPROB, QTLEFF, SE, CHRMSM and POSN to the list of **Output data Variables**. Click **Ok** to view the results as in Figure 5.

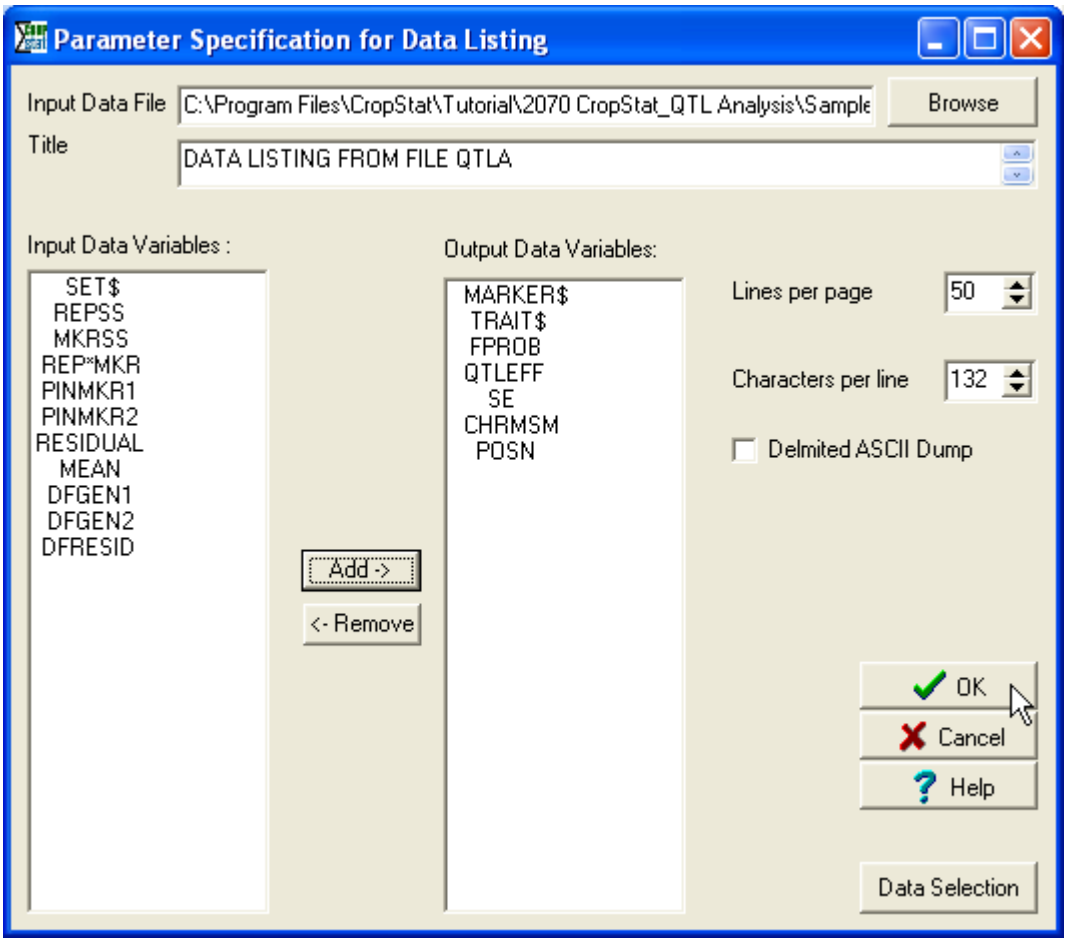


Figure 5. Selected records from single marker ANOVA output

```
 HEADINGS AND VARIATE DESCRIPTIONS  FILE QTLA   5/10/ 4 16: 9
 ------------------------------------------------------------------ :PAGE   1
                          DATA LISTING FROM FILE QTLA

 H SINGLE MARKER QTL LOCATION FOR DATA IN FILE QTLMRG
 V001 SET$ SELECTION SET ID (CMD FILE NAME / NNN)
 V002 MARKER$ MARKER NAME
 V003 TRAIT$ NAME OF THE PHENOTYPIC TRAIT ANALYSED
 V004 REPSS BLOCKS SS OR ZERO IF NO REP VARIATE
 V005 MKRSS SS BETWEEN MARKER CLASSES
 V006 REP*MKR REP BY MARKER SS
 V007 PINMKR1 SS BETWEEN PLANTS WITHIN MARKER CLASS 1
      ZERO IF NO PLANT IDENTIFIER SPECIFIED
 V008 PINMKR2 SS BETWEEN PLANTS WITHIN MARKER CLASS 2
 V009 RESIDUAL RESIDUAL SS
 V010 MEAN MEAN VALUE OF THE PHENOTYPIC TRAIT
 V011 GENEFF LS DEVIATION OF MARKER CLASS 1 FROM MEAN
 V012 SE STANDARD ERROR OF THE GENOTYPE DEVIATION
 V013 FPROB F TEST OF GEN-MS/(PN-GEN1+PN_GEN2)
 V014 DFGEN1 DF WITHIN MARKER CLASS 1
 V015 DFGEN2 DF WITHIN MARKER CLASS 2
 V016 DFRESID RESIDUAL DF
 V017 CHRMSM CHROMOSOME NUMBER OF MARKER
 V018 POSN POSITION OF MARKER ON CHROMOSOME

DATA LISTING  FILE QTLA   5/10/ 4 16: 9
 ------------------------------------------------------------------ :PAGE   2
                          DATA LISTING FROM FILE QTLA

  MARKER$      TRAIT$      FPROB      QTLEFF          SE      CHRMSM        POSN
 RG331       DUR          0.38165     0.74971     0.84556      1.0000        0.00000
 RG331       TIL45        0.90873E-01 0.33534     0.19839      1.0000        0.00000
 RG331       HGT          0.14764E-05  11.369      2.0646      1.0000        0.00000
 RG331       EXS          0.38309E-02  1.1602     0.38859      1.0000        0.00000
 RG331       PAN          0.71459E-03  1.1082     0.31016      1.0000        0.00000

 RG331       NBPAN        0.19819E-01-0.59547     0.25262      1.0000        0.00000
 RG331       TILMAT       0.36753E-02-0.88686     0.29555      1.0000        0.00000
 RG331       PANWGT       0.39862     0.81937E-01 0.95658E-01  1.0000        0.00000
 RG331       NBG          0.59025     -2.5558      4.6419      1.0000        0.00000
 RG331       STR          0.35473      1.0510      1.1231      1.0000        0.00000

 RG331       WGT10        0.56162      3.9686      6.6846      1.0000        0.00000
 RG331       TGW          0.34110E-03  1.3545     0.35459      1.0000        0.00000
 RG331       WGTCOR       0.62821      3.6210      7.3273      1.0000        0.00000
```

ANALYSIS OF CATEGORICAL DATA

I. Introduction

Statistical methodology for categorical data has only recently reached the level of sophistication achieved early in the century by methodology for continuous data. The recent development of methods for categorical data was stimulated by the increasing sophistication of techniques used in the social and biomedical sciences. Though categorical scales are most common in the social and biomedical sciences, they are by no means restricted to those areas. They also occur frequently in other areas including agriculture.

II. Definition of Categorical Data

A categorical variable is one for which the measurement scale consists of a set of categories. For instance, gender can be classified as male or female; plant type can be classified as traditional or modern; and plant variety can be resistant, moderately resistant or susceptible to a particular disease; farmers can be classified as owner-operator, share-rent farmer or fixed-rent farmer.

III. Levels of measurement

There are four levels of measurement - nominal, ordinal, interval, and ratio.

Nominal scale

Measurement at its weakest level exists when numbers or other symbols are used simply to classify observations. The levels or values of nominal variables do not have a natural ordering. The statistical analysis should be in variant to the order of listing of categories.

Example: Brand of fertilizer
Weeding method (hand-weeding, mechanical weeding)
Method of planting (direct-seeded, transplanted)

Ordinal scale

Categorical variables which have ordered values are called ordinal. However, the distances between categories are not defined or are unknown.

Example: Social class (upper, middle, lower)
Maturity class (early, medium, late)
Severity of infection (0, 1, 3, 5, 7, 9)

Interval scale

When a variable has all the characteristics of an ordinal scale but in addition, the distances between any two values on the scale are known then the variable is said to have an interval scale. Interval measurement is considerably stronger than ordinal, however, the zero point and the unit of measurement are arbitrary.

Example: Temperature
IQ

Ratio scale

When a scale has all the characteristics of an interval scale and in addition has a true zero point as its origin, it is called a ratio scale. A variable with this characteristic is called continuous variable.

Example: Annual Income
Grain yield
Number of years in farming

IV. Contingency Tables

A contingency table displays the joint distribution of two or more variables. They are usually presented in a matrix format. Whereas a frequency distribution provides the distribution of one variable, a contingency table describes the distribution of two or more variables simultaneously. Each cell shows the number of respondents that gave a specific combination of responses.

Contingency tables are frequently used because:

1. They are easy to understand. They appeal to people that do not understand the more sophisticated measures.
2. They can be used with any level of data: nominal, ordinal, interval, or ratio – contingency tables treat all data as if they are nominal
3. A table can provide greater insight than single statistics

The following is an example of a 2 × 2 contingency table. The variable "Level of education" has two categories: low and high. The other variable "Adoption of Nitrogen fertilizer" has also two categories: No and Yes. Each cell gives the number of farmers falling under the combination of categories.

Level of Education	Adoption		Total
	Yes	No	
High	21	6	27
Low	22	51	73
Total	43	57	100

These data are relevant to address the question: 'Does the level of education of the farmers affect their decision to adopt nitrogen fertilizer?' To answer such question, the most frequently used statistical tool is the Chi-square test for independence. The Chi-square test calculates, for each cell in the table, the frequency that is expected for each category, assuming no difference between low and high level of education. A comparison is then made between the observed frequencies and the expected frequencies; and the further these two are apart, the more convincing evidence there is to reject the hypothesis.

Totals are included in the table above, to show how expected values can be calculated. The table of expected values is given below – calculated as follows: If there is no effect of education, then the proportion of non-adopters is $\frac{57}{100}$. Hence we would expect that of the 73 farmers with low level of education, $73 \times \frac{57}{100}$ would be non-adopters.

Expected frequencies:

Level of education	Adoption		Total
	Yes	No	
High	$\frac{43*27}{100}=11.61$	$\frac{57*27}{100}=15.39$	27
Low	$\frac{43*73}{100}=31.39$	$\frac{57*73}{100}=41.61$	73
Total	43	57	100

The Chi-square value is then computed as follows:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$
$$=18.25$$

This value is compared with the tabular Chi-square value with (r-1)(c-1) degrees of freedom, where r and c are the number of rows and columns, respectively. For this example, the tabular Chi-square value is 3.84 which is very much less than the computed Chi-square value. Hence, we may conclude that there is some association between level of education and adoption of nitrogen fertilizer. However, aside from this conclusion, the Chi-square statistic does not tell as anything more. Most likely what we really want to know about is the nature of the association. Presenting the data in percentages may help as see the nature of this association.

Note: One limitation of the chi-square test is that the expected value of a particular cell in the contingency table should not be too small. As a rule, the test will be valid provided that fewer than 20% of the cells have an expected count below 5, and none are below 1.

Percentage data:

Level of education	Adoption		Total
	Yes	No	
High	$\frac{21*100}{27}=77.8$	$\frac{6*100}{27}=22.2$	27
Low	$\frac{22*100}{73}=30.1$	$\frac{51*100}{73}=69.9$	73
Total	43	57	100

From the table above we could see easily that there are more non-adopters among farmers with low level of education and more adopters among farmers with high level of education. However, we still do not know how strong the association is.

What we would like to have for categorical data analysis are additional measures that tell us

1. Is there evidence of association?
2. If so, how strong is it?

V. Measuring the Strength of Association

In the case of continuous variables we use r and r^2 to measure the strength of the relationship between variables. Measuring the strength of association between categorical variables is not quite simple.

Odds and Odds Ratios

As an example of using odds and odds ratio to measure the strength of association, we look back at the two-way table for level of education and adoption of nitrogen fertilizer.

Level of education	Adoption	
	Yes	No
High	a	b
Low	c	d

For high level of education, the estimated probability of adoption is $\frac{a}{a+b}$ and the estimated probability of non-adoption is $\frac{b}{a+b}$. Therefore for high level of education, the odds favoring adoption are given by

$$\text{odds} = \frac{\text{probability of success}}{\text{probability of failure}} = \frac{\left(\frac{a}{a+b}\right)}{\left(\frac{b}{a+b}\right)} = \frac{a}{b} = \frac{\text{number of successes}}{\text{number of failures}}$$

where "success" represents adoption. For low level of education, the odds favoring adoption are similarly calculated to be $\frac{c}{d}$. Generally we look at the ratio of odds from two rows, which we denote by $\hat{\theta}$. In this example,

$$\text{odds ratio} = \hat{\theta} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

For our example

$$\hat{\theta} = \frac{\text{odds(High)}}{\text{odds(Low)}} = \frac{21/26}{22/51} = \frac{3.5}{0.43} = 8.13$$

This indicates that the odds in favor of adoption if the farmer has high level of education are eight times the odds in favor of adoption if the farmer has low level of education.

Significance of the Odds Ratio

To get information about the strength of association between two categorical variables we would like to test the significance of the odds ratio $\hat{\theta}$. First we must know the distribution of $\hat{\theta}$ when the true odds ratio is 1 (null hypothesis). We can work out the asymptotic distribution of the odds ratio, but it is much easier to work out the distribution of the natural logarithm of the odds ratio. Note that the values of the odds ratio range from zero to very large positive numbers, that is $0 \le \hat{\theta} \le \infty$, and the values of $\hat{\theta}$ are highly skewed to the right. When the odds of success are equal in the two rows being compared, $\hat{\theta}$=1, indicating no association between the variables. By taking the natural logarithm of $\hat{\theta}$, we pull in the tail. It turns out that, for large samples, the logarithm of the odds ratio, that is $\ln(\hat{\theta})$, is approximately normally distributed with standard error

$$SE\left[\ln\left(\hat{\theta}\right)\right] = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}},$$

Hence the quantity $\frac{\ln\left(\hat{\theta}\right)-0}{SE\left[\ln\left(\hat{\theta}\right)\right]}$ has a distribution that is approximately standard normal for large values of n.

For our example, we determine whether the odds ratio $\hat{\theta}$=8.13 differs significantly from what we would expect if there were no difference in the odds of adoption for farmers with high level of education and for farmers with low level of education. Note that if there were no difference in the odds for the two groups, the odds ratio would be exactly equal to one. Thus, under the null hypothesis of no difference, the natural logarithm of the odds ratio would be equal to zero. Based on our example $\ln\left(\hat{\theta}\right) = \ln(8.13) = 2.096$. The standard error of $\ln\left(\hat{\theta}\right)$ is given by $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{21}+\frac{1}{6}+\frac{1}{22}+\frac{1}{51}} = 0.528534$. Hence, we can do a z-test using z as our test statistic

$$z = \frac{\ln\left(\hat{\theta}\right)-0}{SE\left[\ln\left(\hat{\theta}\right)\right]} = \frac{2.096-0}{0.528534} = 3.96$$

Since this is a two-sided test z-values larger than 2 are generally considered significant. For our example, we could conclude that the odds ratio is significantly greater than one. This would indicate that the odds in favor of adoption if the farmer

has high level of education are significantly higher than the odds in favor of adoption if the farmer has low level of education.

Effect of Sample Size

Consider the following table. In each table, 49% of group X and 51% of group Y favor the adoption of a certain agricultural practice.

	A			B			C		
	Yes	No	Total	Yes	No	Total	Yes	No	Total
X	49	51	100	98	102	200	4900	5100	10000
Y	51	49	100	102	98	200	5100	4900	1000
Total	100	100	200	200	200	400	10000	10000	20000
χ^2	0.08			0.16			8.0		
P	0.78			0.69			0.005		
$\hat{\theta}$	0.92			0.92			0.92		
Z	-0.29			-0.42			-2.95		

Note that the chi-square value increases from a value that is not statistically significant to a value that is highly significant as the sample size increases from 200 to 400 to 20,000. Since the P-value for this test of association is very sensitive to the sample size, we need other measures to describe the strength of the relationship. In every table above, the odds ratio is 0.92. Since the odds ratio is very close to one, this indicates that there is not very much going on in any of these situations. The association is very weak. This example illustrates that we should not simply depend on a test of significance. Increasing sample size can always make a test statistically significant.

ANALYSIS OF CATEGORICAL DATA USING LOGISTIC REGRESSION

At the end of the, tutorial the user should be able to

- Analyze categorical data using logistic regression

I. Introduction

The problem with the Chi-square test is that they are only applicable for simple two-way tables involving categorical data. However, most often survey data involve more than 2 variables. When one of these variables could be regarded as a response variable and that this variable is categorical then logistic regression could be used. Logistic regression is used to predict the outcome of the dependent variable on the basis of the independent variables. In logistic regression, the independent variables can be categorical or quantitative or a mixture of both. When the response variable involves only 2 categories it is called a binary logistic regression. On the other hand, if the response variable involves more than 2 categories it is called multinomial logistic regression. When the multiple categories of the dependent variable can be ranked then it is called ordinal logistic regression. We will only discuss binary logistic regression in this course. But one approach to analyzing multinomial response variables is to analyze each category against all the others pooled together with binary logistic regression.

II. Binary Logistic Regression

To illustrate the use of binary logistic regression, we use the 2×2 table for level of education and adoption of nitrogen fertilizer given below:

Level of Education	Adoption		Total
	Yes	No	
High	21	6	27
Low	22	51	73
Total	43	57	100

The odds in favor of adoption can be modeled as follows

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

where p is the probability of adoption and x=1 for high level of education and x=0 for low level of education. This equation is called the logit model. If we solve for p in this equation we arrive at the following logistic model.

$$p = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \text{ or } p = \frac{1}{1 + \exp(-\alpha - \beta x)}$$

α and β are estimated using logistic regression.

III. Running binary logistic regression in CropStat

Data for Logistic Regression can be entered as counts of individuals.

A. Input data are frequency counts

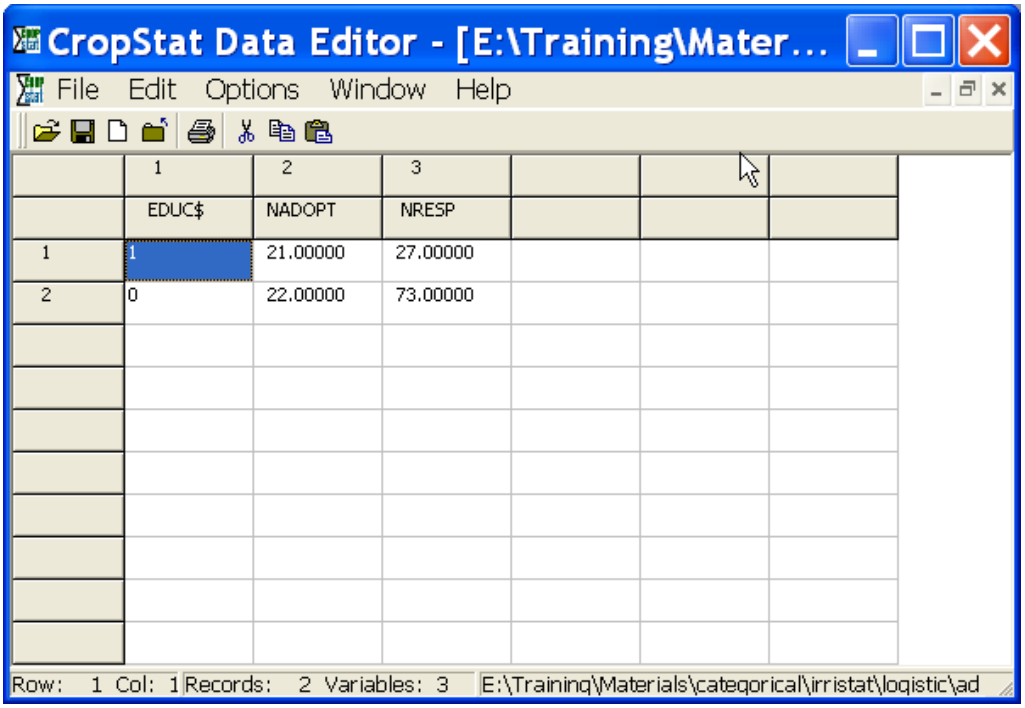
Prepare input data file as follows:

a. a column for the independent variable. In this example the independent variable is level of education which has two levels. Hence our file should have two rows.

b. a column for the frequency count of the favorable response of the dependent variable for each level of the independent variable. In this example, since we are more interested on the adopters then we have to input the number of responses for the adopters for each education level.

c. a column for the total number of responses for each level of the independent variable.

The CropStat file should look as follows:



Note:

1. Always define the independent variable as character.

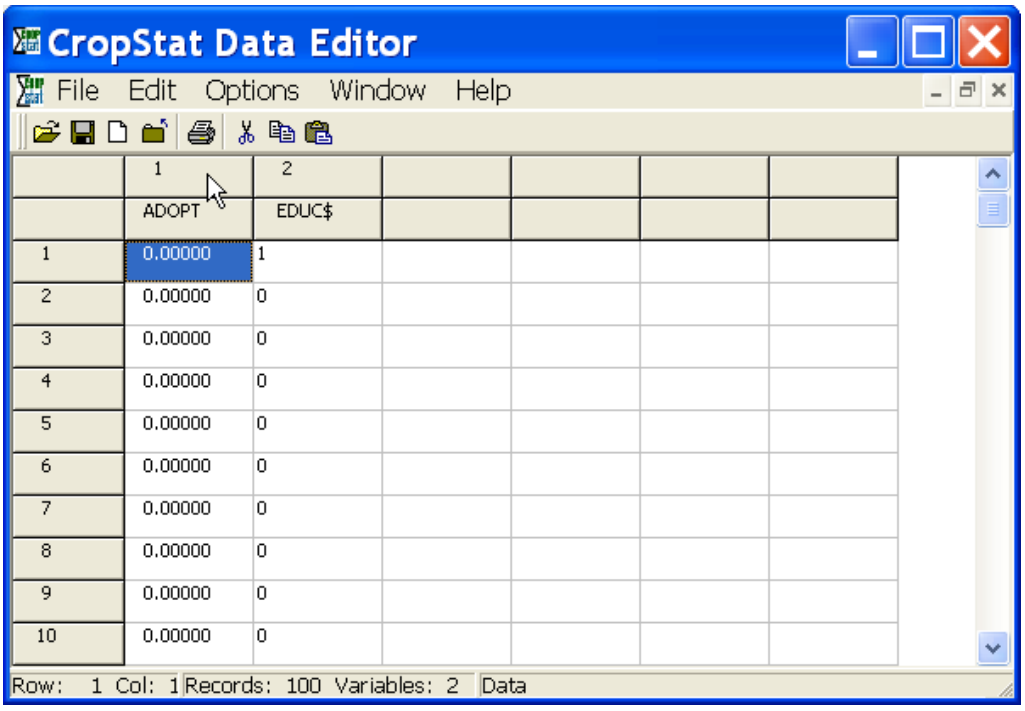
2. Remember that CropStat estimates coefficients relative to the last level. Hence, if you want to make the low (0) level of education as the reference level then you have to input this level in the second row. The idea here is similar to a control treatment in a controlled experiment.

B. Input raw data file is on a per response basis

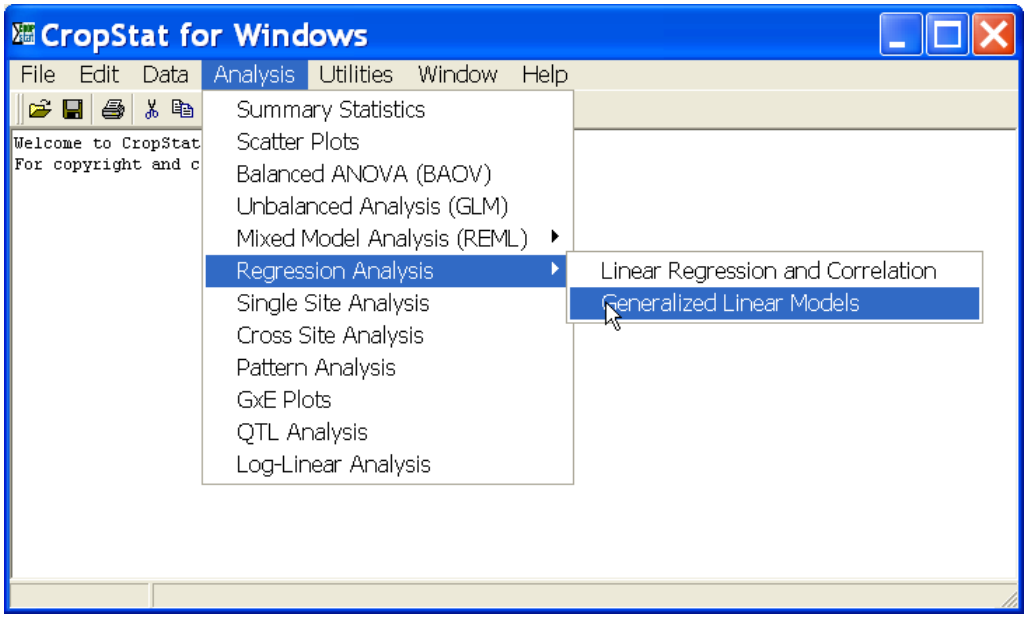
Prepare input data file as follows:

a. a column for the independent variable. Define this variable as character.

b. a column for the response variable



- We use the first input file to illustrate the use of CropStat for logistic regression.
- Open the data file *LOGREGEX.SYS* from the *CROPSTAT7.2\TUTORIAL\ TUTORIAL DATASETS* folder.
- Select **File ⇒ Save-as**. Click the **Save in** box and go go inside working folder *C:\MY CROPSTAT*. Create a subfolder LOGISTIC REGRESSION then click **Save.**
- Choose **Analysis/Regression Analysis/Generalized Linear Models** from the Analysis menu.



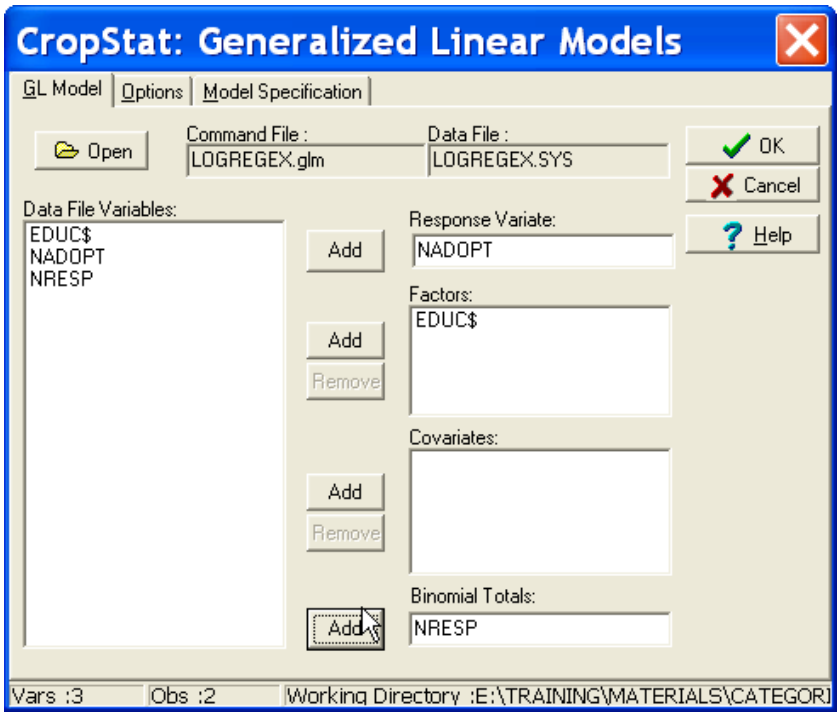
- The **Open** dialog box will prompt you to enter a name for the command file. Click the **Look In** box to go to your working *drive C:\MY CROPSTAT\ LOGISTIC REGRESSION.*

- Enter *LOGREGEX* in the **File name** box. Click **Open** button.

- Since *LOGREGEX.GLM* does not exist, a message box will appear confirming if you want to create the file. Click **Yes** to create new Command File.

- Enter the name of the data file to be used. Enter *LOGREGEX.SYS* in the **File name** box.

- Click **Open**. The **Generalized Linear Models** dialog box will appear for you to fill-in the details of the analysis.



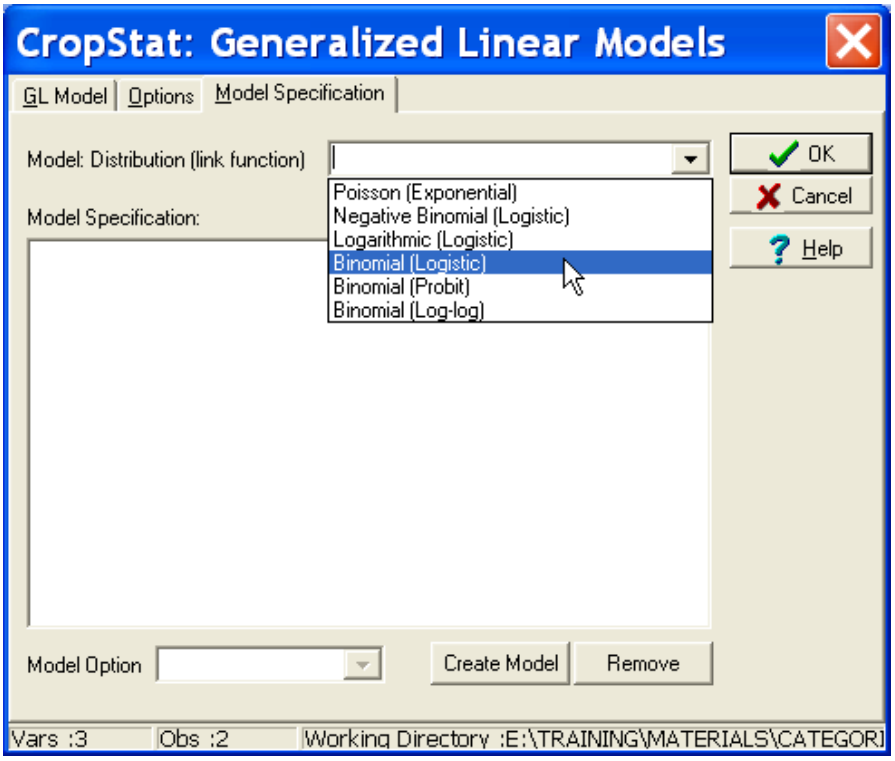
- Click the **Model Specification** tab. The **Generalized linear models model specification** window will appear.

- From the **Data File Variable** list, highlight the response variable then add to the **Response Variate** box; highlight the independent variable then add to the **Factors** box.

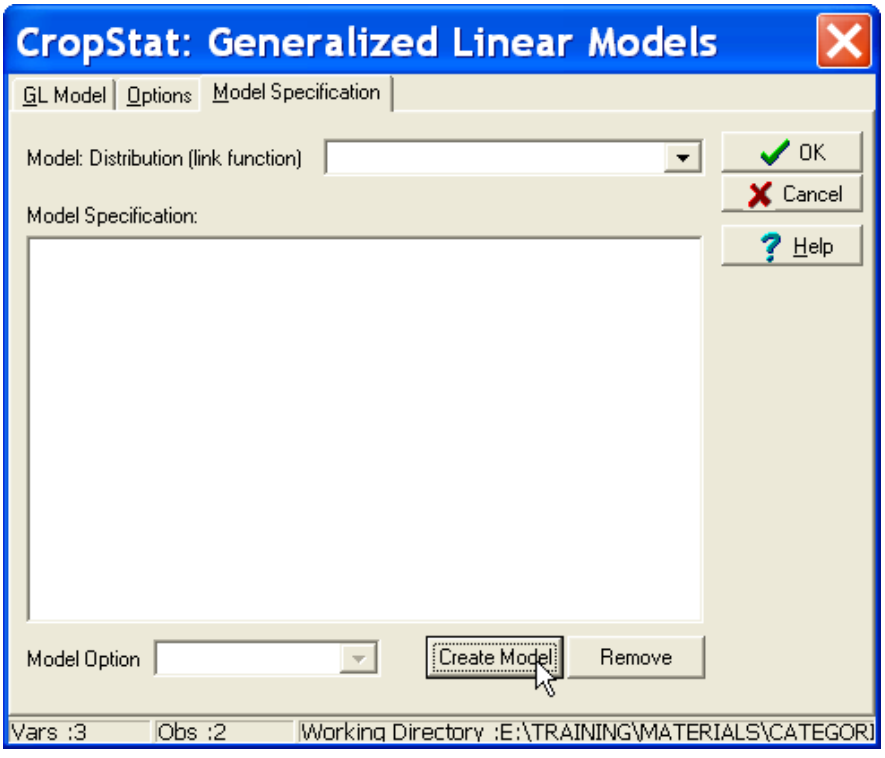
- If your data file contains the frequency counts of the favorable response for each level of the independent variable do the following:

    - From the **Data File Variable** list, highlight the variable *NRESP,* which contains the total number of responses for each level of the independent variable then click 'Add' to add the **Binomial Totals**. Otherwise, go to the next step.

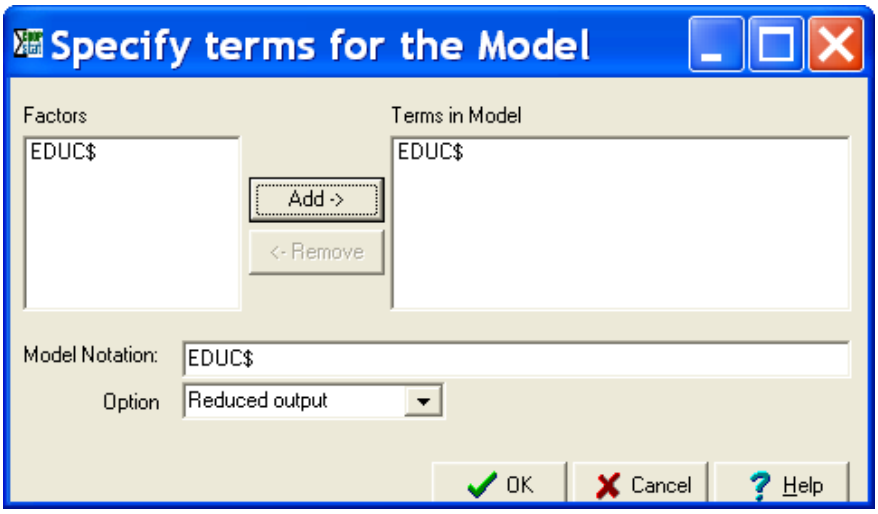
- Click the down-arrow key then choose **Binomial (Logistic)**.



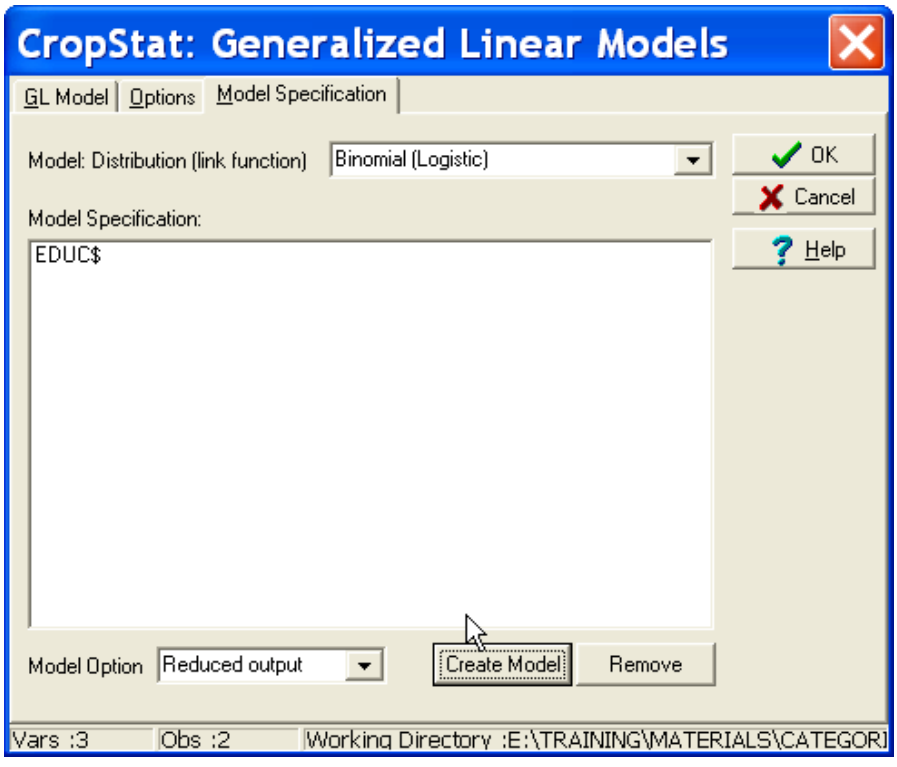
- Click **Create Model**.



- The **Specify terms for the Model** dialog box will appear.
- Select EDUC$ and add into the model.



- Click the **Ok** and the **Generalized linear model specification** window will appear again.



- Click **OK**.

CropStat Output

```
GENERALIZED LINEAR MODEL ANALYSIS  FILE LOGREGEX   13/ 6/ 7  9: 5
 ------------------------------------------------------------------- :PAGE   1

 FACTOR     EDUC$    HAS    2 LEVELS:
 1            0
 ------------------------------------------------------------

 MODEL FOR VARIATE  2 (NADOPT) WITH  2 COEFFICIENTS:
 Binomial with Logistic link, exp(g)/(1+exp(g)) where
 g = EDUC$

 Log-likelihood       -252.5181

                  Coefficient Statistics
                      Standard    Asymptotic    Asymptotic
     Coefficient         Error   Z-statistic       P-value
 1        -0.841         0.255        -3.296         0.001
 2         2.094         0.529         3.961         0.000
```

From the output the logistic regression model can be written as:

$$p = \frac{e^{-0.841+2.094X}}{1+e^{-0.841+2.094X}}$$

This equation simplifies to

$$p = \frac{(0.43128)(8.1773)^x}{1+(0.43128)(8.1773)^x}$$

If we want to predict the probability of adoption when farmer has low education level then we just substitute zero for x and we get p=.3013. The same way if we want to predict the probability of adoption when farmer has high education level then we just substitute one for x and we get p=.7778. Hence, we can conclude that the probability of adoption is higher if farmer has high level of education. The odds ratio could be estimated as the ln(β) which is ln(2.094)=8.1773. This value is the same as what we have gotten earlier.

IV. Binary Logistic Regression with More Than One Independent Variable

Consider the study on the effect of insecticide usage (low, high) and smoking history on the incidence of pulmonary ailments in farm workers.

Data for the 10,919 farmers who had never smoked are provided in the following table.

Insecticide Rate	Pulmonary Ailment		Total
	Yes	No	
High	96	5392	5488
Low	55	5376	5431

A chi-square test to determine whether there is association between having a pulmonary ailment and using high rate of insecticide results in χ^2=10.86 and p-value less than .001. Thus we can conclude that there is strong evidence that high rate of insecticide has an effect on pulmonary ailment.

The highly significant chi-square value does not give information about the direction or strength of the association. The odds ratio of the group who had never smoked is $\hat{\theta}$=1.74. Thus these data indicate that the odds in favor of having a pulmonary ailment if using high rate of insecticide are 1.74 times the odds in favor of having a pulmonary ailment if using low rate of insecticide for the group who have never smoked.

Data for farmers who were past smokers are provided in the following table.

Insecticide Rate	Pulmonary Ailment		Total
	Yes	No	
High	105	4276	4381
Low	63	4310	4373

The odds ratio for this group is $\hat{\theta}$=1.68.

Data for farmers who were current smokers are provided in the following table.

Insecticide Rate	Pulmonary Ailment		Total
	Yes	No	
High	37	1188	1225
Low	21	1192	1213

The odds ratio for this group is $\hat{\theta}=1.77$. Note that the odds ratios are similar for all groups, indicating a negative effect of high rate of insecticide to farmers' health, but not indicating an interaction with smoking.

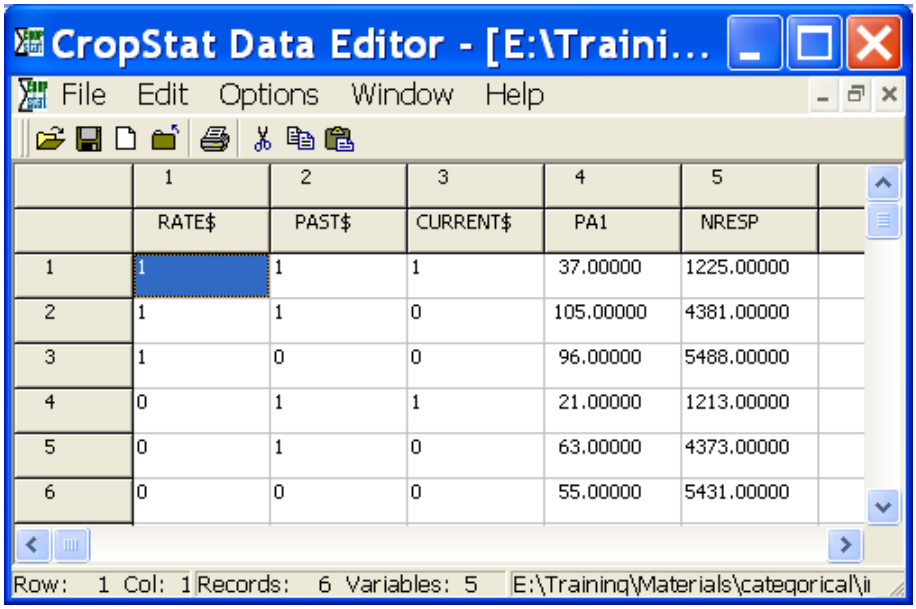
To get more information from these data, we can look at the logistic regression model for the entire group of 22,000 participating farmers. In this model there were two explanatory variables, high or low rate of insecticide and smoking status. The model will have the form

$$\ln(\text{odds}) = b_0 + b_1 R + b_2 P + b_3 C$$

The response variable is whether or not the farmer had pulmonary ailment. All variables are coded as 0 or 1. For example, letting R represent the rate of insecticide, R=1 indicates high rate and R=0 indicate low rate. There are two variables for smoking status. We let P represent past smoker and code P=1 to indicate that the farmer is a past smoker. If P=0 then the farmer is not a past smoker. Similarly C=1 indicates that the farmer is a current smoker, while C=0 indicates that the subject is not a current smoker. A farmer who has never smoked would be indicated by having the values of C and P both zero.

Running binary logistic regression with more than 1 independent variable in CropStat

- For our example, we use the file *PULMONRY.SYS*.



- Open the data file *PULMONRY.SYS* from the *CROPSTAT7.2\TUTORIAL\ TUTORIAL DATASETS* folder.

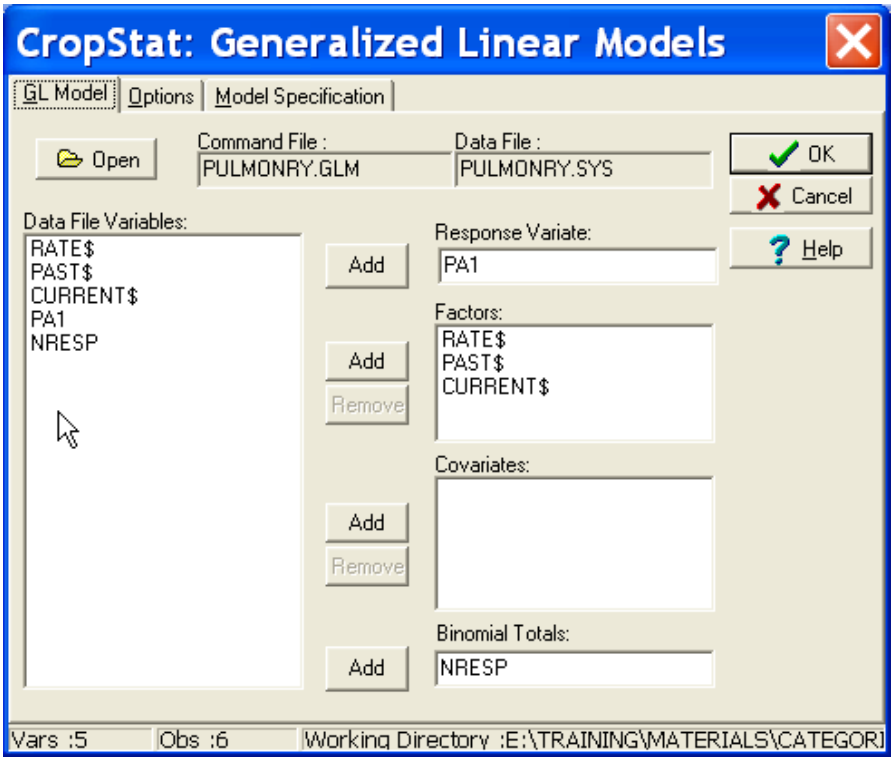
- Select on **File ⇒ Save as.** Click the **Save in** box and go inside the directory *C:\MY CROPSTAT\ LOGISTIC REGRESSION* and save *LOGREGEX*.SYS.

  Note:

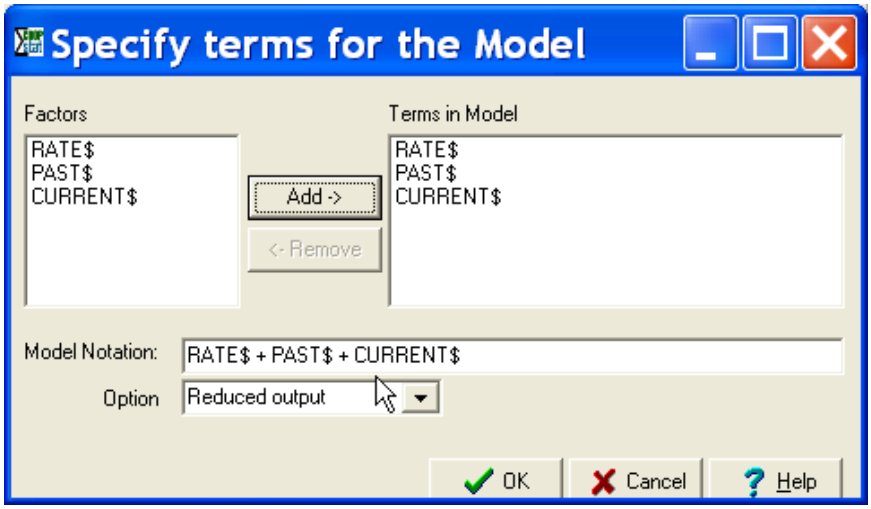
  1. the independent variables, Rate, Past, and Current are defined as character variables.

  2. The zero levels of all independent variables are to be used as reference and hence were inputted as the last levels.

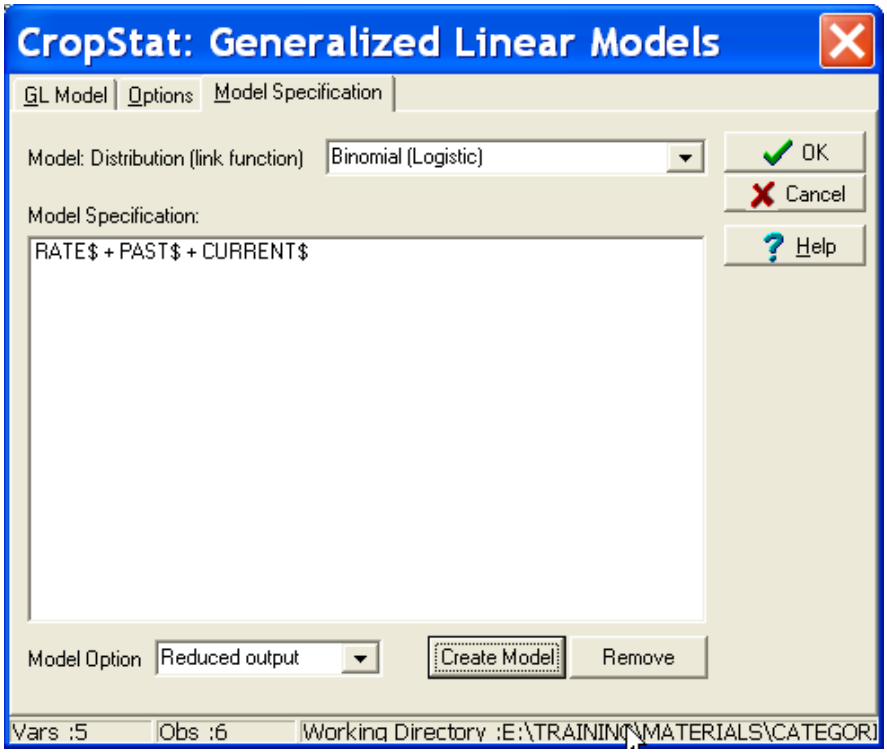
- Follow the same steps as in the previous section until you reach the **Generalized Linear Models** dialog box. Provide the necessary information.



- Choose Binomial (Logistic) for the Model: Distribution (link function).
- Enter Model as follows:



- **The Generalized Linear Models** window should be as follows:



CropStat Output

```
GENERALIZED LINEAR MODEL ANALYSIS  FILE PULMONRY   13/ 6/ 7 13:51
 ------------------------------------------------------ :PAGE   1

 FACTOR     RATE$    HAS    2 LEVELS:
 1            0

 FACTOR     PAST$    HAS    2 LEVELS:
 1            0

 FACTOR  CURRENT$    HAS    2 LEVELS:
 1            0
 ------------------------------------------------------------

 MODEL FOR VARIATE  4 (PA1) WITH  4 COEFFICIENTS:
 Binomial with Logistic link, exp(g)/(1+exp(g)) where
 g = RATE$ +PAST$ +CURRENT$

 Log-likelihood       -396.7880

                  Coefficient Statistics
                      Standard    Asymptotic    Asymptotic
     Coefficient         Error   Z-statistic       P-value
 1         -4.57          0.11        -42.96          0.00
 2          0.54          0.11          5.02          0.00
 3          0.33          0.11          2.96          0.00
 4          0.22          0.15          1.42          0.16
```

From the output the logistic regression model can be written as:

$$p = \frac{e^{-4.57+0.54R+0.33P+0.22C}}{1+e^{-4.57+0.54R+0.33P+0.22C}}$$

This equation simplifies to

$$p = \frac{(0.0103)(1.7177^{R})(1.397^{P})(1.2448^{C})}{1+(0.0103)(1.7177^{R})(1.397^{P})(1.2448^{C})}$$

Hence if we want to predict only for the non-smokers, we substitute zero to P and C and arrive at the following equation

$$p = \frac{(0.0103)(1.7177^{R})}{1+(0.0103)(1.7177^{R})}$$

The odds ratio for non-smokers is 1.7177. The probability of having pulmonary ailment for non-smokers using low rate of insecticide is .0102. Comparing this value

to the actual probability $\frac{55}{5431} = .0101$ which is very close, we can say that the model fits well for this particular cell. If we do the same thing for the other cells and summarize the result we get the following table.

Conditions	R	P	C	Est. p	Actual p
High Rate Never Smoked	1	0	0	0.0174	0.0175
High Rate Past Smoker	1	1	0	0.0242	0.0240
High Rate Current Smoker	1	0	1	0.0299	0.0302
Low Rate Never Smoked	0	0	0	0.0102	0.0101
Low Rate Past Smoker	0	1	0	0.0142	0.0144
Low Rate Current Smoker	0	0	1	0.0126	0.0173

From the result, we can say that on the average, the expected probability of having pulmonary ailment is higher for those who use high rate of insecticide. The probability is further increased if the farmer is a past or current smoker.

V. Binary Logistic Regression with Independent Variable with More Than Two Categories

The examples we had so far had independent variables with only two categories. Consider the data below taken from a study on the effect of two types of treatment on disease incidence. Here we want to estimate the probabilities of getting the disease for each of the treatments.

Treatment	With Disease	No Disease	Total
A	5	15	20
B	6	14	20
Control	17	3	20
Total	28	32	60

CropStat will create two dummy independent variables. The number of dummy variables is one less than the number of categories of that particular independent variable. For this example, the dummy variables may be:

T_1=1 if Treatment=A and 0 otherwise

T_2=1 if Treatment=B and 0 otherwise

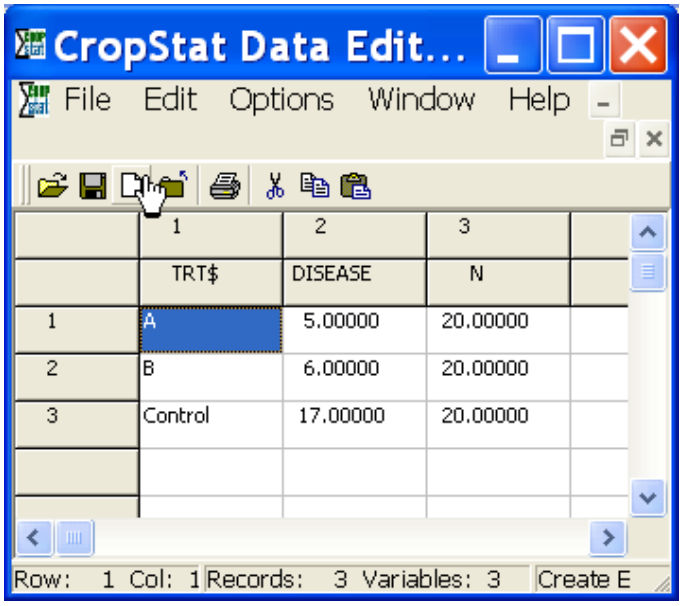
T_1=T_2=0 corresponds to the Control treatment.

With T_1 and T_2 as the independent variables we now have the logit model

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 T_1 + b_2 T_2$$

Running binary logistic regression with categorical independent variable(s) in CropStat

- As an example we use the CropStat file *DISEASE.SYS*

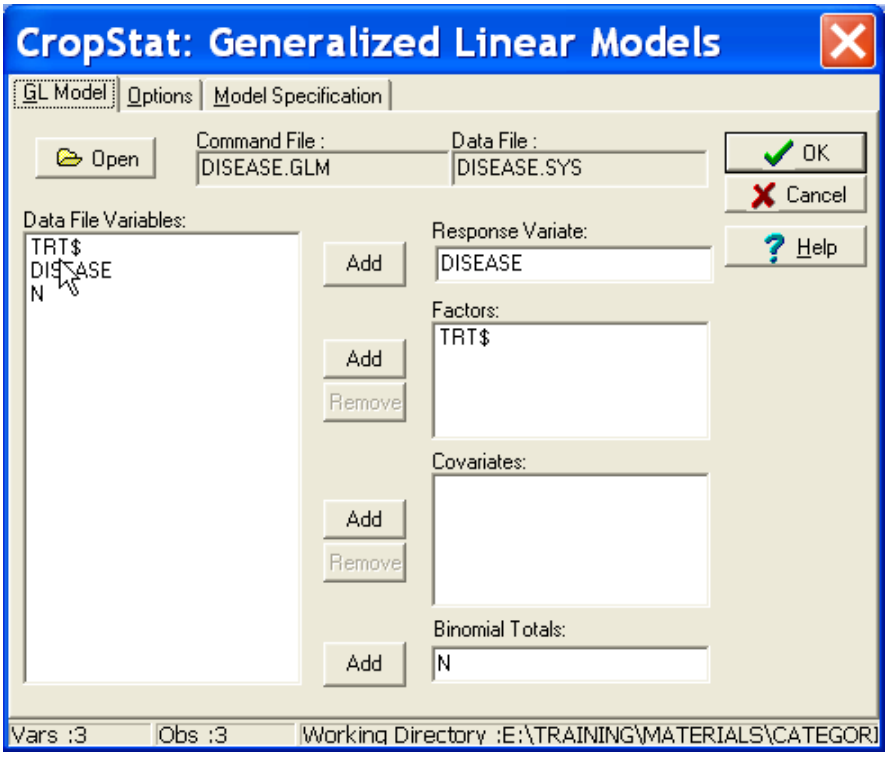


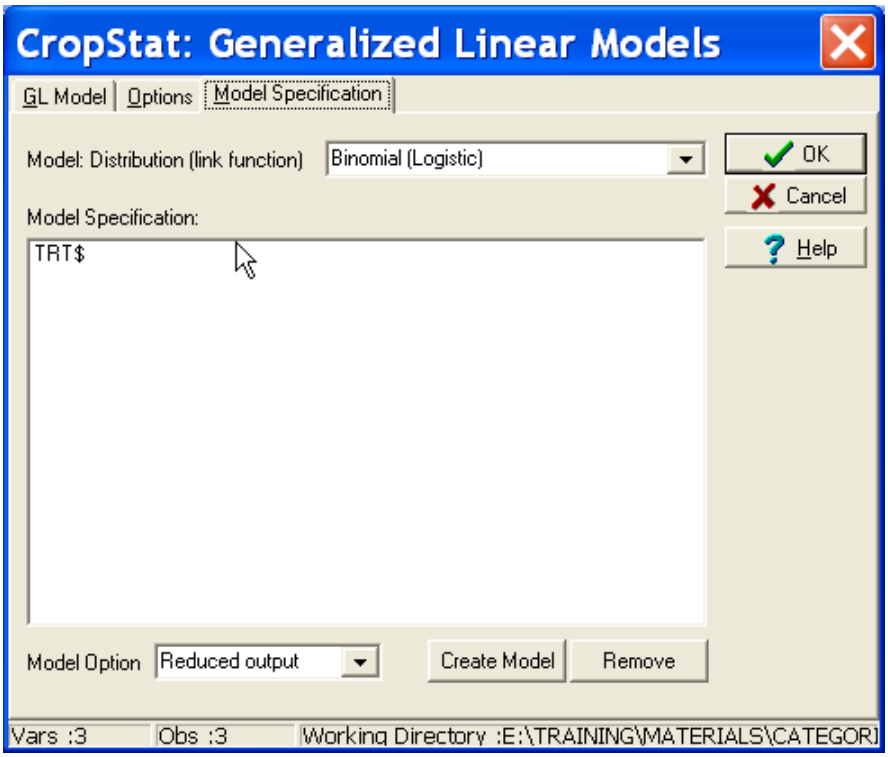
- Open the data file *DISEASE.SYS* from the CROPSTAT7.2\TUTORIAL\ TUTORIAL DATASETS folder.

- Save *DISEASE*.SYS inside your working folder C:\MY CROPSTAT\LOGISTIC REGRESSION by selecting **File ⇒ Save-as**

  Note: CropStat will create dummy variables for TRT before doing the logistic regression. The dummy variable is as discussed above wherein the reference level is the last level inputted.

- Perform logistic regression as discussed in previous sections. The **Generalized Linear Models** dialog box and **Model Specification** window should be as follows:


CropStat: Generalized Linear Models
GL Model
Options
Model Specification
Open
Command File :
DISEASE.GLM
Data File :
DISEASE.SYS
OK
Cancel
Help
Data File Variables:
TRT$
DISEASE
N
Add
Response Variate:
DISEASE
Factors:
TRT$
Add
Remove
Covariates:
Add
Remove
Binomial Totals:
N
Add
Vars :3
Obs :3
Working Directory :E:\TRAINING\MATERIALS\CATEGORI


CropStat: Generalized Linear Models
GL Model
Options
Model Specification
Model: Distribution (link function)
Binomial (Logistic)
OK
Cancel
Help
Model Specification:
TRT$
Model Option
Reduced output
Create Model
Remove
Vars :3
Obs :3
Working Directory :E:\TRAINING\MATERIALS\CATEGORI

CropStat Output

```
GENERALIZED LINEAR MODEL ANALYSIS  FILE DISEASE   13/ 6/ 7 16:11
 ------------------------------------------------------ :PAGE   1

 FACTOR      TRT$    HAS    3 LEVELS:
 A            B           Control
 ------------------------------------------------------------

 MODEL FOR VARIATE  2 (DISEASE) WITH  3 COEFFICIENTS:
 Binomial with Logistic link, exp(g)/(1+exp(g)) where
 g = TRT$

 Log-likelihood       -418.3134

                  Coefficient Statistics
                      Standard    Asymptotic    Asymptotic
     Coefficient         Error   Z-statistic       P-value
 1         1.658         0.610         2.718         0.007
 2        -2.738         0.798        -3.433         0.001
 3        -2.540         0.783        -3.243         0.001
```

From the output the logistic regression model can be written as:

$$p = \frac{e^{1.658-2.738T_1-2.540T_2}}{1+e^{1.658-2.738T_1-2.540T_2}}$$

This equation simplifies to

$$p = \frac{(5.2488)(0.0647^{T_1})(0.0789^{T_2})}{1+(5.2488)(0.0647^{T_1})(0.0789^{T_2})}$$

If we estimate the probability of getting the disease for each treatment we have the following result.

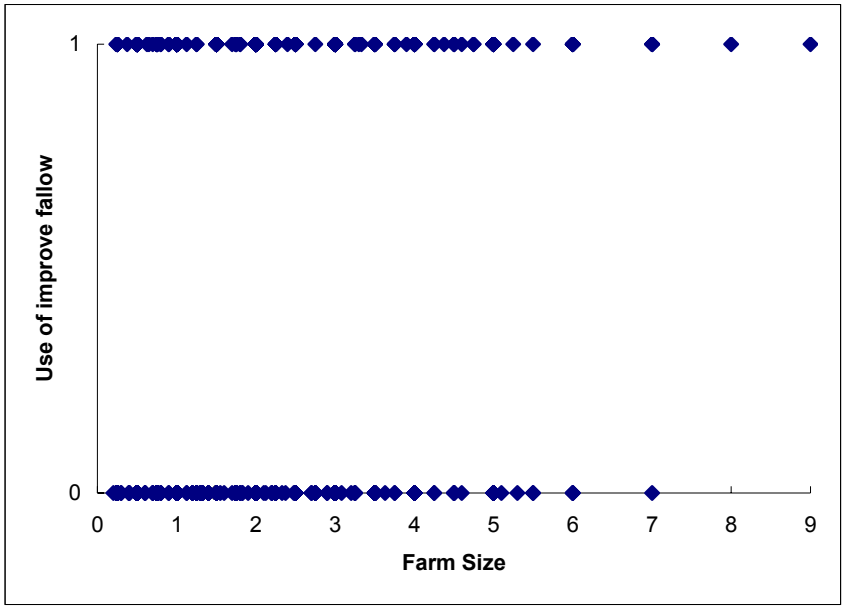
Treatment	T_1	T_2	Expected Probabilities
A	1	0	0.25
B	0	1	0.29
Control	0	0	0.84

From the result we say that the probability of a disease incidence is higher if no treatment is applied. However, probability of a disease incidence is slightly higher if treatment B was used than if treatment A was used.

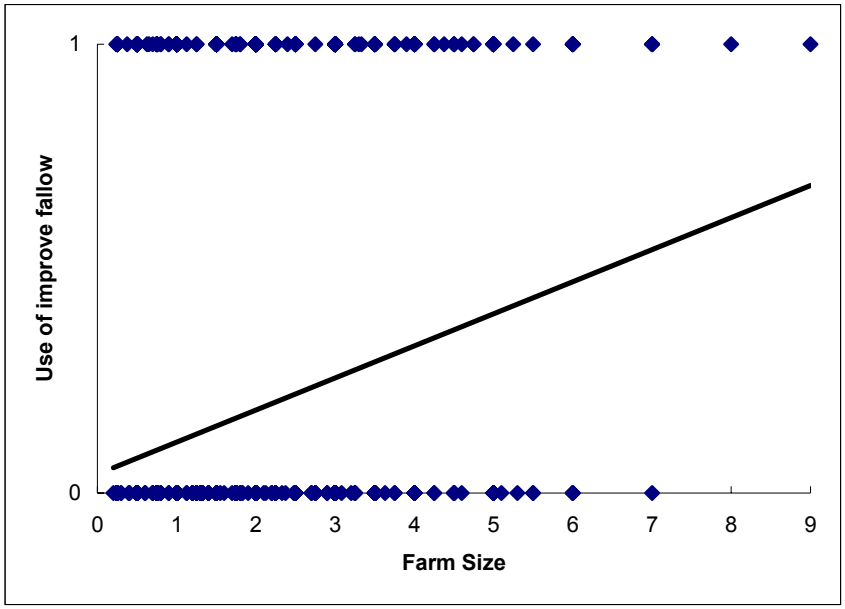
VI. Binary Logistic Regression with Quantitative Independent Variable

In all our previous examples, we have looked at models containing categorical independent variables, but there is no reason to restrict the independent variables to just categorical ones. Quantitative variables can also be included. To show this, consider the study on the effect of farm size on the adoption of improved fallow where the response variable is adoption with values 1 or 0 and independent variable farm size.

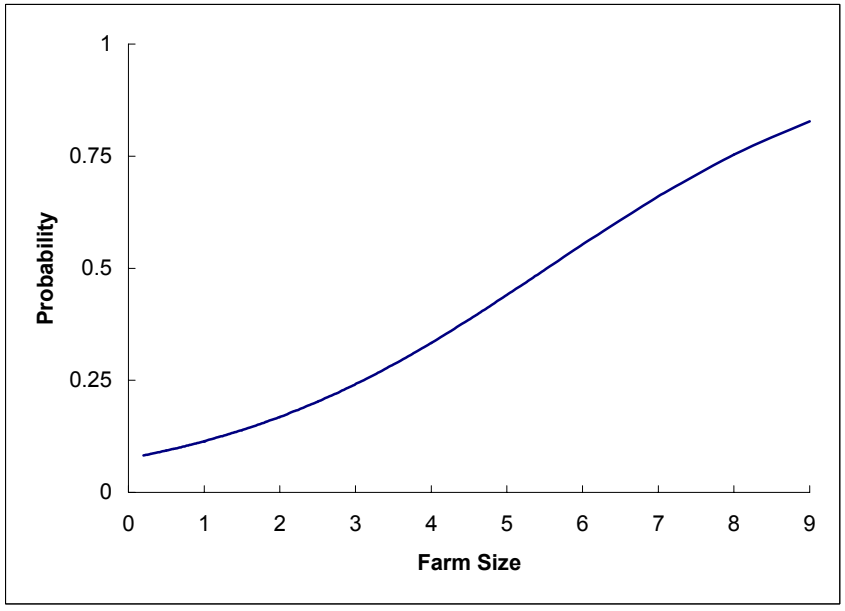
A point plot of the raw data is not informative, as demonstrated below. The points come on two horizontal lines, but there is too much overlap to get a good impression of any relationship, or lack of relationship.



How can we model this to draw some firm conclusions? Our raw data are binary and so simple linear regression is clearly inappropriate. A straight line would not fit the points of the graph of the raw data at all as shown below.



Moreover, we do not want the model to give values that are negative or are greater than 1, which would be meaningless in the context. The model should be such that for very small farms the probability goes asymptotically towards 0, whereas for very large farms it reaches asymptotically 1, in other words a flattened S-shaped curve, confined between 0 and 1 as shown below.



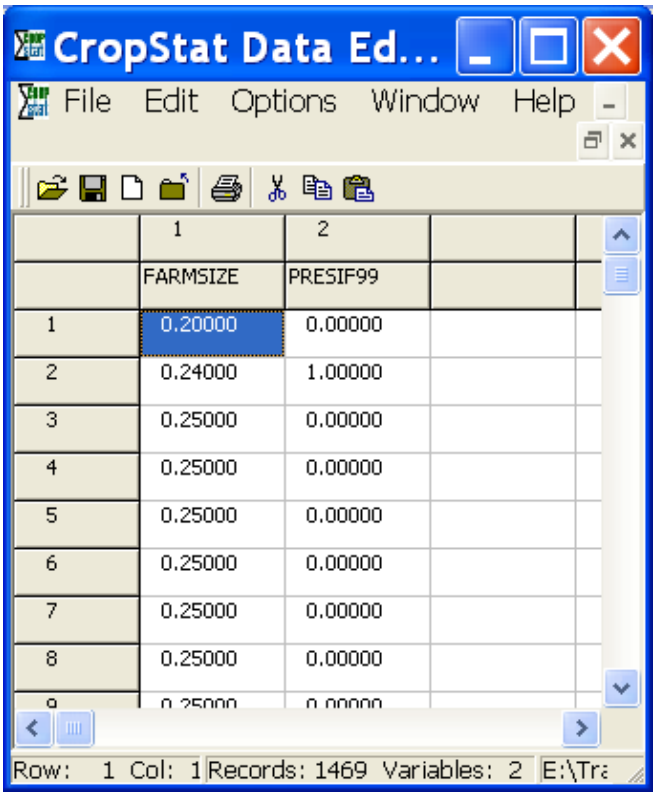
This model could be represented by the following equation

$$p = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

which is the same logistic model we have used before. Hence the logit model could be used to estimate the parameters.

Running binary logistic regression with quantitative independent variable in CropStat

- For our example we will use the CropStat file *FARMSIZE.SYS*.

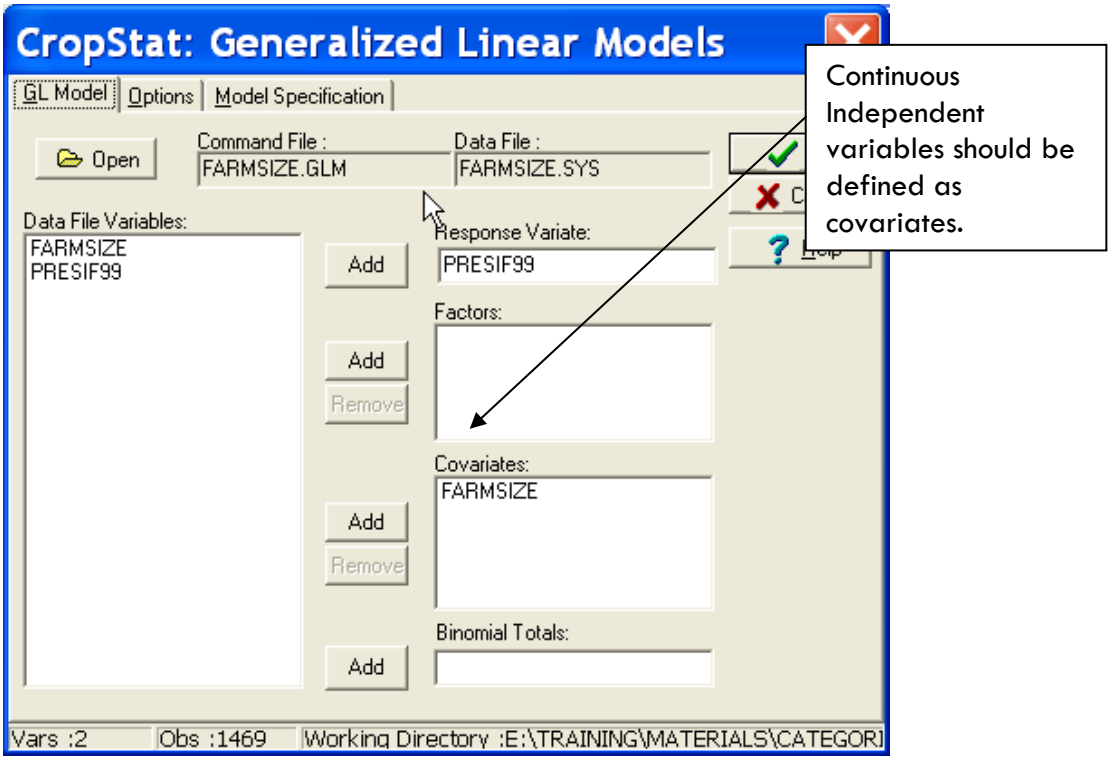


- Open the data file *FARMSIZE.SYS* from the CROPSTAT7.2\TUTORIAL\ TUTORIAL DATASETS folder.

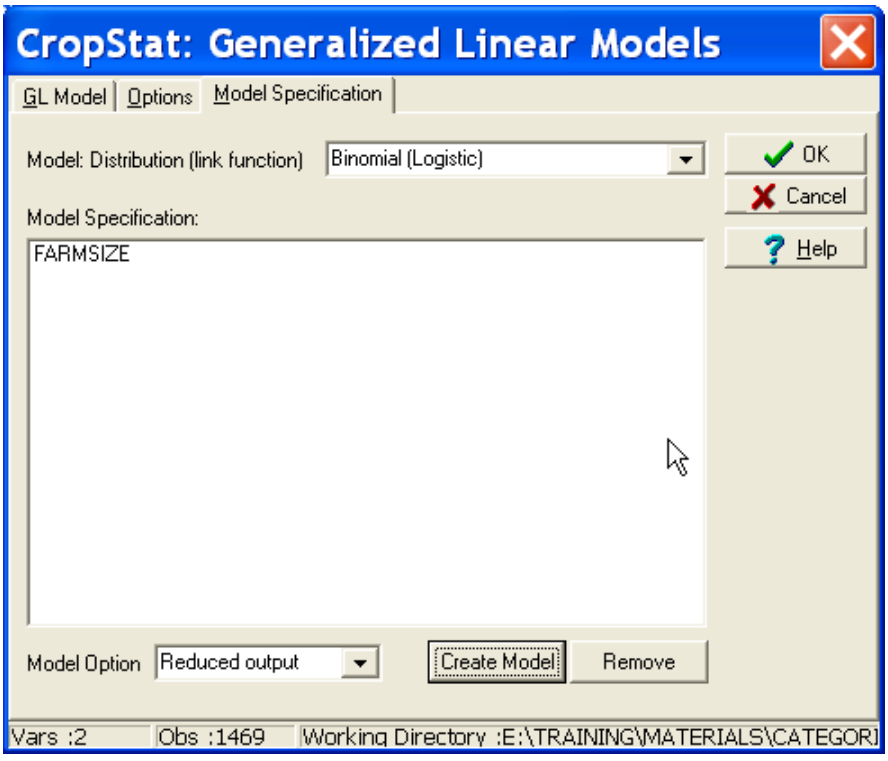
- Select on **File ⇒ Save as.** Click the **Save in** box and go inside the directory *C:\MY CROPSTAT\ LOGISTIC REGRESSION* and save *FARMSIZE*.SYS.

- Follow the same steps as in previous sections until you reach the **Generalized Linear Models** dialog box.

- Provide the inputs of the analysis as follows:



- The **Model Specification** window should be as follows:



CropStat Output

```
GENERALIZED LINEAR MODEL ANALYSIS  FILE FARMSIZE   14/ 6/ 7  8:51
 ------------------------------------------------------ :PAGE   1
 ------------------------------------------------------------

 MODEL FOR VARIATE  2 (PRESIF99) WITH  2 COEFFICIENTS:
 Binomial with Logistic link, exp(g)/(1+exp(g)) where
 g = FARMSIZE

 Log-likelihood       -608.5726

                  Coefficient Statistics
                      Standard    Asymptotic    Asymptotic
     Coefficient         Error   Z-statistic       P-value
 1         -2.50          0.13        -18.93          0.00
 2          0.45          0.05          8.40          0.00
```

From the CropStat output the logistic regression model could be presented as follows:

$$p = \frac{(e^{-2.50})(e^{0.45x})}{1+(e^{-2.50})(e^{0.45x})}$$

$$p = \frac{(0.0821)(1.568^{x})}{1+(0.0821)(1.568^{x})}$$

Hence if we want to estimate the probability of adoption if farm size=5 acres, then we just substitute x=5 and get p=0.44. That is, the probability that the farmer is an adopter is 0.44 when farm size is 5 acres.

VII. Modeling Responses with More than Two Categories

We have seen that contingency tables could only handle two-way tables and that the chi-square test provided a way of making some inference from these tables. While this is suitable for a response variable with any number of categories, a limitation to the method was the rather limited applicability. We could only deal with one single independent variable.

Next we learned that binary logistic models can deal with more complex models containing several independent variables, and they can have both categorical and quantitative independent variables. A drawback of the binary logistic model in the overall discussion of analyzing categorical response data is that the model is only suitable for responses that contain only two categories.

A simple, and often satisfactory, way of dealing with several categories of response is to reorganize the categories or ignore some of them temporarily to reduce it to a binary problem. This is often a logical way of dealing with the data where there is some hierarchy in the responses, e.g. “poor/good/very good” could be reorganized to “poor/(very)good”. Lumping or ignoring of categories could also be a satisfactory solution when very few cases have been recorded for a certain category. Sometimes it could also be informative to look at only the two extreme categories, e.g. in understanding an answer that can be either “negative”, “neutral” or “positive”, most informative cases will be the one that answered “negative” or “positive”.

Another strategy is to divide a response variable with several categories into a series of binary categories or use the multinomial or ordinal logistic regression, which unfortunately are not covered in this tutorial.

In the particular case of all independent variables being categorical, so that we can build a multi-way frequency table with all the data, a solution is to use log-linear models. This is also part of the larger family of generalized linear models, just like the logistic model.

References:

Allan, E., R.D. Stern, R. Coe and J. De Wolf (2002), *Data Analysis of Agroforestry Experiments* (workshop handout), World Agroforestry Centre

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland (1975), *Discrete Multivariate Analysis*, The MIT Press, Cambridge, Mass.

Garson, David G. (2006), *Logistic Regression*. From the website http://www2.chass.ncsu.edu/garson/pa765/logistic.html

Howell, David C. (2002), *Statistical Methods for Psychology*, Duxbury, USA.

Scheaffer, Richard L. (1999), *Categorical Data Analysis*. From the website http://courses.ncssm.edu/math/Stat_Inst/PDFS/Categorical%20Data%20Analysis.pdf

ANALYSIS OF CATEGORICAL DATA USING LOG-LINEAR MODELS

At the end of the, tutorial the user should be able to

- Use log-linear models in the analysis of categorical data

I. Introduction

Log linear models describe association patterns among categorical variables. They are introduced to bring the familiar linear model structure to the analysis of categorical data and to extend the chi-square analysis to higher dimensions in a more structured way than is otherwise possible. With the log linear approach, the cell counts are modeled in a contingency table in terms of linear functions of cell probabilities where coefficients explain the association among the variables. In this sense log linear models analysis is like correlation analysis for continuous variables.

Three sampling models are commonly used to describe cross-classified frequency or count data. These are the Poisson, Multinomial and Product Multinomial models.

Cross classified count data comprises a series of counts or frequencies corresponding to level combinations of one or more explanatory factors. These level combinations define cells of a contingency table for example, counts of insects may be classified according to species and position on a plant where they are found.

The Poisson model results from assuming that the count in each cell is the realization of an independent Poisson process observed for some fixed time without any prior knowledge of the total number of possible observations.

The Multinomial model arises when we have a fixed number of subjects which are classified into the different cells according to their values of the explanatory factors.

The Product Multinomial distribution arises when we have several sets of subjects, each belonging, a priori, to some sub-classification of the contingency table, one set for each row for example. Members of each set are then classified according to their values of the remaining explanatory factors according to a Multinomial model.

The important result concerning these different sampling situations is that they all lead to the same expected values for cell frequencies and the same tests for association between classifying factors because the tests are conditional on the marginal frequencies whether fixed in advance or randomly observed.

II. Two-way Contingency Tables

We will begin with the simplest example of a 2×2 contingency table. As an example we will use the 2×2 table for level of education and adoption of nitrogen fertilizer given below.

Level of Education	Adoption		Total
	Yes	No	
High	21 (11.61)	6 (15.39)	27
Low	22 (31.39)	51 (41.61)	73
Total	43	57	100

Note: figures in () are expected frequencies assuming the classifying factors are independent.

The standard Pearson Chi-square test on these data gives the following result with P<.001.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(f_{ij}-F_{ij})^2}{F_{ij}} = 18.25$$

The second formula represents a change in notation, substituting f_{ij} for the observed frequency in cell ij and F_{ij} for the expected frequency in that cell. This is instituted to bring the notation in line with the standard notation used with log-linear analysis.

If we calculate the likelihood ratio statistic based on the Poisson distribution with independent and dependent cell frequencies instead of the Pearson's Chi-square, we would have

$$X^2 = 2\sum f_{ij} \ln\left(\frac{f_{ij}}{F_{ij}}\right)$$

$$= 2\left[21\times\ln\left(\frac{21}{11.61}\right) + 6\times\ln\left(\frac{6}{15.39}\right) + 22\times\ln\left(\frac{22}{31.39}\right) + 51\times\ln\left(\frac{51}{41.61}\right)\right]$$

$$= 18.704$$

This is also approximated by the χ^2 distribution with 1 degree of freedom. Again we would reject the null hypothesis of independence of rows and columns. We would conclude that the level of farmer's education is associated with the decision to use nitrogen fertilizer or not.

The use of the chi-square test, whether Pearson's statistic or the likelihood ratio statistic, focuses directly on hypothesis testing. But we can look at these data from a different perspective – the perspective of model building.

To do this we consider possible models for a two-way table.

Equiprobability Model

At the simplest level, we might hypothesize that respondents distribute themselves among the four cells at random. In other words, p(low, adopter)=p(low, non-adopter)=p(high, adopter)=p(high, non-adopter)=0.25. This model basically says that nothing interesting is going on in this study and one-quarter of the subjects $(0.25 \times 100 = 25)$ would be expected to fall in each cell.

Using the likelihood ratio χ^2 to test this model, we have

Level of Education	Adoption		Total
	Yes	No	
High	21 (25)	6 (25)	27
Low	22 (25)	51 (25)	73
Total	57	43	100

$$\text{Likelihood ratio } \chi^2 = 2\left[21 \times \ln\left(\frac{21}{25}\right) + 6 \times \ln\left(\frac{6}{25}\right) + 22 \times \ln\left(\frac{22}{25}\right) + 51 \times \ln\left(\frac{51}{25}\right)\right]$$
$$= 42.648$$

This can be evaluated as χ^2 on 4-1=3 df (we lose one degree of freedom due to the restriction that the cell totals must sum to N). From the χ^2 table we find that $\chi^2_{0.05}(3) = 7.82$. Clearly, we can reject H_0 and conclude that this model does not fit the data. In other words, the individual cell frequencies cannot be fit by a model in

which all cells are considered equally probable. Notice that rejection of H_0 is equivalent to rejection of the underlying model.

Conditional Equiprobability Model

A second model might hold that the individual cell frequencies represent differences due to level of education alone, because there were more low-educated than high-educated farmers. Under this model 73/100 =73% of the observations fall in row 1 and 27% fall in row 2. Beyond that, however, observations are assumed to be equally likely to fall in columns 1 and 2. In other words, the null hypothesis states that once we have adjusted for the fact that more farmers are low-educated than high-educated, adoption and non-adoption is equally probable. Put another way, adoption is equiprobable, conditional on level of education. By this model, we would have the expected frequencies (shown in parenthesis) in the table below.

Level of Education	Adoption		Total
	Yes	No	
High	21 (13.5)	6 (13.5)	27
Low	22 (36.5)	51 (36.5)	73
Total	43	57	100

$$\text{Likelihood ratio } \chi^2 = 2\left[51\times\ln\left(\frac{51}{36.5}\right)+22\times\ln\left(\frac{22}{36.5}\right)+6\times\ln\left(\frac{6}{13.5}\right)+21\times\ln\left(\frac{21}{13.5}\right)\right]$$
$$= 20.67$$

This model has 4-2=2 degrees of freedom because we have imposed two restrictions – the cell frequencies in each row must sum to the expected frequency for that row. Since $\chi^2_{0.05}(2) = 5.99$, we will again reject H_o and conclude that the model does not fit the observed data either.

A second conditional equiprobability model could be created by assuming that cell frequencies are affected only by differences in levels of adoption. In this case probabilities are equal within each adoption level but different between them. The expected frequencies are given below.

Level of Education	Adoption		Total
	Yes	No	
High	21 (21.5)	6 (28.5)	27
Low	22 (21.5)	51 (28.5)	73
Total	43	57	100

$$\text{Likelihood ratio}\ \chi^2 = 2\left[21\times\ln\left(\frac{21}{21.5}\right)+6\times\ln\left(\frac{6}{28.5}\right)+22\times\ln\left(\frac{22}{21.5}\right)+51\times\ln\left(\frac{51}{28.5}\right)\right]$$
$$= 40.682$$

This χ^2 again has 2 degrees of freedom and is significant. Thus, we have so far concluded that the data cannot be explained by assuming that observations fall in the four cells are `at random. Nor can they be explained by positing differences due simply to an unequal distribution across either level of education or adoption.

In fact these marginal tests of individual factors are not usually interesting – whether significant or not because they either reflect natural frequencies in the sampled population or they are chosen by the sampling scheme. In either case interest usually centers on whether the marginal frequencies alone account for difference or not.

The next step would be to propose a model involving both level of education and adoption operating independently of one another. This is the standard null model routinely assumed for a chi-square test on a contingency table.

Mutual Independence Model

We now assume that the two factors operate jointly, but independently, to produce expected cell frequencies. If the two variables are independent then

$$F_{ij} = \frac{RT\times CT}{GT} = \frac{f_{i.}\times f_{.j}}{f_{..}}$$

where RT stands for the row total, CT for the column total, GT for the grand total, and the "dot notation" is used to show that we have summed frequencies across that dimension. This is the same formula for expected frequencies that we saw in the case of Pearson's χ^2 test.

We began this section by testing this hypothesis of independence. The expected frequencies and the likelihood ratio were then given. From those calculations we found that

$$\chi^2 = 3.84$$

which is significant on 1 df. Thus, we can further conclude, and importantly so in this case, that a model that posits independence between level of education and adoption also does not fit the data. The only conclusion remaining is that the probability of adoption is associated with his/her level of education. That is, there is an interaction between level of education and adoption of nitrogen fertilizer.

Saturated Model

This is a model in which every expected frequency is forced to be exactly equal to every obtained frequency, and χ^2 will be exactly 0. A saturated model always fits the data perfectly. In log-linear analysis, this model is not tested directly. The saturated model in the $R \times C$ case is basically the model that we adopt if the mutual independence model is rejected.

III. Introduction to Log-linear models

The models discussed above can be represented algebraically and can be compared with the analysis of variance model. In an analysis of variance we may have a model like:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

In the simplest equiprobability model, all cell frequencies are explained by a single parameter $\tau_{..}$ where τ is estimated by the geometric mean of the expected cell frequencies given by the model. That is,

$$F_{ij} = \hat{\tau}$$

A geometric mean is the nth root of the product of n terms, so in this case the geometric mean of the four expected frequencies is:

$$\sqrt[4]{(25)(25)(25)(25)} = 25$$

For the first conditional equiprobable model we again define $\hat{\tau}$ as the geometric mean of the expected cell frequencies in that model:

$$\hat{\tau} = \sqrt[4]{(13.5)(13.5)(36.5)(36.5)} = 22.19797$$

We also define $\hat{\tau}_1^E$ (where the superscript "E" stands for Education) as the ratio of the geometric mean of the expected frequencies for the high level of education to the geometric mean of all the cells ($\hat{\tau}$). Hence,

$$\hat{\tau}_1^E = \frac{\sqrt{(13.5)(13.5)}}{\hat{\tau}} = \frac{13.5}{22.19797} = 0.608164$$

$\hat{\tau}_1^E$ is similar to treatment effect (β_j) in the analysis of variance. However in the analysis of variance, β_j is the amount that is added to the grand mean to obtain the row mean. In log-linear models $\hat{\tau}_1^E$ is the amount by which we multiply $\hat{\tau}$ to obtain the row's expected frequency.

For the low level of education,

$$\hat{\tau}_2^E = \frac{\sqrt{(36.5)(36.5)}}{\hat{\tau}} = \frac{36.5}{22.19797} = 1.644294$$

Then we show that for this model

$$F_{ij} = \hat{\tau}\hat{\tau}_i^E$$

For cell 1,1 we would have $22.19797 \times 0.608164 = 13.5$, which is the same expected frequency as what we have in the table.

If we go a little bit further, we can consider the independence model which contained both education and adoption effects but not their interaction. Here we need both $\hat{\tau}_i^E$ and $\hat{\tau}_i^A$ to account for both education and adoption effects.

$$\hat{\tau} = \sqrt[4]{(11.61)(15.39)(31.39)(41.61)} = 21.97936$$
$$\hat{\tau}_1^E = \frac{\sqrt{(11.61)(15.39)}}{\hat{\tau}} = \frac{13.36705}{21.97936} = 0.608164$$
$$\hat{\tau}_2^E = \frac{\sqrt{(31.39)(41.61)}}{\hat{\tau}} = \frac{36.14053}{21.97936} = 1.644294$$
$$\hat{\tau}_1^A = \frac{\sqrt{(11.61)(31.39)}}{\hat{\tau}} = \frac{19.09026}{21.97936} = 0.868554$$
$$\hat{\tau}_2^A = \frac{\sqrt{(15.39)(41.61)}}{\hat{\tau}} = \frac{25.30569}{21.97936} = 1.151339$$

Thus for our example,

$$F_{11} = \hat{\tau}\hat{\tau}_1^E\hat{\tau}_1^A = 21.97936 \times 0.608164 \times 0.868554 = 11.61$$

which agrees with the actual expected value for the independence model. In the general case, for the independence model, the expected frequency for cell ij is

$$F_{ij} = \hat{\tau}\hat{\tau}_i^E\hat{\tau}_j^A$$

This is a multiplicative model, unlike in the analysis of variance wherein the model is additive. However, if we transform the preceding equation using the natural logarithm, we have

$$\ln(F_{ij}) = \ln(\hat{\tau}) + \ln(\hat{\tau}_i^E) + \ln(\hat{\tau}_j^A)$$

If we substitute λ for $\ln(\hat{\tau})$ we have

$$\ln(F_{ij}) = \lambda + \lambda_i^E + \lambda_j^A$$

which is an additive linear expression directly analogous to the analysis of variance model. This model is linear in the logs, hence the name log-linear models.

Given this notation, we can now express the different models using log-linear models:

1. Equiprobability model: $\ln(F_{ij}) = \lambda$
2. Conditional equiprobability model 1: $\ln(F_{ij}) = \lambda + \lambda_i^E$
3. Conditional equiprobability model 2: $\ln(F_{ij}) = \lambda + \lambda_j^A$
4. Mutual independence model: $\ln(F_{ij}) = \lambda + \lambda_i^E + \lambda_j^A$
5. Saturated model: $\ln(F_{ij}) = \lambda + \lambda_i^E + \lambda_j^A + \lambda_{ij}^{EA}$

Testing Models

The central issue in log-linear analysis is the issue of choosing an optimal model to fit the data. Within the example we have been discussing we have five possible models. We have computed the χ^2 value, degrees of freedom and test of significance of the different models and this is summarized below.

Model	χ^2	df	Test
1. $\ln(F_{ij}) = \lambda$	42.648	3	*
2. $\ln(F_{ij}) = \lambda + \lambda_i^E$	20.670	2	*
3. $\ln(F_{ij}) = \lambda + \lambda_j^A$	40.682	2	*
4. $\ln(F_{ij}) = \lambda + \lambda_i^E + \lambda_j^A$	18.704	1	*
5. $\ln(F_{ij}) = \lambda + \lambda_i^E + \lambda_j^A + \lambda_{ij}^{EA}$	0.00	0	-

We have seen that the first four models all have significant χ^2 values. This means that for each of these models there is a significant difference between observed and expected values; none of them fits the obtained data. From such results we must conclude that only a model that incorporates the interaction term can account for the results. Thus, as we have previously concluded, level of education and adoption of nitrogen fertilizer interact and hence we cannot model the data without taking this interaction into account**.** In view of the above results, we may conclude th**at**

$$\ln(F_{ij}) = \lambda + \lambda_i^E + \lambda_j^A + \lambda_{ij}^{EA}$$

fits the obtained data.

However, as in the analysis of variance we may want to know whether the main effects of education and adoption are significant. We can do this in two ways:

1. Perform chi-square tests on the marginal totals.

    a. Effect of education

	Level of Education	
	High	Low
f_{ij}	27	73
F_{ij}	50	50

$$\chi^2 = 2\sum f_{ij} \ln\left(\frac{f_{ij}}{F_{ij}}\right)$$
$$= 2\left(27\ln\frac{27}{50} + 73\ln\frac{73}{50}\right)$$
$$= 21.978$$

    b. Effect of adoption

	Adoption	
	Yes	No
f_{ij}	43	57
F_{ij}	50	50

$$\chi^2 = 2\sum f_{ij} \ln\left(\frac{f_{ij}}{F_{ij}}\right)$$
$$= 2\left(43\ln\frac{43}{50} + 57\ln\frac{57}{50}\right)$$
$$= 1.966$$

2. Get difference of χ^2 values of concerned models

    a. Effect of education

$\chi^2 = \chi^2$(Model 1) - χ^2(Model 2)

$\chi^2 = 42.648 - 20.670 = 21.978$

b. Effect of adoption

$\chi^2 = \chi^2$(Model 1) - χ^2(Model 3)

$\chi^2 = 42.648 - 40.682 = 1.966$

We have found that the simplest model [$\ln(F_{ij}) = \lambda$] produces a χ^2=42.648. When we added λ^E to this model,. χ^2 dropped to 20.670, reflecting the variation in cell frequencies attributable to level of education. This drop (42.648-20.670) is the χ^2 for level of education, and its degrees of freedom equal the difference between the degrees of freedom in the two models (3 – 2 = 1). This is exactly the same value we obtained in Method 1a, when we compare the marginal frequencies.

By a similar line of reasoning, we can note that taking adoption into account and going from $\ln(F_{ij}) = \lambda$ to $\ln(F_{ij}) = \lambda + \lambda_j^A$ reduces χ^2 from 42.648 to 40.682, for a decrease of 1.966. This is the same as the marginal χ^2 on adoption that we obtained in Method 1b.

Finally, we should note that when we go from a model of $\ln(F_{ij}) = \lambda + \lambda_i^E + \lambda_j^A$ to $\ln(F_{ij}) = \lambda + \lambda_i^E + \lambda_j^A + \lambda_{ij}^{EA}$, χ^2 drops from 18.704 to 0. This drop (18.704) is the same as the χ^2 for the interaction based on marginal frequencies. This equality will not generally hold for more complex designs unless we are looking at the highest-order interaction.

One other feature of log-linear models should be mentioned. The minimal model $\ln(F_{ij}) = \lambda$ produced χ^2=42.648. The individual components of the saturated model had χ^2 values of 21.978, 1.966, and 18.704. These sum to 42.648. In other words, these likelihood ratio χ^2 values are additive. This would not have been the case had we computed the Pearson chi-square statistic instead, which is one good reason to concentrate on likelihood ratio χ^2.

IV. Three-way Tables

We now have all the concepts that are necessary to move to more complex designs. Log-linear models come into their own once we move to contingency tables of more than two dimensions. These are the situations in which standard chi-square analyses are not able to reveal a full understanding of the data. In this section we will concentrate on three-way tables because they illustrate all of the essential points. Extrapolation to tables of higher dimensionality is direct.

One of the pleasant things about log-linear models is the relative absence of assumptions. Like the more traditional chi-square test, log-linear analysis does not make assumptions about population distributions, although it does assume, as does

Pearson's chi-square, that observations are independent. You may apply log-linear analysis in a wide variety of circumstances, including even the analysis of badly distributed (ill-behaved) continuous variables that have been classified into discrete categories.

The major problem with log-linear analysis is the same problem that we encountered with traditional chi-square: the expected frequencies have to be sufficiently large to allow the assumption that frequencies in each cell are approximately normally distributed over repeated sampling. In the case of chi-square, we set the rule that all (or at least most) of the expected frequencies should be at least 5. We have a similar situation with log-linear analysis. Once again we require at least that all cells have expected frequencies greater than 1 and that no more than 20% of the cells have expected frequencies less than 5. The biggest problem comes with what are called sparse matrices, which are contingency tables with a large number of empty cells. In these cases you may wish to combine categories on the basis of some theoretical rationale, increase sample sizes, collapse across variables, or do whatever you can to increase the expected frequencies. Regardless of the effects such small cells have on the level of Type I errors, you are virtually certain to have very low levels of power.

Hierarchical and Nonhierarchical Models

Most, but not all, analyses of log-linear models involve what are called hierarchical models. You can think of a hierarchical model as one for which the presence of an interaction term requires the presence of all lower-order interactions and main effects involving the components of that higher-order interaction. For example, suppose that we had four variables, A, B, C, and D. If you included in the model the three-way interaction ACD, a hierarchical model would also have to include A, C, D, AC, AD, and CD, because each of these terms is a subset of ACD. Similarly, if your model included ABC and ABD, the model would actually include A, B, C, D, AB, AC, BC, AD and BD. It need not include CD, ACD, BCD, or ABCD, because those are not components of either of the three-way interactions.

One of the convenient things about hierarchical designs is that they allow us to specify models very clearly and simply. Assume that we have four variables (A, B, C, and D). The notation ABC specifies a model that includes the ABC interaction, and, because we are speaking about hierarchical models, also includes A, B, C, AB, AC, and BC. We do not have to write out the latter to specify the model – ABC will suffice. Similarly, the label AB stands for a model that includes A, B, and AB, but not C or any interactions involving C. Finally a model written as AB, ACD is really the model that involves A, B, C, D, AB, AC, AD, CD, and ACD, but not BC, BD, ABC, ABD, or BCD. We will characterize models by the interactions that define them (sometimes called their defining set, or generating class).

V. Three-Way Example

As an example, we consider a study whose objective is to determine the relationship between level of education, income and adoption of a new technology.

Adoption of New Tech.	Level of Education	Income			Total
		High	Medium	Low	
Yes	High	42	79	32	153
	Low	23	65	17	105
	Total	65	144	49	258
No	High	4	12	8	24
	Low	11	41	24	76
	Total	15	53	32	100
	Column Total	80	197	81	358

Examining the Saturated Model

In considering two-way tables, we defined a saturated model as one that includes all possible effects. The same holds for three-way and higher-order tables. Consider the model that can be designated as AEI or written as

$$\ln(F_{ij}) = \lambda + \lambda_i^A + \lambda_j^E + \lambda_k^I + \lambda_{ij}^{AE} + \lambda_{ik}^{AI} + \lambda_{jk}^{EI} + \lambda_{ijk}^{AEI}$$

This is the saturated model for our data. It includes all possible effects and exhausts the degrees of freedom available in the data. (One degree of freedom goes to estimating λ, one each for estimating the λs associated with effects A, E, and AE, and two each to estimating those associated with I, AI, EI, and AEI. I has three levels and thus two degrees of freedom (independent λs) for it and its interactions). There are 12 λs to estimate and since we have 12 cells there are no degrees of freedom left. If we knew the values of the various lambdas, and eventually we will, the resultant expected frequencies would equal the observed frequencies, leaving nothing else to be explained. For this reason we know without even looking at the data that the likelihood ratio χ^2 for this model will be exactly 0. We should not be happier with this perfect fit than we are when we draw a straight line to fit perfectly any two points, and for the same reason – the model exhausts the degrees of freedom.

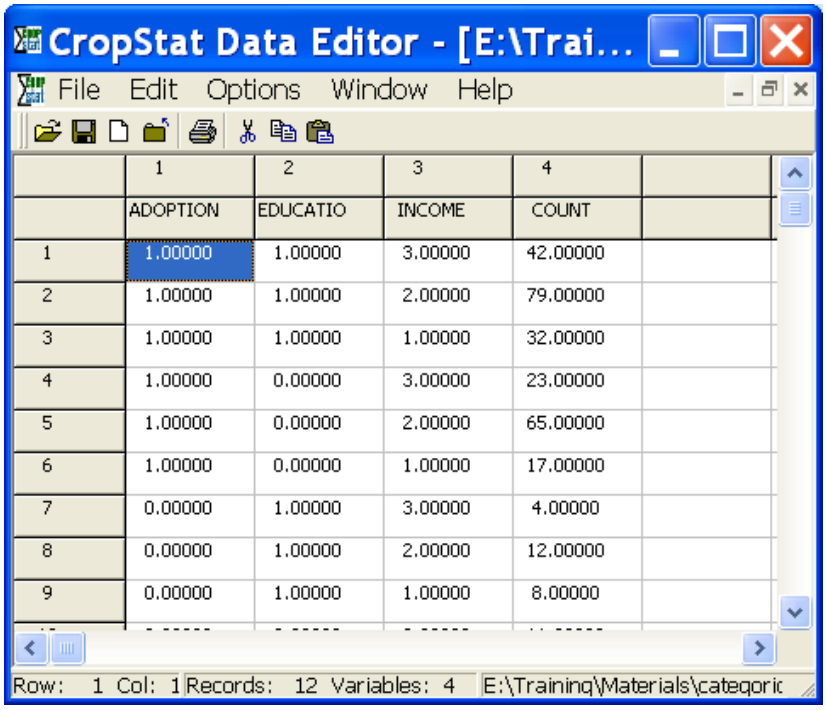
We do not fit a saturated model to data just because we hope that it will fit – we know that before we start. We usually fit it hoping that it will help us identify simpler

models by revealing non-significant effects. If we could show, for example, that we could do about as well by eliminating the three-way interaction and two of the two-way interactions, we would be well on our way to representing the data by a relatively simple model.

One of the reasons for starting with the saturated model is that you can then ask whether various levels of interaction are needed in the model.

Choosing the Most Parsimonious Log-linear Model in CropStat

To illustrate how to perform log-linear analysis in CropStat, we use the sample data set *LOGEX.SYS*.



	1	2	3	4
	ADOPTION	EDUCATIO	INCOME	COUNT
1	1.00000	1.00000	3.00000	42.00000
2	1.00000	1.00000	2.00000	79.00000
3	1.00000	1.00000	1.00000	32.00000
4	1.00000	0.00000	3.00000	23.00000
5	1.00000	0.00000	2.00000	65.00000
6	1.00000	0.00000	1.00000	17.00000
7	0.00000	1.00000	3.00000	4.00000
8	0.00000	1.00000	2.00000	12.00000
9	0.00000	1.00000	1.00000	8.00000

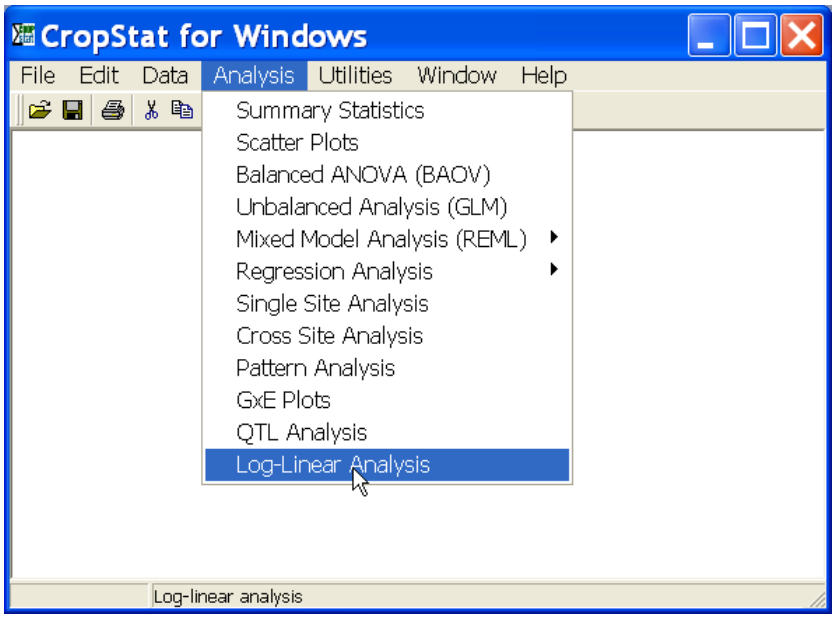
- Open the data file *LOGEX.SYS* from the *CROPSTAT7.2\TUTORIAL\ TUTORIAL DATASETS* folder.

- Select **File ⇒ Save-as**. Click the **Save in** box and go inside your working folder *C:\MY CROPSTAT*. Create a subfolder LOG LINEAR then click **Save**.

- Choose **Log-Linear Analysis** from the Analysis menu.

  A data file for log-linear analysis must contain factors for each categorical variable in the contingency table. These factors may be character variables or numeric but contain discrete levels for each categorical variable. The combination of level values over all factors in a row of the data file identifies a cell in the

contingency table. The data file may also contain a numeric variable containing frequencies of observations in each cell – a count variable. If not, the frequencies are derived by counting the number of rows with the same factor level combinations. Even when there is a count variable the frequencies in the table are obtained by summing the counts over rows with the same level combinations. This allows tables to be collapsed by simply omitting factors defining the categories to be collapsed without having to re-compute the counts.



- The **Open** dialog box will prompt you to enter a name for the command file. Click the **Look In** box to go to your working drive C:\MY CROPSTAT\LOG LINEAR.

- Enter *LOGEX* in the **File name** box. Click **Open** button.

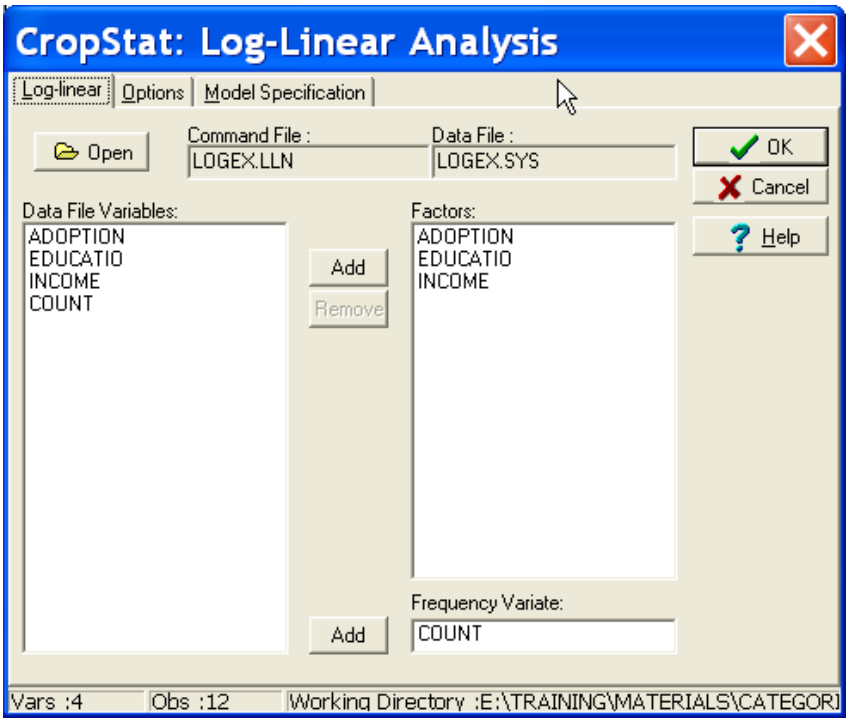
- Since *LOGEX.LLN* does not exist, a message box will appear confirming if you want to create the file. Click **Yes** to create new Command File.

- Enter the name of the data file to be used. Enter *LOGEX.SYS* in the **File name** box.

- Click **Open**. The **Log-Linear Analysis** dialog box will appear for you to fill-in the details of the analysis.

- From the **Data File Variable** list, highlight all factors included in the analysis then add to the Factors box.

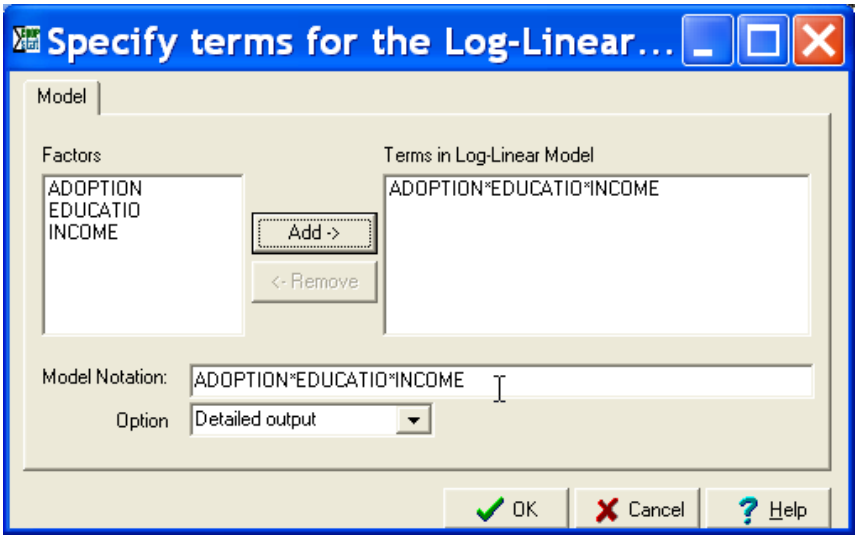
- If your data file contains the frequency counts for each cell of your contingency table, highlight the variable in the **Data File Variables** then click 'Add' to add the **Frequency Variate**. Otherwise, go to the next step.



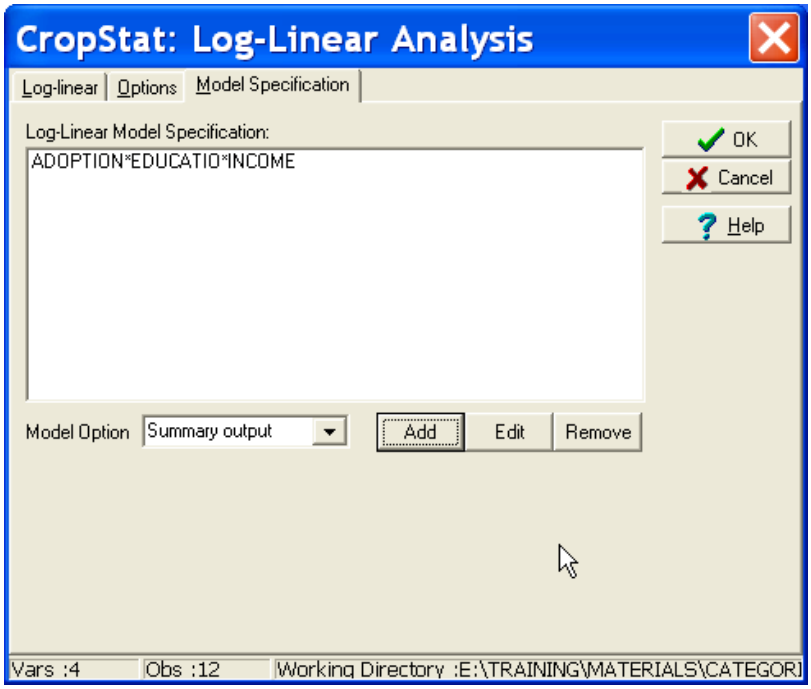
- Click the **Model Specification** tab. The **Log-linear model specification** window will appear.

- Click **Add**. The **Specify terms for the Log-Linear Model** dialog box will appear.

- Select all factors and add into the model. This defines the saturated model.



- Click the **Ok** and the **Log-linear model specification** window will appear again.



- You may add another model by clicking **Add**. But for now click **Ok** to analyze the saturated model.

CropStat Output

Table A. Factor Description

```
FACTOR A    INCOME      HAS     3 LEVELS:
1             2             3

FACTOR B EDUCATIO       HAS     2 LEVELS:
0             1

FACTOR C ADOPTION       HAS     2 LEVELS:
0             1
```

CropStat assigns the letter A to the first factor, B to the second, etc.

Table B. Observed Counts

TABLE OF OBSERVED COUNTS

	EDUCATIO	0	0	1	1
	ADOPTION	0	1	0	1
	INCOME				
1	COUNT	24.0	17.0	8.0	32.0
2	COUNT	41.0	65.0	12.0	79.0
3	COUNT	11.0	23.0	4.0	42.0

Table C. Partial Association Statistics

Partial Association Statistics

Omitted Effect	Chi-Square	Degrees of Freedom	P-value	Marginal Zeros
A	70.75	2.0	0.0000	0.0
B	0.04	1.0	0.8326	0.0
C	72.19	1.0	0.0000	0.0
A*B	2.56	2.0	0.2785	0.0
A*C	8.41	2.0	0.0149	0.0
B*C	36.99	1.0	0.0000	0.0
A*B*C	0.26	2.0	0.8801	0.0

Table D. Test for all k and higher interactions are zero

Chi-square statistics for testing that all k and higher interactions are zero.

k	Likelihood Ratio	P-Value	Degrees of Freedom	Pearson	P-Value
1	191.92	0.0000	11.0	200.91	0.0000
2	48.93	0.0000	7.0	49.02	0.0000
3	0.26	0.8801	2.0	0.26	0.8802

Table E. Test for all k-factor interactions are simultaneously zero

Chi-square statistics for testing that all k-factor interactions are simultaneously zero.

k	Likelihood Ratio	P-Value	Degrees of Freedom	Pearson	P-Value
1	142.99	0.0000	4.0	151.88	0.0000
2	48.68	0.0000	5.0	48.77	0.0000
3	0.26	0.8801	2.0	0.26	0.8802

The first item of interest in the output is the section showing the simultaneous test on main effects and interactions (Tables D and E). Either of the two tables is quite valuable in determining how complex a model is needed to fit the data. Pooled chi-square tests (both Pearson and Likelihood ratio) are performed to inform you whether any k-way (one-way, two-way, etc.) effects are significant. Of the two tables, Table E provides a more direct summary, although the same conclusions can be derived from Table D. For example, in Table E, the row where k is 2 tests whether any two-way interactions are significant. The corresponding row in Table D tests whether any two-way or three-way interactions are significant. From either table we would conclude that there is no significant three-way interaction, but there is at least one significant two-way interaction.

The next step in identifying a model is shown in Table C. From Tables D and E we have concluded that the 3-factor interaction is not significant. Hence the next model to consider is a model containing all three 2-factor interactions. This model as shown in Tables D and E has a χ^2 value 0.26 which indicates that the model fits the data well. But the question is: Are all two-factor interactions significant? From Table C we can see that removing A*B ($I \times E$) from the model the χ^2 value 2.56 is not significant indicating that the model without the $I \times E$ term also fits the data. But this is not true with the $A \times I$ and $A \times E$ interactions. Excluding any of these terms from the model will result to significant χ^2 values. Since the objective of log-linear analysis is to fit the most parsimonious model, then we conclude that

$$\ln(F_{ij}) = \lambda + \lambda_i^A + \lambda_j^E + \lambda_k^I + \lambda_{ij}^{AE} + \lambda_{ik}^{AI}.$$

The defining set of effects for this model is AE and AI.

Estimating Predicted Frequencies and Residuals

After choosing the most parsimonious model, we have to estimate the predicted frequencies and examine residuals.

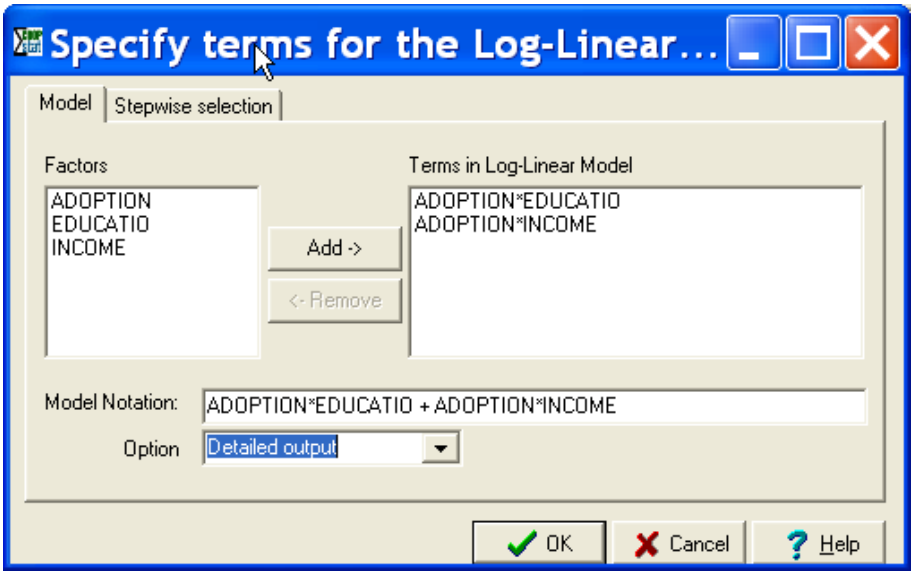
- Choose **Log-Linear Analysis** from the Analysis menu.
- Open the *LOGEX.LLN* command file we have created earlier.

- Click the **Model Specification** tab. Highlight the model defined earlier then click **Remove**.

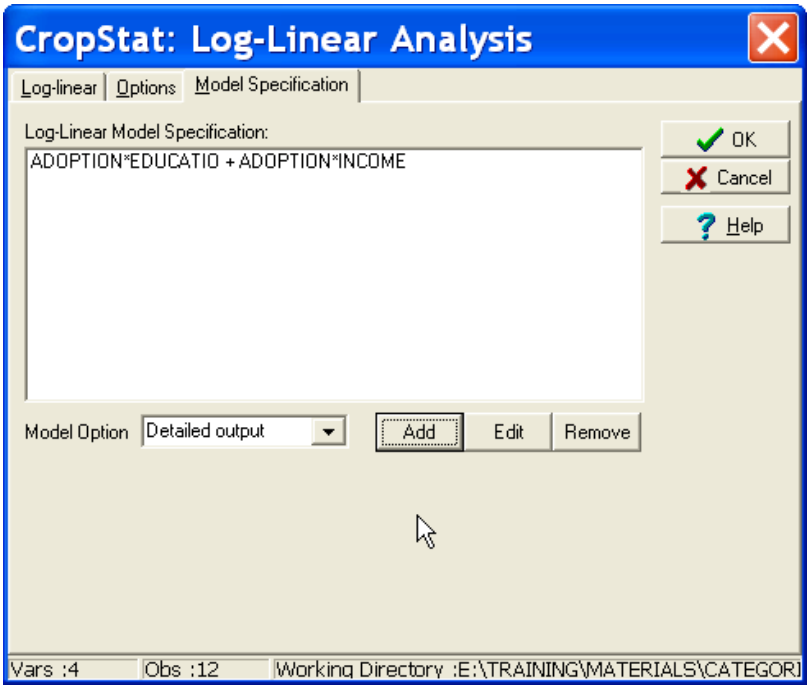
- Click **Add**. The **Specify terms for the Log-Linear Model** dialog box will appear. We need only to specify the defining set of effects to specify the model we want.

- Highlight Adoption and Education then click **Add**. Highlight Adoption and Income then click **Add**.

- Click the down arrow key for option then choose *Detailed output*.



- Click the **Ok** and the **Log-linear model specification** window will appear again.



- Click **Ok** to run the model.

CropStat Output

Goodness of Fit Statistics

```
MODEL: EDUCATIO*ADOPTION +INCOME*ADOPTION HAS   8 COEFFICIENTS

 FITTING STATISTICS

 Log-Likelihood    3.660

 Likelihood ratio   2.812

 Degrees of Freedom     4.

 p-Value           0.5898
```

Parameter Estimates

PARAMETER ESTIMATES

NO.	PARAMETER	COEFFICIENT	S.E	ASYMPTOTIC Z	P-VALUE
1	INTERCEPT	3.083	0.7343E-01	41.98	0.0000
2	INCOME(1)	-0.1846	0.9347E-01	-1.97	0.1195
3	INCOME(2)	0.6067	0.8113E-01	7.48	0.0017
4	EDUCATIO	0.1941	0.6656E-01	2.92	0.0434
5	ADOPTION	-0.5520	0.7343E-01	-7.52	0.0017
6	INCOME.ADOPTION(1)	0.2689	0.9347E-01	2.88	0.0451
7	INCOME.ADOPTION(2)	-0.1777E-01	0.8113E-01	-0.22	0.8373
8	EDUCATIO.ADOPTION	0.3823	0.6656E-01	5.74	0.0046

For each effect the parameter estimate, standard error, z statistic, and P-value are given. For income there are two dummy variables created with high income as the reference income. That is the variable Income(1) compares low with high income and Income(2) compares medium with high income. Result shows that Income(1) is not significant indicating that a farmer is equally likely to be in the low and high income bracket. On the other hand, Income(2) is significant and the sign of the coefficient is positive. This result indicates that a farmer is more likely to be from the middle than high income bracket. Both education and adoption are significant. However, the signs of their coefficients are different. This means that a farmer is more likely to have a low than a high education level and more likely to be an adopter than a non-adopter.

Table of Predicted Frequencies

TABLE OF FITTED FREQUENCIES

INCOME	EDUCATIO	ADOPTION		0	1
	1	0	OBSERVED	24.	17.
			FITTED	24.32	19.94
	1	1	OBSERVED	8.	32.
			FITTED	7.68	29.06
	2	0	OBSERVED	41.	65.
			FITTED	40.28	58.60
	2	1	OBSERVED	12.	79.
			FITTED	12.72	85.40
	3	0	OBSERVED	11.	23.
			FITTED	11.40	26.45
	3	1	OBSERVED	4.	42.
			FITTED	3.60	38.55

Table of Residuals

TABLE OF FOUR FORMS OF RESIDUALS

INCOME	EDUCATIO	ADOPTION		0	1
1		0	ROOT CHISQ	-0.06	-0.66
			LIKELIHOOD	-0.64	-5.43
			F-T RESIDS	-0.01	-0.62
			RESIDUALS	-0.32	-2.94
1		1	ROOT CHISQ	0.12	0.55
			LIKELIHOOD	0.65	6.17
			F-T RESIDS	0.20	0.57
			RESIDUALS	0.32	2.94
2		0	ROOT CHISQ	0.11	0.84
			LIKELIHOOD	1.45	13.46
			F-T RESIDS	0.15	0.84
			RESIDUALS	0.72	6.40
2		1	ROOT CHISQ	-0.20	-0.69
			LIKELIHOOD	-1.40	-12.30
			F-T RESIDS	-0.13	-0.68
			RESIDUALS	-0.72	-6.40
3		0	ROOT CHISQ	-0.12	-0.67
			LIKELIHOOD	-0.79	-6.44
			F-T RESIDS	-0.05	-0.64
			RESIDUALS	-0.40	-3.45
3		1	ROOT CHISQ	0.21	0.56
			LIKELIHOOD	0.84	7.21
			F-T RESIDS	0.31	0.58
			RESIDUALS	0.40	3.45

CropStat outputs four types of residuals

- Root ChiSq is the standardized residuals. Here the raw residual is divided by the estimated standard deviation of observed counts $\left(\frac{f_i - F_i}{\sqrt{F_i}}\right)$. Significant standardized residuals have absolute values greater than 1.96 and such cells

may be considered "model outliers." As a rule of thumb, more than one model outlier per 20 table cells may cause the researcher to seek a different model.

- Likelihood residuals are the deviance residuals and are also called "studentized deviance residuals". They indicate how much each cell contributes to the likelihood ratio. Likelihood ratio chi-square is the sum of squared cell deviances. These residuals also have a mean of 0 and a standard deviation of 1 for large samples. They are computed as:

$$2f_i \log\left(\frac{f_i}{F_i}\right)$$

- F-T Resids are the Freeman-Tukey deviates. They are computed as

$$\sqrt{f_i} + \sqrt{f_i + 1} - \sqrt{4F_i + 1}\,.$$

This is derived by considering the variance-stabilizing transformation

$$y_i = \sqrt{f_i} + \sqrt{f_i + 1}\,.$$

In the event that f_i follows a Poisson distribution with mean F_i, we use the result that y_i is approximately normally distributed, with approximate mean

$$\sqrt{4F_i + 1}$$

and variance 1.

- Residuals: : Observed minus expected frequency $(f_i - F_i)$

References:

Allan, E., R.D. Stern, R. Coe and J. De Wolf (2002), *Data Analysis of Agroforestry Experiments* (workshop handout), World Agroforestry Centre

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland (1975), *Discrete Multivariate Analysis*, The MIT Press, Cambridge, Mass.

Garson, David G. (2006), *Log-Linear, Logit, and Probit Models*. From the website http://www2.chass.ncsu.edu/garson/pa765/logit.html

Howell, David C. (2002), *Statistical Methods for Psychology*, Duxbury, USA.

Scheaffer, Richard L. (1999), *Categorical Data Analysis*. From the website http://courses.ncssm.edu/math/Stat_Inst/PDFS/Categorical%20Data%20Analysis.pdf