Developing GenStat into the 21st Century

Roger Payne & Simon Harding,
VSN International, Hemel Hempstead, Herts, UK
email: Roger.Payne@vsn-intl.com

RSS/ASC 25th January 2006
History - development

- Rothamsted pioneered the use of computers for statistics in agriculture from 1954 onwards
- Genstat development began at RES in 1968
- Batch on ICL system 4’s from 1970
- Revised, interactive Genstat 5 on VAX in 1987
  ..
History - development

- first PC DOS version in 1989
- GenStat *for Windows*
  - 8th Edition, 2005

..
History - commercialisation

- Originally written for Rothamsted
- Distributed more widely in 1970's
  - Initially by Rothamsted
  - 1979 by Numerical Algorithms Group
- Now developed & marketed by VSN International
  - Joint venture: Rothamsted + NAG
  - Providing stronger collaboration between R&D and sales & marketing
  - Close focus on user needs
  - But still strong links to the research community

..
Design principles

• it's easier if you can get it right first time...

• external
  • flexible and extendable command language
  • programming features
  • standard (and varied) data structures
  • general algorithms (customise by macro / procedure)

• internal
  • avoid non-standard language features (standard Fortran, C &c)
  • identify and centralise machine dependencies
  • standardised use of internal workspace
  • shared utilities
  • documented code
  • centralised output
  • reusable components

...
Fundamental issues

• as far as possible use only standard features of the implementation environment
  • standard Fortran
  • extensions identified and localised
  • machine dependencies parameterised
  • conversion comments for unavoidable alternatives (e.g. i/o)

• in 1970's and 80's
  • crucial to allow implementation on a range of platforms: e.g.
    • IBM / ICL 32-bit reals and integers (storing 4 characters)
    • CDC 6000 & 7000 series 60-bit reals and integers (10 characters, so e.g. no lower case)
    • keyboards, character sets ...

• in 1990's
  • IEEE standard representations (32-bit) and arithmetic
  • fewer viable systems

• in 2000's
  • same concepts facilitate 64-bit implementations

..
Fundamental issues

- reusable components
- in 1970's and 80's
  - library of accessing functions and programming utilities
- in 1990's and 2000's
  - library of software components
    - dll's
    - output viewers
    - graphics viewer
    - statistics engine
    - dataset converter
    - database engine
    - spreadsheet
    - etc ... ?
- advantages
  - mixed language
  - convenient cooperation
  - ability to construct new systems reliably and efficiently
  ..
Command language

  - Genstat 4.03e (first PC DOS version)
  - example – probit lines from Finney (1971) p. 103

```
> 'REFERENCE' probit
1 'REFERENCE' probit
> 'UNIT' $ 14
2 'UNIT' $ 14
> 'define the data'
3 'define the data'
> 'NAME' Dname = Morphine, Amidone, Phenadoxone, Pethidine
4 'NAME' Dname = Morphine, Amidone, Phenadoxone, Pethidine
> 'FACTOR' Drug $ Dname = 3(1...4), 4, 4
5 'FACTOR' Drug $ Dname = 3(1...4), 4, 4
> 'VARIATE' Dose = (1.5, 3, 6)2, 0.75, 1.5, 3, 5, 7.5, 10, 15, 20
6 'VARIATE' Dose = (1.5, 3, 6)2, 0.75, 1.5, 3, 5, 7.5, 10, 15, 20
> : Ntest = 103, 120, 123, 60, 110, 100, 90, 80, 90, 60, 85, 60, 90, 60
7 : Ntest = 103, 120, 123, 60, 110, 100, 90, 80, 90, 60, 85, 60, 90, 60
> : Nrespond = 19, 53.83, 14, 54.81, 31, 54, 80, 13, 27, 32, 55, 44
8 : Nrespond = 19, 53.83, 14, 54.81, 31, 54, 80, 13, 27, 32, 55, 44
> 'RUN'
```
Genstat 4.03e

- print the data
• transform dose, define glm and fit constant

```plaintext
Phenadox  3.0000E 0 9.0000E 1 8.0000E 1
Pethidin   5.0000E 0 6.0000E 1 1.3000E 1
Pethidin 7.5000E 0 8.5000E 1 2.7000E 1
Pethidin  1.0000E 1 6.0000E 1 3.2000E 1
Pethidin 1.5000E 1 9.0000E 1 5.5000E 1
Pethidin  2.0000E 1 6.0000E 1 4.4000E 1

'> transform dose to log base 10 ''
12 '' transform dose to log base 10 ''
> 'CALCULATE' Logdose = LOG10(Dose)
13 'CALCULATE' Logdose = LOG10(Dose)
>'' define the maximal model and binomial totals ''
14 '' define the maximal model and binomial totals ''
> 'TERMS/TOTAL=Ntest' Nrespond + Drug*Logdose
15 'TERMS/TOTAL=Ntest' Nrespond + Drug*Logdose
>'' define the glm and y-variate ''
16 '' define the glm and y-variate ''
> 'Y/ERROR=BINOMIAL,LINK=PROBIT' Nrespond
17 'Y/ERROR=BINOMIAL,LINK=PROBIT' Nrespond
>'' fit the constant ''
18 '' fit the constant ''
> 'FIT/PRINT=Z,ANODEV=I'
19 'FIT/PRINT=Z,ANODEV=I'
> 'RUN'
```
Genstat 4.03e

- add a common line

```plaintext
45 ' add a common regression line '
'ADD/PRINT=C' Logdose
46 'ADD/PRINT=C' Logdose
'RUN'
```
Genstat 4.03e

- common line

```
Command Prompt - g

46

***** REGRESSION ANALYSIS *****
ERROR DISTRIBUTION: BINOMIAL LINK FUNCTION: PROBIT

*** REGRESSION COEFFICIENTS ***
Y-VARIATE: Nrespond

<table>
<thead>
<tr>
<th></th>
<th>ESTIMATE</th>
<th>S.E.</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>-0.29757</td>
<td>0.06629</td>
<td>-4.49</td>
</tr>
<tr>
<td>Logdose</td>
<td>0.59721</td>
<td>0.09589</td>
<td>6.23</td>
</tr>
</tbody>
</table>

*** STANDARD ERRORS BASED ON SCALE PARAMETER WITH VALUE 1.000

>'' fit parallel lines ''
48 '' fit parallel lines ''
>ADD/PRINT=C' Drug
49 'ADD/PRINT=C' Drug
>''RUN''
```
**Genstat 4.03e**

- parallel lines

---

```
***** REGRESSION ANALYSIS *****
ERROR DISTRIBUTION: BINOMIAL LINK FUNCTION: PROBIT

*** REGRESSION COEFFICIENTS ***
Y-VARIATE: Nrespond

<table>
<thead>
<tr>
<th></th>
<th>ESTIMATE</th>
<th>S.E.</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>-1.37926</td>
<td>0.11421</td>
<td>-12.08</td>
</tr>
<tr>
<td>Logdose</td>
<td>2.46849</td>
<td>0.17260</td>
<td>14.30</td>
</tr>
<tr>
<td>Drug Amidone</td>
<td>0.23787</td>
<td>0.10835</td>
<td>2.20</td>
</tr>
<tr>
<td>Drug Phenadox</td>
<td>1.35955</td>
<td>0.12964</td>
<td>10.49</td>
</tr>
<tr>
<td>Drug Pethidin</td>
<td>-1.17991</td>
<td>0.13303</td>
<td>-8.97</td>
</tr>
</tbody>
</table>

*** STANDARD ERRORS BASED ON SCALE PARAMETER WITH VALUE 1.000

> ' try different slopes '
> 27 ' try different slopes '
> ADD/PRINT=C' Logdose.Drug
> 28 ' ADD/PRINT=C' Logdose.Drug
> 'RUN'
```
Genstat 4.03e

- different lines for each drug
- andev table

```
*** SUMMARY ANALYSIS OF DEVIANCE ***

Y-VARIATE: Nrespond

<table>
<thead>
<tr>
<th>TERMS</th>
<th>RESIDUAL DEVIANCE</th>
<th>CHANGE DEVIANCE</th>
<th>MEAN CHANGE</th>
<th>MN RA</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEV. TION</td>
<td>DF</td>
<td>DF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INITIAL MODEL</td>
<td>CONSTANT</td>
<td>13</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
| MODIFICATIONS TO MODEL
| +Logdose       | 12                | 1               | 39.4078     | 39.4078 |
| +Drug          | 9                 | 3               | 206.6821    | 68.8940 |
| +Logdose.Drug  | 6                 | 3               | 1.5336      | 0.5112 |
|                | 6                 | 0               | 0.0000      | *     |
| **DENOMINATOR OF RATIO IS RES.DEV./RES.DF FROM LINE ABOVE,=** | | | | 0.3891 |

> 'STOP'
```
Genstat 5 Release 1.0 <VAX/VMS>
Copyright 1987, Lawes Agricultural Trust <Rothamsted Experimental Station>

```plaintext
> FACTOR [LABELS="Morphine,Amidone,Phenadoxone,Pethidine"] Drug
> READ [PRINT=data] Drug,Dose,Ntest,Nrespond
> 1 1.50 103 19 1 3.00 120 53 1 6.00 123 83
> 2 1.50 60 14 2 3.00 110 54 2 6.00 100 81
> 3 0.75 90 31 3 1.50 80 54 3 3.00 90 80
> 4 5.00 60 13 4 7.50 85 27 4 10.00 60 32
> 4 15.00 90 55 4 20.00 60 44 :
> CALCULATE Logdose = LOG10(Dose)
> MODEL [DISTRIBUTION=binomial; LINK=probit] Nrespond; NBINOMIAL=Ntest
> TERMS Logdose*Drug
> "Fit a model ignoring the types of drug used."
> FIT [NOMESSAGE=leverage,residual] Logdose
```

***** Regression Analysis *****

Response variate: Nrespond
Binomial totals: Ntest
Distribution: Binomial
Link function: Probit
Fitted terms: Constant + Logdose

*** Summary of analysis ***

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>deviance</th>
<th>mean deviance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>39.4</td>
<td>39.41</td>
</tr>
<tr>
<td>Residual</td>
<td>12</td>
<td>210.6</td>
<td>17.55</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>250.0</td>
<td>19.23</td>
</tr>
<tr>
<td>Change</td>
<td>-1</td>
<td>-39.4</td>
<td>39.41</td>
</tr>
</tbody>
</table>

* MESSAGE: ratios are based on dispersion parameter with value 1

*** Estimates of regression coefficients ***

<table>
<thead>
<tr>
<th></th>
<th>estimate</th>
<th>s.e.</th>
<th>t&lt;*&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.2976</td>
<td>0.0663</td>
<td>-4.49</td>
</tr>
<tr>
<td>Logdose</td>
<td>0.5972</td>
<td>0.0959</td>
<td>6.23</td>
</tr>
</tbody>
</table>

* MESSAGE: s.e.s are based on dispersion parameter with value 1
Genstat 5 Release 1

```plaintext
> ADD Drug

***** Regression Analysis *****

Response variate: Nrespond
Binomial totals: Ntest
  Distribution: Binomial
  Link function: Probit
  Fitted terms: Constant + Logdose + Drug

*** Summary of analysis ***

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>deviance</th>
<th>deviance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>4</td>
<td>246.090</td>
<td>61.52</td>
</tr>
<tr>
<td>Residual</td>
<td>9</td>
<td>3.868</td>
<td>0.427</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>249.958</td>
<td>19.225</td>
</tr>
<tr>
<td>Change</td>
<td>-3</td>
<td>-206.682</td>
<td>68.890</td>
</tr>
</tbody>
</table>

* MESSAGE: ratios are based on dispersion parameter with value 1

*** Estimates of regression coefficients ***

<table>
<thead>
<tr>
<th></th>
<th>estimate</th>
<th>s.e.</th>
<th>t(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.379</td>
<td>0.114</td>
<td>-12.08</td>
</tr>
<tr>
<td>Logdose</td>
<td>2.468</td>
<td>0.173</td>
<td>14.30</td>
</tr>
<tr>
<td>Drug Amidone</td>
<td>0.238</td>
<td>0.108</td>
<td>2.20</td>
</tr>
<tr>
<td>Drug Phenadoxone</td>
<td>1.360</td>
<td>0.130</td>
<td>10.49</td>
</tr>
<tr>
<td>Drug Pethidine</td>
<td>-1.180</td>
<td>0.133</td>
<td>-8.87</td>
</tr>
</tbody>
</table>

* MESSAGE: s.e.s are based on dispersion parameter with value 1
```
Genstat 5 Release 1

> ADD Le.al Logdose.Drug

***** Regression Analysis *****

*** Estimates of parameters ***

<table>
<thead>
<tr>
<th></th>
<th>estimate</th>
<th>s.e.</th>
<th>t(*)</th>
<th>t pr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.255</td>
<td>0.171</td>
<td>-7.34</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>LogDose</td>
<td>2.226</td>
<td>0.304</td>
<td>7.32</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Drug Amidone</td>
<td>-0.006</td>
<td>0.272</td>
<td>-0.02</td>
<td>0.983</td>
</tr>
<tr>
<td>Drug Phenadoxone</td>
<td>1.205</td>
<td>0.197</td>
<td>6.11</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Drug Pethidine</td>
<td>-1.194</td>
<td>0.402</td>
<td>-2.97</td>
<td>0.003</td>
</tr>
<tr>
<td>LogDose.Drug Amidone</td>
<td>0.475</td>
<td>0.485</td>
<td>0.98</td>
<td>0.328</td>
</tr>
<tr>
<td>LogDose.Drug Phenadoxone</td>
<td>0.480</td>
<td>0.475</td>
<td>1.01</td>
<td>0.313</td>
</tr>
<tr>
<td>LogDose.Drug Pethidine</td>
<td>0.134</td>
<td>0.464</td>
<td>0.29</td>
<td>0.772</td>
</tr>
</tbody>
</table>

* MESSAGE: s.e.s are based on dispersion parameter with value 1

*** Accumulated analysis of deviance ***

<table>
<thead>
<tr>
<th>Change</th>
<th>d.f.</th>
<th>deviance</th>
<th>mean deviance</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Logdose</td>
<td>1</td>
<td>39.4079</td>
<td>39.4079</td>
<td>39.41</td>
</tr>
<tr>
<td>+ Drug</td>
<td>3</td>
<td>206.6821</td>
<td>68.8940</td>
<td>68.89</td>
</tr>
<tr>
<td>+ Logdose.Drug</td>
<td>3</td>
<td>1.5336</td>
<td>0.5112</td>
<td>0.51</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.3344</td>
<td>0.3891</td>
<td></td>
</tr>
</tbody>
</table>

Total 13       249.9579  19.2275

* MESSAGE: ratios are based on dispersion parameter with value 1

- revised (simplified & consistent) syntax
- revised execution strategy and output – interactive use
- procedure structure – like a standard command (directive)
- procedure libraries – extendability
- structured programming
- enhanced data structures
- designed to last ..
Regression analysis

Estimates of parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>estimate</th>
<th>s.e.</th>
<th>t(*)</th>
<th>t pr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.255</td>
<td>0.171</td>
<td>-7.34</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>LogDose</td>
<td>2.226</td>
<td>0.304</td>
<td>7.32</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Drug Amidone</td>
<td>-0.006</td>
<td>0.272</td>
<td>0.02</td>
<td>0.983</td>
</tr>
<tr>
<td>Drug Phenadoxone</td>
<td>1.205</td>
<td>0.197</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug Pethidine</td>
<td>-1.194</td>
<td>0.402</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogDose:Drug Amidone</td>
<td>0.475</td>
<td>0.485</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogDose:Drug Phenadoxone</td>
<td>0.480</td>
<td>0.475</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogDose:Drug Pethidine</td>
<td>0.134</td>
<td>0.464</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Message: s.e.s are based on dispersion parameter with value 1.

Parameters for factors are differences compared with the reference level:
Factor    Reference level
Drug      Morphine

Accumulated analysis of deviance

<table>
<thead>
<tr>
<th>Change</th>
<th>d.f.</th>
<th>deviance</th>
<th>deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ LogDose</td>
<td>1</td>
<td>39.4079</td>
<td>39.4079</td>
</tr>
<tr>
<td>+ Drug</td>
<td>3</td>
<td>206.6821</td>
<td>68.8210</td>
</tr>
<tr>
<td>+ LogDose:Drug</td>
<td>3</td>
<td>1.5336</td>
<td>1.5336</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.3344</td>
<td>0.1289</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>249.9579</td>
<td>19.8579</td>
</tr>
</tbody>
</table>

- same (scripting) language, but menus too
Input and output

- standard output-stream utilities

- utilities for standard output components

- centralised i/o statements

- advantages
  - 1970's
    - convenient implementation of algorithms
    - consistent formatting and precision
  - 1980's
    - interactive working
    - transcript files of input and/or output
  - 1990's
    - Windows
  - 2000's
    - formatted output (RTF, HTML, Latex)
    - maths and Greek symbols
Composite link functions

Composite link example (Thompson & Baker, 1981, Appl. Statist.)

Frequency table for O,A,B blood groups

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>O</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>p</td>
<td>p</td>
<td>q</td>
</tr>
<tr>
<td>O</td>
<td>r</td>
<td>p</td>
<td>r</td>
</tr>
<tr>
<td>B</td>
<td>q</td>
<td>p</td>
<td>q</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>O</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>p</td>
<td>A</td>
<td>AB</td>
</tr>
<tr>
<td>O</td>
<td>r</td>
<td>A</td>
<td>O</td>
</tr>
<tr>
<td>B</td>
<td>q</td>
<td>AB</td>
<td>B</td>
</tr>
</tbody>
</table>
Composite link functions

\[
E(y) = \mu = C \gamma ; \quad \gamma = h(\eta) ; \quad \eta = X \beta
\]

\( C \) is a matrix, so the effect is that each element of \( \mu \) can be the sum of several elements of the linear predictor.

<table>
<thead>
<tr>
<th>( X )</th>
<th>( p )</th>
<th>( r )</th>
<th>( q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p^2 )</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( pr )</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( pq )</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( rp )</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( r^2 )</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>( rq )</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( qp )</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( qr )</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( q^2 )</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\beta & \log (p \times \sqrt{n}) \\
& \log (r \times \sqrt{n}) \\
& \log (q \times \sqrt{n}) \\
\end{align*}
\]

\[
C = \begin{array}{cccccccccccc}
p^2 & pr & pq & rp & r^2 & rq & qp & qr & q^2 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\
\end{array}
\]
Composite link functions

Nonlinear regression analysis

Estimates of parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>estimate</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_p_root_n</td>
<td>1.6424</td>
<td>0.0686</td>
</tr>
<tr>
<td>log_r_root_n</td>
<td>2.6636</td>
<td>0.0346</td>
</tr>
<tr>
<td>log_q_root_n</td>
<td>0.027</td>
<td>0.154</td>
</tr>
</tbody>
</table>
Composite link functions

Estimates of functions of parameters

Estimates and standard errors

<table>
<thead>
<tr>
<th>Parameter</th>
<th>estimate</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>422.0</td>
<td>20.5</td>
</tr>
<tr>
<td>p</td>
<td>0.2516</td>
<td>0.0161</td>
</tr>
<tr>
<td>q</td>
<td>0.6984</td>
<td>0.0171</td>
</tr>
<tr>
<td>r</td>
<td>0.05001</td>
<td>0.00761</td>
</tr>
</tbody>
</table>

Message: s.e.s are based on dispersion parameter with value 1
Data structures


Table 1. Structures provided

<table>
<thead>
<tr>
<th></th>
<th>Scalar</th>
<th>Variate</th>
<th>Multi-Response</th>
<th>Variate</th>
<th>Data Matrix</th>
<th>Hierarchical Data matrix</th>
<th>Rectangular matrix</th>
<th>Symmetric matrix</th>
<th>Diagonal matrix</th>
<th>SSP matrix</th>
<th>Standard Table</th>
<th>Table with margins</th>
<th>Compound Table</th>
<th>Special structures</th>
<th>Macros</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMDP</td>
<td>R</td>
<td>RC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cents Aid II</td>
<td>R</td>
<td>R*</td>
<td>(+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genstat</td>
<td>R</td>
<td>RCI</td>
<td>(+)</td>
<td></td>
<td>R</td>
<td>R</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cmstart II</td>
<td>R</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osiris III</td>
<td>RC</td>
<td>RC</td>
<td>(+)</td>
<td></td>
<td>R</td>
<td>R</td>
<td>(+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fstat</td>
<td>R</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGSP</td>
<td>R</td>
<td>R</td>
<td>(+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAS75</td>
<td>R</td>
<td>RC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPSS</td>
<td>RC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPL</td>
<td>RDIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Key to Table
* – documentation provided not sufficiently detailed for information to be discovered.
+ – The structure occurs in the package.
(+) – The structure occurs in the package but…. (see the appropriate section in the text).
Data structures


<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Name &amp; Suffix</th>
<th>Direct</th>
<th>Assignment</th>
<th>Equating</th>
<th>Algorithms</th>
<th>Display from</th>
<th>Store/Retr.</th>
<th>Used in</th>
<th>Calculations</th>
<th>Output from</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMDP</td>
<td>V</td>
<td>VP</td>
<td>D(T)</td>
<td>VDT</td>
<td>V</td>
<td>DP(T)</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>P</td>
</tr>
<tr>
<td>Cents Aid II</td>
<td>V</td>
<td>V</td>
<td>D</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genstat</td>
<td>SVM</td>
<td>SVMT</td>
<td>VMT</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ommitab II</td>
<td>V</td>
<td>(S)</td>
<td>S</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osiris III</td>
<td>V</td>
<td>IMP</td>
<td>D</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pstat</td>
<td>V</td>
<td>VIM</td>
<td>IM</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGSP</td>
<td>V</td>
<td>VIMP</td>
<td>V</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAS75</td>
<td>V</td>
<td>VD</td>
<td>D</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPSS</td>
<td>V</td>
<td>VD</td>
<td>D</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPL</td>
<td>V</td>
<td>VT</td>
<td>D</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Key to Table
- S = Scalar
- V = Variate
- D = Data matrix
- M = Rectangular matrix
- P = SSP matrix
- T = Table
Data structures

• statistical data
  • often best represented in compound data structures
    • Guide to GenStat, Part 1, Chapter 4

• examples
  • factor (grouping variable)
    • integer vector values
    • integer scalar #levels
    • {real vector levels}
    • {text vector labels}
  • ssp
    • real symmetric matrix sums of squares & products
    • real vector means
    • real scalar (weighted) replication
  • lrv
    • real rectangular matrix latent vectors
    • real diagonal matrix latent roots
    • scalar trace
User-defined compound structures

- multi-response factor
  - integer scalar \#levels ("responses")
  - integer vectors indicator variable (for each "response")
  - {real vector} levels
  - {text vector} labels

- design definition structures
  - Payne & Franklin (1994). Data structures and algorithms for an open system to design and analyse generally balanced designs. COMPSTAT 94, 429-434.
  - e.g. using a design key
    - integer rectangular matrix design key
    - integer scalars #levels of block pseudo-factors
    - integer scalars #levels of treat. pseudo-factors
    - integer vector mapping: block factors from pseudo-factors
    - integer vector mapping: treatment factors from pseudo-factors
    - pointer blocking factors
    - pointer treatment factors

...
• dhglm save / specification structure
  • a DHGLM is an HGLM where the model for one of the dispersion terms is itself an HGLM (rather than an ordinary generalized linear model)
  • recursive definition structure
  • elements RESDISPMODEL and RANDDISPMODEL[] can be HGLM structures
  • allows fitting by recursive calls of the HGFIT procedure
User-defined compound structures

- multitiered analysis structure
- tree representing the design structure
- pointer at the node for each analysis term
multitiered analysis structure

- tree structure to represent the design
- example Brien (1983) Biometrics
  - Tasters / Sittings
  - Blocks / Plots
  - Treatments

![Output](image)

```
Analysis of variance
Source d.f. e.f.
tasters 4 1.000
tasters.sittings
 . blocks 2 1.000
 . blocks.plots
  . treatments 3 1.000
 . Residual 6 1.000
 . Residual 44 1.000
Total 59 1.000
```
Data structures

- implementation
  - 1970 onwards (Fortran IV)
    - list-processing within arrays in blank common
    - standard routines to access attributes and values-blocks
    - standard routines to create new data structures
    - workspace allocated from top and bottom of array
    - much expertise required for use, and much scope for mistakes
  - 1983 onwards (Fortran 77)
    - character handling
    - workspace allocated in "cells"
    - ability to resize workspace dynamically
    - user-defined compound structures

..
Data structures

• future implementation
  • directly in the implementation language
  • now possible in C++ and Fortran 90 onwards
    • standard accessing utilities isolate the algorithmic code (that uses the data structures) from the GenStat kernel (that implements them)
    • less expertise required to use, and some protection against mistakes (overwriting &c)

• actions
  • revise the implementation
  • update the existing utilities
  • provide new utilities / direct access for future algorithms (e.g. to enable vector processing)
  • update existing algorithms as required / feasible
Algorithmic interface

- **in 1970's**
  - standard utilities to access options and parameters of a command

- **in 1980's**
  - syntax of each directive defined in a "bootstrap" file of commands (similar to definition of a procedure)

- **in future**
  - compound data structure to represent a command
  - utilities revised to keep existing code operational
  - new utilities / direct access in future algorithms
  - less "folk-knowledge" required by implementors
  - better adapted to collaborative development, the flexible job market &c &c

...
GenStat aims

• to promote good statistical practice
  • all standard analyses accessible by menus
  • advanced menus build on the concepts in the simple ones
  • general algorithms – provide the right analysis not the least inappropriate
  • context-sensitive help
  • all documentation on-line (Introduction, Ref Manual, Guides, Tutorials)
  • full audit trail (recorded in the Input Log)
  • output directly exportable into reports, scientific papers or web pages

• to encourage innovation
  • can write programs in the command language to develop new ideas
  • can incorporate own code
  • programs can be formed into procedures (i.e. new commands) and distributed in libraries for automatic access by anyone

• to promote new ideas
  • annual releases
  • standard procedure library including refereed user contributions