Answers to exercises

Chapter 1

Dioecious trees

The question is whether the number of flowers produced depends upon the sex of the tree. The variable SEX is categorical with two levels. The commands to execute the ANOVA are as before. We have assumed that the dataset *dioecioustrees* has been downloaded into the gandh library.

SAS COMMANDS FOR BOX 15.1 **Analysis for dioecious trees**

Commands

```
proc glm data=gandh.dioecioustrees;
  class SEX;
  model FLOWERS = SEX;
  run;
```

Menu route

Statistics > Anova > Factorial Anova…

FLOWERS → Dependent

SEX → Independent

This would produce the following output:

SAS OUTPUT FOR BOX 15.1 **Analysis for dioecious trees**

```
                              The GLM Procedure

Dependent Variable: FLOWERS

                                        Sum of
      Source                  DF       Squares     Mean Square    F Value    Pr > F

      Model                    1    171841.333      171841.333       1.18    0.2837

      Error                   48   7017255.167      146192.816

      Corrected Total         49   7189096.500

                      R-Square     Coeff Var      Root MSE    FLOWERS Mean

                      0.023903      99.70057      382.3517        383.5000

      Source                  DF     Type I SS     Mean Square    F Value    Pr > F

      SEX                      1   171841.3333     171841.3333       1.18    0.2837

      Source                  DF   Type III SS     Mean Square    F Value    Pr > F

      SEX                      1   171841.3333     171841.3333       1.18    0.2837
```

There are a number of ways in which these results could be illustrated graphically, one of which is a boxplot. Notice that in the command line route we need to sort the dataset by SEX in the first place, as SAS requires this for the grouping variable in the plot. The menu method handles this automatically.

SAS COMMANDS FOR FIGURE 15.1 **Boxplot for dioecious trees**

Commands

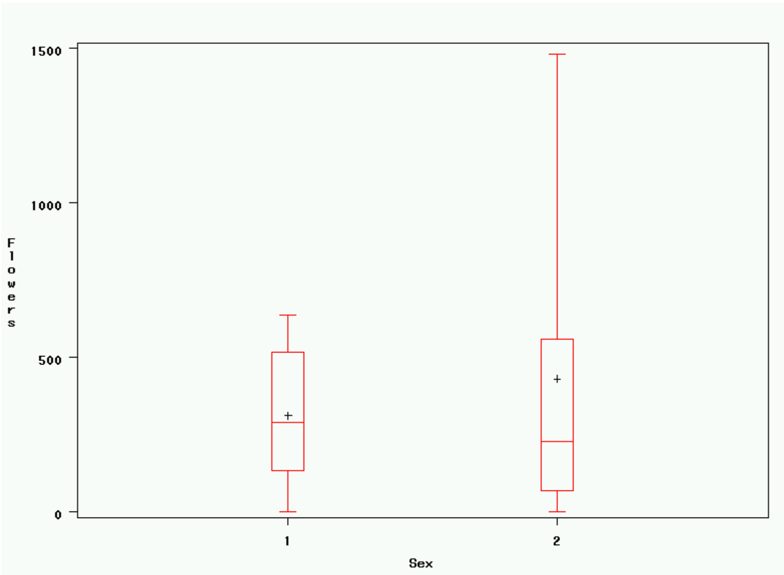
```
proc sort;
  by SEX;

proc boxplot;
  plot FLOWERS*SEX;
  run;
```

Menu route

Graphs > Box Plot
FLOWERS → Analysis
SEX → Class

This would produce the following graph:

SAS OUTPUT FOR FIGURE 15.1



Chapter 2

Dioecious trees

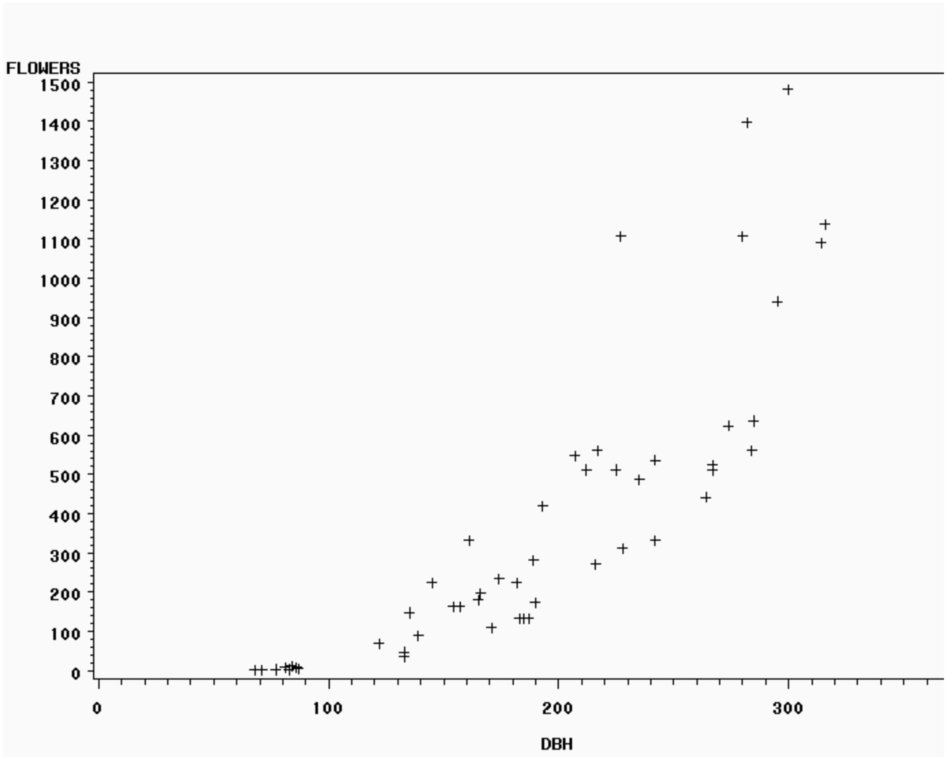
The variables `FLOWERS` and `DBH` are both continuous, so the best illustration of the relationship between them would be a scatter plot.

SAS COMMANDS FOR FIGURE 15.2 **Graph of FLOWERS versus DBH**

Commands	`proc gplot data=gandh.Chapter2;`
	`  plot FLOWERS * DBH;`
	`  run;`

Menu route	Graphs > Scatter Plot > Two-Dimensional... `FLOWERS` → Y Axis `DBH` → X Axis.

This would produce the following graph:



The regression analysis would be executed using the following commands:

SAS COMMANDS FOR BOX 15.2 Analysis for dioecious trees

Commands	`proc reg data=gandh.Chapter2;`
	`  model FLOWERS = DBH;`
	`  run;`

Menu route	Statistics > Regression > Linear… `FLOWERS` → Dependent `DBH` → Explanatory

which would give the following output:

SAS OUTPUT FOR BOX 15.2 **Analysis for dioecious trees**

```
                              The REG Procedure
                                Model: MODEL1
                         Dependent Variable: FLOWERS

                              Analysis of Variance

                                     Sum of           Mean
 Source                   DF        Squares         Square    F Value    Pr > F

 Model                     1        5060723        5060723     114.13    <.0001
 Error                    48        2128374          44341
 Corrected Total          49        7189097

              Root MSE            210.57330    R-Square     0.7039
              Dependent Mean      383.50000    Adj R-Sq     0.6978
              Coeff Var            54.90829

                              Parameter Estimates

                                   Parameter       Standard
Variable     Label         DF       Estimate          Error    t Value    Pr > |t|

Intercept    Intercept      1     -481.16042       86.24085      -5.58      <.0001
DBH          DBH            1        4.51284        0.42242      10.68      <.0001
```

Chapter 3

How variability in the population will influence our analysis

To create data sets for this exercise, just follow the procedure outlined in SAS Box 3.3a 'Creating your own datasets', but alter the standard deviation specified when you draw the random numbers from the Normal distribution (the ‘sigma=’ step in the second group of commands in SAS Box 3.3). We reproduce the commands in the following box, because here we have named the output file ‘results’. One advantage of using the command line method is that you can repeatedly execute the whole lot by pressing the 'Run' button, just noting down the grand mean produced in the output box between runs. Then you can edit the number you want to change and repeat the whole exercise.

```
data withdums;
   set gandh.Chapter3sim;
   dum1=(FERTIL=1);
   dum2=(FERTIL=2);
   run;
data withparameters;
   set withdums;
   k3=13.5;
   k1=1.2;
   k2=3.4;
   sigma=8;
   run;
data results;
   set withparameters;
   noise=normal(0);
   Y=k3+k1*dum1+k2*dum2+sigma*noise;
   run;
proc glm data=work.results;
  class FERTIL;
  model Y=FERTIL / solution;
  run;
```

The final step may also be carried out by menu route as shown below:

Menu route Statistics > Anova > Linear Models…

`Y` → Dependent

`FERTIL` → Class

Statistics

☑ Parameter estimates

Each analysis will produce one estimate of the grand mean, which may then be entered into a new column in a new worksheet. To create a new worksheet, make sure you are in the Analyst environment. Use File > New. The column headings may then be named `SIGMA2`, `SIGMA4`, `SIGMA8` etc. The values obtained from each analysis may then be entered into this new worksheet. In the example below, this worksheet has been saved in the work library and called 'SIGS'. (An alternative way of creating this new dataset is explained for the menu route in the box below). When ten estimates have been obtained, a histogram may be drawn for the column `SIGMA8` (for example) using the commands:

SAS COMMANDS FOR FIGURE 15.3

Commands

```
proc capability data=work.SIGS;
  var SIGMA8;
  histogram;
  run;
```

Menu route

Edit > Insert Columns > Numeric

Data > Column Properties

SIGMA8 → Name

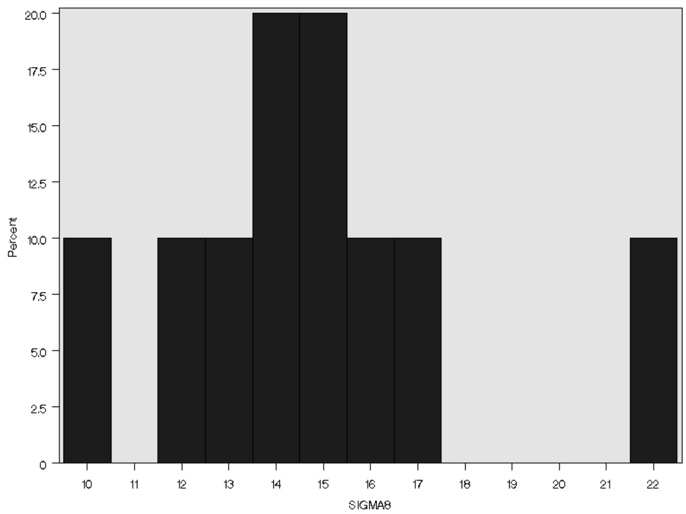
Now type the observed values of the grand mean into the column

Graphs > Histogram…

SIGMA8 → Analysis

This would produce a graph similar to the one below:

SAS OUTPUT FOR FIGURE 15.3



Comparison of the four graphs with the error standard deviation set at 2, 4, 8 and 16 will illustrate how the error variance will influence the reliability of the grand mean estimate (see main text for an example).

Chapter 4

Investigating Obesity

First you need to conduct two separate analyses to explain FOREARM using HT or WT.

SAS COMMANDS FOR BOX 15.4 (A) **First analysis of FOREARM**

Commands
```
proc glm data=gandh.Chapter4;
  model FOREARM = HT;
  run;
```

Menu route: Statistics > Anova > Linear Models…

FOREARM → Dependent

HT → Quantitative

SAS OUTPUT FOR BOX 15.4 (A) **First analysis of FOREARM**

```
                                    The GLM Procedure

Dependent Variable: FOREARM

                                            Sum of
      Source                      DF          Squares     Mean Square    F Value    Pr > F

      Model                        1        0.9439390       0.9439390       0.18    0.6778

      Error                       37      199.0940026       5.3809190

      Corrected Total             38      200.0379416

                      R-Square     Coeff Var      Root MSE    FOREARM Mean

                      0.004719      45.74012      2.319681        5.071436

      Source                      DF       Type I SS     Mean Square    F Value    Pr > F

      HT                           1      0.94393898      0.94393898       0.18    0.6778

      Source                      DF     Type III SS     Mean Square    F Value    Pr > F

      HT                           1      0.94393898      0.94393898       0.18    0.6778

                                                 Standard
               Parameter         Estimate           Error    t Value    Pr > |t|

               Intercept     0.9587990575      9.82623490       0.10      0.9228
               HT            0.0260479467      0.06219132       0.42      0.6778
```

SAS COMMANDS FOR BOX 15.4 (B) Second analysis of FOREARM

Commands
```
proc glm data=gandh.Chapter4;
  model FOREARM = WT;
  run;
```

Menu route: Statistics > Anova > Linear Models…

FOREARM → Dependent

WT → Quantitative

SAS OUTPUT FOR BOX 15.4 (B) Second analysis of FOREARM

```
                                    The GLM Procedure

Dependent Variable: FOREARM

                                           Sum of
      Source                      DF      Squares     Mean Square    F Value    Pr > F

      Model                        1   59.1372267      59.1372267      15.53    0.0003

      Error                       37  140.9007148       3.8081274

      Corrected Total             38  200.0379416

                      R-Square     Coeff Var      Root MSE    FOREARM Mean

                      0.295630      38.47909      1.951442        5.071436

      Source                      DF     Type I SS     Mean Square    F Value    Pr > F

      WT                           1   59.13722674     59.13722674      15.53    0.0003

      Source                      DF   Type III SS     Mean Square    F Value    Pr > F

      WT                           1   59.13722674     59.13722674      15.53    0.0003

                                                    Standard
               Parameter          Estimate             Error    t Value    Pr > |t|

               Intercept      -4.594757196        2.47272868      -1.86      0.0711
               WT              0.152978988        0.03882014       3.94      0.0003
```

Then the joint analysis using both explanatory variables

SAS COMMANDS FOR BOX 15.5 Analysis of FOREARM using both explanatory variables

Commands
```
proc glm data=gandh.Chapter4;
  model FOREARM = HT WT;
  run;
```

Menu route: Statistics > Anova > Linear Models…

FOREARM → Dependent

HT WT → Quantitative

Statistics

☑ Type 1

SAS OUTPUT FOR BOX 15.5 Analysis of FOREARM using both explanatory variables

```
                                  The GLM Procedure

Dependent Variable: FOREARM

                                        Sum of
      Source                  DF       Squares     Mean Square    F Value    Pr > F

      Model                    2    83.9143634      41.9571817      13.01    <.0001

      Error                   36   116.1235782       3.2256550

      Corrected Total         38   200.0379416


                      R-Square     Coeff Var      Root MSE    FOREARM Mean

                      0.419492      35.41425      1.796011        5.071436


      Source                  DF     Type I SS     Mean Square    F Value    Pr > F

      HT                       1    0.94393898      0.94393898       0.29    0.5919
      WT                       1   82.97042439     82.97042439      25.72    <.0001


      Source                  DF   Type III SS     Mean Square    F Value    Pr > F

      HT                       1   24.77713663     24.77713663       7.68    0.0088
      WT                       1   82.97042439     82.97042439      25.72    <.0001
```

Chapter 5

The dorsal crest of the male smooth newt

In the first analysis you are asked to investigate the relationship between LCREST and the body size of the newt. The response variable you are given is logged, therefore the body size variable should also be logged. The commands and menus are given below:

SAS COMMANDS FOR BOX 15.6 Dorsal crest analysis

Commands

```
proc glm data=gandh.Chapter5;
  model LCREST = LSVL / solution;
  run;
```

Menu route: Statistics > Anova > Linear Models…

LCREST → Dependent

LSVL → Quantitative

Statistics

☑ Parameter estimates

This would result in the following output:

SAS OUTPUT FOR BOX 15.6 Dorsal crest analysis

```
                                 The GLM Procedure

Dependent Variable: LCREST

                                             Sum of
      Source                      DF        Squares     Mean Square    F Value    Pr > F

      Model                        1     2.38939159      2.38939159      45.81    <.0001

      Error                       85     4.43372434      0.05216146

      Corrected Total             86     6.82311593

                      R-Square     Coeff Var      Root MSE    LCREST Mean

                      0.350191      29.30913      0.228389       0.779241

      Source                      DF      Type I SS     Mean Square    F Value    Pr > F

      LSVL                         1     2.38939159      2.38939159      45.81    <.0001

      Source                      DF    Type III SS     Mean Square    F Value    Pr > F

      LSVL                         1     2.38939159      2.38939159      45.81    <.0001

                                                Standard
               Parameter         Estimate          Error    t Value    Pr > |t|

               Intercept     -9.380587570     1.50132730      -6.25      <.0001
               LSVL           5.086989483     0.75160915       6.77      <.0001
```

In the second analysis, POND is also included.

SAS COMMANDS FOR BOX 15.7 Further dorsal crest analysis

Commands

```
proc glm data=gandh.Chapter5;
  class POND;
  model LCREST = POND LSVL / solution;
  run;
```

Menu route

Statistics > Anova > Linear Models…

LCREST → Dependent

POND → Class

LSVL → Quantitative

Statistics

☑ Parameter estimates

SAS OUTPUT FOR BOX 15.7 **Further dorsal crest analysis**

```
                                The GLM Procedure

Dependent Variable: LCREST

                                          Sum of
      Source                      DF         Squares     Mean Square    F Value    Pr > F

      Model                       10      2.63001908      0.26300191       4.77    <.0001

      Error                       76      4.19309685      0.05517233

      Corrected Total             86      6.82311593

                    R-Square     Coeff Var      Root MSE    LCREST Mean

                    0.385457      30.14315      0.234888       0.779241


      Source                      DF       Type I SS     Mean Square    F Value    Pr > F

      POND                         9      0.32519109      0.03613234       0.65    0.7466
      LSVL                         1      2.30482798      2.30482798      41.78    <.0001


      Source                      DF     Type III SS     Mean Square    F Value    Pr > F

      POND                         9      0.24062749      0.02673639       0.48    0.8807
      LSVL                         1      2.30482798      2.30482798      41.78    <.0001


                                                   Standard
             Parameter           Estimate             Error    t Value    Pr > |t|

             Intercept       -9.432039841 B      1.58678711      -5.94      <.0001
             POND      1      0.054134365 B      0.15467248       0.35      0.7273
             POND      2      0.007033449 B      0.11165526       0.06      0.9499
             POND      3     -0.174302587 B      0.12868697      -1.35      0.1796
             POND      4     -0.018521732 B      0.11235126      -0.16      0.8695
             POND      5     -0.126419116 B      0.10805039      -1.17      0.2457
             POND      6     -0.026660389 B      0.10516802      -0.25      0.8006
             POND      7     -0.033509202 B      0.09099004      -0.37      0.7137
             POND      8     -0.086771999 B      0.11144842      -0.78      0.4386
             POND      9     -0.033014208 B      0.12162730      -0.27      0.7868
             POND      10     0.000000000 B       .                .         .
             LSVL             5.134435845        0.79439065       6.46      <.0001


NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
      the normal equations.  Terms whose estimates are followed by the letter 'B' are not
      uniquely estimable.
```

Chapter 6

Determinants of the Grade Point Average

This exercise uses the *grades* dataset.

SAS COMMANDS FOR BOX 15.8 Analysis of the Grades dataset

Commands

```
proc glm data=gandh.Chapter6;
  class YEAR;
  model GPA = YEAR VERBAL MATH / solution;
  run;
```

Menu route Statistics > Anova > Linear Models…

GPA → Dependent

YEAR → Class

VERBAL MATH → Quantitative

Statistics

☑ Parameter estimates

SAS OUTPUT FOR BOX 15.8 Analysis of the Grades dataset

```
                                The GLM Procedure

Dependent Variable: GPA

                                            Sum of
     Source                      DF         Squares     Mean Square    F Value    Pr > F

     Model                        3      8.62390974      2.87463658       9.65    <.0001

     Error                      196     58.39609026      0.29793924

     Corrected Total            199     67.02000000

                    R-Square     Coeff Var      Root MSE      GPA Mean

                    0.128677      20.75430      0.545838      2.630000

     Source                      DF       Type I SS     Mean Square    F Value    Pr > F

     YEAR                         1      1.15520000      1.15520000       3.88    0.0504
     VERBAL                       1      6.75953191      6.75953191      22.69    <.0001
     MATH                         1      0.70917782      0.70917782       2.38    0.1245

     Source                      DF     Type III SS     Mean Square    F Value    Pr > F

     YEAR                         1      0.84599265      0.84599265       2.84    0.0936
     VERBAL                       1      5.15998884      5.15998884      17.32    <.0001
     MATH                         1      0.70917782      0.70917782       2.38    0.1245

                                                  Standard
            Parameter           Estimate             Error    t Value    Pr > |t|

            Intercept       0.5930373541 B      0.43943013       1.35      0.1787
            YEAR      1     0.1304169627 B      0.07739531       1.69      0.0936
            YEAR      2     0.0000000000 B       .                .         .
            VERBAL          0.0022879692        0.00054978       4.16      <.0001
            MATH            0.0009374861        0.00060765       1.54      0.1245

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
      the normal equations.  Terms whose estimates are followed by the letter 'B' are not
      uniquely estimable.
```

Chapter 7

Weight, fat and sex

This question requires a factorial analysis, with an interaction between a categorical and continuous explanatory variable.

SAS COMMANDS FOR BOX 15.9 **Analysis for weight, fat and sex**

Commands

```
proc glm data=gandh.Chapter7;
  class SEX;
  model FAT = SEX|WEIGHT / solution;
  run;
```

Menu route

Statistics > Anova > Linear Models…

FAT → Dependent

SEX → Class

WEIGHT → Quantitative

Model

Independent → Effects in Model

SEX WEIGHT → Add

(SEX & WEIGHT) → Cross

Statistics

☑ Parameter Estimates

Tests

☑ Type 1

SAS OUTPUT FOR BOX 15.9 **Analysis for weight, fat and sex**

```
                                The GLM Procedure

Dependent Variable: FAT

                                          Sum of
      Source                      DF     Squares     Mean Square    F Value    Pr > F

      Model                        3     188.2834083      62.7611361      31.24    <.0001

      Error                       15      30.1376444       2.0091763

      Corrected Total             18     218.4210526

                      R-Square     Coeff Var      Root MSE      FAT Mean

                      0.862020      4.996592      1.417454      28.36842

      Source                      DF       Type I SS     Mean Square    F Value    Pr > F

      SEX                          1     90.32105263     90.32105263      44.95    <.0001
      WEIGHT                       1     87.10492012     87.10492012      43.35    <.0001
      WEIGHT*SEX                   1     10.85743552     10.85743552       5.40    0.0345

      Source                      DF     Type III SS     Mean Square    F Value    Pr > F

      SEX                          1      2.10773879      2.10773879       1.05    0.3220
      WEIGHT                       1     79.54187398     79.54187398      39.59    <.0001
      WEIGHT*SEX                   1     10.85743552     10.85743552       5.40    0.0345

                                                Standard
             Parameter           Estimate          Error    t Value    Pr > |t|

             Intercept        11.57095793 B     2.86807619       4.03      0.0011
             SEX        1     -6.33128851 B     6.18148377      -1.02      0.3220
             SEX        2      0.00000000 B      .               .          .
             WEIGHT            0.18550431 B     0.03567800       5.20      0.0001
             WEIGHT*SEX 1      0.21738825 B     0.09351504       2.32      0.0345
             WEIGHT*SEX 2      0.00000000 B      .               .          .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
      the normal equations.  Terms whose estimates are followed by the letter 'B' are not
      uniquely estimable.
```

Chapter 8

Combining data from different experiments

We assume that you use Excel or some other software to stack the columns on top of each other into a new dataset. (A rather laborious way in SAS is to go into edit mode and edit the data table directly. Beware that there is another variable called YEAR in the chapter 8 dataset. If you find an easy way in SAS, do let us know!)

SAS COMMANDS FOR BOX 15.11 **Bird data combined**

Commands

```
proc glm data=gandh.Chapter8;
  class YEAR;
  model YOUNG = YEAR SONGDAY /solution;
  run;
```

Menu route Statistics > Anova > Linear Models…

YOUNG → Dependent

YEAR → Class

SONGDAY → Quantitative

Statistics

☑ Type I

☑ Parameter estimates

SAS OUTPUT FOR BOX 15.11 **Bird data combined**

```
                                 The GLM Procedure

Dependent Variable: YOUNG

                                            Sum of
     Source                      DF        Squares     Mean Square    F Value    Pr > F

     Model                        5    22.90003809      4.58000762       3.76    0.0073

     Error                       38    46.28178009      1.21794158

     Corrected Total             43    69.18181818

                    R-Square     Coeff Var      Root MSE    YOUNG Mean

                    0.331012      38.53855      1.103604      2.863636

     Source                      DF      Type I SS     Mean Square    F Value    Pr > F

     YR                           4    10.69769120      2.67442280       2.20    0.0879
     SONGDAY                      1    12.20234690     12.20234690      10.02    0.0030

     Source                      DF    Type III SS     Mean Square    F Value    Pr > F

     YR                           4    19.78164888      4.94541222       4.06    0.0077
     SONGDAY                      1    12.20234690     12.20234690      10.02    0.0030

                                                   Standard
           Parameter           Estimate               Error    t Value    Pr > |t|

           Intercept        3.952056192 B      0.68156014       5.80      <.0001
           YR        1      1.717708704 B      0.77197644       2.23      0.0321
           YR        2      2.949080600 B      0.78840872       3.74      0.0006
           YR        3      1.048516981 B      0.61473929       1.71      0.0962
           YR        4      1.985456395 B      0.80929746       2.45      0.0189
           YR        5      0.000000000 B       .                .         .
           SONGDAY         -0.109127384        0.03447667      -3.17      0.0030

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
      the normal equations.  Terms whose estimates are followed by the letter 'B' are not
      uniquely estimable.
```

Chapter 9

Stabilising the variance in a blocked experiment

The commands required for the first part of this exercise a) transform the data; b) perform the analysis on the transformed data, storing the residuals and plotting the residuals against the fitted values; c) display descriptive statistics for the residuals by treatment. We do all three transformations and analyses at once, illustrating SAS's flexibility in this respect.

However, there is one way in which SAS is not flexible. The chapter nine worksheet contains datasets of many different lengths. If you use the command line route for this exercise, SAS will not provide you with summary statistics for the residuals of treatment five—the mixed lengths of variables causing SAS some confusion. Therefore, for this exercise, we have imported the *cotyledons* dataset separately.

SAS COMMANDS FOR BOXES 15.12 TO 15.17, AND FIGURE 15.6 TO 15.8

Commands

```
data;
  set gandh.cotyledons;
  SQRTNCOT =sqrt(NCOT);
  LOGNCOT=log(NCOT);
  INVNCOT=1/ NCOT;

proc glm;
  class BLOCK TRMNT;
  model SQRTNCOT LOGNCOT INVNCOT = BLOCK TRMNT;
  output p= SQRTfv LOGfv INVfv STUDENT= SQRTres LOGres
      INVres;

proc gplot;
  plot SQRTres * SQRTfv;
  plot LOGres * LOGfv;
  plot INVres * INVfv;

proc means;
  by TRMNT;
  var SQRTres LOGres INVres;
  run;
```

(Contd.)

SAS COMMANDS FOR BOXES 15.12 TO 15.17, AND FIGURE 15.6 TO 15.8 (*Contd.*)

Menu route Edit > Mode > Edit…

Data > Transform > Compute…

Type "SQRTNCOT" into top left box

sqrt(NCOT) → Main pane

Data > Transform > Compute…

Type "LOGNCOT" into top left box

log(NCOT) → Main pane

Data > Transform > Compute…

Type "INVNCOT" into top left box

1/NCOT → Main pane

Statistics > Anova > Linear Models…

SQRTNCOT LOGNCOT INVNCOT → Dependent

BLOCK TRMNT → Class

Plots with Residual tab

☑ Plot residuals vs variables

☑ Residuals: Studentized

☑ Variables: Predicted Y

Save Data

☑ Create and save diagnostic data

STUDENT → Add

Now click on the most recent 'Diagnostic Table' icon in left hand pane of Analyst, and choose File > Save as by SAS Name…, saving it under some name. Then open that file within Analyst using File > Open by SAS Name.

Statistics > Descriptive > Summary Statistics…

_STUDENT1 _STUDENT2 _STUDENT3 → Analysis

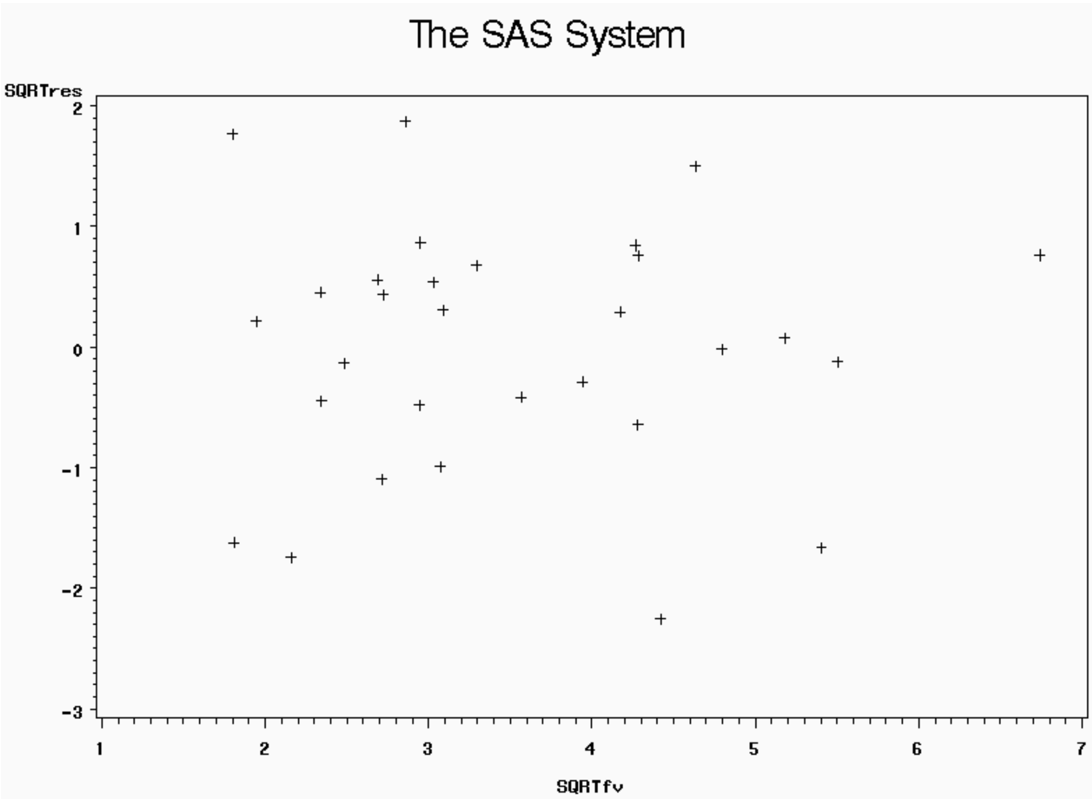
TRMNT → Class

Just the first ANOVA table and residual plots is illustrated below:

SAS OUTPUT FOR BOX 15.12 **Analysis of blocked experiment**

```
                                 The GLM Procedure

Dependent Variable: SQRTNCOT

                                         Sum of
     Source                      DF         Squares     Mean Square    F Value    Pr > F

     Model                        9     44.44384413      4.93820490      55.19    <.0001

     Error                       20      1.78947158      0.08947358

     Corrected Total             29     46.23331571

                  R-Square     Coeff Var      Root MSE    SQRTNCOT Mean

                  0.961295      8.508592      0.299121         3.515521

     Source                      DF       Type I SS     Mean Square    F Value    Pr > F

     BLOCK                        5     23.90980329      4.78196066      53.45    <.0001
     TRMNT                        4     20.53404083      5.13351021      57.37    <.0001

     Source                      DF     Type III SS     Mean Square    F Value    Pr > F

     BLOCK                        5     23.90980329      4.78196066      53.45    <.0001
     TRMNT                        4     20.53404083      5.13351021      57.37    <.0001
```

SAS OUTPUT FOR FIGURE 15.6 **Standardised residual plot**



SAS OUTPUT FOR BOX 15.13, 15.15 AND 15.17 **Examining the residuals**

```
                                   The SAS System

-------------------------------------------- TRMNT=1 --------------------------------------------

                                 The MEANS Procedure

         Variable     N            Mean         Std Dev         Minimum         Maximum
         ------------------------------------------------------------------------------
         SQRTres      6    3.145632E-16       0.9364663      -1.6633021       0.8675659
         LOGres       6    1.082467E-15       0.7729623      -1.1574905       0.9904145
         INVres       6    3.666049E-16       0.6576565      -0.7034478       1.0938756
         ------------------------------------------------------------------------------


-------------------------------------------- TRMNT=2 --------------------------------------------

         Variable     N            Mean         Std Dev         Minimum         Maximum
         ------------------------------------------------------------------------------
         SQRTres      6    -3.33067E-16       1.0912566      -1.0959101       1.8692103
         LOGres       6    8.881784E-16       0.8702716      -1.1118747       1.3163716
         INVres       6    9.251859E-18       0.6243825      -0.7704611       0.9619229
         ------------------------------------------------------------------------------


-------------------------------------------- TRMNT=3 --------------------------------------------

         Variable     N            Mean         Std Dev         Minimum         Maximum
         ------------------------------------------------------------------------------
         SQRTres      6    -1.21662E-15       0.4247971      -0.4433083       0.5573088
         LOGres       6    7.031412E-16       0.4116188      -0.4166356       0.6015208
         INVres       6    -1.75785E-16       0.3131422      -0.3915120       0.5031805
         ------------------------------------------------------------------------------


-------------------------------------------- TRMNT=4 --------------------------------------------

         Variable     N            Mean         Std Dev         Minimum         Maximum
         ------------------------------------------------------------------------------
         SQRTres      6    -4.57967E-16       1.4024633      -1.7435106       1.7666929
         LOGres       6     -6.4763E-17       1.8411957      -2.5521149       1.9012582
         INVres       6    -7.17019E-17       2.0836062      -3.4601347       1.9890440
         ------------------------------------------------------------------------------


-------------------------------------------- TRMNT=5 --------------------------------------------

         Variable     N            Mean         Std Dev         Minimum         Maximum
         ------------------------------------------------------------------------------
         SQRTres      6     -5.9952E-15       1.3359768      -2.2548799       1.4984861
         LOGres       6     7.21645E-16       1.0419813      -1.0269795       1.3153124
         INVres       6    -1.06396E-15       0.8591633      -1.0726851       1.2975456
         ------------------------------------------------------------------------------
```

The commands also produce the output for the log and inverse transformations.

Checking the 'perfect' model

The commands to produce the histograms of the original and transformed data are reproduced below. We specify 'midpoints' so that SAS will use the same scale on the *x*-axes, making the histograms easier to compare by eye.

SAS COMMANDS TO PRODUCE FIGURE 15.9 A AND B

Commands

```
data;
  set gandh.chapter9;
proc capability noprint;
  var MALE FEMALE;
  histogram / midpoints = 0.2 to 1.1 by 0.1;
  run;
```

Menu route

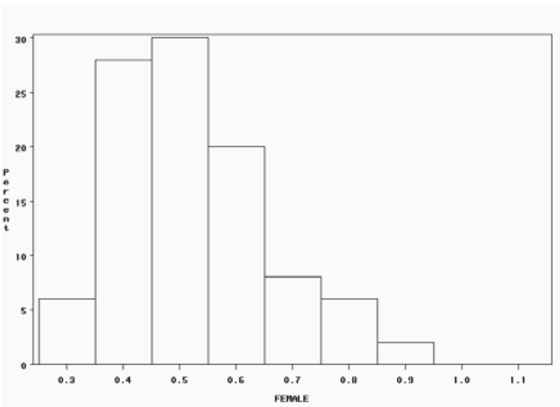
Graphs > Histogram…

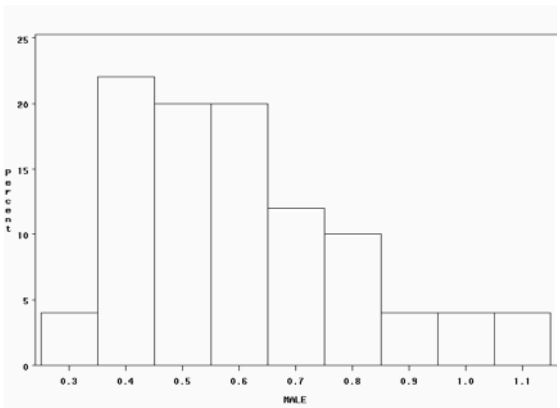
MALE FEMALE → Analysis

Display

Type 0.2 , 1.1, 0.1 into: Midpoints of histogram intervals

SAS OUTPUT FOR FIGURE 15.9 A AND B **Histograms of squirrel body weights**





SAS COMMANDS TO PRODUCE FIGURE 15.9 C AND D

Commands

```
data ;
  set gandh.Chapter9;
  LMALE =log(MALE);
  LFEM =log(FEMALE);
  run;

proc capability noprint;
  var LMALE LFEM;
  histogram / midpoints = -1.3 to 0.1 by 0.1;
  run;
```

Menu route

Edit > Mode > Edit

Data > Transform > Compute…

Type "LMALE" into top left box

log(MALE) → Main pane

Data > Transform > Compute…

Type "LFEM" into top left box

log (FEMALE) → Main pane

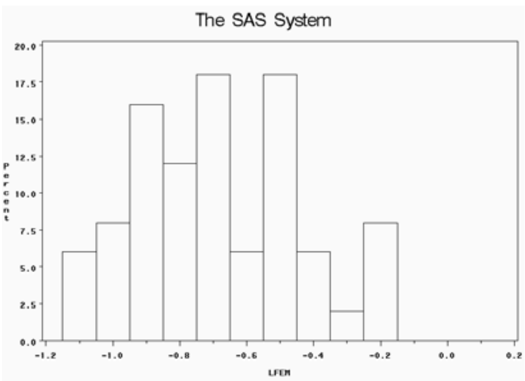
Graphs > Histogram…

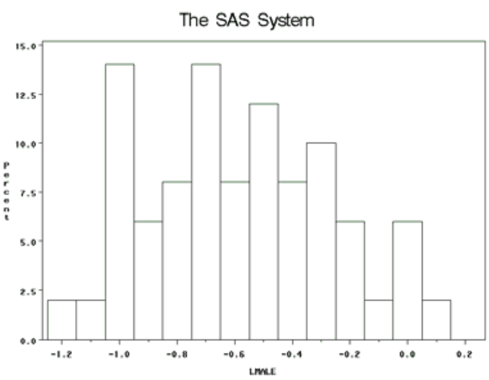
LMALE LFEM → Analysis

$\boxed{\text{Display}}$

Type –1.3, 0.1, 0.1 into: Midpoints of histogram intervals

SAS OUTPUT FOR FIGURE 15.9 C AND D **Histograms of squirrel log (body weights)**





We now produce the Normal probability plots. We use the option 'normaltest' so that SAS will produce tests for Normality in among all the other printed output. These include the Anderson-Darling test shown in the main text.

SAS COMMANDS FOR FIGURE 15.10

Commands

```
data;
  set gandh.chapter9;
  LMALE =log(MALE);
  LFEM =log(FEMALE);

proc capability normaltest;
  var MALE FEMALE LMALE LFEM;
  probplot / normal(mu=est sigma=est);
  run;
```

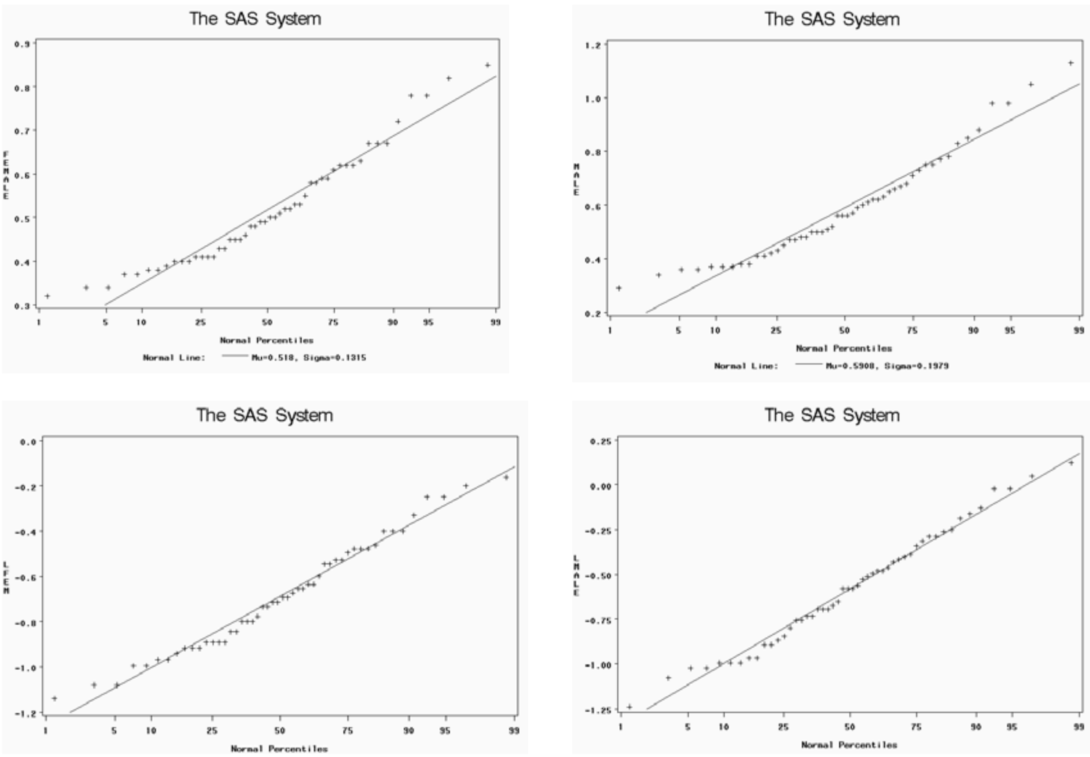
Menu route

Graphs > Probability Plot…

`MALE FEMALE LMALE LFEM` → Analysis

This produces the graphs below.

SAS OUTPUT FOR FIGURE 15.10 **Normal probability plots for the squirrels dataset**



When comparing these graphs with those in the main text, note that the axes are reversed. The Anderson-Darling Normality test provides us with a criterion for rejecting the assumption of Normality. The *p*-value is below 0.05 for the raw data, but not for the transformed data.

EDITED SAS OUTPUT SHOWING THE ANDERSON-DARLING TEST FOR NORMALITY

```
        The CAPABILITY Procedure
            Variable:  MALE
          Tests for Normality

Test                 --Statistic---    -----p-value-----

Shapiro-Wilk         W     0.940648    Pr < W      0.014
Kolmogorov-Smirnov   D     0.099719    Pr > D     >0.150
Cramer-von Mises     W-Sq  0.115478    Pr > W-Sq   0.071
Anderson-Darling     A-Sq  0.812115    Pr > A-Sq   0.035

        The CAPABILITY Procedure
           Variable:  FEMALE
          Tests for Normality

Test                 --Statistic---    -----p-value-----

Shapiro-Wilk         W     0.941389    Pr < W      0.015
Kolmogorov-Smirnov   D     0.103651    Pr > D     >0.150
Cramer-von Mises     W-Sq  0.128795    Pr > W-Sq   0.045
Anderson-Darling     A-Sq  0.858657    Pr > A-Sq   0.025

        The CAPABILITY Procedure
           Variable:  LMALE

          Tests for Normality

Test                 --Statistic---    -----p-value-----

Shapiro-Wilk         W     0.982436    Pr < W      0.658
Kolmogorov-Smirnov   D     0.064466    Pr > D     >0.150
Cramer-von Mises     W-Sq  0.029189    Pr > W-Sq  >0.250
Anderson-Darling     A-Sq  0.236910    Pr > A-Sq  >0.250

        The CAPABILITY Procedure
            Variable:  LFEM
          Tests for Normality

Test                 --Statistic---    -----p-value-----

Shapiro-Wilk         W     0.974443    Pr < W      0.347
Kolmogorov-Smirnov   D     0.096994    Pr > D     >0.150
Cramer-von Mises     W-Sq  0.055795    Pr > W-Sq  >0.250
Anderson-Darling     A-Sq  0.370725    Pr > A-Sq  >0.250
```

The following commands allow you to create the null datasets designed to follow a Normal distribution, with a mean and standard deviation equal to that of the female squirrel weights (which can be obtained using PROC MEANS, as used earlier).

SAS COMMANDS TO SIMULATE DATASETS

Commands

```
data;
  do I=1 to 50;
  noise=normal(0);
  SIMFEM1=0.518+0.1315*noise;
  output;
  end;
  run;
```

Menu route

File > New (*necessary only the first time*)

Data > Random Variates > Normal

50 → Number of values to generate (*necessary only the first time*)

`SIMFEM1` → New column name (*change the name each time*)

0.518 → mean

0.1315 → standard deviation

This can be repeated to create as many columns of simulated female (or male) data as your require. This is particularly easy using the written commands in the session window, as each time you press the submit button, a new dataset is produced in the work folder (called data1, data2 etc.). These datasets may then be examined using the histogram and Normal probability plot commands used earlier.

Chapter 10

Partitioning a sum of squares into polynomial components

Note that in the command language the analysis is conducted in two parts. In the first part, the 'interaction means' are stored at the same time as conducting the PROC GLM and printing out a table of means. In the second part, we set up some formats using PROC GOPTIONS, and then plot using PROC GPLOT. There are three SYMBOLi statements because there are three lines.

Notice that in the menu route we first remove `BSPACE` and `BVARIETY` from the model. This is necessary in order to control the order of variables in the model. The menu route allows us to put error bars on the plot, whereas the command route does not.

SAS COMMANDS FOR BOX 15.18 AND FIGURE 10.8

Commands

```
proc glm data=gandh.Chapter10;
  class BBLOCK BSPACE BVARIETY;
  model BYIELD = BBLOCK BSPACE|BVARIETY;
  lsmeans BSPACE*BVARIETY / out=WORK.INDIAG2;

goptions reset=SYMBOL;

proc goptions;
  SYMBOL1 LINE=1 CI=BLACK INTERPOL=JOIN;
  SYMBOL2 LINE=2 CI=BLUE INTERPOL=JOIN;
  SYMBOL3 LINE=3 CI=RED INTERPOL=JOIN;

proc gplot data=WORK.INDIAG2;
  label lsmean = "Predicted YIELD";
  plot lsmean * BSPACE = BVARIETY;
  run;
```

Menu route

Statistics > Anova > Linear Models…

BYIELD → Dependent

BBLOCK BSPACE BVARIETY → Class

Model

Effects in Model

BSPACE BVARIETY → Remove

Select 2 as factorial order by clicking on ▲

Independent → Effects in model

(BSPACE & BVARIETY) → Factorial

Plots with Means tab

☑ Plot dependent means for two-way effects

⊙ Predicted means

⊙ 2 se

Means with LSMeans tab

BSPACE*BVARIETY → LS Mean

SAS OUTPUT FOR BOX 15.18 **Analysis for barley yield**

The GLM Procedure

Dependent Variable: BYIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	2203.527778	200.320707	11.34	<.0001
Error	24	424.111111	17.671296		
Corrected Total	35	2627.638889			

R-Square	Coeff Var	Root MSE	BYIELD Mean
0.838596	7.272180	4.203724	57.80556

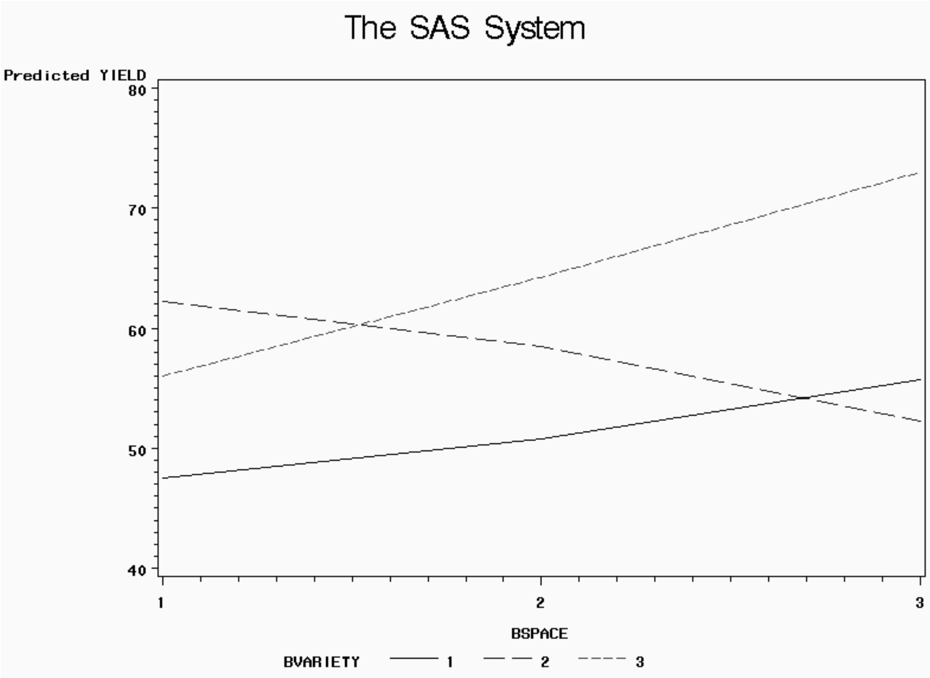
Source	DF	Type III SS	Mean Square	F Value	Pr > F
BBLOCK	3	255.638889	85.212963	4.82	0.0091
BSPACE	2	155.055556	77.527778	4.39	0.0238
BVARIETY	2	1027.388889	513.694444	29.07	<.0001
BSPACE*BVARIETY	4	765.444444	191.361111	10.83	<.0001

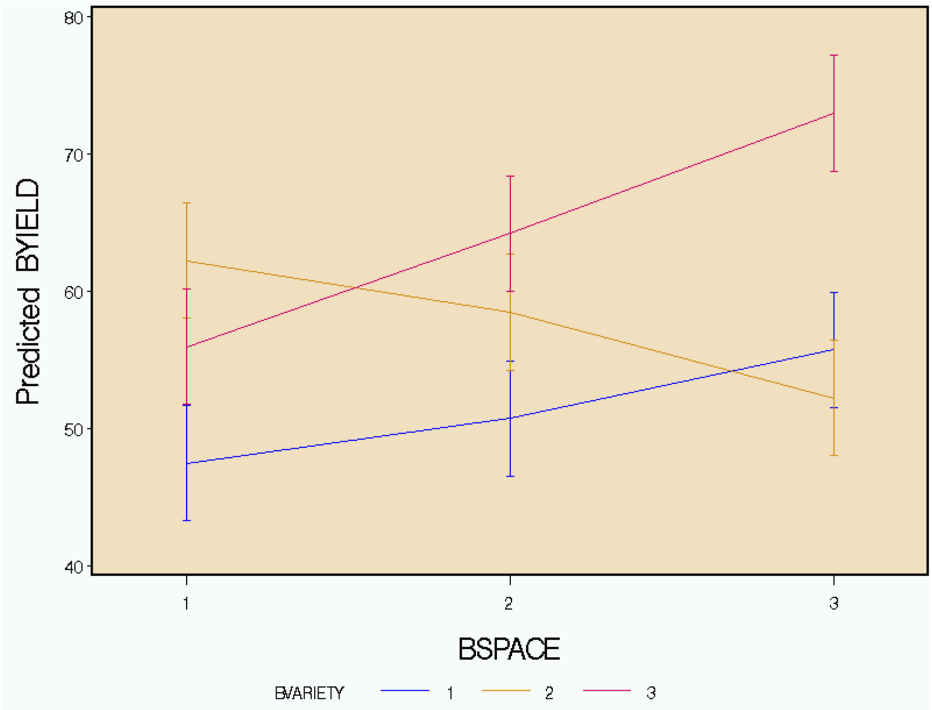
The GLM Procedure
Least Squares Means

BSPACE	BVARIETY	BYIELD LSMEAN
1	1	47.5000000
1	2	62.2500000
1	3	56.0000000
2	1	50.7500000
2	2	58.5000000
2	3	64.2500000
3	1	55.7500000
3	2	52.2500000
3	3	73.0000000

SAS OUTPUT FOR FIGURE 10.8

Interaction diagram from the *barley* dataset. Above by the command line route, and below by the menu route





A second analysis is then conducted on the *barley* dataset, involving a polynomial decomposition. We choose to include the interaction of BVARIETY with the quadratic term BSPACE*BSPACE so that the polynomial analysis is an exact decomposition of the non-polynomial analysis. This means that the error SS, MS and DF will be the same in the two analysis, allowing us to compare them more easily. It also allows us to investigate whether the response of yield to spacing is curvilinear for some varieties but not others. Note the use of Type 1 (that is, sequential) sums of squares, as the use of BSPACE as a covariate has destroyed some of the orthogonality of the design. For example, knowing BSPACE certainly helps in guessing the value of BSPACE*BSPACE!

SAS COMMANDS FOR BOX 15.19 **Second analysis for barley yield**

Commands

```
proc glm data=gandh.Chapter10;
  class BBLOCK BSPACE BVARIETY;
   model BYIELD = BBLOCK BVARIETY|BSPACE|BSPACE / SS1;
  run;
```

Menu route

Statistics > Anova > Linear Models…

BYIELD → Dependent

BBLOCK BVARIETY → Class

BSPACE → Quantitative

Model

Effects in Model

BSPACE BVARIETY → Remove

Select 2 as Polynomial order by clicking on ▲

Independent → Effects in model

BVARIETY → Add

BSPACE → Add

(BSPACE & BVARIETY) → Cross

BSPACE → Polynomial

(BSPACE & BVARIETY) → Cross

Select BVARIETY in 'Independent' pane, and BSPACE* BSPACE in 'Effects in model' pane, then click Cross

Statistics

-☑ Type 3

☑ Type 1

SAS OUTPUT FOR BOX 15.19 **Second analysis for barley yield**

```
                              The GLM Procedure

Dependent Variable: BYIELD

                                     Sum of
      Source                 DF     Squares     Mean Square    F Value    Pr > F

      Model                  11   2203.527778    200.320707      11.34    <.0001

      Error                  24    424.111111     17.671296

      Corrected Total        35   2627.638889

              R-Square     Coeff Var      Root MSE    BYIELD Mean

              0.838596      7.272180      4.203724       57.80556

      Source                 DF     Type I SS   Mean Square    F Value    Pr > F

      BBLOCK                  3    255.638889     85.212963       4.82    0.0091
      BVARIETY                2   1027.388889    513.694444      29.07    <.0001
      BSPACE                  1    155.041667    155.041667       8.77    0.0068
      BSPACE*BVARIETY         2    759.083333    379.541667      21.48    <.0001
      BSPACE*BSPACE           1      0.013889      0.013889       0.00    0.9779
      BSPACE*BSPACE*BVARIE    2      6.361111      3.180556       0.18    0.8364
```

Chapter 11

Finding the best treatment for cat fleas

The final model is:

SAS COMMANDS FOR BOX 15.20 **Analysis for cat fleas**

Commands

```
proc glm data=gandh.Chapter11;
  class TRTMT CARPET;
  model LOGFLEAS = TRTMT NCATS CARPET / SS3;
  run;
```

Menu route

Statistics > Anova > Linear Models…

LOGFLEAS → Dependent

TRTMT CARPET → Class

NCATS → Quantitative

SAS OUTPUT FOR BOX 15.20 **Analysis for cat fleas**

```
                                 The GLM Procedure

Dependent Variable: LOGFLEAS

                                          Sum of
     Source                      DF      Squares     Mean Square    F Value    Pr > F

     Model                        3    31.44462221    10.48154074     16.57    <.0001

     Error                       85    53.75863885     0.63245457

     Corrected Total             88    85.20326106

                  R-Square     Coeff Var      Root MSE    LOGFLEAS Mean

                  0.369054      19.76603      0.795270         4.023419


     Source                      DF    Type III SS    Mean Square    F Value    Pr > F

     TRTMT                        1     5.75019579     5.75019579      9.09    0.0034
     NCATS                        1    22.96168694    22.96168694     36.31    <.0001
     CARPET                       1     6.10297060     6.10297060      9.65    0.0026
```

Multiplicity of p-values

To create the random data set you can either write a short program in the command line route, or make use of the Random Variates option in menus.

SAS COMMANDS TO PRODUCE THE NULL DATASET

Commands

```
data WORK.randata;
  array COL{11};
  drop I J;
  do I=1 to 30;
    do J=1 to 11;
      COL{J}=10.5+2.4*normal(0);
    end;
  output;
  end;
  run;
```

Menu route

File > New
Data > Random Variates > Normal
    30 → Number of values to generate
    COL1 → New column name
    10.5 → mean
    2.4 → standard deviation
Edit > Mode > Edit
Data > Random Variates > Normal
    COL2 → New column name
    10.5 → mean
    2.4 → standard deviation

and repeat this final step for COL3, COL4 etc up to COL11

PROC REG will provide output giving the *p*-value for the regression as a whole, with the *p*-values for each individual X variable in the analysis being found in the table of coefficients. (The *p*-value resulting from the *t* test for the slope against zero for a particular variable is the same as the *p*-value for the *F* ratio test for the same explanatory variable in the analysis of variance table. This arises from the fact that $t^2 = F$; see Appendix 3).

SAS COMMANDS TO PRODUCE OUTPUT FOR TABLE 15.6

Commands

```
proc reg data=gandh.Chapter11;
  model COL1= COL2-COL11;
  run;
```

Menu route: Statistics > Regression > Linear …

COL1 → Dependent

COL2 COL3 COL4 COL5 COL6 COL7 COL8 COL9 COL10 COL11 → Explanatory

EXAMPLE SAS OUTPUT FOR TABLE 15.6 Multiple regression with unrelated variables

```
                              The REG Procedure
                                Model: MODEL1
                          Dependent Variable: COL1

                             Analysis of Variance

                                       Sum of           Mean
Source                    DF          Squares         Square    F Value    Pr > F

Model                     10         68.31104        6.83110       1.36    0.2349
Error                     39        195.95819        5.02457
Corrected Total           49        264.26922


          Root MSE              2.24156    R-Square     0.2585
          Dependent Mean       10.23035    Adj R-Sq     0.0684
          Coeff Var            21.91082


                             Parameter Estimates

                            Parameter       Standard
     Variable     DF         Estimate          Error    t Value    Pr > |t|

     Intercept     1          6.90922        4.71496       1.47      0.1508
     COL2          1          0.18923        0.14159       1.34      0.1891
     COL3          1         -0.00625        0.14082      -0.04      0.9648
     COL4          1         -0.26071        0.14593      -1.79      0.0818
     COL5          1          0.07864        0.12610       0.62      0.5365
     COL6          1         -0.13006        0.14481      -0.90      0.3746
     COL7          1          0.30202        0.16153       1.87      0.0690
     COL8          1         -0.06789        0.14262      -0.48      0.6367
     COL9          1          0.25457        0.14110       1.80      0.0789
     COL10         1          0.01550        0.14639       0.11      0.9162
     COL11         1         -0.08342        0.14385      -0.58      0.5653
```

Chapter 12

How a nested analysis can solve problems of non-independence

These commands first calculate the variable LUPRATE, and then conduct a nested analysis.

SAS COMMANDS FOR BOX 15.21: Nested analysis of *sheep* dataset

Commands

```
data;
  set gandh.Chapter12;
  LUPRATE=NLOOKUPS/DURATION;

proc mixed method=type1 covtest;
  class SEX SHEEP;
  model LUPRATE = SEX / htype=1 ddfm=satterth;
  random SHEEP(SEX);
  run;
```

Menu route

Edit > Mode > Edit

Data > Transform > Compute…

Type "LUPRATE" into top left box

NLOOKUPS / DURATION → Main pane

Statistics > Anova > Mixed Models…

LUPRATE → Dependent

SEX SHEEP → Class

Model

SEX → Add

☑ Random Effects

SHEEP → Add

Select SHEEP in Random effects box, and SEX in Class box, and then click 'Nest'

Tests

-☑ Type III

☑ Type I

☑ Tests of variance components

Options

⊙ Type 1

SAS OUTPUT FOR BOX 15.21 **Nested analysis of *sheep* dataset**

```
                         The Mixed Procedure

                          Model Information

               Data Set                      WORK.DATA1
               Dependent Variable            LUPRATE
               Covariance Structure          Variance Components
               Estimation Method             Type 1
               Residual Variance Method      Factor
               Fixed Effects SE Method       Model-Based
               Degrees of Freedom Method     Satterthwaite

                       Class Level Information

                 Class    Levels    Values

                 SEX           2    1 2
                 SHEEP         3    1 2 3

                             Dimensions

                        Covariance Parameters             2
                        Columns in X                      3
                        Columns in Z                      6
                        Subjects                          1
                        Max Obs Per Subject             120
                        Observations Used               120
                        Observations Not Used             0
                        Total Observations              120

                     Type 1 Analysis of Variance

                              Sum of
Source           DF          Squares    Mean Square   Expected Mean Square

SEX               1         0.132816       0.132816   Var(Residual) + 20 Var(SHEEP(SEX)) + Q(SEX)

                     Type 1 Analysis of Variance

                                                              Error
        Source       Error Term                                  DF    F Value    Pr > F

        SEX          MS(SHEEP(SEX))                               4       1.80    0.2506


                         The Mixed Procedure

                     Type 1 Analysis of Variance

                              Sum of
Source           DF          Squares    Mean Square   Expected Mean Square

SHEEP(SEX)        4         0.294774       0.073693   Var(Residual) + 20 Var(SHEEP(SEX))
Residual        114         0.312387       0.002740   Var(Residual)

                     Type 1 Analysis of Variance

                                                              Error
        Source       Error Term                                  DF    F Value    Pr > F

        SHEEP(SEX)   MS(Residual)                               114      26.89    <.0001
        Residual     .                                            .        .        .

                   Covariance Parameter Estimates

                                        Standard         Z
             Cov Parm       Estimate       Error     Value         Pr Z

             SHEEP(SEX)     0.003548    0.002606      1.36      0.1733
             Residual       0.002740    0.000363      7.55      <.0001

                           Fit Statistics

                   -2 Res Log Likelihood          -339.9
                   AIC (smaller is better)        -335.9
                   AICC (smaller is better)       -335.8
                   BIC (smaller is better)        -336.4

                    Type 1 Tests of Fixed Effects

                             Num      Den
                  Effect      DF       DF    F Value    Pr > F

                  SEX          1        4       1.80    0.2506
```

Chapter 13

Fig trees in Costa Rica

Descriptive statistics are obtained using proc means, or via summary statistics as follows:

SAS COMMANDS FOR TABLE 15.7 **Means and variances for fig tree sites**

Commands

```
proc means data=gandh.Chapter13 mean std median q1 q3
     min max;
  class SITE;
  var NINDIVS;

proc means data=gandh.Chapter13 mean std median q1 q3
     min max;
  var NINDIVS;
  run;
```

Menu route

Statistics > Descriptive > Summary Statistics…

NINDIVS → Analysis

SITE → Class

Statistics > Descriptive > Summary Statistics…

NINDIVS → Analysis

SAS OUTPUT FOR TABLE 15.7 **Means and variances for fig tree sites**

The MEANS Procedure

Analysis Variable : NINDIVS

SITE	N Obs	N	Mean	Std Dev	Minimum	Maximum
1	100	100	9.5100000	3.2176753	2.0000000	17.0000000
2	100	100	18.9300000	4.4090243	8.0000000	28.0000000
3	100	100	41.4300000	6.2607564	27.0000000	56.0000000

The MEANS Procedure

Analysis Variable : NINDIVS

N	Mean	Std Dev	Minimum	Maximum
300	23.2900000	14.2394443	2.0000000	56.0000000

The appropriate glm analysis is then as follows:

SAS COMMANDS FOR BOX 15.22

Commands

```
data;
  set gandh.Chapter13;
  SQRTN =sqrt(NINDIVS);

proc glm;
  class SITE;
  model SQRTN = SITE;
  run;
```

Menu route

Data > Transform > Compute…

Type “SQRTN” into top left box

sqrt(NINDIVS) → Main pane

Statistics > Anova > Linear Models…

SQRTN → Dependent

SITE → Class

SAS OUTPUT FOR BOX 15.22

```
                                  The GLM Procedure

Dependent Variable: SQRTN

                                            Sum of
     Source                      DF        Squares     Mean Square    F Value    Pr > F

     Model                        2    582.6257220     291.3128610    1091.99    <.0001

     Error                      297     79.2310695       0.2667713

     Corrected Total            299    661.8567915

                    R-Square     Coeff Var      Root MSE    SQRTN Mean

                    0.880290      11.24851      0.516499      4.591711

     Source                      DF      Type I SS     Mean Square    F Value    Pr > F

     SITE                         2    582.6257220     291.3128610    1091.99    <.0001

     Source                      DF    Type III SS     Mean Square    F Value    Pr > F

     SITE                         2    582.6257220     291.3128610    1091.99    <.0001
```