

Statistical Methods AEMA-610

R. I. Cue

Department of Animal Science
Faculty of Agricultural and Environmental Sciences
McGill University

Roger.Cue@McGill.ca

Fall 2021

Contents

1	Why Statistics ?	9
1.1	Introduction	9
1.2	What is this course about and for whom?	10
1.3	A brief review of matrices	11
1.4	Matrices STD Ch 12,13	11
1.5	Addition and Subtraction of Matrices	13
1.6	Multiplication of Matrices	14
1.7	Matrix Inversion	14
1.8	Solving Simultaneous Equations	16
1.9	Rank	17
2	Regression - STD: Ch. 14	18
3	Multiple Regression - STD: Ch. 14	19
3.1	Assumptions	19
3.2	Linear Model	19
3.3	Parameters of the model	20
3.4	Hypotheses	20
3.5	Matrix Notation	23
3.6	Parameter Estimates	24
4	Least-squares	24
4.1	Derivation of Least Squares	24
4.2	The Normal Equations	25
4.3	Obtaining a Solution	25
4.4	Example 1	26
4.5	Estimated Model Equation, Prediction Equation	29
4.6	Sampling Variance-Covariance Matrix	29
4.7	Parameter estimates and Standard Errors	30
4.8	Hypotheses	31
4.9	Sums of Squares	31
4.10	Example analysis using SAS	32
4.11	Hypotheses revisited	35
4.12	Analysis of Variance	35
4.13	F-tests, t-tests and Chi-squared	37

4.13.1	F-values	38
4.13.2	t-values	38
4.13.3	Chi-squared values	39
4.13.4	F, t and Chi-squared values from the datastep	40
4.14	Reflections	40
4.15	Order of Equations	44
4.16	Standardized regression coefficients	44
4.17	Partial R ²	45
5	t-test	46
5.1	Testing a group of regression parameters simultaneously	47
6	Confidence Intervals	49
6.1	C.I. for a fixed effect estimate	49
6.2	C.I. for a random effect variance component	49
7	Predicted/Estimated/Fitted Values	51
7.1	Predicting the value of an observation.	51
7.2	Using SAS PROC GLM+IML to estimate a fitted value	52
7.3	Predicting the value of some future observation	52
7.4	Using SAS PROC GLM+IML to predict a future value	53
8	Linear and Quadratic Regressions	54
8.1	Linear Model	54
8.2	Matrix Equations	54
8.3	Testing the quadratic effect	56
8.4	Using SAS PROC GLM to fit a quadratic	57
9	Interactions amongst regression effects	59
9.1	Linear Model	59
9.2	Using SAS PROC GLM to fit an interaction	60
10	Correlations	61
10.1	Variance-Covariance matrix	61
10.2	Simple correlations	63
10.3	Partial correlations	63
10.4	Partial correlations, adjusting for only some variables	64
10.5	Statistical significance of an estimate of a correlation coefficient	64

10.6	Sampling Distribution of an estimate of a correlation coefficient	66
10.7	Statistical Significance	67
10.8	Confidence Interval	67
10.9	Using SAS CORR	69
10.10	Correlations accounting for the effects of fixed effects	70
11	1-Way Classification. STD. Ch.7	76
11.1	Linear Model	76
11.2	Parameters of the Model	77
11.3	Hypotheses to be tested	77
11.4	Matrix Equations	81
11.5	Example, adapted from STD, Ch.7	83
11.6	The Normal Equations	83
11.7	Generalised inverses from GLM and IML	84
11.8	Analysis of Variance	86
11.9	Expectations of Mean Squares	87
11.10	Using SAS/IML	87
11.11	Using SAS/GLM	91
12	Fitted Values	95
13	Assumptions for the Model	97
14	Treatment Differences	98
14.1	Estimation of Treatment Differences	98
14.2	Sampling Variance of Treatment Differences	99
14.3	Test Whether Difference is Statistically Significant	101
14.4	Testable Hypotheses	101
14.5	Using the SAS CONTRAST statement	103
14.6	Using the SAS ESTIMATE statement	105
14.7	Equation Order	107
15	Random effects models	108
15.1	Parameters	108
15.2	Example	108
15.3	Terms in the model	109
15.4	Parameters of the model	109
15.5	Expectations of Mean squares	109

15.6	Estimation of variance components from E(MS)	110
15.7	Using SAS/GLM with a random effect	111
15.8	Using SAS/MIXED	111
15.9	Reasons for our interest in random effects	112
16	Multiple Comparisons	113
16.1	Issues related to multiple comparisons	113
16.2	Bonferroni's Test and Sidak's Inequality	114
16.3	Scheffé's Test, STD Ch 8.5	116
17	Partitioning Sums of Squares, Linear, Quadratic, etc	119
17.1	Hypotheses	120
17.2	Analysis of Variance	121
17.3	SAS code for Classifications	122
17.4	Fitted values	123
17.5	Least Squares Means	124
17.6	SAS Output, classification model	125
17.7	SAS CONTRAST code	129
17.8	Coefficients for Orthogonal Polynomials	130
17.9	Over-parameterized model (GLM)	131
17.10	SAS Output, over-parameterized model	133
17.11	Interpretation of over-parameterized model	134
17.12	Simplification of over-parameterized model	134
17.13	SAS Output from regression model	136
17.14	Conclusion	137
18	Normality and Homogeneity of Variance	138
18.1	Requirements for Analysis of Variance	138
18.2	Normality	138
18.3	Homogeneity of Variance	145
18.4	SAS code for Homogeneity of Variance	145
18.5	SAS output from PROC MIXED, homogeneous variances	148
18.6	SAS output from PROC MIXED, heterogeneous variances (6)	149
18.7	SAS code for 2 Residual Variances	152
18.8	SAS output from PROC MIXED, heterogenous variances (2)	153

19 Homogeneity of Variance - STD Ch 19.3	157
19.1 Basic Formulae	157
19.2 What to use for Homogeneity test	158
19.3 Rationale for the use of residuals	162
19.4 Exercises	162
20 Multiway Classification - STD Ch. 9	163
20.1 Linear model	164
20.2 Parameters of the Model	165
20.3 Parameters	166
20.4 Observations	166
20.5 Matrix Equations	167
20.6 Normal Equations	168
20.7 A solution vector (GLM)	168
20.8 Analysis using SAS/IML	169
20.9 Analysis using SAS/GLM	173
20.10 Analysis of Variance	175
21 Hypotheses to be tested	177
22 Partitioning Sums of Squares for the Model	180
22.1 Sums of Squares for Factor 1	180
22.2 SAS CONTRAST statement for Factor 1	182
22.3 Sums of Squares for Factor 2	182
22.4 SAS CONTRAST statement for Factor 2	182
22.5 Expectations of Mean Squares	184
23 Gains in Efficiency STD 9.7	185
24 Expectation of Mean Squares	186
25 Least Squares Means	188
25.1 LSMEANS using SAS/GLM	190
26 Multiway Classification - Fixed effect and Random Effect, STD Ch. 9	191
26.1 Parameters	192
26.2 Observations	192
26.3 Analysis using SAS/MIXED	192

26.4 Results	195
27 Subsamples, or Nested Models	198
27.1 Linear model	198
27.2 Parameters for a Nested Model	198
27.3 Hypotheses	198
27.4 Matrix Equations	201
27.5 Normal Equations	202
27.6 Analysis of Variance	203
27.7 Expectations of Mean Squares	204
27.8 Computing Sums of Squares	204
27.9 Comparisons amongst treatment means	209
27.10 Analysis using SAS	209
27.11 SAS output	212
27.12 Using Group Means	220
27.13 Expectation of Mean Squares	220
28 Factorial Experiments	222
28.1 Observations	223
28.2 SAS code for Data Step	223
28.3 Linear model	224
28.4 Parameters of the Model	225
28.5 Linear model in Matrix Notation	225
28.6 Normal Equations	226
28.7 Solution to the linear model	227
28.7.1 SAS code for Factorial model	227
28.8 Basic Analysis of Variance	228
28.9 Hypotheses of Interest	228
28.10 Derivation of Testable Hypotheses	229
28.10.1 SAS code for Fitted values	229
28.10.2 Factor Store	230
28.10.3 SAS CONTRAST code for Store	232
28.10.4 Factor Treatment	232
28.10.5 SAS CONTRAST code for Treatments	234
28.10.6 Store*Treatment Interaction	234
28.10.7 Degrees of freedom for the A*B Interaction	235
28.10.8 Sums of Squares for the A*B Interaction	236

28.10.9 SAS CONTRAST code for Interaction	238
28.11 Analysis of Variance, partitioned	239
28.12 Expectation of Mean Squares	240
28.12.1 SAS code for LSMeans	241
29 Latin Square	243
29.1 Linear Model	245
29.2 Analysis of Variance	250
29.3 Fixed or Random ?	250
29.4 Analysis using SAS/GLM and MIXED	251
29.5 Gains in efficiency	252
30 Covariance	254
30.1 Linear model	255
30.2 Matrix Equations	255
30.3 Analysis of Variance	256
30.4 Analysis using SAS	256
30.5 Least Squares Means	264
31 Split Plot	265
31.1 Linear Model	267
31.2 SAS/PROC GLM Model	268
31.3 SAS/PROC MIXED Model	268
31.4 Analysis of Variance	273
31.5 Analysis using SAS	273
31.6 Expectation of Mean Squares	276
32 Cross Over Design	278
32.1 General comments	278
32.2 Description	279
32.3 Linear Model	280
32.4 Parameters of the model	280
32.5 Matrix Equations	281
32.6 Example data set	282
32.7 Derivation of CONTRASTS	282
32.8 Analysis using SAS/MIXED	283
32.9 Parameter Estimates And Significance	284

32.10	Output from PROC MIXED, including animal effect	285
32.11	Output from PROC MIXED, omitting animal effect	288
32.12	Interpretation and comparison of Model 1 and 2	289
33	Repeated measurements	290
33.1	Background	290
33.2	Linear model	291
33.3	Specifying the covariance structure	294
33.4	Common covariance structures	294
33.5	Results	298
33.6	References	299

1 Why Statistics ?

1.1 Introduction

This is, in my opinion, quite a good time to work in research; there are lots of new things continually being discovered as well as many new and powerful techniques (in the laboratory, methodological improvements, detection, precision, etc). So, in essence we are continually developing lots of 'knowledge'. But all this comes at a price, and the price of our greater knowledge is the increased complexity that goes hand in hand with our greater knowledge of biological forces and processes. There are often many things that we do not control or that we cannot control; this is where variability enters into our experiments.

As an analogy, in mathematics, $2 + 2$ **always** equals 4, whereas in statistics $2 + 2$ **on average** equals 4, but sometimes it might equal 3, sometimes it might equal 5, more of the time it will be closer to 4, so that on average it adds to 4, but not in any one particular experiment.

With modern experimental techniques we often have the capacity to generate lots of 'data', which we have to analyse to try and find the underlying common average consistent effect. This is where computers, computer programmes and statistical programmes come in to play. Modern computers mean that we have lots of 'information' and lots of computer 'power' to do analyses. There is no need to do statistical analyses 'by hand'. This means that much of the mechanical drudgery (and risk of errors) has been eliminated. However, it has not removed the 'GIGO: Garbage In Garbage Out' factor. Neither has it removed the potential for doing the wrong statistical analysis at lightening speed! We still need to think about the statistical methods.

What statistical methods? What sort of data? Are we dealing with Normally distributed data (as known as Gaussian distribution, or the bell curve)? We will briefly mention Binomial/Multinomial and Poisson distributions.

It is assumed that you know Statistical Methods I (AEMA-310) and have at least an introduction to matrices. As an aside there is a very good introduction to matrices with an agricultural, biological orientation by Shayle Searle.

What is the purpose of statistics? It is to describe and explain things, and to allow us to summarize experiments and results and to indicate which results are likely to be real.

1.2 What is this course about and for whom?

This course is designed for graduate students who have already taken Statistical Methods 1 (or an equivalent) and who need to learn more about the statistical assumptions and methods. It is intended to introduce you to SAS, entering data, running analyses, interpreting SAS output. It should get you started on understanding statistical notation so that you can continue your statistical reading and education in your own field of research.

We will refer to various other texts and sources:

Steel, Torrie and Dickey

Searle - Matrix Algebra

SAS System for Linear Models

SAS System for Mixed Models

Analysis of Binary Data - Collett

Cochran and Cox - Experimental Design

SAS language manual

SAS/STAT manual

Lucas - Design and Analysis of Feeding Experiments with milking dairy cattle

My Web Site for Stats II

I cannot hope to cover everything, but you should have the basics and know what sort of questions to ask, e.g. is this a Normal distribution, independence of the observation, are there fixed effects and/or random effects, do we have repeated observations, etc?

1.3 A brief review of matrices

1.4 Matrices STD Ch 12,13

This brief introduction to, and overview of, matrices is not supposed to be a complete coverage of matrices. Neither is it simply a beginner's guide to matrices. Anybody who knows nothing about matrices is referred to the excellent book by S.R.Searle:

- Matrix Algebra Useful for Statistics
- S.R. Searle

This introduction is more an aide-memoire to refresh your knowledge of matrices and matrix operations: addition, subtraction, multiplication, inversion, etc.

Matrices are nothing miraculous, they are simply a convenient way of grouping sets of similar equations.

Suppose that we have:

$$2x + 6y = 10 \tag{1}$$

$$4x + 7y = 10 \tag{2}$$

Can we solve for x and y ? How?

If we multiple equation (1) by 2 we get

$$4x + 12y = 20 \tag{3}$$

If we subtract equation (3) from equation (2)

$$4x + 12y = 20 \tag{4}$$

$$4x + 7y = 10 \tag{5}$$

Subtract:

$$0x + 5y = 10 \tag{6}$$

$$5y = 10 \tag{7}$$

$$y = 2 \tag{8}$$

Then substituting back into equation (1)

$$2x + 6y = 10 \tag{9}$$

$$2x + 6(2) = 10 \tag{10}$$

$$2x + 12 = 10 \tag{11}$$

$$2x = -2 \tag{12}$$

$$x = -1 \tag{13}$$

Thus solutions are $x = -1$ and $y = 2$.

This is how we solved simultaneous equations in highschool, by repeatedly substituting and solving for one variable (in this case y) and then back-substituting to solve for the other(s) (x in this case).

We can also write the equation (1) and (2) using a matrix notation.

Equation (1) is $2 \times x + 6 \times y = 10$

In matrix format we would separate out the x and y variables (unknowns):

$$\begin{bmatrix} 2 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 10 \end{bmatrix}$$

The first matrix is a matrix with 1 row and 2 columns, a 1×2 matrix (also a row vector). The second matrix is a matrix with 2 rows and 1 column, a 2×1 matrix (also a column vector). To be 'conformable' for matrix multiplication the number of columns of the first matrix must equal the number of rows of the second matrix. In general if we have 2 matrices, $A_{m \times n}$ (m rows and n columns) and $B_{n \times p}$ (n rows and p columns) then we can write their product as $C_{m \times p} = A_{m \times n} B_{n \times p}$, with $C_{ij} = \sum_{k=1}^{k=n} A_{ik} B_{kj}$. Note also that unlike scalar multiplication, matrix multiplication is not commutative.

1.5 Addition and Subtraction of Matrices

To be 'conformable' for addition and/or subtraction 2 matrices must have the same dimension; that is to say that the two matrices must have the same number of rows and they must also have the same number of columns. Thus if:

$$A = \begin{bmatrix} 2 & 6 \\ 4 & 7 \\ 0 & 1 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 1 & 0 \\ 6 & -5 \\ 2 & 1 \end{bmatrix}$$

then

$$A + B = \begin{bmatrix} 2 & 6 \\ 4 & 7 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 6 & -5 \\ 2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 6 \\ 10 & 2 \\ 2 & 2 \end{bmatrix}$$

Note that the corresponding elements of the 2 matrices are added (or subtracted).

1.6 Multiplication of Matrices

To be 'conformable' for multiplication 2 matrices must have 'conformable' dimensions, as noted previously. There is no division defined for matrices.

1.7 Matrix Inversion

If we take the matrix of coefficients relating to x and y from our original problem of 2 simultaneous equations we have

$$A = \begin{bmatrix} 2 & 6 \\ 4 & 7 \end{bmatrix}$$

For scalar arithmetic we have multiplication and division, for example $2 \div 2 = 1$. We can look at this as $2/2$, or alternatively as $2 \times \frac{1}{2}$, where $\frac{1}{2}$ is the reciprocal of 2, i.e. 2^{-1} . For matrices we use the 'Inverse' which is like a reciprocal in scalar arithmetic. The Inverse of a matrix is another matrix, such that the product is the 'Identity' matrix, a matrix with 1's on the diagonal and 0's elsewhere.

The Identity matrix is analogous to the scalar 1; multiplying a matrix by the identity matrix leaves it unchanged.

Our matrix $\begin{bmatrix} 2 & 6 \\ 4 & 7 \end{bmatrix}$, what is its inverse?

The inverse will be such that

$$\begin{bmatrix} 2 & 6 \\ 4 & 7 \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

We can write this inverse as:

$$\begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 4 & 7 \end{bmatrix}^{-1}$$

How do we calculate the elements of this inverse (e, f, g and h)? For a general methodology see Searle, Matrix Algebra Useful for Statistics, or any other introductory book about matrices.

In this course we will use SAS to compute inverses of matrices; to do so by hand is long, tedious and error-prone; computers are ideal for this task, leaving us to think about our biological and statistical models. However, a 2×2 matrix is easy to invert and will allow us to demonstrate solving simultaneous equations.

Let the 2×2 matrix be

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Then the inverse is

$$\frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$\begin{bmatrix} 2 & 6 \\ 4 & 7 \end{bmatrix}^{-1} = \frac{1}{2 \times 7 - 4 \times 6} \begin{bmatrix} 7 & -6 \\ -4 & 2 \end{bmatrix}$$

$$= \frac{1}{14 - 24} \begin{bmatrix} 7 & -6 \\ -4 & 2 \end{bmatrix}$$

$$= \frac{1}{-10} \begin{bmatrix} 7 & -6 \\ -4 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} -0.7 & 0.6 \\ 0.4 & -0.2 \end{bmatrix}$$

1.8 Solving Simultaneous Equations

If we return to our 2 simultaneous equations (1) and (2) and write them as matrices we have:

$$\begin{bmatrix} 2 & 6 \\ 4 & 7 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$$

we can pre-multiply both sides by the inverse:

$$\begin{bmatrix} -0.7 & 0.6 \\ 0.4 & -0.2 \end{bmatrix} \begin{bmatrix} 2 & 6 \\ 4 & 7 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -0.7 & 0.6 \\ 0.4 & -0.2 \end{bmatrix} \begin{bmatrix} 10 \\ 10 \end{bmatrix}$$

$$\text{N.B. } \begin{bmatrix} -0.7 & 0.6 \\ 0.4 & -0.2 \end{bmatrix} \begin{bmatrix} 2 & 6 \\ 4 & 7 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This gives

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -0.7 & 0.6 \\ 0.4 & -0.2 \end{bmatrix} \begin{bmatrix} 10 \\ 10 \end{bmatrix} \quad (14)$$

Taking the left hand side of the above equation (14):

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} \quad (15)$$

i.e. pre-multiplying by the identity matrix leaves the vector of x and y unchanged.

Looking at the right hand side of equation 14 we have

$$\begin{bmatrix} -0.7 & 0.6 \\ 0.4 & -0.2 \end{bmatrix} \begin{bmatrix} 10 \\ 10 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \quad (16)$$

Thus, from 15 and 16 we have

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \quad (17)$$

This solution is exactly the same as that which we obtained initially solving by hand. For 2 simultaneous equations solving by hand is less complicated than using a matrix inverse, but if we had 10, 20, or 30 or more simultaneous equations it would be much easier to use a matrix approach and use a computer to solve the equations in the above manner.

1.9 Rank

The column rank of a matrix is the number of linearly independent columns of the matrix. If we can form a column as a linear function of some other columns then this column is not linearly independent of the others.

2 Regression - STD: Ch. 14

- equations in matrix form
- assumptions
- Normal Equations
- using SAS/IML to obtain estimates
- sums of squares, sampling variances and standard errors
- Analysis of Variance
- F-test, t-test and χ^2 for parameters
- Partitioning the Sums of Squares for the Model (SSR)
- Testing several parameters simultaneously
- Confidence Intervals
- Linear and Quadratic Regressions
- Predicted (fitted) Values and Sampling variances
- Non-linear regression and curve fitting

3 Multiple Regression - STD: Ch. 14

3.1 Assumptions

- 1. The model is appropriate to the data
- 2. Homogeneity of variances
- 3. Independence of the observations

Simple Linear Regression

Review STD, Chapter 10, Linear Regression.

Review STD, Chapter 12, Matrix Notation.

Review STD, Chapter 13, Linear Regression in Matrix Notation.

$$Y = a + bx + e \tag{18}$$

Multiple Regression

In a multiple regression context we can extend (18) to more than simply 1 single (simple) regression coefficient; in the example below we have 2 regression coefficients:

$$Y = a + b_1X_1 + b_2X_2 + e \tag{19}$$

Re-write as

3.2 Linear Model

$$Y = b_0 + b_1X_1 + b_2X_2 + e$$

thus

$$Y_1 = b_0 + b_1X_{11} + b_2X_{12} + e_1$$

$$Y_2 = b_0 + b_1X_{21} + b_2X_{22} + e_2$$

· · ·

· · ·

$$Y_i = b_0 + b_1X_{i1} + b_2X_{i2} + e_i$$

· · ·

· · ·

$$Y_n = b_0 + b_1X_{n1} + b_2X_{n2} + e_n$$

3.3 Parameters of the model

$b_0, b_1, b_2, \sigma_e^2$

- definitions of terms in the model

Y_i = the observed, measured value of Y, our dependent variable for the i^{th} observation,
 $i = 1, \dots, N$

b_0 = the expected value of Y when all of X_1 and X_2 are equal to zero

X_{1i} = the observed, measured value of X_1 for the i^{th} observation

b_1 = the regression coefficient of the regression of Y on X_1

X_{2i} = the observed, measured value of X_2 for the i^{th} observation

b_2 = the regression coefficient of the regression of Y on X_2

e_i = the random residual associated with the i^{th} observation,

$$e_i \sim N(0, \sigma_e^2)$$

3.4 Hypotheses

The Null Hypothesis (for the Model) (H_o): that the fixed effects parameters of the model (b_0, b_1 and b_2) do not explain variation in the dependent variable (Y), *i.e.* b_0, b_1 and b_2 are all equal to zero. Our Alternative Hypothesis (H_A) is that the fixed effects parameters of the model DO explain variation in the dependent variable, *i.e.* b_0, b_1 and b_2 are not all equal to zero.

The simplest model would thus be that b_1 and b_2 have no effect, *i.e.* that they are both equal to Zero. If that were the case our model would be $Y = (\bar{Y}) + e$

The Null Hypothesis (for the Mean) (H_o): that $\bar{Y} = 0$, our Alternative Hypothesis (H_A) is that $\bar{Y} \neq 0$.

The Null Hypothesis (for the Model over and above the Mean) (H_o): that the fixed effects regression parameters of the model over and above the Mean (b_1 and b_2 over and

above the Mean) do not explain variation in the dependent variable (Y), *i.e.* b_1 and b_2 are all equal to zero. Our Alternative Hypothesis (H_A) is that these fixed effects parameters of the model, over and above the mean, DO explain variation in the dependent variable, *i.e.* b_1 and b_2 are not all equal to zero, see overleaf, pun intended :-).

Suppose we think that leaf burn is a linear function of N% and CI%

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + e_i$$

Thus

$$\begin{aligned} 0.34 &= 1b_0 + 3.05b_1 + 1.45b_2 + e_1 \\ 0.11 &= 1b_0 + 4.22b_1 + 1.35b_2 + e_2 \\ 0.38 &= 1b_0 + 3.34b_1 + 0.26b_2 + e_3 \\ &\cdot &&\cdot &&\cdot \\ &\cdot &&\cdot &&\cdot \\ &\cdot &&\cdot &&\cdot \\ 0.23 &= 1b_0 + 2.94b_1 + 2.22b_2 + e_{30} \end{aligned}$$

$$\begin{bmatrix} 0.34 \\ 0.11 \\ 0.38 \\ \cdot \\ \cdot \\ \cdot \\ 0.23 \end{bmatrix} = \begin{bmatrix} 1 & 3.05 & 1.45 \\ 1 & 4.22 & 1.35 \\ 1 & 3.34 & 0.26 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 2.94 & 2.22 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_{30} \end{bmatrix}$$

Table 1: Example STD Ch 14. Table 14.1 P. 330

Sample	X_1 (N%)	X_2 (Cl%)	X_3 (K%)	Ln leaf burn, Y
1	3.05	1.45	5.67	0.34
2	4.22	1.35	4.86	0.11
3	3.34	0.26	4.19	0.38
4	3.77	0.23	4.42	0.68
5	3.52	1.10	3.17	0.18
6	3.54	0.76	2.76	0.00
7	3.74	1.59	3.81	0.08
8	3.78	0.39	3.23	0.11
9	2.92	0.39	5.44	1.53
10	3.10	0.64	6.16	0.77
11	2.86	0.82	5.48	1.17
12	2.78	0.64	4.62	1.01
13	2.22	0.85	4.49	0.89
14	2.67	0.90	5.59	1.40
15	3.12	0.92	5.86	1.05
16	3.03	0.97	6.60	1.15
17	2.45	0.18	4.51	1.49
18	4.12	0.62	5.31	0.51
19	4.61	0.51	5.16	0.18
20	3.94	0.45	4.45	0.34
21	4.12	1.79	6.17	0.36
22	2.93	0.25	3.38	0.89
23	2.66	0.31	3.51	0.91
24	3.17	0.20	3.08	0.92
25	2.79	0.24	3.98	1.35
26	2.61	0.20	3.64	1.33
27	3.74	2.27	6.50	0.23
28	3.13	1.48	4.28	0.26
29	3.49	0.25	4.71	0.73
30	2.94	2.22	4.58	0.23

3.5 Matrix Notation

thus

$$\begin{aligned}
 Y_1 &= b_0 + b_1 X_{11} + b_2 X_{12} + e_1 \\
 Y_2 &= b_0 + b_1 X_{21} + b_2 X_{22} + e_2 \\
 &\cdot \quad \quad \cdot \quad \quad \cdot \\
 &\cdot \quad \quad \cdot \quad \quad \cdot \\
 Y_i &= b_0 + b_1 X_{i1} + b_2 X_{i2} + e_i \\
 &\cdot \quad \quad \cdot \quad \quad \cdot \\
 &\cdot \quad \quad \cdot \quad \quad \cdot \\
 Y_n &= b_0 + b_1 X_{n1} + b_2 X_{n2} + e_n
 \end{aligned}$$

rewriting we have

$$\begin{aligned}
 Y_1 &= 1 * b_0 + X_{11} * b_1 + X_{12} * b_2 + e_1 \\
 Y_2 &= 1 * b_0 + X_{21} * b_1 + X_{22} * b_2 + e_2 \\
 &\cdot \quad \quad \cdot \quad \quad \cdot \\
 &\cdot \quad \quad \cdot \quad \quad \cdot \\
 Y_i &= 1 * b_0 + X_{i1} * b_1 + X_{i2} * b_2 + e_i \\
 &\cdot \quad \quad \cdot \quad \quad \cdot \\
 &\cdot \quad \quad \cdot \quad \quad \cdot \\
 Y_n &= 1 * b_0 + X_{n1} * b_1 + X_{n2} * b_2 + e_n
 \end{aligned}$$

In matrix notation

$$Y = Xb + e$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & X_{n1} & X_{n2} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

3.6 Parameter Estimates

Estimates of the parameters b , $\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$ are \hat{b} , $\begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix}$

σ_e^2 will be estimated by the Mean Square Error, $\widehat{\sigma_e^2}$

Deviations,

$$Y = Xb + e$$

$$e = Y - Xb$$

$$\hat{e} = Y - X\hat{b}$$

4 Least-squares

4.1 Derivation of Least Squares

Based on minimising the Sums of Squares of Errors (SSE), where

$$\text{SSE} = \sum_{i=1}^{i=n} \hat{e}_i^2$$

$$\hat{e}'\hat{e} = (Y - X\hat{b})'(Y - X\hat{b})$$

n.b. $\hat{e}' = Y' - \hat{b}'X'$

$$\hat{e}'\hat{e} = (Y' - \hat{b}'X')(Y - X\hat{b})$$

$$= Y'Y - \hat{b}'X'Y - Y'X\hat{b} + \hat{b}'X'X\hat{b}$$

n.b. $\hat{b}'X'Y = Y'X\hat{b}$

$$\hat{e}'\hat{e} = Y'Y - 2\hat{b}'X'Y + \hat{b}'X'X\hat{b} \tag{20}$$

We wish to obtain parameter estimates (\hat{b}) such that $\hat{e}'\hat{e}$ (SSE) is minimised.

Therefore we can differentiate (20), set to zero and solve.

$$\begin{aligned} \frac{\partial}{\partial b}(\hat{e}'\hat{e}) &= \frac{\partial}{\partial b} (Y'Y - 2\hat{b}'X'Y + \hat{b}'X'X\hat{b}) \\ &= -2X'Y + 2X'X\hat{b} \end{aligned}$$

$$-2X'Y + 2X'X\hat{b} = 0$$

$$2X'X\hat{b} = 2X'Y$$

4.2 The Normal Equations

$$X'X\hat{b} = X'Y$$

4.3 Obtaining a Solution

$$(X'X)^{-1}X'X\hat{b} = (X'X)^{-1}X'Y$$

$$\hat{b} = (X'X)^{-1}X'Y$$

4.4 Example 1

$$Y = Xb + e$$

Analyse the effect of N% and Cl% on leaf burn

Thus

$$\begin{aligned} 0.34 &= 1b_0 + 3.05b_1 + 1.45b_2 + e_1 \\ 0.11 &= 1b_0 + 4.22b_1 + 1.35b_2 + e_2 \\ &\cdot &&\cdot &&\cdot \\ &\cdot &&\cdot &&\cdot \\ &\cdot &&\cdot &&\cdot \\ 0.23 &= 1b_0 + 2.94b_1 + 2.22b_2 + e_{30} \end{aligned}$$

We could write the model as

$$Y_i = b_0 + b_1X_{i1} + b_2X_{i2} + plant_i + \epsilon_i$$

However, with only 1 observation per plant we cannot separate $plant_i$ from ϵ_i , thus we are forced to combine them into one term, *i.e.* we have pooled $plant_i$ and ϵ_i into the one single 'error' or residual e_i , where $e_i = plant_i + \epsilon_i$.

Thus the model becomes

$$Y_i = b_0 + b_1X_{i1} + b_2X_{i2} + e_i$$

$$Y = Xb + e$$

$$\begin{bmatrix} 0.34 \\ 0.11 \\ \cdot \\ \cdot \\ \cdot \\ 0.23 \end{bmatrix} = \begin{bmatrix} 1 & 3.05 & 1.45 \\ 1 & 4.22 & 1.35 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 2.94 & 2.22 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_{30} \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 3.05 & 4.22 & 3.34 & \dots & 2.94 \\ 1.45 & 1.35 & 0.26 & \dots & 2.22 \end{bmatrix} \begin{bmatrix} 1 & 3.05 & 1.45 \\ 1 & 4.22 & 1.35 \\ 1 & 3.34 & 0.26 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 2.94 & 2.22 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 30 & 98.36 & 24.23 \\ 98.36 & 332.3352 & 81.5834 \\ 24.23 & 81.5834 & 30.1907 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 20.58 \\ 61.6502 \\ 12.4103 \end{bmatrix}$$

Note, the elements of $X'X$ and $X'Y$ are :

$$X'X = \begin{bmatrix} N & \sum X_{1i} & \sum X_{2i} \\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{1i}X_{2i} \\ \sum X_{2i} & \sum X_{1i}X_{2i} & \sum X_{2i}^2 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \sum Y_i \\ \sum X_{1i}Y_i \\ \sum X_{2i}Y_i \end{bmatrix}$$

$$\begin{pmatrix} 30 & 98.36 & 24.23 \\ 98.36 & 332.3352 & 81.5834 \\ 24.23 & 81.5834 & 30.1907 \end{pmatrix} \hat{b} = \begin{pmatrix} 20.58 \\ 61.6502 \\ 12.4103 \end{pmatrix}$$

So the Normal Equations are:

$$\begin{array}{rclcl}
30\hat{b}_0 & +98.36\hat{b}_1 & +24.23\hat{b}_2 & = & 20.58 \\
98.36\hat{b}_0 & +332.3352\hat{b}_1 & +81.5834\hat{b}_2 & = & 61.6502 \\
24.23\hat{b}_0 & +81.5834\hat{b}_1 & +30.1907\hat{b}_2 & = & 12.4103
\end{array}$$

(Unique) Inverse of $X'X$ is $(X'X)^{-1}$

$$= \begin{pmatrix} 1.12604 & -0.33098 & -0.00931 \\ -0.33098 & 0.10623 & -0.02142 \\ -0.00931 & -0.02142 & 0.09847 \end{pmatrix}$$

$$\hat{b} = (X'X)^{-1}X'Y$$

$$= \begin{pmatrix} 1.12604 & -0.33098 & -0.00931 \\ -0.33098 & 0.10623 & -0.02142 \\ -0.00931 & -0.02142 & 0.09847 \end{pmatrix} \begin{pmatrix} 20.58 \\ 61.6502 \\ 12.4103 \end{pmatrix}$$

$$\hat{b} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} = \begin{pmatrix} 2.6531 \\ -0.5285 \\ -0.2900 \end{pmatrix} \text{ n.b. rounded to 4 decimal places}$$

To obtain any particular value from \hat{b} , or any combination thereof, we use a matrix k' , which we shall use throughout, in many different forms:

$$k' = [0 \quad 1 \quad 0]$$

so that

$$k'\hat{b} = k' = [0 \quad 1 \quad 0] \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} \text{ i.e. "pulls out" } \hat{b}_1$$

i.e. "pulls out" \hat{b}_1

If we want to extract \hat{b}_0 we would use

$$k' = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

so that

$$k'\hat{b} = k' \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} \text{ i.e. "pulls out" } \hat{b}_0$$

4.5 Estimated Model Equation, Prediction Equation

$$\hat{Y} = 2.6531 - 0.5285 * N\% - 0.2900 * Cl\%$$

(log of leaf burn)

$$\text{Total Sums of Squares} = Y'Y$$

$$\text{Sums of Squares for the model} = \hat{b}'X'Y$$

$$\text{Sums of Squares Error (Residual)} = Y'Y - \hat{b}'X'Y$$

$$\text{Mean Square Error} = \hat{\sigma}_e^2 = \frac{Y'Y - \hat{b}'X'Y}{N - r(X)}$$

number of observations = 30 rank of X = 3

4.6 Sampling Variance-Covariance Matrix

$$V(\hat{b}) = (X'X)^{-1}\hat{\sigma}_e^2$$

$$= \begin{bmatrix} \underline{v(\hat{b}_0)} & \text{cov}(\hat{b}_0, \hat{b}_1) & \text{cov}(\hat{b}_0, \hat{b}_2) \\ \text{cov}(\hat{b}_1, \hat{b}_0) & \underline{v(\hat{b}_1)} & \text{cov}(\hat{b}_1, \hat{b}_2) \\ \text{cov}(\hat{b}_2, \hat{b}_0) & \text{cov}(\hat{b}_2, \hat{b}_1) & \underline{v(\hat{b}_2)} \end{bmatrix}$$

We again use our k' matrix to obtain the sampling variance of our estimate $k'\hat{b}$:

$$V(k'\hat{b}) = k' V(\hat{b}) k$$

$$(X'X)^{-1}\sigma_e^2 = \begin{pmatrix} 1.12604 & -0.33098 & -0.00931 \\ -0.33098 & 0.10623 & -0.02142 \\ -0.00931 & -0.02142 & 0.09847 \end{pmatrix} * \sigma_e^2$$

$$(X'X)^{-1}\sigma_e^2 = \begin{pmatrix} 0.099663 & -0.02929 & -0.000824 \\ -0.02929 & 0.009402 & -0.0018958 \\ -0.000824 & -0.0018958 & 0.008715 \end{pmatrix}$$

$$k'(X'X)^{-1}k\sigma_e^2 = [0 \ 1 \ 0] \begin{pmatrix} 0.099663 & -0.02929 & -0.000824 \\ -0.02929 & 0.009402 & -0.0018958 \\ -0.000824 & -0.0018958 & 0.008715 \end{pmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$k'V(\hat{b})k = (0.009402)$$

The standard error, s.e., is then simply the square root of the sampling variance.

And likewise for any other estimate which we write using k' . Compare and check this against STD Ch14.5, Page 332, or against the classical formulae (*i.e.* non-matrix) in most statistics textbooks.

4.7 Parameter estimates and Standard Errors

Thus we have

\hat{b}_0 and s.e. (\hat{b}_0)

\hat{b}_1 and s.e. (\hat{b}_1)

\hat{b}_2 and s.e. (\hat{b}_2)

Thus we can test their significance using a t-test

Assumptions

- 1. Normally distributed errors
- 2. Finite and homogeneous variance
- 3. Independence of observations

4.8 Hypotheses

We have specified hypotheses for the model and the model over and above the mean; we have specified a model, equations and obtained solutions/estimates. What about b_1 and b_2 ? We want to continue to subdivide and partition our variability (Sums of Squares) into the component factors/effects. Our Null Hypothesis for b_1 will be, that there is no effect of b_1 , *i.e.* $b_1 = 0$, *i.e.* $R(b_1 | b_0, b_2)$ is not statistically significant, and our corresponding Alternative Hypothesis will be that $b_1 \neq 0$, *i.e.* that $R(b_1 | b_0, b_2)$ explains a statistically significant amount of the variation of the dependent variable. Our Null Hypothesis for b_2 will be, that there is no effect of b_2 , *i.e.* $b_2 = 0$, *i.e.* $R(b_2 | b_0, b_1)$ is not statistically significant, and our corresponding Alternative Hypothesis will be that $b_2 \neq 0$, *i.e.* that $R(b_2 | b_0, b_1)$ explains a statistically significant amount of the variation of the dependent variable.

4.9 Sums of Squares

Total sums of squares (TSS) = $Y'Y = 20.8074$

Correction Factor, for the Mean (CF) = $N\bar{y}^2 = 14.11788$

Sums of Squares for the model = $\hat{b}'X'Y = 18.417732$

Sums of Squares Error (Residual) = $Y'Y - \hat{b}'X'Y$

$$= 20.8074 - 18.417732$$

$$= 2.3896675$$

Number of observation (N) = 30

$$r(X) = 3$$

Note: We sometimes run into problems if the numbers are very large or very small and/or if there are many numbers, such that the elements of $X'X$ and $X'Y$ are very large or very small, or a mixture of the two. This leads to what is known as numerical instability and is due to the fact that computers have only a finite numerical precision (usually about 14 significant digits). We can alleviate this problem by scaling our numbers to be all in the range 1 to 10, and/or expressing them as deviations from their respective means; this for both the dependent variable (Y) and also for the independent regression covariates (the X's). This practise is often called "Centring Variables".

4.10 Example analysis using SAS

Examples using SAS, with both PROC REG and GLM and explicitly using PROC IML

```
USING SAS/PROC IML
```

```
proc iml;  
reset print;
```

```
x = { 1 3.05 1.45,  
      1 4.22 1.35,  
      1 3.34 0.26,  
      .  
      .  
      1 2.94 2.22}; /* create X matrix */
```

```
y = {0.34,  
      0.11,  
      0.38,  
      .  
      .  
      0.23}; /* create Y matrix */
```

```
xtx = x' * x; /* create X transpose X matrix */  
xty = x' * y; /* create X transpose Y matrix */
```

```
invxtx = inv(xtx); /* obtain the inverse of X'X */
```

```

bhat = invxtx * xty; /* estimate of b */

tss = y' * y; /* Total Sums of Squares */
ssr = bhat' * xty; /* Reductions Sums of Squares, for the model */
sse = tss - ssr; /* Residual Sums of Squares */
nobs = nrow(x); /* N = number of observations */
rx = 3; /* rank of X */
dfe = nobs - rx; /* residual degrees of freedom, N - r(X) */

sumy = sum(y); /* sum the Ys */
ybar = sumy/nobs; /* average of Y */
cf = nobs * ybar * ybar; /* Correction Factor for the mean */
ssrm = ssr - cf; /* Sums of Squares for the model
                  corrected for the mean */

mse = sse/dfe; /* Residual Mean Square, Mean Square Error */
covb = invxtx * mse; /* sampling variance-covariance matrix */

yhat = x * bhat; /* estimated value for each observation */
ehat = y - yhat; /* estimates/predictions of the errors */

/* k' matrices to generate Sums of Squares */
/* SS b1 */
kp = {0 1 0};
kb = kp * bhat;
kinvk = kp * invxtx * kp';
invkk = inv(kinvk);
ss1 = kb' * invkk * kb;

/* SS b2 */
kp = {0 0 1};
kb = kp * bhat;
kinvk = kp * invxtx * kp';
invkk = inv(kinvk);
ss2 = kb' * invkk * kb;

quit;

USING SAS/PROC GLM

data reg1;
input x1 x2 x3 y;
cards;

```

```

3.05 1.45 5.67 0.34
4.22 1.35 4.86 0.11
.      .      .      .
.      .      .      .
.      .      .      .
2.94 2.22 4.58 0.23
;

/* again, in GLM we are using the options / XPX I
   to request that GLM print out X prime X, X prime Y,
   and the inverse
*/
proc glm data=reg1;
model y = x1 x2/XPX I;
output out=reg1out p=yhat r=ehat stdp=se;
/* output to a new SAS data set (reg1out), y x1 x2 yhat ehat se */
run;

/* plot, using high quality graphics SAS/GRAPH */
proc gplot data=reg1out;
plot ehat*x1;
run;

```

4.11 Hypotheses revisited

Our initial Null Hypothesis was that the fixed effects parameters of the model did not explain variation in the dependent variable. This was the Null Hypothesis for the Model. We now need to extend our hypotheses to consider the various parts of the Analysis of Variance (ANOVA) table that we are going to produce.

For the Correction Factor for the Mean, our Null Hypothesis is that the Mean of the dependent variable, \bar{Y} , is equal to Zero, and our Alternative Hypothesis is that the Mean of the dependent variable, \bar{Y} , is not equal to Zero.

For the Model over and above the Mean, our Null Hypothesis is that the regression coefficients, b_1 and b_2 , do not explain variation in the dependent variable, *i.e.* that $b_1 = 0$ and $b_2 = 0$. This we can write statistically as:

$$H_o \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{vs } H_A \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

4.12 Analysis of Variance

Note, the Correction Factor for the mean is testing whether the model $Y = b_0 + e$, *i.e.* just a single overall mean, would be adequate. This is why we must be careful; in this case then b_0 would be equivalent to μ , but it is not the same b_0 as in our model of b_0, b_1, b_2 , in our model of b_0, b_1, b_2 , the term b_0 in the model is the **INTERCEPT**.

$$V(\hat{b}) = (X'X)^{-1}\hat{\sigma}_e^2$$

$$= \begin{pmatrix} 1.12604 & -0.33098 & -0.00931 \\ -0.33098 & 0.10623 & -0.02142 \\ -0.00931 & -0.02142 & 0.09847 \end{pmatrix} * 0.0885062$$

Table 2: Initial ANOVA, Regression

Source	df	SS	MS	E(MS)
Total	N=30	$Y'Y = 20.8074$		
Model R(b_0, b_1, b_2)	$r(X) = 3$	$\hat{b}'X'Y$ $= 18.417732$	$\frac{\hat{b}'X'Y}{r(X)}$ $= 18.417732/3$ $= 6.1392442$	
C.F.	1	$N\bar{y}^2 = 14.11788$	14.11788	
S.S.R. _m R(b_1, b_2 Mean)	$r(X)-1=2$	18.417732 $- 14.11788$ $= 4.2998525$	$4.2998525/2$ $= 2.149926$	$\sigma_e^2 + f(b_1^2, b_2^2)$
Residual	$N-r(X)$ $30 - 3$ $= 27$	$Y'Y - \hat{b}'X'Y$ $20.8074 - 18.417732$ $= 2.3896675$	$2.3897/27$ $= 0.0885$	σ_e^2

$$V(\hat{b}) = \begin{pmatrix} 0.0996618 & -0.029294 & -0.000824 \\ -0.029294 & 0.0094017 & -0.001895 \\ -0.000824 & -0.001895 & 0.0087153 \end{pmatrix}$$

$$s.e.(\hat{b}_0) = \sqrt{0.0996618} = 0.3156926$$

$$s.e.(\hat{b}_1) = \sqrt{0.0094017} = 0.0969625$$

$$s.e.(\hat{b}_2) = \sqrt{0.0087153} = 0.0933558$$

$$\text{t-test of } \hat{b}_0 = \left| \frac{\hat{b}_0}{s.e.(\hat{b}_0)} \right| = \left| \frac{2.6531}{0.3156926} \right| = 8.40406$$

$$\text{t-test of } \hat{b}_1 = \left| \frac{\hat{b}_1}{s.e.(\hat{b}_1)} \right| = \left| \frac{-0.5285}{0.0969625} \right| = 5.45056$$

$$\text{t-test of } \hat{b}_2 = \left| \frac{\hat{b}_2}{s.e.(\hat{b}_2)} \right| = \left| \frac{-0.2900}{0.0933558} \right| = 3.10601$$

NOTE: Compare these values and how we obtained them with those from STD P333.

Table A.3 Values of t, from STD, P611.

d.f. (residual) = 27

read down columns	0.05,	0.01,	0.001
at d.f. = 27 we have	2.052,	2.771,	3.690

Thus all 3 values (\hat{b}_0 , \hat{b}_1 and \hat{b}_2) are significant at the 1% level

It means that there is a less than 1% chance of obtaining such a value due to random chance of sampling with 30 observations (and 27 d.f.e.) when there is no real effect!!!

4.13 F-tests, t-tests and Chi-squared

See:

SAS documentation,

Base SAS Software,

SAS Language Reference Concepts,

SAS System Concepts,

Functions and CALL Routines

For the analysis of variance (ANOVA) we can look up the tabulated F value for the Model. The Model has 3 degrees of freedom for the numerator (**ndf**) and 27 degrees of freedom for the denominator (**ddf**). If we are using a 5% probability level for accepting or rejecting our Null Hypothesis we can look up the critical, tabulated value and find that it is 2.96. Similarly, for our t-test above, with ddf (dfe) of 27, and a 5% probability level then our [2-tailed] t-test critical value is 2.052. This is straightforward when the probabilities and degrees of freedom correspond to those we have in our statistical tables, or when we can easily use linear interpolation to find the values we need. However, often we may have degrees of freedom beyond what are given in the usual published tables. We can make use of SAS (either the procedure IML, or with the datastep) and the functions **finv()**, **tinv()**, and **cinv()**. There are also corresponding functions to give us the probability corresponding to a given calculated F, t or χ^2 value, **probf()**, **probt()**, **prochi()**.

4.13.1 F-values

The function `finv(pupto,ndf,ddf)` gives us the critical, tabulated F-value for a probability less than or equal to `pupto`, *i.e.* prob up to. Note, this is usually the reverse (`pupto`) of what we want, therefore `pupto=1-probability`.

If we use the example data from our regression problem, we recall that the Model has 3 degrees of freedom and 27 residual degrees of freedom (`dfe`, or `ddf`).

```
ndf = 3
```

```
ddf = 27
```

```
probability = 0.05    ≡ 5%
```

Then the SAS code (in IML) would be

```
USING SAS/PROC IML
```

```
proc iml;
reset print;

ndf = 3;
ddf = 27;
probability = 0.05;
ftab = finv(1-probability,ndf,ddf);

quit;
```

4.13.2 t-values

We can similarly use `tinv(pupto,ddf)`. Note, `pupto` is the probability of obtaining up to a given t-value, *i.e.* **NOT** a 2-tailed t-value, but rather a cumulative probability. Therefore

we want 1-probability. In addition, since we are considering 2-tailed tests, we must divide our probability by 2. Thus, if we were carrying out a [2-tailed] t-test, with a probability level of 5% and our 27 residual degrees of freedom, we could write (in PROC IML):

```
USING SAS/PROC IML
```

```
proc iml;
reset print;

ddf = 27;
probability = 0.05;
ttab = tinv(1-probability/2,ddf);

quit;
```

4.13.3 Chi-squared values

We can similarly use `cinv(pupto,ndf)`. Note, `pupto` is the probability of obtaining up to a given χ^2 , i.e. a cumulative probability. Therefore we want 1-probability.

```
USING SAS/PROC IML
```

```
proc iml;
reset print;

ddf = 27;
probability = 0.05;
chitab = cinv(1-probability,ddf);

quit;
```


4.13.4 F, t and Chi-squared values from the datastep

We can also use the SAS datastep to obtain tabulated values; the datastep has similar functions, finv(), tinv() and cinv().

```
/* Using the SAS datastep */

data tabulated;
input pr ndf ddf;
ftab = finv(1-pr,ndf,ddf);
ttab = tinv(1-pr/2,ddf);
chitab = cinv(1-pr,ndf);
cards;
  0.05 3 27
  0.01 3 27
;

proc print data=tabulated;
var pr ndf ddf ftab ttab chitab;
run;
```

4.14 Reflections

N.B.

$X'X$ is symmetric

$(X'X)^{-1}$, diagonals are positive

TSS, SSR, CF, SSR_m , SSE are positive

If the off-diagonals of $(X'X)^{-1}$ are non-zero then the estimates of \hat{b} are not independent.

We can subdivide SSR_m for the regression parameters (the regression on N% and CI%).

To obtain the Marginal, Type III, Sums of Squares for b_1 and b_2

$$\text{SS}(b_1) \quad k' = [0 \ 1 \ 0]$$

$$\text{SS}_i = (k'\hat{b})'[k'(X'X)^{-1}k]^{-1}(k'\hat{b})$$

$$\text{So } k = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$k'\hat{b} = [0 \ 1 \ 0] \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} \quad i.e. \text{ "pulls out" } \hat{b}_1$$

$$C = (X'X)^{-1} = \begin{pmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{pmatrix}$$

$$= \begin{pmatrix} 1.12604 & -0.33098 & -0.00931 \\ -0.33098 & 0.10623 & -0.02142 \\ -0.00931 & -0.02142 & 0.09847 \end{pmatrix}$$

$$k'(X'X)^{-1}k = [0 \ 1 \ 0] \begin{pmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{pmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$(X'X)^{-1}k = \begin{pmatrix} 1.12604 & -0.33098 & -0.00931 \\ -0.33098 & 0.10623 & -0.02142 \\ -0.00931 & -0.02142 & 0.09847 \end{pmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$(X'X)^{-1}k = \begin{pmatrix} -0.33098 \\ 0.10623 \\ -0.02142 \end{pmatrix}$$

$$k'(X'X)^{-1}k = [0 \ 1 \ 0] \begin{pmatrix} -0.33098 \\ 0.10623 \\ -0.02142 \end{pmatrix}$$

$$k'(X'X)^{-1}k = (0.10623)$$

$$= C_{11} - \text{i.e. "pulls out" the part of } (X'X)^{-1} \text{ corresponding to } b_1 \\ = 0.10623$$

$$SS_1 = R(b_1 | b_0, b_2) = -0.5285 * [0.10623]^{-1} * -0.5285 \\ = 2.6293$$

Likewise, we can do the same thing for b_2 .

$$SS(b_2) \quad k' = [0 \ 0 \ 1]$$

$$SS_i = (k'\hat{b})'[k'(X'X)^{-1}k]^{-1}(k'\hat{b})$$

$$\text{So } k = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$k'\hat{b} = [0 \ 0 \ 1] \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} \text{ i.e. "pulls out" } \hat{b}_2$$

$$C = (X'X)^{-1} = \begin{pmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{pmatrix}$$

$$= \begin{pmatrix} 1.12604 & -0.33098 & -0.00931 \\ -0.33098 & 0.10623 & -0.02142 \\ -0.00931 & -0.02142 & 0.09847 \end{pmatrix}$$

$$k'(X'X)^{-1}k = [0 \ 0 \ 1] \begin{pmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{pmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$= C_{22} - \text{i.e. "pulls out" the part of } (X'X)^{-1} \text{ corresponding to } b_2$$

Table 3: Complete Regression ANOVA

ANOVA Source	df	SS	MS	F - ratio	E(MS)
Total, TSS	$N = 30$	$Y'Y$ 20.8074			
Model, SSR	$r(X)$ $= 3$	$\hat{b}'X'Y$ 18.417732	6.13924	69.37	$\sigma_e^2 + f(b_0^2, b_1^2, b_2^2)$
Mean, C.F.	1	$N\bar{y}^2$ 14.11788	14.11788	159.513	$\sigma_e^2 + f(\bar{y})$
Model, after the mean, SSR_m $R(b_1b_2 \text{Mean})$	$r(X) - 1$ $= 2$	$\hat{b}'X'Y - N\bar{y}^2$ 4.29985	2.149926	24.291	$\sigma_e^2 + f(b_1^2, b_2^2)$
SS_{b_1} $R(b_1 b_0b_2)$	1	$\hat{b}_1^2 C_{11}^{-1}$	2.6293	29.708	$\sigma_e^2 + f(b_1^2)$
SS_{b_2} $R(b_2 b_0b_1)$	1	$\hat{b}_2^2 C_{22}^{-1}$	0.85406	9.650	$\sigma_e^2 + f(b_2^2)$
Error, Residual	$N - r(X)$ $30 - 3$	$Y'Y - \hat{b}'X'Y$ 2.3896675	.0885062		σ_e^2

$$= 0.09847$$

$$SS_2 = R(b_2 | b_0, b_1) = -0.2900 * [0.09847]^{-1} * -0.2900$$

$$= 0.85406$$

The tabulated F values are :

$$F_{pr=5\%, ndf=1, ddf=27} = 4.210$$

$$F_{pr=5\%, ndf=2, ddf=27} = 3.354$$

$$F_{pr=5\%,ndf=3,ddf=27} = 2.960$$

4.15 Order of Equations

It is important to note that the order in which we set up the equations does not matter. That is to say we can just as easily have X_2 first, as:

$$Y_i = b_0 + b_2 X_{2i} + b_1 X_{1i} + e_i$$

Try re-doing the above problem but with X_2 first, and note that it makes not one iota of difference.

Note, the sum of the Type III Sums of Squares (the Marginal Sums of Squares [what we are dealing with]) do not necessarily add up to the Sums of Squares of the Model corrected for the mean! See STD Page 333 for a discussion. There can be 2 main reasons, co-linearity (a real problem), or the fact that we have an unbalanced design (your problem!).

4.16 Standardized regression coefficients

The regression coefficient estimates that we have computed tell us by how much the dependent variable (what we have called Y) is expected to change, for each unit change in your independent variable (X1, X2, etc). Consequently the regression estimates are dimensioned in terms of the variables. This can make it difficult to compare one regression estimate with another (if they refer to variables which have quite difference variances). Thus it is not uncommon to see standardized regression coefficients presented, such that all variables, dependent and independent have unit variance. We can obtain standardized regression coefficients by dividing the regression parameter estimate (e.g. \hat{b}_1) by the ratio of the sample standard deviation of the dependent variable ($\hat{\sigma}_Y$) to the sample standard deviation of the regressor (the independent variable, $\hat{\sigma}_{X_1}$)

$$\text{standardized regression coefficient} = \frac{\hat{b}_1}{\frac{\hat{\sigma}_Y}{\hat{\sigma}_{X_1}}} \quad \text{N.B. } \hat{\sigma}_Y = \hat{\sigma}_e$$

4.17 Partial R^2

Another quite useful statistic that we can compute is the partial R^2 . This gives us a measure of how much of the 'variation' is explained by an effect. We compute this R^2 as the Sums of Squares for an effect divided by the Corrected Total Sums of Squares.

$$CTSS = TSS - CF$$

$$= SSR_m + SSE$$

Example:

$$R_{b_1}^2 = SS_{b_1}/CTSS$$

$$= 2.6293/6.6895 = 0.393$$

$$R_{b_2}^2 = SS_{b_2}/CTSS$$

$$= 0.85406/6.6895 = 0.1277$$

5 t-test

see STD, Ch.3, P56 and Ch.14, P333, also STD Ch10.6 P269-271

$$\begin{array}{l} H_o \quad \text{parameter} = \text{constant} \\ \text{vs} \\ H_A \quad \text{parameter} \neq \text{constant} \end{array}$$

which we can re-write in a standard form as

$$\begin{array}{l} H_o \quad \text{parameter} - \text{constant} = 0 \\ \text{vs} \\ H_A \quad \text{parameter} - \text{constant} \neq 0 \end{array}$$

Thus

$$\begin{array}{l} H_o \quad b_1 = -0.5 \\ \text{vs} \\ H_A \quad b_1 \neq -0.5 \end{array}$$

which becomes

$$\begin{array}{l} H_o \quad b_1 - -0.5 = 0 \\ \text{vs} \\ H_A \quad b_1 - -0.5 \neq 0 \end{array}$$

Formally we can consider a t-test to be a statistical test with an appropriate hypothesis to be accepted or rejected.

Test a regression parameter e.g.

$$\frac{\text{parameter} - \text{constant (null hypothesis)}}{s.e. \text{ parameter}}$$

Test a regression parameter

$$\left| \frac{b_i - constant}{s.e.b_i} \right| = \text{t-value}$$

to test whether b_i is significantly different from constant. Compare against the tabulated t-values at a given level of probability, 5%, 1%, 0.1%

e.g. test whether b_1 is significantly different from -0.5.

$$\left| \frac{-0.5285 - -0.5}{0.0969625} \right|$$

$$= \left| \frac{-0.0285}{0.0969625} \right|$$

$$= 0.2941 \text{ n.s.s.}$$

5.1 Testing a group of regression parameters simultaneously

- test if b_1 & b_2 are significant, jointly

$$SS_{12} = (k'\hat{b})'[k'(X'X)^{-1}k]^{-1}(k'\hat{b})$$

$$k' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ n.b. 2 rows}$$

$$k'\hat{b} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} -0.5285 \\ -0.2900 \end{bmatrix}$$

$$C = (X'X)^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}$$

$$k'(X'X)^{-1}k = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

i.e. "pulls out" the subcell of C pertaining to b_1 and b_2

$$= \begin{bmatrix} 0.10623 & -0.02142 \\ -0.02142 & 0.09847 \end{bmatrix}$$

Then compute the inverse, and pre-multiply by $(k'\hat{b})'$, and post-multiply by $(k'\hat{b})$.

This will give us the joint effect of b_1 and b_2 , *i.e.* $R(b_1 \ b_2 \mid b_0) \equiv SSR_m$

$$\begin{bmatrix} 0.10623 & -0.02142 \\ -0.02142 & 0.09847 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 9.84537 & 2.14165 \\ 2.14165 & 10.62125 \end{bmatrix}$$

$$(k'\hat{b})'[k'(X'X)^{-1}k]^{-1}(k'\hat{b})$$

$$= [-0.5285 \ -0.2900] \begin{bmatrix} 9.84537 & 2.14165 \\ 2.14165 & 10.62125 \end{bmatrix} \begin{bmatrix} -0.5285 \\ -0.2900 \end{bmatrix}$$

$$= 4.29985$$

6 Confidence Intervals

6.1 C.I. for a fixed effect estimate

- Parameter \pm standard error

- e.g. b_1 and $s.e.b_1$

- to construct the 95% confidence interval

$$t_{dfe,5\%} = 2.052$$

from S & T Table A.3. Values of t

$$dfe = 27$$

$$\hat{b}_1 - 2.052 * s.e.b_1$$

to

$$\hat{b}_1 + 2.052 * s.e. b_1$$

$$= -0.5285 - 2.052 * 0.0969625$$

to

$$-0.5285 + 2.052 * 0.0969625$$

$$= -0.5285 - 0.19897$$

to

$$-0.5285 + 0.19897$$

$$= -0.72747 \text{ to } -0.32953$$

Thus 95% of the time the real b_1 will lie within our lower and upper limits.

See STD P333

6.2 C.I. for a random effect variance component

See STD Ch 19.2. In STD the example uses a data set on the variability amongst sheep, there are 7 sheep, so $N = 7$, and $N-1$ corresponds to the degrees of freedom for the estimation of the variance component (σ^2). This is correct for this example, it is the degrees of freedom for the variance component. So, in general, we can use this methodology, but bearing in mind that we should use the degrees of freedom for the variance component. Thus, in our example, with tobacco plants, we estimated the residual variance (from our ANOVA table) as 0.0885, with 27 degrees of freedom. Note: the 27 degrees of freedom are **NOT** the residual degrees of freedom, rather they are the degrees of freedom for the variance component, *i.e.* that line of our ANOVA.

Thus we have, d.f. = 27, and $\hat{\sigma}_e^2 = 0.08851$.

$$\text{therefore } \Pr\left(\frac{(d.f.)\hat{\sigma}_e^2}{\chi_{df=27,pr=.025}^2} \leq \sigma^2 \leq \frac{(d.f.)\hat{\sigma}_e^2}{\chi_{df=27,pr=.975}^2}\right) = 0.95$$

We have to obtain the χ^2 value for d.f.=27, for Pr=0.975 and Pr=0.025, these are 14.6 and 43.2 respectively.

$$\text{thus } \frac{27 * 0.08851}{\chi_{df=27,pr=.025}^2} = \frac{2.3897}{43.2} = 0.05532$$

$$\text{and } \frac{27 * 0.08851}{\chi_{df=27,pr=.975}^2} = \frac{2.3897}{14.6} = 0.1637$$

Therefore we can compute a 95% C.I. for our variance component; it is from 0.05532 to 0.1637.

7 Predicted/Estimated/Fitted Values

See STD, Ch 13.5 and 14.8

7.1 Predicting the value of an observation.

The value of any observation (Y_i) is estimable, (\hat{Y}_i)

$$\hat{Y} = X\hat{b}$$

So, to predict the value of Y for sample N°5, $N = 3.52$ and $Cl = 1.10$, take line 5 from X, *i.e.* X_5

$$\hat{Y}_5 = X_5\hat{b}$$

$$\hat{Y}_5 = (1 \ 3.52 \ 1.10) \begin{bmatrix} 2.6531 \\ -0.5285 \\ -0.2900 \end{bmatrix}$$

$$\hat{Y}_5 = 0.47378$$

The sampling variance, $V(X_5\hat{b}) = X_5V(\hat{b})X_5'$, is simply

$$\begin{aligned} & X_5(X'X)^{-1}X_5'\hat{\sigma}_e^2 \\ &= (1 \ 3.52 \ 1.10) \begin{pmatrix} 1.12604 & -0.33098 & -0.00931 \\ -0.33098 & 0.10623 & -0.02142 \\ -0.00931 & -0.02142 & 0.09847 \end{pmatrix} \begin{pmatrix} 1 \\ 3.52 \\ 1.10 \end{pmatrix} * 0.0885 \end{aligned}$$

$$= V(\hat{Y}_5) = s.v. = 0.0039795$$

$$\text{s.e. of estimate, } (\hat{Y}_5) = \sqrt{s.v.}$$

$$= \sqrt{0.0039795}$$

$$\Rightarrow \text{s.e.} = 0.06308$$

$$(\mu_y | X_1 = 3.52 \ X_2 = 1.10)$$

$$= 0.47378 \pm 2.052 * 0.06308$$

$$= 0.47378 \pm 0.1294471$$

n.b. $t_{5\%}$ for 27 d.f. from Table A.3 (STD, Page 611)

7.2 Using SAS PROC GLM+IML to estimate a fitted value

```
proc glm data=reg1;
model y = x1 x2/XPX I;
output out=reg1out p=yhat r=ehat stdp=se;
/* output to a new SAS data set (reg1out), y x1 x2 yhat ehat se */
estimate 'obs 1' intercept 1 x1 3.05 x2 1.45;
estimate 'obs 5' intercept 1 x1 3.52 x2 1.10;
run;
```

USING SAS/PROC IML

previous IML matrices and code here

```
x5 = {1 3.52 1.10};
y5 = x5 * b;
sv = x5 * invvtx * x5' * mse;
se = sqrt(sv);
```

7.3 Predicting the value of some future observation

i.e. one not in the data set, see STD, Ch. 13.5 and Ch.14.8.

e.g. $N\% = 3.11$ $CI\% = 1.0$

The formula for estimating the Y value is the same

$$\hat{Y}_0 = X_0 \hat{b}$$

$$\hat{Y} = (1 \ 3.11 \ 1.0) \begin{pmatrix} 2.6531 \\ -0.5285 \\ -0.2900 \end{pmatrix}$$

= 0.719465

However, the sampling variance must reflect the fact that there is a sampling variance due to our prediction, and another due to the random error associated with any new observation.

Thus sampling variance =

$$[1 + X_0(X'X)^{-1}X_0']\hat{\sigma}_e^2$$

Compute the fitted value for each observation, calculate the sampling variance for each observation and hence the 95 % confidence limits for each observation. Plot!

7.4 Using SAS PROC GLM+IML to predict a future value

```
proc glm data=reg1;
model y = x1 x2/XPX I;
output out=reg1out p=yhat r=ehat stdp=se;
/* output to a new SAS data set (reg1out), y x1 x2 yhat ehat se */
estimate 'obs new' intercept 1 x1 3.11 x2 1.0;
run;
```

USING SAS/PROC IML

previous IML matrices and code here

```
xnew = {1 3.11 1.0};
ynew = xnew * b;
sv = (1 + xnew * invxtx * xnew') * mse;
se = sqrt(sv);
```

8 Linear and Quadratic Regressions

The relationship between Y and some or all, of the X 's may not be linear, but quadratic in nature, with an intermediate optimum.

This can be investigated in exactly the same manner as previously, except that we shall add a column for the X^2 terms. For example, if we wish to see whether there is a quadratic effect of $N\%$ on leaf burn (returning to our example data from STD, P333) then we might propose a new model;

8.1 Linear Model

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + e$$

Thus

$$Y = Xb + e$$

8.2 Matrix Equations

$$\begin{bmatrix} 0.34 \\ 0.11 \\ \cdot \\ \cdot \\ \cdot \\ 0.23 \end{bmatrix} = \begin{bmatrix} 1 & 3.05 & 1.45 & 9.3025 \\ 1 & 4.22 & 1.35 & 17.8084 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 2.94 & 2.22 & 8.6436 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_{30} \end{bmatrix}$$

and proceed as before

N.B. Fit linear and quadratic terms and test the significance of the quadratic component, e.g. Fit $N\%$ and $N\%^2$. Test $R(b_3 \mid b_0, b_1, b_2)$.

If $N\%^2$ (quadratic) is significant then retain both the linear and quadratic components, since the covariance between the estimates is not likely to be zero.

If $N\%$ (quadratic) is not significant then it can be dropped from the model.

$$X'X = \begin{bmatrix} 30 & 98.36 & 24.23 & 332.3352 \\ 98.36 & 332.3352 & 81.5834 & 1156.3337 \\ 24.23 & 81.5834 & 30.1907 & 282.13909 \\ 332.3352 & 1156.3337 & 282.13909 & 4137.7125 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 20.58 \\ 61.6502 \\ 12.4103 \\ 189.26515 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 32.172314 & -19.2762 & 0.2069003 & 2.7888251 \\ -19.2762 & 11.667078 & -0.153356 & -1.701812 \\ 0.2069003 & -0.153356 & 0.0999773 & 0.0194221 \\ 2.7888251 & -1.701812 & 0.0194221 & 0.2505146 \end{bmatrix}$$

$$\hat{b} = (X'X)^{-1}X'Y = \begin{bmatrix} 4.1195217 \\ -1.423378 \\ -0.279752 \\ 0.1317229 \end{bmatrix}$$

$$\text{TSS} = Y'Y = 20.8074$$

$$\text{SSR (Model)} = \hat{b}'X'Y$$

$$= (4.1195 \ -1.42338 \ -0.27975 \ 0.131723) \begin{bmatrix} 20.58 \\ 51.6502 \\ 12.4103 \\ 189.26515 \end{bmatrix}$$

$$= 18.486994$$

$$\begin{aligned} \text{SSE} &= \text{TSS} - \text{SSR} \\ &= 20.8074 - 18.486994 \\ &= 2.3204 \end{aligned}$$

$$\begin{aligned}
SSR_m &= SSR - CF \\
&= 18.486994 - 14.11788 \\
&= 4.3691136 \\
MSR &= SSR/4 \\
&= 18.486994/4 \\
&= 4.6217 \\
MSR_m &= SSR_m/3 \\
&= 4.3691136/3 \\
&= 1.4564 \\
MSE &= \hat{\sigma}_e^2 \\
&= 2.3204/(30 - 4) \\
&= 2.3204/26 \\
&= 0.0892
\end{aligned}$$

8.3 Testing the quadratic effect

$$SS = (k'\hat{b})'[k'(X'X)^{-1}k]^{-1}(k'\hat{b})$$

$$k' = [0 \ 0 \ 0 \ 1]$$

$$k'\hat{b} = [0 \ 0 \ 0 \ 1] \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{bmatrix} = [\hat{b}_3] = [0.1317229]$$

The effect (marginal, over and above other effects) of the quadratic is: .06926

Using our standard, ubiquitous formula for the Sums of Squares of an effect,

$$(k'\hat{b})'[k'(X'X)^{-1}k]^{-1}(k'\hat{b})$$

we have:

$$\begin{aligned}
k'(X'X)^{-1}k &= 0.2505146 \\
(k'\hat{b})'[k'(X'X)^{-1}k]^{-1}(k'\hat{b}) &= .131723 * [.2505146]^{-1} * .131723 \\
&= .06926
\end{aligned}$$

Thus we can conclude that there is no quadratic effect of N% on log leaf burn.

Therefore the term $b_3 (N^2)$ can be dropped from the model.

8.4 Using SAS PROC GLM to fit a quadratic

```

proc glm data=reg1;
model y = x1 x2 x1*x1/XPX I;
run;

```

Table 4: ANOVA, including a quadratic

ANOVA <i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F - ratio</i>	<i>E(M.S.)</i>
Total, TSS	$N = 30$	$Y'Y$ 20.8074			
Model, SSR	$r(X)$ $= 4$	$\hat{b}'X'Y$ 18.486994	4.6217		
Mean, C.F.	1	$N\bar{y}^2$ 14.11788	14.11788	158.2722	
Model, after the mean, SSR_m $R(b_1b_2b_3 \text{Mean})$	$r(X) - 1$ $= 3$	$\hat{b}'X'Y - N\bar{y}^2$ 4.3691136	1.4564	16.327	
SS_{b_1} $R(b_1 b_0b_2b_3)$	1	$\hat{b}_1^2 C_{11}^{-1}$			
SS_{b_2} $R(b_2 b_0b_1b_3)$	1	$\hat{b}_2^2 C_{22}^{-1}$			
SS_{b_3} $R(b_3 b_0b_1b_2)$	1	$\hat{b}_3^2 C_{33}^{-1}$	0.06926	0.776 ^{n.s.s.}	$\sigma_e^2 + f(b_3^2)$
Error, SSE	$N - r(X)$ $30 - 4$	$Y'Y - \hat{b}'X'Y$ 2.3204	.0892		σ_e^2

9 Interactions amongst regression effects

There may be interactions between the regression effects, that is to say, the effect of one factor may not be independent of another regression effect; so far our assumed model has **ASSUMED** that they are independent.

This can be investigated in exactly the same manner as previously, except that we shall add a column for the interaction term. For example, if we wish to see whether there is an interaction between N% and Cl% on leaf burn (returning to our example data from STD, P333) then we might propose a new model;

9.1 Linear Model

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_1X_2 + e$$

We would add a column for $X_1 * X_2$ and proceed with our analysis exactly as before. We would use exactly the same logic as for a quadratic; that is to say, that with our (two-way) interaction between X_1 and X_2 we would have the linear terms (for X_1 and X_2) in our model as well. We would first of all test the interaction between X_1 and X_2 to determine whether it needs to be retained in the model. Much as for a quadratic, if it does then the linear, lower-order effects of X_1 and X_2 must remain in our model, and testing their statistical significance is pointless. If on the otherhand the interaction term is not statistically significant, then we should drop it from the model, re-run the analysis and then test the linear effects of X_1 and X_2 .

We can have more than one interaction term in the model, we might have $X_1 * X_2$, and $X_2 * X_3$, etc. We can even have a 3-way interaction (if we have also included all the possible 2-way interactions). We can also have quadratics as well as interactions between the linear effects all in the same model!

9.2 Using SAS PROC GLM to fit an interaction

```
proc glm data=reg1;  
model y = x1 x2 x1*x2/XPX I;  
run;
```

10 Correlations

Only when both (all) variables are random, separate, in the sense of being different traits, and in the sense of coming from experimental units which are independent of one another, and where the residual errors are normally distributed, *i.e.* bivariate normal, and where there are no fixed effects affecting either of the traits.

i.e. where we can describe each observation as $Y_{ji} = \mu_j + e_{ji}$

where Y_{ji} = the observation on the i^{th} experimental unit for trait j

μ_j = overall mean for trait j

and e_{ji} = the random effect of the i^{th} experimental unit for trait j

10.1 Variance-Covariance matrix

Consider that we have measurements on 3 traits X_1 , X_2 and X_3 .

Table 5: Example data for correlations

Obs	X_1	X_2	X_3
1	.	.	.
2	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
N	.	.	.

Then, using the conventional simple formulae for variances and covariances:

$$V(X_1) = \sigma_{X_1}^2 = \frac{\sum_{i=1}^{i=N} X_{1i}^2 - \frac{(\sum_{i=1}^{i=N} X_{1i})^2}{n}}{n-1}$$

$$V(X_2) = \sigma_{X_2}^2 = \frac{\sum_{i=1}^{i=N} X_{2i}^2 - \frac{(\sum_{i=1}^{i=N} X_{2i})^2}{n}}{n-1}$$

$$cov(X_1 X_2) = \sigma_{X_1 X_2} = \frac{\sum_{i=1}^{i=N} X_{1i} X_{2i} - \frac{\sum_{i=1}^{i=N} X_{1i} \sum_{i=1}^{i=N} X_{2i}}{n}}{n-1}$$

etc.

Thus:

$$\begin{bmatrix} \frac{\sum X_{1i}^2 - (\sum X_{1i})^2}{n-1} & \frac{\sum X_{1i} X_{2i} - \frac{\sum X_{1i} \sum X_{2i}}{n}}{n-1} & \frac{\sum X_{1i} X_{3i} - \frac{\sum X_{1i} \sum X_{3i}}{n}}{n-1} \\ \frac{\sum X_{2i} X_{1i} - \frac{\sum X_{2i} \sum X_{1i}}{n}}{n-1} & \frac{\sum X_{2i}^2 - (\sum X_{2i})^2}{n-1} & \frac{\sum X_{2i} X_{3i} - \frac{\sum X_{2i} \sum X_{3i}}{n}}{n-1} \\ \frac{\sum X_{3i} X_{1i} - \frac{\sum X_{3i} \sum X_{1i}}{n}}{n-1} & \frac{\sum X_{3i} X_{2i} - \frac{\sum X_{3i} \sum X_{2i}}{n}}{n-1} & \frac{\sum X_{3i}^2 - (\sum X_{3i})^2}{n-1} \end{bmatrix}$$

$$\equiv \begin{pmatrix} V(X_1) & cov(X_1 X_2) & cov(X_1 X_3) \\ cov(X_1 X_2) & V(X_2) & cov(X_2 X_3) \\ cov(X_1 X_3) & cov(X_2 X_3) & V(X_3) \end{pmatrix}$$

We can then scale these to unit variance to give simple correlations.

10.2 Simple correlations

$$\equiv R = \begin{bmatrix} 1 & \frac{\text{cov}(X_1 X_2)}{\sqrt{V(X_1)V(X_2)}} & \frac{\text{cov}(X_1 X_3)}{\sqrt{V(X_1)V(X_3)}} \\ \frac{\text{cov}(X_1 X_2)}{\sqrt{V(X_1)V(X_2)}} & 1 & \frac{\text{cov}(X_2 X_3)}{\sqrt{V(X_2)V(X_3)}} \\ \frac{\text{cov}(X_1 X_3)}{\sqrt{V(X_1)V(X_3)}} & \frac{\text{cov}(X_2 X_3)}{\sqrt{V(X_2)V(X_3)}} & 1 \end{bmatrix}$$

$$\equiv R = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{bmatrix}$$

10.3 Partial correlations

$$R^{-1} = C = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}$$

Then the partial correlation between two variables i and j , adjusting for the other variables in our variance-covariance matrix, is

$$r_{ij} = \frac{-C_{ij}}{\sqrt{C_{ii}C_{jj}}}$$

e.g. partial correlation of 1 and 3, adjusting for 2, is

$$r_{13|2} = \frac{-C_{13}}{\sqrt{C_{11}C_{33}}}$$

10.4 Partial correlations, adjusting for only some variables

Suppose that we measure 5 traits and compute the matrix of (simple) correlations amongst them:

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{12} & 1 & r_{23} & r_{24} & r_{25} \\ r_{13} & r_{23} & 1 & r_{34} & r_{35} \\ r_{14} & r_{24} & r_{34} & 1 & r_{45} \\ r_{15} & r_{25} & r_{35} & r_{45} & 1 \end{bmatrix}$$

If we are interested in the partial correlation between 1 and 3, adjusting for 2 and 5, then we simply 'pull out' the correlations amongst 1, 2, 3 and 5, and then proceed:

$$R_{1235} = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{15} \\ r_{12} & 1 & r_{23} & r_{25} \\ r_{13} & r_{23} & 1 & r_{35} \\ r_{15} & r_{25} & r_{35} & 1 \end{bmatrix}$$
$$R^{-1} = C = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{15} \\ C_{21} & C_{22} & C_{23} & C_{25} \\ C_{31} & C_{32} & C_{33} & C_{35} \\ C_{51} & C_{52} & C_{53} & C_{55} \end{bmatrix}$$

10.5 Statistical significance of an estimate of a correlation coefficient

Test significance via a t-test (see Stats I notes of Hypothesis Testing of a correlation).

$$t = \frac{r\sqrt{N-k}}{\sqrt{1-r^2}}$$

but only for testing the Null Hypothesis (H_0) that the real, true parameter $\rho = 0$

where r = correlation (estimate)
 N = number of observations
 k = number of parameters

so for a simple correlation r_{12} $k = 2$

and for a partial correlation $r_{13|2}$ $k = 3$

and for a partial correlation $r_{13|24}$ $k = 4$

and we would compare this calculated t value against a tabulated t value with d.f. -
 $(N-k)$

A numerical example e.g. from STD, P 337

The variables are X1, X2 and Y respectively

$$R = \begin{bmatrix} 1 & .20940 & -.717729 \\ .20940 & 1 & -.499638 \\ -.717729 & -.499638 & 1 \end{bmatrix}$$

$$R^{-1} = C = \begin{bmatrix} 2.1968526 & .4368302 & 1.7950017 \\ .4368302 & 1.4195512 & 1.0227874 \\ 1.7950017 & 1.0227874 & 2.7993483 \end{bmatrix}$$

$$r_{y1|2} = \frac{-1.7950017}{\sqrt{2.1968526 * 2.7993483}} = -.723829$$

Compare $r_{y1} = -.717729$

with $r_{y1|2} = -.723829$

not much difference, here, but it can be very substantial, possibly even changing sign.

e.g. perhaps $r_{12} = + 0.56$ and $r_{12} \dots = -0.4!!!$

10.6 Sampling Distribution of an estimate of a correlation coefficient

STD, Chapter 4, Ch. 11.4, P292; Stats I notes

Fisher's (1921) transformation is

$$z = .5 \ln \left(\frac{1+r}{1-r} \right) \quad (21)$$

$$\text{standard deviation} = \frac{1}{\sqrt{n-3}} \quad (22)$$

which has a normal distribution, not a t distribution => use ∞ d.f. We can use a table of the t values with an infinite number of degrees of freedom (because the t-distribution tends to the normal as the sample size increases), or we can use a table of normal distribution values.

See STD, Page 292-293 for an example and for more details about correlations and their statistical significance.

e.g. correlation of % resin and % rubber was 0.527 from 50 plants

$$n - 3 = 47$$

Using (21) we calculate the z value as

$$z = .5 \ln \left[\frac{1+r}{1-r} \right]$$

$$= .5 \ln \left[\frac{1.527}{.473} \right]$$

$$= .5 \ln[3.23] = .5 * 1.172$$

$$= .586$$

Thus, using (22), we can obtain the standard deviation of z

$$\sigma = \sqrt{\frac{1}{47}} = .146$$

10.7 Statistical Significance

To test the statistical significance we use the estimate divided by its standard deviation and compare this to the NORMAL distribution, not a t-distribution.

$$\frac{.586}{.146} = 4.01$$

Compare this value (4.01) against the tabulated cutoff values of the Normal distribution tables to determine the probability. We find that it is statistically significant at 1%.

10.8 Confidence Interval

Confidence Interval, CI ($z \pm K \text{ s.d.}_z$) (n.b. on the transformed scale), where K is the appropriate cutoff values from the Normal distribution tables. Note, a 5% probability corresponds to a cutoff of 1.96 standard deviations. Therefore to determine a 95% Confidence Interval:

$$\begin{aligned} &= .586 \pm 1.96 * (.146) \\ &= .300 \quad \text{to} \quad .872 \end{aligned}$$

How can we 'back-transform' from our z scale to the 'observed' scale? We need to find the inverse function of (21). It is

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \tag{23}$$

If we therefore back transform .300 by substituting in (23)

$$= 0.290$$

and similarly backing transform .872

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$
$$= 0.703$$

estimate is 0.527

95% C.I. is 0.290 to 0.703

- Non-symmetric

Note, if we want to compare our correlation estimate with a value other than Zero, we would need to transform (using the z-transformation) this hypothesized value too.

10.9 Using SAS CORR

```
data reg1;
input x1 x2 x3 y;
cards;
3.05 1.45 5.67 0.34
4.22 1.35 4.86 0.11

.
.
.
2.94 2.22 4.58 0.23
;

proc corr cov;
/* the cov option to proc corr provides the variances and covariances
too */
var x1 x2 y x3;
run;

/* produce the correlations amongst x1 y and x3 adjusting for x2,
i.e. the partial correlations adjusted for x2 */
proc corr cov;
var x1 y x3;
partial x2;
run;

/* produce the correlation between x1 and y adjusting for x2 and x3,
i.e. the partial correlation adjusted for x2 and x3 */
proc corr cov;
var x1 y;
partial x2 x3;
run;
```

Compute the simple correlations amongst X1, X2, X3 and Y.

Compute the partial correlations and confidence intervals.

10.10 Correlations accounting for the effects of fixed effects

If the original requirement (of no fixed effects) is not met what can we do? Well, we could analyse each trait separately, output the residuals, match up the residuals (for each trait) from each experimental unit and then compute the correlation. Or, we could use a multivariate model (see the appropriate section on MyCourses).

An example will demonstrate the problem. We carry out an experiment to look at weight gain (X_6) and two blood parameters (X_1, X_2) in a group of humans who were fed various levels of energy (X_4) and protein (X_5). We have hypothesised that the energy and protein will/may affect the individuals' weightgain and blood parameters. We obtain the following data:

```
data coleman;
input pid x1 x2 x3 x4 x5 x6;
cards;
1 3.83 28.87 7.20 26.60 6.19 37.01
2 2.89 20.10 -11.71 24.40 5.17 26.51
3 2.86 69.05 12.32 25.70 7.04 36.51
4 2.92 65.40 14.28 25.70 7.10 40.70
5 3.06 29.59 6.31 25.40 6.15 37.10
6 2.07 44.82 6.16 21.60 6.41 33.90
7 2.52 77.37 12.70 24.90 6.86 41.80
8 2.45 24.67 -0.17 25.01 5.78 33.40
9 3.13 65.01 9.85 26.50 6.51 41.01
10 2.44 9.99 -0.05 28.01 5.57 37.20
11 2.09 12.20 -12.86 23.51 5.62 23.30
12 2.52 22.55 0.92 23.60 5.34 35.20
13 2.22 14.30 4.77 24.51 5.80 34.90
14 2.67 31.79 -0.96 25.80 6.19 33.10
15 2.71 11.60 -16.04 25.20 5.62 22.70
16 3.14 68.47 10.62 25.01 6.94 39.70
17 3.54 42.64 2.66 25.01 6.33 31.80
18 2.52 16.70 -10.99 24.80 6.01 31.70
19 2.68 86.27 15.03 25.51 7.51 43.10
```

Table 6: Correlations when fixed effects are present

Pid	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆
1	3.83	28.87	7.20	26.60	6.19	37.01
2	2.89	20.10	-11.71	24.40	5.17	26.51
3	2.86	69.05	12.32	25.70	7.04	36.51
4	2.92	65.40	14.28	25.70	7.10	40.70
5	3.06	29.59	6.31	25.40	6.15	37.10
6	2.07	44.82	6.16	21.60	6.41	33.90
7	2.52	77.37	12.70	24.90	6.86	41.80
8	2.45	24.67	-0.17	25.01	5.78	33.40
9	3.13	65.01	9.85	26.50	6.51	41.01
10	2.44	9.99	-0.05	28.01	5.57	37.20
11	2.09	12.20	-12.86	23.51	5.62	23.30
12	2.52	22.55	0.92	23.60	5.34	35.20
13	2.22	14.30	4.77	24.51	5.80	34.90
14	2.67	31.79	-0.96	25.80	6.19	33.10
15	2.71	11.60	-16.04	25.20	5.62	22.70
16	3.14	68.47	10.62	25.01	6.94	39.70
17	3.54	42.64	2.66	25.01	6.33	31.80
18	2.52	16.70	-10.99	24.80	6.01	31.70
19	2.68	86.27	15.03	25.51	7.51	43.10
20	2.37	76.73	12.77	24.51	6.96	41.01


```

20 2.37 76.73 12.77 24.51 6.96 41.01
;

/* analyse each trait seperately, outputting residuals, use id = pid
   so we can be certain of matching up the residuals from the same
   patients
*/
proc glm data=coleman;
id pid;
model x6 = x4 x5;
output out=residx6 p=yhatx6 r=ehatx6;
run;

proc glm data=coleman;
id pid;
model x1 = x4 x5;
output out=residx1 p=yhatx1 r=ehatx1;
run;

proc glm data=coleman;
id pid;
model x2 = x4 x5;
output out=residx2 p=yhatx2 r=ehatx2;
run;

proc print data=residx6;
run;

/* sort by Patient id (pid), to be absolutely certain
   that we can merge correctly
*/
proc sort data=residx6;
by pid;
run;

proc sort data=residx1;
by pid;
run;

proc sort data=residx2;
by pid;
run;

```

```

/* merge residuals from x6 and x2, by pid */
data resids61;
  merge residx6 residx1;
  by pid;
run;

/* merge with residuals from x2 */
data resids;
  merge resids61 residx2;
  by pid;
run;

proc print data=resids;
run;

proc print data=resids;
var pid ehatx6 ehatx1 ehatx2;
run;

proc corr data=resids;
  var ehatx6 ehatx1 ehatx2;
run;

/* multivariate GLM, only suitable if each obs is
   independent, if all independent variables are
   fixed effects, and if the same independent
   variables are a suitable model for each
   dependent variable, AND if there are no missing
   values for any of the observations
*/
proc glm data=coleman;
id pid;
model x6 x1 x2 = x4 x5;
output out=residx612 p=yhatx6 yhatx1 yhatx2 r=ehatx6 ehatx1 ehatx2;
manova /printe;
run;

proc print data=residx612;
var pid ehatx6 ehatx1 ehatx2;
run;

/* (much) more sophisticated model using proc mixed
   in multivariate mode, however this is only suitable

```

when the only random effect in a model is the
observational unit

```
*/  
data coleman1;  
set coleman;  
  y = x6;  
  trait = 6;  
  output;  
  y = x1;  
  trait = 1;  
  output;  
  y = x2;  
  trait = 2;  
  output;  
run;  
  
proc print data=coleman1;  
run;  
  
data coleman2;  
  set coleman1;  
  keep pid trait y x4 x5;  
run;  
  
proc print data=coleman2;  
var pid trait y x4 x5;  
run;  
  
/* data layout will look like :  
  
  pid trait  y      x4      x5  
  1   6      37.01  26.60  6.19  
  1   1       3.83  26.60  6.19  
  1   2      28.87  26.60  6.19  
  2   6      26.51  24.60  5.17  
  2   1       2.89  24.60  5.17  
  2   2      20.10  24.60  5.17  
  3   6      36.51  25.70  7.04  
  etc  
*/  
  
proc mixed data=coleman2;  
class pid trait;
```

```

model y = trait trait*x4 trait*x5;
repeated trait/type=un subject=pid rcorr;
run;

/* INCORRECT correlations, because we have ignored the
   fixed effects of X4 and X5
*/
proc corr data=coleman;
var x6 x1 x2;
run;

```

Results.

Good, from the residuals, or the proc GLM multivariate ANOVA (MANOVA option), or the proc mixed multivariate mode. NOTE, the layout of the traits is 6, 1, 2

$$\begin{bmatrix} 1 & -0.1266 & 0.3844 \\ . & 1 & 0.0979 \\ . & . & 1 \end{bmatrix}$$

Bad, from the simple corr

$$\begin{bmatrix} 1 & 0.1930 & 0.7534 \\ . & 1 & 0.1811 \\ . & . & 1 \end{bmatrix}$$

11 1-Way Classification. STD. Ch.7

- 1-way classification, completely random design (CRD)
- allocation of experimental units to treatments must be completely at random
- or, the classifications must be 'levels'/groupings which are mutually exclusive, e.g. sex (Male, Female, Castrate), and then within each of these groups the experimental units are able to be considered as a random sample of such.

Nitrogen content of red clover plants, or different diets being fed to dairy cows.

The random assignment to treatments, or the experimental units within each group being a true random sample **IS** a **CRITICAL** assumption; ignore it at your peril!

5 inoculants + a composite = 6 treatments

Table 7: Rhizobium example

3Dok1 Diet 1	3Dok5 Diet 2	3Dok4 Diet 3	3Dok7 Diet 4	3Dok13 Diet 5	composite Diet 6
19.4	17.7	17.0	20.7	14.3	17.3
32.6	24.8	19.4	21.0	14.4	19.4
27.0	27.9	9.1	20.5	11.8	19.1
32.1	25.2	11.9	18.8	11.6	16.9
33.0	.	.	18.6	14.2	20.8

11.1 Linear Model

We can write the linear model as :

$$Y_{ij} = \mu + trt_i + plot_{ij} + \epsilon_{ij}$$

However, we have only 1 measurement on each of the classification effect experimental units (plots). Thus we CANNOT separate the effect of $plot_{ij}$ from ϵ_{ij} , and they will be both combined into a term $e_{ij} = plot_{ij} + \epsilon_{ij}$. Note also that we can, and indeed should, consider that plot (or whatever is the experimental unit) is nested within treatment. This is an important point as it should aid you in understanding when an effect is considered 'nested' in more complicated models!

$$\sigma_e^2 = \sigma_{plot}^2 + \sigma_\epsilon^2$$

$$Y_{ij} = \mu + trt_i + e_{ij}$$

11.2 Parameters of the Model

$$\mu, trt_1, trt_2, trt_3, trt_4, trt_5, trt_6, \sigma_e^2$$

11.3 Hypotheses to be tested

The first hypothesis to test is, as per our test in the multiple regression models, whether the model explains variation in the dependent variable. Our Null Hypothesis (H_o) will be that the Model does not explain variation in Y (our dependent variable), and our Alternate Hypothesis (H_A) will be that the Model does explain variation in Y. Thus:

$$H_o \quad \begin{bmatrix} \mu \\ trt_1 \\ trt_2 \\ trt_3 \\ trt_4 \\ trt_5 \\ trt_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Note: although there are 7 lines to this H_o the rank is 6, 1 for the mean and 5 amongst the 6 treatments (see below). We could also write the hypothesis about the model as:

$$H_o \begin{bmatrix} \mu + trt_1 \\ \mu + trt_2 \\ \mu + trt_3 \\ \mu + trt_4 \\ \mu + trt_5 \\ \mu + trt_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The next hypothesis is about the Mean; as before it is: $H_o: \bar{Y} = 0$, and $H_A: \bar{Y} \neq 0$

Continuing our subdivision of the source of variation, we have the Model over and above the Mean. An obvious hypothesis to be tested that flows directly and immediately from the very reason for the experiment is to test whether there are differences between the treatments (over and above the Mean). This we can describe (in words) in the form of a 'Null Hypothesis' that the treatments are all equal, vs an 'Alternative Hypothesis' that the treatments are not all equal; *i.e.*

$$H_o : trt_1 = trt_2 = trt_3 = trt_4 = trt_5 = trt_6$$

H_A : treatments are not all equal

The Null Hypothesis we can re-write as a series of comparisons:

6 Treatments, 5 separate comparisons

i) $trt_1 = trt_2$

ii) $trt_1 = trt_3$

iii) $trt_1 = trt_4$

iv) $trt_1 = trt_5$

v) $trt_1 = trt_6$

which we can re-write as a series of comparisons with Null Hypotheses of Zero:

- i) $\text{trt}_1 - \text{trt}_2 = 0$
- ii) $\text{trt}_1 - \text{trt}_3 = 0$
- iii) $\text{trt}_1 - \text{trt}_4 = 0$
- iv) $\text{trt}_1 - \text{trt}_5 = 0$
- v) $\text{trt}_1 - \text{trt}_6 = 0$

Which we can express statistically (as one hypothesis) as:

$$H_o \begin{bmatrix} \text{trt}_1 - \text{trt}_2 \\ \text{trt}_1 - \text{trt}_3 \\ \text{trt}_1 - \text{trt}_4 \\ \text{trt}_1 - \text{trt}_5 \\ \text{trt}_1 - \text{trt}_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and

$$H_A \begin{bmatrix} \text{trt}_1 - \text{trt}_2 \\ \text{trt}_1 - \text{trt}_3 \\ \text{trt}_1 - \text{trt}_4 \\ \text{trt}_1 - \text{trt}_5 \\ \text{trt}_1 - \text{trt}_6 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The Null Hypothesis we can also re-write as another series of comparisons:

e.g. 6 Treatments, 5 separate comparisons

- i) $\text{trt}_6 = \text{trt}_1$
- ii) $\text{trt}_6 = \text{trt}_2$
- iii) $\text{trt}_6 = \text{trt}_3$
- iv) $\text{trt}_6 = \text{trt}_4$
- v) $\text{trt}_6 = \text{trt}_5$

which we can re-write as a series of comparisons with Null Hypotheses of Zero:

- i) $\text{trt}_6 - \text{trt}_1 = 0$
- ii) $\text{trt}_6 - \text{trt}_2 = 0$
- iii) $\text{trt}_6 - \text{trt}_3 = 0$
- iv) $\text{trt}_6 - \text{trt}_4 = 0$
- v) $\text{trt}_6 - \text{trt}_5 = 0$

Which we can express statistically as:

$$H_o \begin{bmatrix} \text{trt}_6 - \text{trt}_1 \\ \text{trt}_6 - \text{trt}_2 \\ \text{trt}_6 - \text{trt}_3 \\ \text{trt}_6 - \text{trt}_4 \\ \text{trt}_6 - \text{trt}_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and

$$H_A \begin{bmatrix} \text{trt}_6 - \text{trt}_1 \\ \text{trt}_6 - \text{trt}_2 \\ \text{trt}_6 - \text{trt}_3 \\ \text{trt}_6 - \text{trt}_4 \\ \text{trt}_6 - \text{trt}_5 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Note: these are equivalent, and they lead to **EXACTLY** the same test of significance and Sums of Squares.

More completely, we have a hypothesis for the Model, H_o that the model does not explain variation in Y , *i.e.* $R(\mu, \text{trt})$ is not statistically significant, versus the Alternative Hypothesis (H_A) that the Model does explain variation in Y , *i.e.* that the Reduction Sums of Squares due to the Model $R(\mu, \text{trt})$ is statistically significant. We have a hypothesis about the Correction Factor for the mean: our H_o is that the mean of Y equals Zero (note Not μ), and the H_A is that $\bar{Y} \neq 0$. Our Null Hypothesis for the Model over and above the Mean, $R(\text{trt} \mid \mu)$, is that there are no differences amongst the levels of treatment in their

effect on Y . This is what we have described above in our hypothesis for treatments, since treatments is the only effect in the model over and above the mean, so in this case they are synonymous.

11.4 Matrix Equations

Any observation can be written

$$Y_{11} = \mu + trt_1 + e_{11}$$

$$Y_{12} = \mu + trt_1 + e_{12}$$

$$Y_{13} = \mu + trt_1 + e_{13}$$

$$Y_{14} = \mu + trt_1 + e_{14}$$

$$Y_{15} = \mu + trt_1 + e_{15}$$

$$Y_{21} = \mu + trt_2 + e_{21}$$

$$Y_{22} = \mu + trt_2 + e_{22}$$

$$Y_{25} = \mu + trt_2 + e_{25}$$

$$Y_{61} = \mu + trt_6 + e_{61}$$

etc

We can then re-write these as:

$$Y_{11} = \mu + 1trt_1 + 0trt_2 + 0trt_3 + 0trt_4 + 0trt_5 + 0trt_6 + e_{11}$$

$$Y_{12} = \mu + 1trt_1 + 0trt_2 + 0trt_3 + 0trt_4 + 0trt_5 + 0trt_6 + e_{12}$$

$$Y_{13} = \mu + 1trt_1 + 0trt_2 + 0trt_3 + 0trt_4 + 0trt_5 + 0trt_6 + e_{13}$$

$$Y_{21} = \mu + 0trt_1 + 1trt_2 + 0trt_3 + 0trt_4 + 0trt_5 + 0trt_6 + e_{21}$$

$$Y_{22} = \mu + 0trt_1 + 1trt_2 + 0trt_3 + 0trt_4 + 0trt_5 + 0trt_6 + e_{22}$$

$$Y_{25} = \mu + 0trt_1 + 1trt_2 + 0trt_3 + 0trt_4 + 0trt_5 + 0trt_6 + e_{25}$$

$$Y_{31} = \mu + 0trt_1 + 0trt_2 + 1trt_3 + 0trt_4 + 0trt_5 + 0trt_6 + e_{31}$$

$$Y_{32} = \mu + 0trt_1 + 0trt_2 + 1trt_3 + 0trt_4 + 0trt_5 + 0trt_6 + e_{32}$$

$$Y_{61} = \mu + 0trt_1 + 0trt_2 + 0trt_3 + 0trt_4 + 0trt_5 + 1trt_6 + e_{61}$$

We can then write these equations in a matrix notation, much as before:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ \vdots \\ Y_{25} \\ \vdots \\ \vdots \\ Y_{65} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \\ \vdots & & & & & & \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \\ \vdots & & & & & & \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ trt_1 \\ trt_2 \\ trt_3 \\ trt_4 \\ trt_5 \\ trt_6 \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ \vdots \\ e_{25} \\ \vdots \\ \vdots \\ e_{65} \end{bmatrix}$$

- a "design" matrix, X

11.5 Example, adapted from STD, Ch.7

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ \vdots \\ Y_{21} \\ \vdots \\ \vdots \\ Y_{54} \\ \vdots \\ \vdots \\ Y_{65} \end{bmatrix} = \begin{bmatrix} 19.4 \\ 32.6 \\ \vdots \\ \vdots \\ 17.7 \\ \vdots \\ \vdots \\ 11.6 \\ \vdots \\ \vdots \\ 20.8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \\ \vdots & & & & & & \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & \\ \vdots & & & & & & \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ \vdots & & & & & & \\ \vdots & & & & & & \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ trt_1 \\ trt_2 \\ trt_3 \\ trt_4 \\ trt_5 \\ trt_6 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ \vdots \\ e_{21} \\ \vdots \\ \vdots \\ e_{54} \\ \vdots \\ \vdots \\ e_{65} \end{bmatrix}$$

$$Y = Xb + e$$

11.6 The Normal Equations

thus the Normal Equations are :

$$X'X\tilde{b} = X'Y$$

$$\begin{bmatrix} 28 & 5 & 4 & 4 & 5 & 5 & 5 \\ 5 & 5 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 4 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 5 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 5 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} \mu \\ trt_1 \\ trt_2 \\ trt_3 \\ trt_4 \\ trt_5 \\ trt_6 \end{bmatrix} = \begin{bmatrix} 556.5 \\ 144.1 \\ 95.6 \\ 57.4 \\ 99.6 \\ 66.3 \\ 93.5 \end{bmatrix}$$

However, there is no unique inverse, since the 6 (trt_i) treatment columns (2 to 7) add to the mean μ (the first column). The matrix X has 7 columns, but the rank of X , $r(X)$, is only 6. this means that the rank of $X'X$ is also 6. Therefore it does not have a unique inverse, the determinant is zero and the matrix is singular.

Thus we use a generalised inverse, G .

$$G = (X'X)^-$$

$$\tilde{b} = GX'Y$$

n.b. \tilde{b} is a solution.

11.7 Generalised inverses from GLM and IML

The generalised inverse produce by GLM $(X'X)^-$, together with the solution vector \tilde{b} is shown below:

$$(X'X)^- = \begin{bmatrix} 0.2 & -0.2 & -0.2 & -0.2 & -0.2 & -0.2 & 0 \\ -0.2 & 0.4 & 0.2 & 0.2 & 0.2 & 0.2 & 0 \\ -0.2 & 0.2 & 0.45 & 0.2 & 0.2 & 0.2 & 0 \\ -0.2 & 0.2 & 0.2 & 0.45 & 0.2 & 0.2 & 0 \\ -0.2 & 0.2 & 0.2 & 0.2 & 0.4 & 0.2 & 0 \\ -0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\tilde{b}_{GLM} = \begin{bmatrix} 18.7 \\ 10.12 \\ 5.2 \\ -4.35 \\ 1.22 \\ -5.44 \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{\mu} \\ \tilde{trt}_1 \\ \tilde{trt}_2 \\ \tilde{trt}_3 \\ \tilde{trt}_4 \\ \tilde{trt}_5 \\ \tilde{trt}_6 \end{bmatrix}$$

The generalised inverse produce by IML $(X'X)^-$, together with the solution vector \tilde{b} is shown below, note that they are **NOT** the same as that produced by GLM:

$$\begin{bmatrix} 0.0265306 & 0.0020408 & 0.0091837 & 0.0091837 & 0.0020408 & 0.0020408 & 0.0020408 \\ 0.0020408 & 0.1693878 & -0.037755 & -0.037755 & -0.030612 & -0.030612 & -0.030612 \\ 0.0091837 & -0.037755 & 0.205102 & -0.044898 & -0.037755 & -0.037755 & -0.037755 \\ 0.0091837 & -0.037755 & -0.044898 & 0.205102 & -0.037755 & -0.037755 & -0.037755 \\ 0.0020408 & -0.030612 & -0.037755 & -0.037755 & 0.1693878 & -0.030612 & -0.030612 \\ 0.0020408 & -0.030612 & -0.037755 & -0.037755 & -0.030612 & 0.1693878 & -0.030612 \\ 0.0020408 & -0.030612 & -0.037755 & -0.037755 & -0.030612 & -0.030612 & 0.1693878 \end{bmatrix}$$

$$\tilde{b}_{IML} = \begin{bmatrix} 16.992857 \\ 11.827143 \\ 6.9071429 \\ -2.642857 \\ 2.9271429 \\ -3.732857 \\ 1.7071429 \end{bmatrix} = \begin{bmatrix} \tilde{\mu} \\ \tilde{trt}_1 \\ \tilde{trt}_2 \\ \tilde{trt}_3 \\ \tilde{trt}_4 \\ \tilde{trt}_5 \\ \tilde{trt}_6 \end{bmatrix}$$

$\hat{Y} = X\tilde{b}$ is estimable

$$\text{TSS} = Y'Y$$

$$\text{SSR} = \tilde{b}'X'Y$$

$$\text{SSE} = \text{TSS} - \text{SSR}$$

$$\text{CF} = N\bar{y}^2$$

$$\tilde{b}_{IML} = \begin{bmatrix} 16.992857 \\ 11.827143 \\ 6.9071429 \\ -2.642857 \\ 2.9271429 \\ -3.732857 \\ 1.7071429 \end{bmatrix}$$

11.8 Analysis of Variance

ANOVA					
Source	df	SS	MS	F - ratio	E(MS)
Total, TSS	$N = 28$	$Y'Y$ $= 12154.23$			
Model, SSR	$r(X)$ $= 6$	$\tilde{b}'X'Y$ $= 11873.112$	1978.852	154.864	
Mean, C.F.	1	$N\bar{y}^2$ $= 11060.437$	11060.437	865.584	
Model, after the mean SSR_m	$r(X) - 1$ $= 5$	812.675	162.535	12.720**	$\sigma_e^2 + f \sum (trt_i - \bar{trt})^2$ $\equiv \sigma_e^2 + Q(trt)$
Error, SSE	$N - r(X)$ $= 22$	281.118	12.778		σ_e^2

- we test the significance of the model using an F-ratio, see STD. Table A.6.

- Model F-ratio = 154.864 with 6 d.f. and 22 d.f.

Tabulated F values $F_{6,22,5\%} = 2.55$, $F_{6,22,1\%} = 3.76$

$F_{calc} \gg F_{tabulated\ 6,22,5\%} \Rightarrow$ we conclude that the model accounts for a significant amount of variation.

- we test the significance of the Mean using an F-ratio, see STD. Table A.6.

- Mean F-ratio = 865.584 with 1 d.f. and 22 d.f.

Tabulated F values $F_{1,22,5\%} = 4.301$, $F_{1,22,1\%} = 7.945$

$F_{calc} \gg F_{tabulated\ 1,22,5\%} \Rightarrow$ we conclude that the Mean is significantly different from Zero.

- we test the significance of the model over and above the Mean (ie the effect of treatment) using an F-ratio, see STD. Table A.6.

- Model over and above the Mean F-ratio = 12.72 with 5 d.f. and 22 d.f.

Tabulated F values $F_{5,22,5\%} = 2.661$, $F_{5,22,1\%} = 3.988$

$F_{calc} \gg F_{tabulated\ 5,22,5\%} \Rightarrow$ we conclude that the treatments (i.e. the model over and above the mean) accounts for a significant amount of variation.


```

1 0 0 0 1 0 0,
1 0 0 0 1 0 0,
1 0 0 0 1 0 0,
1 0 0 0 1 0 0,
1 0 0 0 1 0 0,
1 0 0 0 0 1 0,
1 0 0 0 0 1 0,
1 0 0 0 0 1 0,
1 0 0 0 0 1 0,
1 0 0 0 0 1 0,
1 0 0 0 0 0 1,
1 0 0 0 0 0 1,
1 0 0 0 0 0 1,
1 0 0 0 0 0 1};
y = {19.4,
      32.6,
      27.0,
      32.1,
      33.0,
      17.7,
      24.8,
      27.9,
      25.2,
      17.0,
      19.4,
      9.1,
      11.9,
      20.7,
      21.0,
      20.5,
      18.8,
      18.6,
      14.3,
      14.4,
      11.8,
      11.6,
      14.2,
      17.3,
      19.4,
      19.1,
      16.9,
      20.8};

```

```

xtx = x' * x;
xty = x' * y;
invxtx = ginv(xtx); /* NOTE: we are using ginv() */
b = invxtx * xty;
tss = y' * y;
sumy = sum(y);
nobs = nrow(x);
dftrt = 5;
dfe = nobs - dftrt - 1;
ssr = b' * xty;
ybar = sumy/nobs;
cf = nobs * ybar * ybar;
ssrm = ssr - cf;
sse = tss - ssr;
mse = sse/dfe;

/* Estimates of fitted values and their standard errors */

k1 = {1, 1, 0, 0, 0, 0, 0};
kb = k1' * b;
kgk = k1' * invxtx * k1;
sv = kgk * mse;
se = sqrt(sv);

k2 = {1, 0, 1, 0, 0, 0, 0};
kb = k2' * b;
kgk = k2' * invxtx * k2;
sv = kgk * mse;
se = sqrt(sv);

k3 = {1, 0, 0, 1, 0, 0, 0};
kb = k3' * b;
kgk = k3' * invxtx * k3;
sv = kgk * mse;
se = sqrt(sv);

k4 = {1, 0, 0, 0, 1, 0, 0};
kb = k4' * b;
kgk = k4' * invxtx * k4;
sv = kgk * mse;
se = sqrt(sv);

k5 = {1, 0, 0, 0, 0, 1, 0};

```

```

kb = k5' * b;
kgk = k5' * invvxtx * k5;
sv = kgk * mse;
se = sqrt(sv);

k6 = {1, 0, 0, 0, 0, 0, 1};
kb = k6' * b;
kgk = k6' * invvxtx * k6;
sv = kgk * mse;
se = sqrt(sv);

/* Estimates of differences between treatments and the standard */
/* errors of the differences */

k12 = {0, 1, -1, 0, 0, 0, 0};
kb = k12' * b;
kgk = k12' * invvxtx * k12;
sv = kgk * mse;
se = sqrt(sv);

k13 = {0, 1, 0, -1, 0, 0, 0};
kb = k13' * b;
kgk = k13' * invvxtx * k13;
sv = kgk * mse;
se = sqrt(sv);

k14 = {0, 1, 0, 0, -1, 0, 0};
kb = k14' * b;
kgk = k14' * invvxtx * k14;
sv = kgk * mse;
se = sqrt(sv);

k15 = {0, 1, 0, 0, 0, -1, 0};
kb = k15' * b;
kgk = k15' * invvxtx * k15;
sv = kgk * mse;
se = sqrt(sv);

k16 = {0, 1, 0, 0, 0, 0, -1};
kb = k16' * b;
kgk = k16' * invvxtx * k16;
sv = kgk * mse;
se = sqrt(sv);

```

```

/* N.B. kp = k' to generate the test for treatments, i.e. SS trt */
kp = {0  1  -1  0  0  0  0,
      0  1  0  -1  0  0  0,
      0  1  0  0  -1  0  0,
      0  1  0  0  0  -1  0,
      0  1  0  0  0  0  -1};
df = nrow(kp);
kb = kp * b;
kgk = kp * invvtx * kp';
ss = kb' * inv(kgk) * kb;
ms = ss/df;
f = ms/mse;

/* N.B. kp = k' to generate the test for treatments,
   using the second example set of contrasts i.e. SS trt */
kp = {0  -1  0  0  0  0  1,
      0  0  -1  0  0  0  1,
      0  0  0  -1  0  0  1,
      0  0  0  0  -1  0  1,
      0  0  0  0  0  -1  1};
df = nrow(kp);
kb = kp * b;
kgk = kp * invvtx * kp';
ss = kb' * inv(kgk) * kb;
ms = ss/df;
f = ms/mse;

/* fitted values and residuals */
yhat = x * b;
ehat = y - yhat;

```

11.11 Using SAS/GLM

USING SAS/PROC GLM

```

data oneway;
input trt y;
cards;
1 19.4
1 32.6

```

```

1 27.0
1 32.1
1 33.0
2 17.7
2 24.8
2 27.9
2 25.2
3 17.0
3 19.4
3 9.1
3 11.9
4 20.7
4 21.0
4 20.5
4 18.8
4 18.6
5 14.3
5 14.4
5 11.8
5 11.6
5 14.2
6 17.3
6 19.4
6 19.1
6 16.9
6 20.8
;
proc glm;
classes trt;
model y = trt;
contrast ' trt 1 vs trt 2' trt 1 -1 0 0 0 0;
contrast ' trt 1 vs trt 3' trt 1 0 -1 0 0 0;
contrast ' trt 1 vs trt 4' trt 1 0 0 -1 0 0;
contrast ' trt 1 vs trt 5' trt 1 0 0 0 -1 0;
contrast ' trt 1 vs trt 6' trt 1 0 0 0 0 -1;
contrast ' trt 2 vs trt 3' trt 0 1 -1 0 0 0;
contrast ' trt 2 vs trt 4' trt 0 1 0 -1 0 0;
contrast ' trt 2 vs trt 5' trt 0 1 0 0 -1 0;
contrast ' trt 2 vs trt 6' trt 0 1 0 0 0 -1;
contrast ' trt 3 vs trt 4' trt 0 0 1 -1 0 0;
contrast ' trt 3 vs trt 5' trt 0 0 1 0 -1 0;
contrast ' trt 3 vs trt 6' trt 0 0 1 0 0 -1;
contrast ' trt 4 vs trt 5' trt 0 0 0 1 -1 0;

```

```

contrast ' trt 4 vs trt 6' trt 0 0 0 1 0 -1;
contrast ' trt 5 vs trt 6' trt 0 0 0 0 1 -1;
estimate ' trt 1 - trt 2' trt 1 -1 0 0 0 0;
estimate ' trt 1 - trt 3' trt 1 0 -1 0 0 0;
estimate ' trt 1 - trt 4' trt 1 0 0 -1 0 0;
estimate ' trt 1 - trt 5' trt 1 0 0 0 -1 0;
estimate ' trt 1 - trt 6' trt 1 0 0 0 0 -1;
estimate ' trt 2 vs trt 3' trt 0 1 -1 0 0 0;
estimate ' trt 2 vs trt 4' trt 0 1 0 -1 0 0;
estimate ' trt 2 vs trt 5' trt 0 1 0 0 -1 0;
estimate ' trt 2 vs trt 6' trt 0 1 0 0 0 -1;
estimate ' trt 3 vs trt 4' trt 0 0 1 -1 0 0;
estimate ' trt 3 vs trt 5' trt 0 0 1 0 -1 0;
estimate ' trt 3 vs trt 6' trt 0 0 1 0 0 -1;
estimate ' trt 4 vs trt 5' trt 0 0 0 1 -1 0;
estimate ' trt 4 vs trt 6' trt 0 0 0 1 0 -1;
estimate ' trt 5 vs trt 6' trt 0 0 0 0 1 -1;
estimate ' mean + trt 1 ' intercept 1 trt 1 0 0 0 0 0;
estimate ' mean + trt 2 ' intercept 1 trt 0 1 0 0 0 0;
estimate ' mean + trt 3 ' intercept 1 trt 0 0 1 0 0 0;
estimate ' mean + trt 4 ' intercept 1 trt 0 0 0 1 0 0;
estimate ' mean + trt 5 ' intercept 1 trt 0 0 0 0 1 0;
estimate ' mean + trt 6 ' intercept 1 trt 0 0 0 0 0 1;

```

```

/* A [needlessly] complete contrast statement to compute the Sums of
Squares for trt, which also happens to be SSRm */

```

```

contrast 'SS trt' intercept 0 trt 1 -1 0 0 0 0,
            intercept 0 trt 1 0 -1 0 0 0,
            intercept 0 trt 1 0 0 -1 0 0,
            intercept 0 trt 1 0 0 0 -1 0,
            intercept 0 trt 1 0 0 0 0 -1;

```

```

/* A contrast statement to compute the Sums of Squares for
trt, which also happens to be SSRm */

```

```

contrast 'SS trt' trt 1 -1 0 0 0 0,
            trt 1 0 -1 0 0 0,
            trt 1 0 0 -1 0 0,
            trt 1 0 0 0 -1 0,
            trt 1 0 0 0 0 -1;

```

```

/* Another contrast statement to compute the Sums of Squares for
trt, which also happens to be SSRm */

```

```

contrast 'SS trt' trt -1 0 0 0 0 1,

```

```

        trt 0 -1 0 0 0 1,
        trt 0 0 -1 0 0 1,
        trt 0 0 0 -1 0 1,
        trt 0 0 0 0 -1 1;

/* A contrast to compare the 5 inoculants vs 6th (mixture)
   note: the coefficients sum to Zero, a contrast,
   note: no need to divide by 5, avoid fractions */
contrast 'SS compare' trt 1 1 1 1 1 -5;

/* An estimate of the first 5 inoculants - mixture.
   note, scaleup to avoid fractions, use /divisor= */
estimate 'Inocs vs mixture' trt 1 1 1 1 1 -5/divisor=5;

lsmeans trt/stderr pdiff;
/* fitted values and residuals */
output out=fred1 p=yhat r=ehat;
run;

proc print data=fred1;
run;

```

Compare this with STD Ch 7.3, P141-147.

12 Fitted Values

It is most important to know that the 'fitted' values' are always estimable. Anything and everything that is statistically estimable can be expressed as a linear function of these fitted values. Thus, if we cannot express something as a linear function of these fitted values then it is said to be 'non-estimable'.

$$\hat{Y} = X\tilde{b}$$

$$\hat{Y}_{11} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 16.992857 \\ 11.827143 \\ 6.9071429 \\ -2.642857 \\ 2.9271429 \\ -3.732857 \\ 1.7071429 \end{bmatrix}$$

$$= 28.82$$

$$\hat{Y}_{23} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 16.992857 \\ 11.827143 \\ 6.9071429 \\ -2.642857 \\ 2.9271429 \\ -3.732857 \\ 1.7071429 \end{bmatrix}$$

$$= 23.9$$

$$\hat{Y}_{31} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 16.992857 \\ 11.827143 \\ 6.9071429 \\ -2.642857 \\ 2.9271429 \\ -3.732857 \\ 1.7071429 \end{bmatrix}$$

$$= 14.35$$

Note that in this Completely Randomized Design, one way ANOVA the estimate or fitted value, for treatment 1 is the mean of the observations for treatment 1, *i.e.* :

$$\frac{1}{n_1} \sum_{j=1}^{j=n_1} Y_{ij}$$

Suggestion : see the estimate of Y_1 above.

Any estimable value (estimate) must have a standard error.

$$\Rightarrow \text{if } \hat{Y} = X\tilde{b}$$

$$\text{then } V(\hat{Y}) = V(X\tilde{b})$$

$$= XV(\tilde{b})X'$$

$$\text{n.b. } V(\tilde{b}) = (X'X)^{-1}\hat{\sigma}_e^2$$

$$\text{thus } V(\hat{Y}) = X(X'X)^{-1}X'\hat{\sigma}_e^2$$

$$V(\hat{Y}_{11}) = X_{11}(X'X)^{-1}X'_{11}\hat{\sigma}_e^2$$

$$\text{where } X_{11} = [1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0]$$

$$\Rightarrow V(\hat{Y}_{11}) = 2.55552$$

$$\Rightarrow s.e._{Y_{11}} = \sqrt{2.55552} = 1.5986$$

Note that this equals $\sqrt{\frac{\hat{\sigma}_e^2}{n_1}}$

$$\text{and } V(\hat{Y}_{23}) = X_{23}(X'X)^{-1}X'_{23}\hat{\sigma}_e^2$$

$$\text{where } X_{23} = [1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0]$$

$$\Rightarrow V(\hat{Y}_{23}) = 3.19444$$

$$\Rightarrow s.e._{Y_{23}} = \sqrt{3.19444} = 1.7873$$

$$\text{and } V(\hat{Y}_{31}) = X_{31}(X'X)^{-1}X'_{31}\hat{\sigma}_e^2$$

$$\text{where } X_{31} = [1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0]$$

$$\Rightarrow V(\hat{Y}_{31}) = 3.19444$$

$$\Rightarrow s.e._{Y_{31}} = \sqrt{3.19444} = 1.7873$$

Question: What is the SAS code to estimate/generate these?

13 Assumptions for the Model

1. The model is assumed to be appropriate.
2. The model is assumed to be linear and additive.
3. Assignment to classes (treatments) is without error.
4. The error terms are independent and uncorrelated.
5. The errors are random with a finite and homogeneous variance, σ_e^2

Normality of the error terms is not required for obtaining solutions. It is only required for statistical tests, hypothesis testing and tests of significance. See STD. 7.10 Page 174.

14 Treatment Differences

14.1 Estimation of Treatment Differences

What is the difference between treatment 1 and 2?

$$\hat{Y}_{11} = X_{11}\tilde{b} = [1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0] \tilde{b}$$

$$\hat{Y}_{21} = X_{21}\tilde{b} = [1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0] \tilde{b}$$

If \hat{Y}_{11} and \hat{Y}_{21} are estimable, then $(\hat{Y}_{11} - \hat{Y}_{21})$ is estimable, *i.e.* treatment₁ - treatment₂.

$$\text{Thus } X_{11}\tilde{b} - X_{21}\tilde{b} = (X_{11} - X_{21})\tilde{b}$$

$$[(1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0) - (1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0)] \tilde{b}$$

$$= [0 \ 1 \ -1 \ 0 \ 0 \ 0 \ 0] \tilde{b}$$

$$= 11.827143 - 6.9071429$$

$$= 4.92$$

$$\equiv 28.82 - 23.9$$

Question: What would be the appropriate SAS statements for this?

14.2 Sampling Variance of Treatment Differences

We estimated the treatment differences from $k'\tilde{b}$, $k' = [0 \ 1 \ -1 \ 0 \ 0 \ 0 \ 0]$

$$\Rightarrow s.v.(k'\tilde{b}) = k' s.v.(\tilde{b}) k$$

$$= k'(X'X)^{-1}k\hat{\sigma}_e^2$$

$$s.e.diff = \sqrt{s.v.}$$

Using a numerical example, with the $(X'X)^{-1}$ and \tilde{b} from GLM, we have

$$(X'X)^{-1} = \begin{bmatrix} 0.2 & -0.2 & -0.2 & -0.2 & -0.2 & -0.2 & 0 \\ -0.2 & 0.4 & 0.2 & 0.2 & 0.2 & 0.2 & 0 \\ -0.2 & 0.2 & 0.45 & 0.2 & 0.2 & 0.2 & 0 \\ -0.2 & 0.2 & 0.2 & 0.45 & 0.2 & 0.2 & 0 \\ -0.2 & 0.2 & 0.2 & 0.2 & 0.4 & 0.2 & 0 \\ -0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\tilde{b}_{GLM} = \begin{bmatrix} 18.7 \\ 10.12 \\ 5.2 \\ -4.35 \\ 1.22 \\ -5.44 \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{\mu} \\ \tilde{trt}_1 \\ \tilde{trt}_2 \\ \tilde{trt}_3 \\ \tilde{trt}_4 \\ \tilde{trt}_5 \\ \tilde{trt}_6 \end{bmatrix}$$

Using our formula, from above,

$$(X'X)^{-1}k = \begin{bmatrix} 0.2 & -0.2 & -0.2 & -0.2 & -0.2 & -0.2 & 0 \\ -0.2 & 0.4 & 0.2 & 0.2 & 0.2 & 0.2 & 0 \\ -0.2 & 0.2 & 0.45 & 0.2 & 0.2 & 0.2 & 0 \\ -0.2 & 0.2 & 0.2 & 0.45 & 0.2 & 0.2 & 0 \\ -0.2 & 0.2 & 0.2 & 0.2 & 0.4 & 0.2 & 0 \\ -0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$(X'X)^{-1}k = \begin{bmatrix} 0.0 \\ 0.2 \\ -0.25 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0 \end{bmatrix}$$

$$\text{and } k'(X'X)^{-1}k = [0 \ 1 \ -1 \ 0 \ 0 \ 0 \ 0] \begin{bmatrix} 0.0 \\ 0.2 \\ -0.25 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0 \end{bmatrix}$$

$$= 0.45$$

Thus, the sampling variance of our estimate of the difference is $(k'(X'X)^{-1}k * MSE) = 0.45 * 12.778 = 5.7501$, and the standard error is the square root of this, i.e. $\sqrt{5.7501} = 2.398$

Thus our estimate of

$$\widehat{t_1 - t_2} = 4.92$$

$$s.e._{\widehat{t_1 - t_2}} = 2.398$$

Is this difference statistically significant ?

What is the confidence interval for our estimate ?

14.3 Test Whether Difference is Statistically Significant

t-test

$$t = \frac{\text{estimate} - \text{constant}}{s.e.}$$

$$\text{estimate} = k'\tilde{b} = 4.92$$

$$\text{constant} = 0 = (H_o \text{ under the null hypothesis})$$

$$s.e. = 2.398$$

$$\Rightarrow t_{calc} = \frac{4.92}{2.398}$$

$$= 2.052$$

$$\text{d.f. error} = 22$$

Tabulated t values (S&T., Table A.3)

$$t_{22,5\%} = 2.074$$

$$t_{22,1\%} = 2.819$$

$t_{calc} < t_{tabulated}$ ($2.052 < 2.074$), therefore we can conclude that the difference is not statistically significant (at the 5% level).

Suppose that we were interested in whether the difference was 4 or not!

14.4 Testable Hypotheses

Let us go back to the 'fitted values', ($\hat{Y} = X\tilde{b}$); n.b. as always!

We have that an observation from the 1st treatment is estimable. It does not matter particularly which one, they all have the same estimated, or fitted, value.

$$\text{thus } \hat{Y}_{11} = \tilde{\mu} + t\tilde{r}t_1 = k'_1\tilde{b}$$

$$k'_1 = (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$\text{similarly } \hat{Y}_{24} = \tilde{\mu} + t\tilde{r}t_2 = k'_2\tilde{b}$$

$$k'_2 = (1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0)$$

and $\widehat{Y}_{35} = \tilde{\mu} + \tilde{t}rt_3 = k'_3\tilde{b}$

$$k'_3 = (1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0)$$

and $\widehat{Y}_{41} = \tilde{\mu} + \tilde{t}rt_4 = k'_4\tilde{b}$

$$k'_4 = (1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0)$$

and $\widehat{Y}_{52} = \tilde{\mu} + \tilde{t}rt_5 = k'_5\tilde{b}$

$$k'_5 = (1 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0)$$

and $\widehat{Y}_{61} = \tilde{\mu} + \tilde{t}rt_6 = k'_6\tilde{b}$

$$k'_6 = (1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1)$$

Therefore differences between fitted values are estimable.

$$\text{So } \widehat{Y}_{11} - \widehat{Y}_{24} = (\tilde{\mu} + \tilde{t}rt_1) - (\tilde{\mu} + \tilde{t}rt_2) = \tilde{t}rt_1 - \tilde{t}rt_2$$

with $k'_{1-2} = (0 \ 1 \ -1 \ 0 \ 0 \ 0 \ 0)$

which corresponds to one of the comparisons, or Contrasts, from our original Null Hypothesis relating to treatments.

$$\text{likewise } \widehat{Y}_{11} - \widehat{Y}_{35} = (\tilde{\mu} + \tilde{t}rt_1) - (\tilde{\mu} + \tilde{t}rt_3) = \tilde{t}rt_1 - \tilde{t}rt_3$$

with $k'_{1-3} = (0 \ 1 \ 0 \ -1 \ 0 \ 0 \ 0)$

$$\text{and } \widehat{Y}_{11} - \widehat{Y}_{41} = (\tilde{\mu} + \tilde{t}rt_1) - (\tilde{\mu} + \tilde{t}rt_4) = \tilde{t}rt_1 - \tilde{t}rt_4$$

with $k'_{1-4} = (0 \ 1 \ 0 \ 0 \ -1 \ 0 \ 0)$

$$\text{and } \widehat{Y}_{11} - \widehat{Y}_{52} = (\tilde{\mu} + \tilde{t}rt_1) - (\tilde{\mu} + \tilde{t}rt_5) = \tilde{t}rt_1 - \tilde{t}rt_5$$

with $k'_{1-5} = (0 \ 1 \ 0 \ 0 \ 0 \ -1 \ 0)$

$$\text{and } \widehat{Y}_{11} - \widehat{Y}_{61} = (\tilde{\mu} + \tilde{t}rt_1) - (\tilde{\mu} + \tilde{t}rt_6) = \tilde{t}rt_1 - \tilde{t}rt_6$$

with $k'_{1-6} = (0 \ 1 \ 0 \ 0 \ 0 \ 0 \ -1)$

Putting the k' rows together we have

$$k'_{trt} = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}$$

This matrix, k'_{trt} , we can use in our general formula for computing Sums of Squares:

$$\text{thus } SS_{trt} = (k'_{trt}\tilde{b})' [k'_{trt}(X'X)^{-1}k_{trt}]^{-1} (k'_{trt}\tilde{b})$$

in fact this will give us exactly the result for our Analysis of Variance, $R(trt|\mu)$. In this case it is slightly redundant since we already have the Sums of Squares for the Model corrected for the Mean, which corresponds to the Sums of Squares for treatments, since there is nothing else in the model apart from treatments (over and above the Mean). However, and it cannot be stressed enough, it illustrates how we can build up Sums of Squares for specified Hypotheses tests; all derived from linear combinations of the fitted values.

14.5 Using the SAS CONTRAST statement

We can use the SAS CONTRAST statement in PROC GLM to build up and compute Sums of Squares and tests of hypothesis in an analagous manner. With SAS the k' matrix is divided into named sections corresponding to the effects in the model. If we have specified our model in GLM as

```
proc glm;
classes trt;
model y = trt;

/* A contrast statement to compute the Sums of Squares for trt */
contrast 'SS trt' trt 1 -1 0 0 0 0,
               trt 1 0 -1 0 0 0,
               trt 1 0 0 -1 0 0,
               trt 1 0 0 0 -1 0,
               trt 1 0 0 0 0 -1/E=MS to be used as Error;
```


Note: We can specify the Mean Square to be used as the “Error” (divisor), it defaults to the MSE if we omit it. At the moment this is what we want, we are using the MSE as our divisor in our F-tests. However, later we will need to over-ride this and specify the MS (as for example in Nested/Sub-sampling models).

This corresponds to our design matrix (X) having the following columns:

(μ trt₁ trt₂ trt₃ trt₄ trt₅ trt₆)

SAS names the corresponding parts of the k' matrix as:

intercept - refers to the first column, corresponding to that for the mean

trt - refers to columns 2 to 7, corresponding to those for treatment levels

So we can write a CONTRAST statement (after the model statement) as:

```
contrast 'SS treatments' intercept 0 trt 1 -1 0 0 0 0,
                        intercept 0 trt 1 0 -1 0 0 0,
                        intercept 0 trt 1 0 0 -1 0 0,
                        intercept 0 trt 1 0 0 0 -1 0,
                        intercept 0 trt 1 0 0 0 0 -1;
```

We have 5 rows to our matrix, analagous to the 5 rows of our k' matrix. At the end of each row we have a comma (,) to indicate to the CONTRAST statement that it is really the end of the row. At the end of the 5th row we end with a semicolon (;), being the normal SAS end-of-statement indicator. Note that we have 6 coefficients for trt, which correspond to the 6 levels for treatments.

There is one simplification that we can make to the above CONTRAST statement. We can see that for the intercept coefficient it is Zero for each level of the intercept (in fact there is only 1 level) for each row, so we can, if we want, omit it. This is simply an abbreviation. Thus we can write:

```
contrast 'SS treatments' trt 1 -1 0 0 0 0,
                        trt 1 0 -1 0 0 0,
                        trt 1 0 0 -1 0 0,
```

```

trt 1 0 0 0 -1 0,
trt 1 0 0 0 0 -1;

```

Thus, to recap we could write our SAS statements as:

```

proc glm;
classes trt;
model y = trt;
contrast 'SS treatments' trt 1 -1 0 0 0 0,
trt 1 0 -1 0 0 0,
trt 1 0 0 -1 0 0,
trt 1 0 0 0 -1 0,
trt 1 0 0 0 0 -1;
contrast 'other SS trt' trt 1 -1 0 0 0 0,
trt 0 1 -1 0 0 0,
trt 0 0 1 -1 0 0,
trt 0 0 0 1 -1 0,
trt 0 0 0 0 1 -1;
run;
quit;

```

14.6 Using the SAS ESTIMATE statement

Another useful feature of the SAS PROC GLM procedure is the ESTIMATE statement, which allows us to compute 'fitted values', or indeed any value that is 'estimable'.

Consider the 'fitted value', or estimate, of $\hat{Y}_{11} = \tilde{\mu} + \tilde{t}rt_1$. A suitable k' matrix to estimate this would be

$$\hat{Y}_{11} = \tilde{\mu} + \tilde{t}rt_1 = k'_1 \tilde{b}$$

$$k'_1 = (1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0)$$

in a manner similar to the CONTRAST statement, we therefore obtain the following SAS ESTIMATE statement:

```
estimate 'mu + trt 1' intercept 1 trt 1 0 0 0 0 0;
```

Consider the 'fitted value', or estimate, of $\hat{Y}_{24} = \tilde{\mu} + \tilde{t}t_2$. A suitable k' matrix to estimate this would be

$$\hat{Y}_{24} = \tilde{\mu} + \tilde{t}t_2 = k'_2 \tilde{b}$$

$$k'_2 = (1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0)$$

and we have the following SAS ESTIMATE statement:

```
estimate 'mu + trt 2' intercept 1 trt 0 1 0 0 0 0;
```

We are not restricted to estimating only fitted values, anything that is a linear function of the fitted values, and hence estimable, can be estimated with the ESTIMATE statement, via a suitably specified k' matrix and ESTIMATE statement. For example, we have seen that $trt_1 - trt_2$ is estimable, and that a suitable k' matrix would be

$$k'_{1-2} = (0 \ 1 \ -1 \ 0 \ 0 \ 0 \ 0)$$

which we could translate into the following SAS ESTIMATE statement:

```
estimate 'trt 1 - trt 2' intercept 0 trt 1 -1 0 0 0 0;
```

and as before we could therefore abbreviate to the following:

```
estimate 'trt 1 - trt 2' trt 1 -1 0 0 0 0;
```

We could also compare the average of the first 5 treatments with the 6th treatment (which is a mix of the first 5).

```
estimate 'trt 6 - 1...5' trt -1 -1 -1 -1 -1 5/divisor=5;
```

14.7 Equation Order

In our CRD example we had 6 diets and we called them Diet 1, Diet 2, Diet 3, Diet 4, Diet 5 and Diet 6. This was also the order in which we arranged them when we were setting up the design matrix (X). It should be noted that this ordering is/was completely arbitrary. We could just have easily put Diet 6 as the first one so that it was the first column related to treatments.

Re-number the data from Section 11.5 so that Diet 6 is now the first diet and Diet 4 is the second diet and re-analyse. Construct the ANOVA table as well as estimating treatment differences and fitted values. Note that they are the same as the values in the notes even though the generalized inverse $(X'X)^-$ and the solution vector (\tilde{b}) are substantially different.

Although not a statistical proof, this does help to illustrate the fact that the various estimable values are invariant to the ordering and to the choice of the generalized inverse.

15 Random effects models

See Steel, Torrie and Dickey, Ch. 7.5

If the effects we consider are fixed effects then we will be interested in the specific differences between the various treatments; these treatments and no other treatments. Our results apply to ONLY these treatments. However, if the effects that we are considering are classed as random effects then it is the variability in the population that we shall be interested in.

15.1 Parameters

Variance components; variance between levels and within levels.

15.2 Example

Consider that we had recorded the weights of apples on each of 6 apple trees at Macdonald Campus of McGill University. The trees were a random sample and are therefore considered representative of this type of apple tree (variety) growing in the region (the St Lawrence river valley). From each tree we randomly sample 4 or 5 apples and weigh each, see table 8.

Table 8: Example data

T1	T2	T3	T4	T5	T6
19.4	17.7	17.0	20.7	14.3	17.3
32.6	24.8	19.4	21.0	14.4	19.4
27.0	27.9	9.1	20.5	11.8	19.1
32.1	25.2	11.9	18.8	11.6	16.9
33.0	.	.	18.6	14.2	20.8

Our statistical model will be

$$Y_{ij} = \mu + tree_i + e_{ij}$$

This looks much like our CRD/One-way Analysis of Variance. It is, except that the parameters of our model are slightly different. We are not interested in estimating differences amongst particular trees, rather we are interested in estimating the variance amongst trees.

15.3 Terms in the model

Y_{ij}	= the weight of the j^{th} apple from the i^{th} tree
μ	= the overall mean apple weight
$tree_i$	= the random effect of the i^{th} tree on the weight of an apple, $tree_i \sim N(0, \sigma_t^2)$
e_{ij}	= the random residual effect specific to the j^{th} apple from the i^{th} tree $e_{ij} \sim N(0, \sigma_e^2)$

15.4 Parameters of the model

μ (fixed effect), and σ_{trees}^2 and σ_e^2 (random effects variance components).

$$tree_i \sim N(0, \sigma_{trees}^2) \quad \text{and} \quad e_{ij} \sim N(0, \sigma_e^2)$$

15.5 Expectations of Mean squares

For Model I, fixed effects, the Mean Square tests whether there are any treatment effects. For Model II, random effects, the Mean Square tests whether there is a statistically significant effect of trees σ_t^2 , *i.e.* whether $\sigma_t^2 > 0$.

Table 9: E(M.S.)

Source of Variation	df	Model I	Model II
Treatments	t-1	$\sigma^2 + r_1 \sum (t_i - \bar{t})^2 / (t - 1)$	$\sigma^2 + r_1 \sigma_t^2$
Residual	t(r-1)	σ^2	σ^2

15.6 Estimation of variance components from E(MS)

We can calculate σ_t^2 from (Mean Square Trees - Mean Square Error) divided by the 'effective' number of apples per tree, r_1 . If the experiment is completely 'balanced' the r_1 is equal to the number of apples per tree, if we have an 'unbalanced' design the r_1 is always a bit less than the actual average number of apples per tree (see Steel, Torrie and Dickey, Ch 7.5). In this example $r_1 = 4.6571$.

For a simple One-way ANOVA, such as the above example, the general formula for computing r_1 is given as

$$r_1 = k_1 = \left(\sum_{i=1}^{i=t} n_i - \frac{\sum_{i=1}^{i=t} n_i^2}{\sum_{i=1}^{i=t} n_i} \right) / (t - 1)$$

where n_i equals the number of 'replicates', or subsamples, for the i^{th} group, in the above example it means the number of apples on the i^{th} tree, and there are t trees.

$$\sigma_t^2 = (162.534 - 12.778) / 4.6571 = 32.156$$

$$\text{Thus } \sigma_t^2 = 32.156 \text{ and } \sigma_e^2 = 12.778, \text{ and } \sigma_p^2 = \sigma_t^2 + \sigma_e^2.$$

$$\text{Thus } \sigma_p^2 = 32.156 + 12.778 = 44.934$$

Therefore the variation between trees σ_t^2 equals $32.156/44.934$, 71.56%, of the total variation in apple weight. Thus we could conclude that there are major differences between trees in the weight of their apples and that within trees the apples are relatively uniform. This would be good news if we were an apple tree breeder; it would suggest that *a priori* there would be possibilities of selecting better trees and improving apple weight.

With a balanced design it is relatively simple to compute the variance components. However, if the design is unbalanced it is more tedious, and more easily left to SAS!

15.7 Using SAS/GLM with a random effect

```
USING SAS/PROC GLM
```

```
data oneway;
input tree y;
cards;
1 19.4
.
.
.
6 20.8
;

proc glm;
classes tree;
model y = tree;
random tree;
run;
```

SAS will compute the coefficient(s) of the expectation of the Mean Squares.

NOTE: Whilst PROC GLM will compute the Expectations for Mean Squares when we give the RANDOM option ALL analyses are carried out assuming a fixed effects model. Thus SAS/PROC GLM is in fact schizophrenic when the RANDOM statement is used. For anything other than simple one way analysis with equal numbers in each and every class such analyses are in fact suboptimal and one should be using mixed models, e.g. PROC MIXED.

15.8 Using SAS/MIXED

```
USING SAS/PROC MIXED
```



```

data oneway;
input tree y;
cards;
1 19.4
.
.
.
6 20.8
;

proc mixed;
classes tree;
model y = ;
random tree;
run;

```

15.9 Reasons for our interest in random effects

Why is it of interest to us to estimate variance components for a random effect?

1) To know the between group and within group variance. This would give us *a priori* evidence as to whether there are differences between groups (trees) and hence whether selection might be possible.

2) We need to know these variances if we are planning experiments, to use as our [appropriate] variance estimate in the formulae for computing sample size.

3) These variances may be biologically interesting; they may suggest interesting future research to look at why the between group differences exist, *i.e.* what are the differences between the trees that results in them producing apples of different weights?

4) We shall, if we have random effects in our statistical model, need to account for these in our tests of fixed effects, both in terms of the structure of the statistical model, and also in terms of the degrees of freedom for our significance tests. We cannot simply say “Oh, it is random, hence we can ignore it”.

16 Multiple Comparisons

16.1 Issues related to multiple comparisons

References

Steel, Torrie and Dickey, Chapter 8

Multiple Comparisons and Multiple Tests using the SAS System, Westfall, P. H., *et al.*

There are two main issues related to multiple comparisons and tests; they are, firstly, the question of whether the comparison(s) is/are pre-planned or not, and secondly the fact that one is making multiple comparisons.

If we have an experiment involving 2 treatments, then there is only 1 comparison to make, and a simple t-test or F-test is appropriate, efficient and sufficient. If we have more than 2 treatments (suppose we have 6 treatments) then there will be $6*(6-1)/2 = 15$ possible comparisons amongst the 6 treatments means. However, there are only 5 degrees of freedom for treatments, that is to say that there are a maximum of 5 comparisons which are linearly independent of one another. Only if we decided beforehand EXACTLY which comparison to make would it have a 'real' probability of accepting/rejecting H_o at the probability we thought we were doing (see STD Ch 8.2 for more discussion on this point).

Suppose that we have 1 test to make; we can use a t-test (or equivalently an F-test). If H_o (the Null Hypothesis, of no real effect) is true, then there is a 0.95 probability of accepting this and a 0.05 probability (chance) of erroneously rejecting it and thinking (mistakenly) that there is a real effect.

Suppose that we have 2 tests to make; further suppose that the Null Hypothesis (H_o) is true in both cases. We should want to accept the H_o for both; the probability of accepting (not rejecting) both is therefore 0.95^2 ; consequently the probability of rejecting at least 1 of the comparisons is $1 - 0.95^2 = 0.0975$. This is not at all our 5% probability that we might have thought that we were testing at! See table 10.

What is being shown here is that the Family Wide Error rate (FWE) (*i.e.* overall) is

Table 10: Multiple comparisons

Comparisons	Probability
1	0.05
2	0.0975
3	0.1426
4	0.1855
5	0.2262
.	
.	
10	0.4013

not at all 5%; even though any one test, if it was pre-planned and linearly independent of all others, has a Comparison Wide Error rate (CWE) of 5 %. Note, however, in STD Ch 8.2, in the discussion of least significant differences (lsd) it is stated that lsd is a valid procedure for pre-planned comparisons; this is passé. It does nothing for the multiple comparison problem and hence FWE rate. So, lsd is just that, a mind-altering drug to be avoided.

16.2 Bonferroni's Test and Sidak's Inequality

$$\Pr (A_1 \text{ or } A_2 \text{ or } A_3 \dots \text{ or } A_l) \leq \Pr(A_1) + \Pr(A_2) + \dots + \Pr(A_l)$$

Thus, if we are making v comparisons, then use an adjusted p-value, \tilde{p} , of α/v ; so if our probability level is $\alpha = 0.05$ (5%) and we have 10 possible tests, then use $0.05/10 = 0.005$ as the adjusted probability level for accepting/rejecting whether an effect is statistically significant (at a FWE rate of 5%).

SAS functions which are useful, `finv()`, `probf()`, `tinvt()` and `probt()`.

So, suppose that we want the critical t-value for contrasts where we have $s=6$ 'treatments' (as per our CRD experiment), *i.e.* $v = 6*(6-1)/2 = s*(s-1)/2 = 15$, for an $\alpha = 0.05$.

$$\tilde{p} = 0.05/15 = 0.003333$$

The numerator degrees of freedom, ndf , = 5, and the denominator degrees of freedom, ddf , = 22. The following SAS code will produce the F-value and t-value corresponding to the (adjusted) probability level of 0.003333. Note, that since `tinvt()` gives the probability of a larger t value (1-tailed test/sign is important) and we want the t-statistics pertaining to a 2-tailed test (absolute difference) then we have to divide the adjusted probability by 2 to get the value that we use in the `tinvt()` function. Thus $0.003333/2 = 0.001666$, and we use $1 -$ the value (*i.e.* $1 - 0.001666$), since we need the critical t-value for the probability of getting a larger (absolute) t-value by chance alone.

```
data critvals;
input pvalue ndf ddf;
F = finv(1-pvalue, 1, ddf);
t = tinvt(1-pvalue/2, ddf);
cards;
0.003333    5    22
;

proc print data=critvals;
run;
```

We obtain the following results, for the F-value and t-value:

$$t_{adj} = 3.29$$

$$F_{adj} = 10.824$$

This compares with the t-value of 2.074 for the 5% probability level tabulated in STD. Thus we need a calculated t-value of at least 3.29 before we should declare the difference between 2 treatment means to be significant if we are going to examine the differences amongst the treatment means. Similarly, we can construct a Confidence Interval (C.I.) as

being this tabulated t-value (of 3.29) \times the standard error (s.e.) of the difference between the two means. Bonferroni's test is valid for comparisons amongst pairs of treatment means, **IFF** the comparisons were pre-planned. For $v \geq 20$ it is inefficient and too conservative, and we would be better off to make use of Scheffé's Test. Note, $v \geq 20$ corresponds to $s = 7$ treatments = $7*6/2 = 21$ comparisons amongst all pairs of means. So, for up to, and including, 6 treatments (levels) we can use Bonferroni's test to compare the treatment means, for 7+ treatments (levels) it will be more efficient to use Scheffé's Test. When using SAS/GLM we can easily make use of Boneferroni's test, even without any complicated hand calculations or even any subsequent use of SAS (either the data step or IML). Recall that using Boneferroni's test we need an adjusted probability level, for example with our 6 treatments (15 tests), then we need an adjusted probability level of $0.05/15 = 0.00333$. If we compute differences between treatments using the ESTIMATE statement in GLM, SAS computes the estimate, its standard error, the calculated t-value, and the probability of obtaining such a large t-value by chance under H_o . Thus, we only have to look whether the probability is less than 0.003333 to decide whether to reject H_o under Boneferroni's test!!!

16.3 Scheffé's Test, STD Ch 8.5

If we want linear combinations, not just differences amongst individual means, then Bonferroni's test is not appropriate, because v is now effectively ∞ ; and hence Bonferroni's test (using a probability level of α/v) is **N.B.G.!!!** However, Scheffé's Test is valid for all and any linear combinations of means.

$$(\bar{y}_i - \bar{y}_j)$$

$$k'\tilde{b} = \text{contrast} = \text{estimate}$$

$$\text{s.e.} = \sqrt{k'(X'X)^{-1}k\sigma^2}$$

$$\text{critical difference} = \text{s.e.}_{k'\tilde{b}} * \sqrt{(s-1)F_{\alpha, s-1, dfe}}$$

e.g. Consider our One-way ANOVA, CRD, experiment (Section 11.5, 14.1 and 14.2) where we had 6 treatments, $\Rightarrow s = 6$.

The d.f.e. = 22.

At the 5% level $F_{5\%,5,22} = 2.66$

Consider that we wish to compare the 5 inoculants against the mixture, $(trt_1 + trt_2 + trt_3 + trt_4 + trt_5)/5 - trt_6$. The estimate = 1.35 and the standard error is 1.77

Thus the critical difference is $1.77 * \sqrt{5 * 2.66} = 6.45$

if the difference is greater than 6.45 we can reject the null hypothesis at the 5% level and accept that there is a real difference between the average of the inoculants and the mixture. Note; this CD is equivalent to a Confidence Interval and can be used in this way; indeed it should be so used when we are faced with a multiple comparison situation.

☞ A simple t-test would have concluded that a critical difference of $2.074 * 1.77$ (3.67) was sufficient.

More conservative!!! Important.

Scheffé's test is not the only multiple comparison test available

SNK Student-Newman-Keul's, dubious

TUKEY

WALLER Waller - if you are (sure) you are a Bayesian

Duncan's test - absolutely verboten, passé, n.b.g.

We can use Scheffé's test with SAS as an option on the Least Squares Means statement (and similarly with Bonferroni's test).

USING SAS/PROC GLM

```
proc glm;
classes trt;
model y = trt;
lsmeans trt/stderr pdiff adjust=scheffe;
lsmeans trt/stderr pdiff adjust=bon;
run;
```

USING SAS/PROC MIXED

```
proc mixed;  
classes trt;  
model y = trt;  
lsmeans trt/pdiff adjust=scheffe;  
lsmeans trt/pdiff adjust=bon;  
run;
```

Note: for proc mixed we do not have the STDERR option, proc mixed automatically produces the standard errors for the Least Squares Means, whereas for proc glm we have to explicitly request them.

17 Partitioning Sums of Squares, Linear, Quadratic, etc

See also the Statistical Methods II Web site section, “Classification or Regression”.

When looking at Multiple Regressions (Section 2) we were examining the effect of Quantitative levels of the X variables on the value of Y. Likewise, when looking at a One-Way, CRD (Section 11.5), we were examining the effect of Qualitative levels of a Factor (Diet) on the value of Y. We had labelled the 6 diets as: Diet 1, Diet 2, Diet 3, Diet 4 , Diet 5 and Diet 6 and this was the order that we used when constructing our design matrix (X); although, as noted in Section 14.7, the order is arbitrary.

Sometimes we have Factors which are Quantitative in nature, but which can be considered as either Quantitative or Qualitative. For example, suppose that we were looking at the effects of 5 different diets which we are feeding to dairy cows. The diets consist of various amounts of a protein supplement (Table 11).

Table 11: Diets being fed to dairy cows

Diet	Amount of protein supplement fed (kg)
1	2
2	3
3	4
4	5
5	6

We have 15 cows and we randomly assign them to the 5 diets, 3 cows per diet. Each cow is individually housed and fed, such that cows cannot interfere with one another; we have a simple, straightforward CRD. We record the milk production of each cow, shown in Table 12.

We could analyse these 5 diets as 5 different Qualitative levels and analyse this as a CRD. Alternatively, we could also treat this as a Regression problem, regressing milk yield (Y) on the amount of protein supplement in the diet. Which model is better, or most appropriate, or most parsimonious? Do we/should we consider the 5 Diets to be

Table 12: Milk production

Diet	Cow	Yield (kg)
1	1	11.1
1	2	16.6
1	3	15.6
2	4	18.1
2	5	22.8
2	6	21.6
3	7	25.2
3	8	30.1
3	9	26.7
4	10	28.6
4	11	33.7
4	12	35.8
5	13	34.5
5	14	31.4
5	15	33.6

quite separate, distinct 'levels' with whatever effect they each have, or are they simply the amounts/levels that we chose and which could best be described by a quantitative relationship??

If we analyse this as a CRD; the statistical model would be:

$$Y_{ij} = \mu + Diet_i + e_{ij}$$

17.1 Hypotheses

For the Model over and above the mean (our Diets), $R(Diet|\mu)$, our Null Hypothesis (H_0) is that there is no difference between the effects of the five diets, *i.e.* they are all equal, vs. the Alternative Hypothesis that there are differences, *i.e.* they are not all equal. Statistically we can write this as:

$$H_o \begin{bmatrix} d_1 - d_2 \\ d_1 - d_3 \\ d_1 - d_4 \\ d_1 - d_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad H_A \begin{bmatrix} d_1 - d_2 \\ d_1 - d_3 \\ d_1 - d_4 \\ d_1 - d_5 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

17.2 Analysis of Variance

<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F - ratio</i>
Total, TSS	$N = 15$	$Y'Y$ $= 10750.54$		
Model, SSR $= R(\mu, Diet)$	$r(X)$ $= 5$	$\tilde{b}'X'Y$ $= 10676.333$	2135.3	287.75
Mean, C.F.	1	$N\bar{y}^2$ $= 9902.21$	9902.21	1334.41
Model, after the mean, SSR_m $R(Diet Mean)$	$r(X) - 1$ $= 4$	774.123	193.531	26.08***
Error, SSE	$N - r(X)$ $= 10$	74.2066	7.42066	

The tabulated F-value for the model over and above the mean, *i.e.* the effects of Diets, is

$$F_{4,10,5\%} = 3.48$$

$$F_{\text{calc}} > F_{\text{tabulated}}$$

$$26.08 > 3.48$$

Therefore we can reject H_0 , and accept H_A , that the effects of the Diets are not all equal.

17.3 SAS code for Classifications

```
data cows;
input id diet kgsup my;
cards;
1 1 2 11.1
2 1 2 16.6
3 1 2 15.6
4 2 3 18.1
5 2 3 22.8
6 2 3 21.6
7 3 4 25.2
8 3 4 30.1
9 3 4 26.7
10 4 5 28.6
11 4 5 33.7
12 4 5 35.8
13 5 6 34.5
14 5 6 31.4
15 5 6 33.6
;

proc glm data=cows;
class diet;
model my = diet/xpx inverse;
estimate 'mu+d1' intercept 1 diet 1 0 0 0 0;
estimate 'mu+d2' intercept 1 diet 0 1 0 0 0;
estimate 'mu+d3' intercept 1 diet 0 0 1 0 0;
estimate 'mu+d4' intercept 1 diet 0 0 0 1 0;
estimate 'mu+d5' intercept 1 diet 0 0 0 0 1;
contrast 'C+Q' diet -1 2 0 -2 1,
           diet 1 -4 6 -4 1;
lsmeans diet/stderr;
run;
```

17.4 Fitted values

To estimate the fitted values for Diet 1 we need,

$$k' = (1 \ 1 \ 0 \ 0 \ 0 \ 0)$$

$$k'\tilde{b} = 14.433 \pm 1.573$$

To estimate the fitted values for Diet 2 we need,

$$k' = (1 \ 0 \ 1 \ 0 \ 0 \ 0)$$

$$k'\tilde{b} = 20.833 \pm 1.573$$

To estimate the fitted values for Diet 3 we need,

$$k' = (1 \ 0 \ 0 \ 1 \ 0 \ 0)$$

$$k'\tilde{b} = 27.333 \pm 1.573$$

To estimate the fitted values for Diet 4 we need,

$$k' = (1 \ 0 \ 0 \ 0 \ 1 \ 0)$$

$$k'\tilde{b} = 32.700 \pm 1.573$$

To estimate the fitted values for Diet 5 we need,

$$k' = (1 \ 0 \ 0 \ 0 \ 0 \ 1)$$

$$k'\tilde{b} = 33.167 \pm 1.573$$

See the SAS statements and output for the ESTIMATE statements and the results for the fitted values of $(\mu+d_i)$; these are commonly called Least Squares means. Note that this term should more correctly be called estimates of means calculated using the method

of least squares!. SAS calls these lsmeans, and LSMEANS is an option statement in PROC GLM (and many other SAS procedures).

We can see that the fitted values for all the observations pertaining to Diet 1 have the same estimate/value, and similarly for the other levels of diets.

We can see that the standard errors of the estimates are the same for all Diets, because all diets have the same number of observations (cows) and hence the same amount of 'information' (in the statistical sense).

17.5 Least Squares Means

The LSMeans for the 5 diets are:

Diet	LSMeans	s.e.
1	14.433	± 1.573
2	20.833	± 1.573
3	27.333	± 1.573
4	32.700	± 1.573
5	33.167	± 1.573

17.6 SAS Output, classification model

The GLM Procedure

Class Level Information		
Class	Levels	Values
diet	5	1 2 3 4 5

Number of observations	15
------------------------	----

The GLM Procedure

Dependent Variable: my

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	774.1226667	193.5306667	26.08	<.0001
Error	10	74.2066667	7.4206667		
Corrected Total	14	848.3293333			

R-Square	Coeff Var	Root MSE	my Mean
0.912526	10.60232	2.724090	25.69333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
diet	4	774.1226667	193.5306667	26.08	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
diet	4	774.1226667	193.5306667	26.08	<.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
C+Q	2	7.77504762	3.88752381	0.52	0.6076

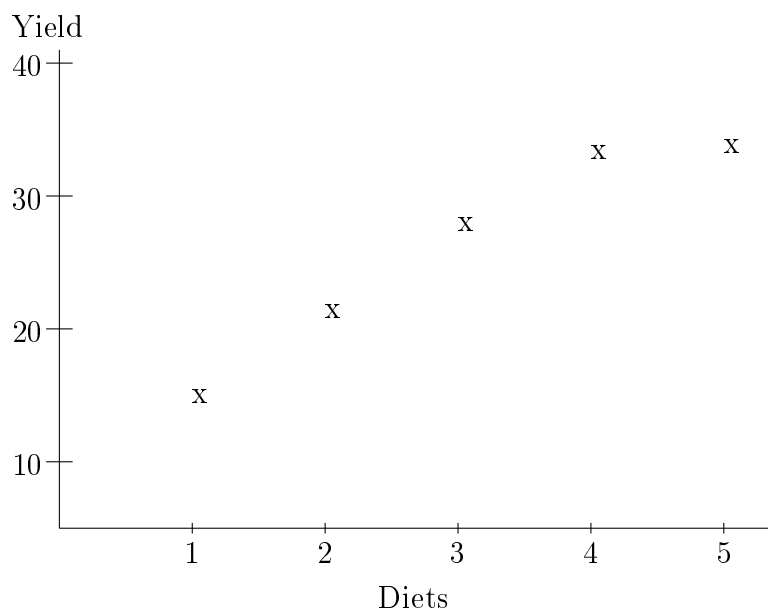
Parameter	Estimate	Standard Error	t Value	Pr > t
mu+d1	14.4333333	1.57275413	9.18	<.0001
mu+d2	20.8333333	1.57275413	13.25	<.0001
mu+d3	27.3333333	1.57275413	17.38	<.0001
mu+d4	32.7000000	1.57275413	20.79	<.0001
mu+d5	33.1666667	1.57275413	21.09	<.0001

The GLM Procedure

Least Squares Means

diet	my LSMEAN	Standard Error	Pr > t
1	14.4333333	1.5727541	<.0001
2	20.8333333	1.5727541	<.0001
3	27.3333333	1.5727541	<.0001
4	32.7000000	1.5727541	<.0001
5	33.1666667	1.5727541	<.0001

If we plot the Least Squares Means we have the following picture:



There is an obvious pattern of an increasing milk yield with increasing Protein supplement in the diet. Is there a 'curvi-linear' trend or just a simple linear trend, or is the classification model significantly better than a regression model? We cannot determine this yet, but that is what we want to address (and know).

With 5 diets we have 4 degrees of freedom for Diets. If we think about the 5 Least Squares Means from a geometry perspective we know that with 5 points we can fit a 4th order polynomial that will provide a perfect fit through the five points. Thus a 4th order polynomial (an equation of the form $Y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_1^3 + b_4x_1^4 + e$) is equivalent to fitting the 5 diets as 5 classes (levels); it would provide exactly the same fitted values and the **SAME** Sums of Squares (SSR_m) (try it and see).

We can therefore partition the Sums of Squares for Diets, and their 4 degrees of freedom, into the 4 effects due to Linear, Quadratic, Cubic and Quartic polynomial regression effects.

Why? We want to see if there is any indication of a Linear effect, or a Quadratic effect, or a [Cubic effect+Quartic effect]. It may be difficult to justify, from a biological viewpoint, anything beyond a quadratic effect. However, if we use orthogonal polynomials, which are independent of one another, then we can test whether there is a statistically

significant improvement in going from (for example) Linear and Quadratic effects to 5 classes (equivalent to L, Q, C and Q regressions). Note, we are NOT interested in Cubic or Quartic effects *per se*, we only want to see if the classification effect has a significant improvement to the fit of the overall model, over and above Linear and/or Quadratic regressions. Basically we are fitting a complete model, with 4 degrees of freedom for Diets and we want to see which path to take, that with 5 levels (4 d.f.) or whether a simpler regression model is the path to explore.

We can use orthogonal contrast to divide up the Sums of squares for Diets. Note, this is true, if and only if, there is an equal 'spacing' between the quantitative factor levels that we are considering, which is true in this example. If it was not true then we would need to use the 'over-parameterized model' approach (see section 17.9).

For a linear regression we can set up the coefficients of the 5 classes effects as:

Diets	D ₁	D ₂	D ₃	D ₄	D ₅
Linear	-2	-1	0	+1	+2

These give us a linear progression from the first diet to the fifth diet. We can set up the coefficients for the Quadratic effect as:

Diets	D ₁	D ₂	D ₃	D ₄	D ₅
Quadratic	2	-1	-2	-1	+2

Similarly, the cubic and quartic effects coefficients are:

Diets	D ₁	D ₂	D ₃	D ₄	D ₅
Cubic	-1	2	0	-2	+1
Quartic	1	-4	6	-4	1

Note that each of these 4 effects (L, Q, C and Q) are orthogonal to one another. We can see that they are orthogonal since $k'_i k_j = 0$.

Since the Cubic and Quartic contrasts are independent of the other effects they serve to test whether the 4 degrees of freedom for the 5 classes (or the 4th order polynomial) provides a better fit than Linear and Quadratic regressions, *i.e.* over and above Linear and Quadratic regressions.

Note, the above statements are true, and only true, if the spacing between diets/treatments is equal, *i.e.* Diet 1 (2 kg), Diet 2 (3 kg), Diet 3 (4 kg), Diet 4 (5 kg) and Diet 5 (6 kg) all differ by increments of 1 kg. If the spacings between levels are not equal then constructing the appropriate SAS CONTRAST coefficients becomes 'difficult'!

Thus, to determine whether a classification model (with 5 levels, 4 d.f.) is necessary or not, or whether a simpler quantitative model (which could have linear and quadratic effects) is sufficient we need to see whether there is any improvement (significant) due to the effects over and above the linear and quadratic, *i.e.* the cubic and quartic, *i.e.* $R(C+Q|\mu, L, Q)$, thus we want to test whether they add statistically significantly to the goodness-of-fit of the model. In words, our Null Hypothesis (H_o) is that the Cubic & Quartic components together have no effect; our Alternate Hypothesis (H_A) is that they do have an effect.

$$H_o \begin{bmatrix} C \\ Q \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \equiv \quad H_A \begin{bmatrix} b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

vs.

$$H_o \begin{bmatrix} C \\ Q \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \equiv \quad H_A \begin{bmatrix} b_3 \\ b_4 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

17.7 SAS CONTRAST code

```
/* SAS code for L+Q regressions */
contrast 'Diets, L+Q' Diet -2 -1 0 1 2,
           Diet 2 -1 -2 -1 2;
```

```

/* SAS code for the C+Q effect, over and above the L+Q */
contrast 'Diets, C+Q' Diet -1 2 0 -2 1,
           Diet 1 -4 6 -4 1;

```

If we use these contrast statements we obtain the following, for the Marginal effect of the Cubic and Quartic effects (combined, since we have no interest in them individually, only as the 'over and above regressions' effect):

Effect	d.f.	S.S.	M.S.	F-ratio	Pr >F
Diets	4	774.123	193.531	26.08	.0001
Diets, C+Q	2	7.775	3.8875	0.52	n.s.s.

The contrast of the Cubic + Quartic is $R(b_3, b_4|b_0, b_1, b_2)$ and is not statistically significant; so we shall accept H_o . Thus we would conclude that fitting Diets as 5 classes does not provide a significantly better fit than a more simple (parsimonious) model with just Linear and Quadratic regressions on the amount of protein in the diet.

Conclusions. We accept H_o , that the Cubic and Quartic components are not statistically significant and hence the model with 5 treatment levels (of Diet) (qualitative model) is not demonstrably better than the model with linear and quadratic regressions (although from these analyses so far we cannot tell if the quadratic component of a quantitative regression model would be significant). So, we would then re-analyse as a multiple regression model with Linear and Quadratic regressions, and proceed.

17.8 Coefficients for Orthogonal Polynomials

Linear, Quadratic & Higher components. Subdivide Sums of Squares for an effect into Linear, Quadratic, etc. See STD Ch 15.7 P386+

treatments	Degree of Polynomial	T ₁	T ₂	T ₃	T ₄	T ₅	
2	1	-1	+1				
3	1	-1	0	+1			
	2	+1	-2	+1			
4	1	-3	-1	+1	+3		
	2	+1	-1	-1	+1		
	3	-1	+3	-3	+1		
5	1	-2	-1	0	+1	+2	
	2	+2	-1	-2	-1	+2	
	3	-1	+2	0	-2	+1	
	4	+1	-4	+6	-4	+1	
6	1	-5	-3	-1	1	3	5
	2	5	-1	-4	-4	-1	5
	3	-5	7	4	-4	-7	5
	4	1	-3	2	2	-3	1
	5	-1	5	-10	10	-5	1

If one does not have equally spaced treatments (and even if one does) it is always possible to fit the linear and quadratic regressions and in addition to fit the classification variable after the regressions. In this way one 'automagically' obtains the additional effect due to the extra classes over and above the regression effects.

17.9 Over-parameterized model (GLM)

To fit an over-parameterized model we have to have a regression variable AND a classification variable to fit together in our model. When we read in our data set we had read in an additional variable, kgsup, which we have not so far used; now we shall;

```
proc glm data=cows;
```

```
class diet;  
model my = kgsup kgsup*kgsup diet;  
run;
```

This is a model with the treatment (Diet effect) fitted as 2 different effects, Diet, as a classification effect, and kgsup as a quantitative regression effect, as Linear and Quadratic regressions.

The statistical model will be

$$Y_{ij} = \mu + b_1X_{ij} + b_2X_{ij}^2 + Diet_i + e_{ij}$$

where X_{ij} is the amount of protein supplement fed to the j^{th} cow on the i^{th} diet, and b_1 and b_2 are, as for our multiple regression type models, the regression coefficients for the regression of milk yield on Kg of supplement and Kg^2 .

Then the Sums of Squares for Diet, over and above the linear and quadratic regressions, will have 2 degrees of freedom ($4 - 2 = 2$) and will correspond to the Sums of Squares obtained using the orthogonal contrasts method for the Marginal effect over and above the Regressions.

17.10 SAS Output, over-parameterized model

The GLM Procedure

Class Level Information		
Class	Levels	Values
diet	5	1 2 3 4 5

Number of observations	15
------------------------	----

The GLM Procedure

Dependent Variable: my

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	774.1226667	193.5306667	26.08	<.0001
Error	10	74.2066667	7.4206667		
Corrected Total	14	848.3293333			

R-Square	Coeff Var	Root MSE	my Mean
0.912526	10.60232	2.724090	25.69333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
kgsup	1	730.1333333	730.1333333	98.39	<.0001
kgsup*kgsup	1	36.2142857	36.2142857	4.88	0.0516
diet	2	7.7750476	3.8875238	0.52	0.6076

Source	DF	Type III SS	Mean Square	F Value	Pr > F
kgsup	0	0.00000000	.	.	.
kgsup*kgsup	0	0.00000000	.	.	.
diet	2	7.77504762	3.88752381	0.52	0.6076

17.11 Interpretation of over-parameterized model

Note several things about these results, as compared with those from the classification model. The Sums of Squares for the Model, Mean and Model over and above the Mean, and the Residual have not changed. There is no new 'information' we are just dividing up the Sums of Squares for the effect of diet differences into a quantitative component, and a qualitative component over and above the quantitative component.

Note that the Type III Sums of Squares, the Marginal Sums of squares, for Diet have 2 degrees of freedom (4 d.f. from the differences between diets - the 2 d.f. accounted for by the Linear and Quadratic regressions). The Marginal Sums of Squares for (Diet| μ, L, Q) are 7.7750, EXACTLY as for the orthogonal contrast for the [Cubic+Quartic].

Note that the Type III degrees of freedom and Sums of Squares for $kgsup$ and $kgsup^2$ are both Zero. This is because

$$R(b_1|\mu, diet) \equiv R(b_1|\mu, b_1, b_2, b_3, b_4)$$

which is attempting to fit b_1 after already fitting b_1 (implicitly in the diet classification effect), and hence there is nothing left to fit, \Rightarrow d.f. = 0 and $SS_{b_1|diet} = 0$. Likewise for the degrees of freedom and Sums of Squares for $kgsup * kgsup$.

Thus our conclusions agree with those from the orthogonal contrasts model. This method of over-parameterizing the model will always work, whereas the orthogonal contrasts model is only valid for the case of equal interval spacings. In addition, if there are a large number of levels it can become quite complex to obtain all the necessary coefficients for orthogonal contrasts. Thus, it is simpler, and always valid, to use the over-parameterized model.

17.12 Simplification of over-parameterized model

Since we have determined that the 2 additional degrees of freedom, for the classification model, did not add significantly to the goodness-of-fit of the model, we shall drop the clas-

sification approach and consider a regression model, with linear and quadratic regressions on the amount of protein supplement fed (kgsup).

Thus we shall next try the following statistical model:

$$Y_i = b_0 + b_1X_i + b_2X_i^2 + e_i$$

The first step will therefore be, just as when we started looking at multiple regressions, to determine whether the quadratic effect, b_2 , is statistically significant, using whatever probability level we are using as our criterion for accepting or rejecting H_0 ; we shall use a 5 % probability level. If we look at the results of the GLM analysis, shown in the section 17.13, we see that the probability for $kgsup^2$ is 0.040 and the calculated F-value is 5.30. Thus, directly using the probability provided by GLM we shall conclude that b_2 is statistically significantly different from Zero. Similarly, if we look up the tabulated F-value we find that it is 4.75; thus since our calculated value exceeds the tabulated value we would reject H_0 .

Thus, since the quadratic effect is statistically significant it is neither necessary, nor sensible to examine the statistical significance of the linear component; we SHALL retain it in the model since it is necessary. So we have determined that the effect of protein supplement has a quantitative effect, and that it is curvi-linear quadratic in nature. Our best estimate of the regression equation will be:

$$\hat{Y} = -7.04 + 12.362 * Kg_{\text{supplement}} - 0.9286 * Kg_{\text{supplement}}^2$$

17.13 SAS Output from regression model

The GLM Procedure

Number of observations	15
------------------------	----

The GLM Procedure

Dependent Variable: my

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	766.3476190	383.1738095	56.09	<.0001
Error	12	81.9817143	6.8318095		
Corrected Total	14	848.3293333			

R-Square	Coeff Var	Root MSE	my Mean
0.903361	10.17296	2.613773	25.69333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
kgsup	1	730.1333333	730.1333333	106.87	<.0001
kgsup*kgsup	1	36.2142857	36.2142857	5.30	0.0400

Source	DF	Type III SS	Mean Square	F Value	Pr > F
kgsup	1	98.13915829	98.13915829	14.37	0.0026
kgsup*kgsup	1	36.21428571	36.21428571	5.30	0.0400

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-7.04000000	5.99840508	-1.17	0.2633
kgsup	12.36190476	3.26161055	3.79	0.0026
kgsup*kgsup	-0.92857143	0.40331394	-2.30	0.0400

17.14 Conclusion

So, in summary, the procedure that we can follow in determining which model is most appropriate can be describing in the following table:

1. Fit a model to decide which path/route to take; a classification model, or a regression model.

Fit a Classification Model
+ orthogonal contrasts.
Partition Sums of Squares
into Linear, Quadratic,
The Rest (i.e. Cubic, Quartic,
Quintic, etc)

Fit an over-parameterized model;
regressions+classification (The Rest,
over and above regressions)

Obtain Sums of Squares and
test of significance
of The Rest

Obtain Sums of Squares and
test of significance
of The Rest

2. If the test of significance of (The Rest) is s.s. **THEN** revert back to a classification model.

If the test of significance of (The Rest) is n.s.s. then a classification model is not a better fit over and above a regression model. So, drop the classification model and proceed with a regression analysis.

3. Report your results, *i.e.* show that you considered and compared Regression trends (Quantitative) and Classification (Qualitative).

18 Normality and Homogeneity of Variance

Bartlett's Test (STD Ch. 19.3, P480)

See also the Stats II Web page section about Normality and Homogeneity of Variances (The Sequel!).

18.1 Requirements for Analysis of Variance

See Steel, Torrie and Dickey, Chapter 7.10. So far we have said that the data we are analysing must have equal variances (*i.e.* the variances of each observation must be homogeneous \equiv lack of heterogeneity) and that the observations must be normally distributed. These are quite important assumptions and requirements. When analysing data from an experiment it is incumbent upon the researcher to check and verify these conditions.

18.2 Normality

We can test for normality by computing the 3rd and 4th moments about the mean. N.B. if you have forgotten about this (from Statistical Methods I) the first moment about the mean is simply the mean itself, the second moment about the mean provides us with the variance (and standard deviation), the third moment about the mean measures the skewness and the fourth moment about the mean measures the kurtosis (Snedecor and Cochran, Statistical Methods, Ch 3.13 and 3.14).

PROC UNIVARIATE in SAS provides a test for normality, the W statistic. As an example, consider that we have a group of pigs that we grow from weaning up to 100 days of age and then we weigh them. The weights are given below, in Table X, for male and female pigs, m and f respectively. However, the test for Normality is only valid when the model $Y_i = \mu + e_i$ applies to ALL experimental units, *i.e.* no fixed effects/treatments, etc!

Our Null Hypothesis is that the **VAR**iable is Normally distributed, and our Alternate Hypothesis is that the trait is not Normally distributed.

sex	Wt.	sex	Wt.	sex	Wt.	sex	Wt.	sex	Wt.	sex	Wt.	sex	Wt.
m	94.3	m	83.3	m	84.0	f	79.9	f	83.8	f	80.8	f	76.8
f	76.5	f	79.7	f	80.9	f	81.1	m	79.0	f	81.4	m	86.9
f	85.0	f	76.4	m	84.5	m	92.1	m	91.6	m	84.8	f	81.0
f	78.6	f	79.1	f	81.2	f	81.3	f	81.2	f	76.7	m	89.2
m	82.9	m	88.4	m	86.7	m	86.4	f	81.7	f	78.2	m	82.1
m	85.2	m	88.0	m	89.8	m	85.2	m	88.6	f	85.8	m	86.9
m	84.4	f	74.0	m	88.2	m	86.2	m	86.1	f	77.5	f	84.0
f	79.4	f	78.2	f	81.1	m	93.1	f	76.5	m	83.9	f	83.7
m	92.0	f	73.1	f	77.8	f	79.4	m	89.1	f	85.3	m	88.7
m	84.4	m	89.1	m	90.7	m	81.5	f	78.3	f	81.4	m	86.2
f	78.7	m	79.0	f	79.8	f	81.2	m	84.3	f	85.2	f	77.8
m	87.9	m	78.0	f	77.0	m	83.2	m	89.1	f	79.0	f	80.8
f	79.0	m	85.0	f	79.7	m	87.9	f	74.0	f	88.6	f	77.1
m	84.3	m	74.7	f	75.7	f	75.0	f	82.2	m	85.6	m	85.6
m	88.2	f	87.1	m	90.7	m	83.7	f	82.0	m	89.9	m	88.0
m	80.9	m	87.7	f	76.7	f	80.5	m	83.4	f	86.6	f	75.9
f	85.4	f	75.5	m	79.0	m	92.6	f	80.1	m	93.9	f	87.8
f	78.2	m	86.5	f	72.7	f	84.3	m	91.2	f	78.5	m	91.1
m	85.4	m	82.8	m	93.6	f	88.7	m	90.1	m	87.2	f	74.5
f	79.9	m	90.1	f	82.0	f	79.9	f	87.1	m	84.8	f	83.4
f	77.0	f	84.4	m	91.5	f	86.3	m	93.1	m	89.0	f	84.2
m	84.0	f	83.5	f	78.9	f	82.1	m	88.6	m	85.9	f	80.4
m	91.0	f	84.0	f	82.6	f	73.0	f	82.0	m	88.3	m	82.9
f	80.4	m	82.9	f	82.4	f	83.9	f	73.2	m	82.9	m	94.0
f	78.4	m	82.0	m	89.0	f	84.4	m	86.8	f	77.2	m	83.7
f	81.7	m	93.3	m	80.4	f	83.4	f	78.2	m	83.2	m	87.0
m	80.9	m	88.1	m	87.8	m	84.8	f	78.7	f	78.9	f	79.4
f	80.9	f	79.6	f	79.3	m	83.9	f	82.4	f	77.3	f	82.2
m	80.7	f	80.8	f	81.8	f	80.9						

The following SAS code shows how to read the data for all pigs into a SAS data set (pigs) and then subset the data into two new data sets, for male and female pigs.

```
USING SAS/PROC UNIVARIATE
```

```
data pigs; /* read weights for all pigs into data set pigs */
```

```

input sex $ wt;
cards;
m 94.3
f 76.5
f 85.0
. .
. .
f 81.8
f 80.9
;

data mpigs; /* new data set to be male pigs only */
set pigs; /* copy from data set pigs */
if (sex eq 'm'); /* keep records only if sex = male */
run;

data fpigs; /* new data set to be female pigs only */
set pigs; /* copy from data set pigs */
if (sex eq 'f'); /* keep records only if sex = female */
run;

proc print data=pigs;
run;

proc print data=mpigs;
run;

proc print data=fpigs;
run;

proc univariate data=mpigs normal; /* test for normality */
var wt;
histogram wt;
run;

proc univariate data=fpigs normal;
var wt;
histogram wt;
run;

/* Alternatively */

```

Table 13: Results of testing for Normality, sexes separately

sex	W statistic	Probability
m	0.979059	0.4974
f	0.97656	0.3390

```
proc sort data=pigs;
by sex;
run;

proc univariate data=pigs normal;
by sex;
var wt;
histogram wt;
run;
```

The W statistic and its associated probability gives us a test of whether we can accept the Null Hypothesis that the data are from a Normal distribution or not. If the Probability is less than our specified (prior) cut-off level then we reject the Null Hypothesis and conclude that the data are **NOT** Normally distributed. For these two data sets, male and female pigs respectively, we find that there is no evidence to reject the Null Hypothesis that the data are Normally distributed, *i.e.* we can accept that the data are Normally distributed.

We can see that both probabilities are greater than, for example, 0.05 (if we are using 5 % as our criterion for accepting/rejecting the Null Hypothesis). Thus we would conclude that these data are Normally distributed; which is a quite acceptable conclusion for a quantitative, growth-related trait.

If weight at 100 days in male and female pigs is Normally distributed when we look at the weight in males and in females separately, then it would be logical to assume that together they would be Normal. Is this the case? If we take the same data and use PROC UNIVARIATE on the original combined data (males and females together) we obtain a W statistic of 0.984 and an associated Probability of 0.0205. Thus the results of a simple

Table 14: Incorrect test for Normality, raw data

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.983706	Pr < W	0.0205
Kolmogorov-Smirnov	D	0.06773	Pr > D	0.0238
Cramer-von Mises	W-Sq	0.130714	Pr > W-Sq	0.0443
Anderson-Darling	A-Sq	0.792007	Pr > A-Sq	0.0411

PROC UNIVARIATE analysis would tell us that it is **UNLIKELY** that this data (the combined male and female weights) come from a Normal distribution! This seems to be at variance with what we would intuitively presume, and in fact this is the case; the data are Normally distributed, BUT the simple PROC UNIVARIATE test for Normality on the combined data set is **INCORRECT**. It is incorrect because of the fact that we have not considered the fixed effects (the sex effect) and this causes the results to be completely invalid; the model for the above-mentioned swine data should include the sex effect, *i.e.* $Y_{ij} = \mu + sex_i + e_{ij}$. Thus we see that using PROC UNIVARIATE (which will ignore any fixed effects) will be completely invalid. What should we do? What we need to test is the distribution of the residuals (the e_{ij} 's); thus we should use the residuals for testing for Normality.

How can we, in general, obtain the residuals? Well we first of all have to fit a model which is appropriate (the appropriate model, not just any old model!), and then output from this the residuals, and then we can use these as the input to PROC UNIVARIATE. So, for the above data, we consider that the appropriate model to describe the weight of the piglets (male and female) would be:

$$Y_{ij} = \mu + sex_i + e_{ij}$$

We could therefore fit the following model, using PROC GLM

```

/* fit model wt = mu + sex + e
   output to SAS dataset pigresids observations + yhat + ehat
*/
proc glm data=pigs;
class sex;
model wt = sex;
output out=pigresids p=yhat r=ehat;
run;

/* print out observations, so we can see what we've got */
proc print data=pigresids;
var sex wt yhat ehat;
run;

/* use dataset pigresids as input to proc univariate *
proc univariate data=pigresids normal; /* test for normality */
var ehat;
histogram ehat;
run;

```

If we want to use PROC MIXED the format for outputting the residuals and fitted values (ehats and yhats respectively) is a little different:

```

/* fit model wt = mu + sex + e
   output to SAS dataset pigresids observations + yhat + ehat
*/
proc mixed data=pigs;
class sex;
model wt = sex/outp=pigresids;
run;

/* print out observations, so we can see what we've got */
proc print data=pigresids;
run;

/* use dataset pigresids as input to proc univariate *
proc univariate data=pigresids normal; /* test for normality */
var Resid;
run;

```


Table 15: Correct test for Normality, residuals, after fitting sex

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.993547	Pr < W	0.5349
Kolmogorov-Smirnov	D	0.03274	Pr > D	0.150
Cramer-von Mises	W-Sq	0.026804	Pr > W-Sq	0.2500
Anderson-Darling	A-Sq	0.225772	Pr > A-Sq	0.2500

From the above PROC GLM model, outputting the residuals and then testing for Normality we obtain the following combined result:

We can see that the combined test of the residuals shows no evidence for the residuals not being Normally distributed, the probability is 0.5349, substantially above our 5% or 1% 'cut-off' level i.e. we can accept our Null Hypothesis, which was that the errors **ARE** Normally distributed, and that hence our ANOVA assumption is valid.

18.3 Homogeneity of Variance

This is where the use of PROC MIXED in SAS really comes into its own. We can use PROC MIXED to test for heterogeneity of variance and, if there is heterogeneity, to make a correct analysis, accounting for the heterogeneous variances. If we look at most statistics text books, such as STD, we find that a suitable test for heterogeneity is Bartlett's test (STD, Ch 19.3). However, it is somewhat limited and is only suitable for quite simple designs, for example One-way CRD models. In addition, if we use Bartlett's test and find that the variances are heterogeneous then the recommendation is that a transformation is necessary (STD Ch 9.16). What transformation and whether the transformation will make the variances homogeneous is skated over. PROC MIXED will often allow us to make a much better analysis accommodating the heterogeneity directly in the model, an altogether desirable approach, since it obviates the necessity to find some (arbitrary) transformation. We shall use the Completely Randomized Design data to illustrate the test of heterogeneity of variance. The method that we shall use is "Maximum Likelihood", a very powerful technique. Basically, Maximum Likelihood says "What estimable parameters would be most likely to give us the data that we have obtained?" We can hypothesise a model, for example our CRD, with a mean μ parameters for the treatment effects and a residual variance. We compute the Likelihood (actually the natural logarithm of the likelihood) associated with this model. We can fit another model, for example a model with a mean, with 6 treatment effects (as before), and now a residual variance specific to observations in each treatment group (*i.e.* 6 residual variances) and compute the likelihood (as before actually the log likelihood). The model with the best [Ln] Likelihood is the best fitting model. Fisher showed that -2 the difference in the Log Likelihood of models has a χ^2 distribution. Thus we can use this for testing the change in goodness-of-fit between models. PROC MIXED, allows us to use Restricted Maximum Likelihood (**REML**) methods.

See STD, Ch 22.3 for some comments about Maximum Likelihood.

18.4 SAS code for Homogeneity of Variance

We can fit a model with 1 common, pooled, homogeneous residual variance, and another model with 6 different residual variances and compared the fit of the models, via REML. We read in the data and then use the **REPEATED** statement of PROC MIXED and the

GROUP= option to fit different residual variances for each **GROUP**.

```
data oneway;
input obs trt y;
sg = 1;
if (trt ge 4) then sg = 2;
cards;
1 1 19.4
2 1 32.6
3 1 27.0
4 1 32.1
5 1 33.0
6 2 17.7
7 2 24.8
8 2 27.9
9 2 25.2
10 3 17.0
11 3 19.4
12 3 9.1
13 3 11.9
14 4 20.7
15 4 21.0
16 4 20.5
17 4 18.8
18 4 18.6
19 5 14.3
20 5 14.4
21 5 11.8
22 5 11.6
23 5 14.2
24 6 17.3
25 6 19.4
26 6 19.1
27 6 16.9
28 6 20.8
;

/* fit 1 common residual */
proc mixed data=oneway;
classes trt;
model y = trt;
```

```

run;

/* for 6 residuals, via the group statement
   NOTE, for the repeated statement, the data MUST be sorted
   in the order of the repeated variable, so run proc sort
   by trt first
*/
proc sort data=oneway;
  by trt;
run;

proc mixed data=oneway;
classes trt;
model y = trt;
repeated /group=trt;
run;

```

We have fitted 2 models, with the same fixed effects. The only difference is that in the first model we fit 1 common error variance for all observations, since we have not specified anything else. Thus there is 1 degree of freedom associated with the 1 parameter estimate of the error variance. For the second model we fit separate error variances for the 6 treatment groups, thus there are 6 d.f. associated with the 6 error variances. We can compare the models to see whether there is an improvement in the goodness of fit of the model by going from the 1 simple common variance to the more complicated model with 6 variances, one for each group. If we look at the [Schwartz's] Bayesian Criteria (BIC, or SBC) for each model then the model with the smallest BIC value is the model which has the best variance structure.¹ This is the way to compare 2 models which have the same fixed effects structure, but which differ in the variance-covariance structure (the Random Effects).

This is all subject to the proviso that a BIC difference of less than 3 provides little evidence for any real, substantive difference. A difference of 3 to 5 may be considered to provide moderate evidence of real differences, whilst a BIC difference of 5 to 8 could be

¹N.B. In versions of SAS up to v8.1 the criterion was the model with the highest SBC/BIC value was the best. In SAS v8.2 and above the criterion has been reversed; lower is better; it is written "(lower is better)".

considered to provide good evidence of differences, and BIC differences of more than 8 could be considered to be strong evidence of differences.

18.5 SAS output from PROC MIXED, homogeneous variances

The Mixed Procedure

Model Information	
Data Set	WORK.ONEWAY
Dependent Variable	y
Covariance Structure	Diagonal
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Residual

Class Level Information		
Class	Levels	Values
trt	6	1 2 3 4 5 6

Dimensions	
Covariance Parameters	1
Columns in X	7
Columns in Z	0
Subjects	1
Max Obs Per Subject	28
Observations Used	28
Observations Not Used	0
Total Observations	28

Covariance Parameter Estimates	
Cov Parm	Estimate
Residual	12.7781

Fit Statistics	
-2 Res Log Likelihood	127.7
AIC (smaller is better)	129.7
AICC (smaller is better)	129.9
BIC (smaller is better)	130.8

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
trt	5	22	12.72	<.0001

We should, at this stage, concentrate on the “Fit Statistics”, the -2LnL (127.7), AIC (129.7) and BIC (130.8) values; note them. In and of themselves they tell us very little, we need to compare them with the corresponding Fit Statistics from the next model; the model with heterogeneous residual variances.

18.6 SAS output from PROC MIXED, heterogeneous variances (6)

The Mixed Procedure

Model Information	
Data Set	WORK.ONEWAY
Dependent Variable	y
Covariance Structure	Variance Components
Group Effect	trt
Estimation Method	REML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Class Level Information		
Class	Levels	Values
trt	6	1 2 3 4 5 6

Dimensions	
Covariance Parameters	6
Columns in X	7
Columns in Z	0
Subjects	28
Max Obs Per Subject	1
Observations Used	28
Observations Not Used	0
Total Observations	28

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	127.69374108	
1	1	111.40787380	0.00000000

Convergence criteria met.

Covariance Parameter Estimates		
Cov Parm	Group	Estimate
Residual	trt 1	33.6420
Residual	trt 2	18.9800
Residual	trt 3	22.0300
Residual	trt 4	1.2770
Residual	trt 5	2.0380
Residual	trt 6	2.5650

Fit Statistics	
-2 Res Log Likelihood	111.4
AIC (smaller is better)	123.4
AICC (smaller is better)	129.0
BIC (smaller is better)	131.4

Table 16: Comparison of Full and Reduced models

Parameters	-2LnL	AIC	BIC
1 common variance, σ_e^2	-2LnL _r = 127.7	129.7	130.8
6 residual variances, $\sigma_{e_i}^2$ i = 1,...,6	-2LnL _f = 111.4	123.4	131.4
2LnL _f - 2LnL _r		16.3	

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
5	16.29	0.0061

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
trt	5	22	19.53	<.0001

We shall consider the first model, with 1 common pooled homogenous residual variance to be the 'Reduced model', and the second model with more parameters (6 residual variances) to be the 'Full model'. If we look at the output (-2LnL's) from the two models (see table 16) we see that the difference in -2 log likelihood is 16.3 (127.7-111.4). From standard maximum likelihood theory (all give thanks to Fisher!) we know that this difference has a χ^2 distribution. Thus, we can see that the degrees of freedom for this χ^2 are equal to the difference in the number of parameters (of the variance components for the two models), *i.e.* 6-1=5. Therefore, we compare our χ^2_{calc} (16.3) against the χ^2_{tab} for 5 degrees of freedom and our specified probability level. If we choose to use our ubiquitous 5% probability level we find that the tabulated χ^2 is 11.1. Our calculated χ^2 of 16.3 is greater than the tabulated value, so we can reject the Null Hypothesis, which is that the variances in the 6 groups are homogeneous. So we shall accept the Alternate Hypothesis that they are heterogeneous.

If we look at the residual variances (see Table 17) we can see that there appears to be a pattern, the first 3 treatment groups have residual variances somewhat higher than those of the other treatment groups. Perhaps we do not need 6 residual variances, but could

Table 17: Residual variances, by treatment

Treatment	Variance
1	33.642
2	18.980
3	22.030
4	1.277
5	2.038
6	2.565

have a model with 2 residual variances, one for treatments 1, 2 and 3, and another residual variance for treatments 4, 5 and 6.

When we read the data in, we created a new variable (**SG**) with a value of 1 or 2 depending upon whether the treatment was less than or equal to 3 or greater. Now we can use this to specify a grouping as:

18.7 SAS code for 2 Residual Variances

```

/* for 2 residuals, via the group statement,
   NOTE, for the repeated statement, the data MUST be sorted
   in the order of the repeated variable, so run proc sort
   by sg first
*/
proc sort data=oneway;
  by sg;
run;

proc mixed data=oneway;
  classes trt sg;
  model y = trt;
  repeated /group=sg;
run;

```

18.8 SAS output from PROC MIXED, heterogenous variances (2)

The Mixed Procedure

Model Information	
Data Set	WORK.ONEWAY
Dependent Variable	y
Covariance Structure	Variance Components
Group Effect	sg
Estimation Method	REML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Class Level Information		
Class	Levels	Values
sg	2	1 2
trt	6	1 2 3 4 5 6

Dimensions	
Covariance Parameters	2
Columns in X	7
Columns in Z	0
Subjects	28
Max Obs Per Subject	1
Observations Used	28
Observations Not Used	0
Total Observations	28

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	127.69374108	
1	1	112.20712088	0.00000000

Table 18: Comparison of Various models

Parameters	-2LnL	AIC	BIC
1 common variance, σ_e^2	127.7	129.7	130.8
2 residual variances, $\sigma_{e_1}^2, \sigma_{e_2}^2$	112.2	116.2	118.9
6 residual variances, $\sigma_{e_i}^2$	111.4	123.4	131.4

Convergence criteria met.

Covariance Parameter Estimates		
Cov Parm	Group	Estimate
Residual	sg 1	25.7598
Residual	sg 2	1.9600

Fit Statistics	
-2 Res Log Likelihood	112.2
AIC (smaller is better)	116.2
AICC (smaller is better)	116.8
BIC (smaller is better)	118.9

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
1	15.49	<.0001

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
trt	5	22	19.42	<.0001

Using the BIC statistic to directly compare the models, we can see quite clearly that the model with 2 residual variances (for treatments 1-3, and 4-6) is the best. Not only is it the lowest, but it is also $130.8 - 118.9 = 11.9$ units different; as we noted before, this

lends strong evidence that there is in fact heterogeneity of the residual variances between the two defined groupings. If we were to use the AIC statistics we would arrive at the same conclusion. If we wish to be 'classical' and use the χ^2 approach we cannot look at the -2LnL values and immediately conclude which model is best, we have to test the difference(s) against the tabulated χ^2 . How?

Consider going from 1 residual variance to 2 residual variances; this is a difference of $(127.7 - 112.2 = 15.5)$. This is a difference of $(2 - 1 = 1)$ parameters, and hence 1 d.f.; therefore we look up our tabulated χ^2 for 1 d.f. and 5 % probability, the critical, tabulated value is 3.84. Our calculated χ^2 exceeds this, thus we can conclude that there is a statistically significant improvement in the fit of the model by going from 1 residual variance to 2 residual variances. Likewise, if we examine the change in fit by going from 2 residual variances to 6 residual variances we obtain a calculated χ^2 of $(112.2 - 111.4 = 0.8)$, for $(6-2=4)$ degrees of freedom. The tabulated χ^2 for 5 % and 4 d.f. is 9.49. Thus, our $\chi^2_{\text{calc}} < \chi^2_{\text{tab}}$, therefore we conclude that there is no more significant improvement in goodness of fit from going from 2 residual variances to 6 residual variances. Conclusion: 2 residual variances provide the best-fitting model. Now we can look at the F-test of the fixed effects, lsmeans, estimates of differences, etc, etc.

Note how, to use the classical χ^2 we had to keep track of the changes of degrees of freedom associated with the number of random effect parameters. It seems much easier to simply use the BIC statistic, SAS has already combined the Log Likelihood and the number of parameters. We can simply fit our various models, and the model with the smallest BIC is the one with the best fitting random effect variance-covariance structure.

Using the log likelihoods and Chi-squares is the standard maximum likelihood approach; however, it requires us to keep track of the number of random effects parameters (to determine degrees of freedom). In addition, it is **ONLY** applicable for comparing models where 1 model is a sub-set of the other; the case in this example. An easier approach is to let SAS/PROC MIXED to the work! PROC MIXED computes the log likelihood and then imposes a penalty (as a function of the number of parameters required for the model). There are various penalty methods: Akaike's Information Criterion [AIC], Schwartz's Bayesian Criterion [BIC/SBC], and others; my preference is to use the BIC value, it is parsimonious in ascribing additional random effects parameters. Thus we can directly compare the BIC values from different models, and choose the model with the lowest BIC value. Note, in examining different variance structures one should always apply a large measure of com-

mon sense, and only consider as possible models those for which you are able to give an explanation!

19 Homogeneity of Variance - STD Ch 19.3

Bartlett's Test (STD Ch. 19.3, P480)

19.1 Basic Formulae

$$\chi^2 = 2.3026 * \left(\left[\sum_{i=1}^{i=a} (n_i - 1) \right] \log_{10} \bar{\sigma}^2 - \sum_{i=1}^{i=a} [(n_i - 1) \log_{10} \sigma_i^2] \right)$$

If we use natural logarithms (ln), we do not need to multiple by 2.3026

where $\bar{\sigma}^2$ equals the pooled variance, from our ANOVA since this accounts for the estimation of any other factors in our model and hence the degrees of freedom. In practice, it would be more desirable to use SAS PROC MIXED and directly test the homogeneity of variance (see the Stats II Web site about Homogeneity of Variance and the SAS code). Using PROC MIXED and fitted an effect to account for homogeneity of variances means that if there is significant heterogeneity (lack of homogeneity) then the tests of statistical significance etc will all be 'correct' and we can largely avoid all the need to find a 'suitable' transformation to stabilise the variance!

Correction factor

$$1 + \frac{1}{3(a-1)} \left[\sum_{i=1}^{i=a} \frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{i=a} (n_i - 1)} \right]$$

a = No groups. => d.f. = a-1 for χ^2 test

Then corrected $\chi^2 = \frac{\chi^2}{\text{correction factor}}$

N.b. the correction factor is always greater than 1

19.2 What to use for Homogeneity test

Recall that one of our requirements for the Analysis of Variance to be valid is that the variances be homogeneous. We can use Bartlett's test to verify this. Basically we need to divide the observations into appropriate groups and then test for homogeneity.

For a One-Way, Completely Randomized Design a reasonable, appropriate grouping would therefore be by the groups, or treatments. Most textbooks use this to illustrate Bartlett's test, and STD is no exception. Steel, Torrie and Dickey say to compute the variances in each group, the logarithm of each variance, etc. This is all-well-and-good in such a simple CRD situation, but becomes impossible to apply in any other more complicated model/design; for example in a Randomized Complete Block design. For a RCB should we group the observations by Block or by Treatment?

Why?

Recall that our linear model, for a CRD was

$$Y_{ij} = \mu + trt_i + e_{ij}$$

It is the homogeneity of the variances (of the e_{ij} 's) that is important, not the homogeneity of the variances of the Y_{ij} 's; even though $Y_{ij} = \mu + trt_i + e_{ij}$.

For a Completely Randomized Design if we compute the variance in each group using the observed data, *i.e.* the Y_{ij} 's, we compute the variance in the i^{th} group as:

$$\hat{\sigma}_e^2 = \left(\sum_{j=1}^{j=n_i} Y_{ij}^2 - n_i \bar{Y}_{ij}^2 \right) / (n_i - 1)$$

But, note that $Y_{ij} = \mu + trt_i + e_{ij}$

So

$$\sum_{j=1}^{j=n_i} Y_{ij}^2 = \sum_{j=1}^{j=n_i} (\mu + trt_i + e_{ij})^2$$

$$\begin{aligned}
&= \sum_{j=1}^{j=n_i} (\mu + trt_i + e_{ij})(\mu + trt_i + e_{ij}) \\
&= \sum_{j=1}^{j=n_i} (\mu^2 + \mu trt_i + \mu e_{ij} + \mu trt_i + trt_i^2 + trt_i e_{ij} + \mu e_{ij} + trt_i e_{ij} + e_{ij}^2) \\
&= \sum_{j=1}^{j=n_i} (\mu^2 + 2\mu trt_i + trt_i^2 + 2trt_i e_{ij} + 2\mu e_{ij} + e_{ij}^2) \\
&= n_i \mu^2 + 2\mu \sum_{j=1}^{j=n_i} trt_i + \sum_{j=1}^{j=n_i} trt_i^2 + 2 \sum_{j=1}^{j=n_i} trt_i e_{ij} + 2 \sum_{j=1}^{j=n_i} \mu e_{ij} + \sum_{j=1}^{j=n_i} e_{ij}^2
\end{aligned}$$

Note that $\sum_{j=1}^{j=n_i} e_{ij} = 0$

Thus the above expression simplifies to

$$\begin{aligned}
&n_i \mu^2 + 2\mu \sum_{j=1}^{j=n_i} trt_i + \sum_{j=1}^{j=n_i} trt_i^2 + \sum_{j=1}^{j=n_i} e_{ij}^2 \\
&= n_i \mu^2 + 2n_i \mu trt_i + n_i trt_i^2 + \sum_{j=1}^{j=n_i} e_{ij}^2
\end{aligned}$$

If we turn to the Correction Factor for the mean of the group, $n_i \bar{Y}_{ij}^2$, we can re-write this as:

$$n_i \bar{Y}_{ij}^2 = n_i \left(\frac{\sum_{j=1}^{j=n_i} (\mu + trt_i + e_{ij})}{n_i} \right)^2$$

$$\begin{aligned}
&= \left[\sum_{j=1}^{j=n_i} (\mu + trt_i + e_{ij}) \right]^2 / n_i \\
&= \left(n_i \mu + \sum_{j=1}^{j=n_i} trt_i + \sum_{j=1}^{j=n_i} e_{ij} \right)^2 / n_i \\
&= \left(n_i \mu + \sum_{j=1}^{j=n_i} trt_i + \sum_{j=1}^{j=n_i} e_{ij} \right) \left(n_i \mu + \sum_{j=1}^{j=n_i} trt_i + \sum_{j=1}^{j=n_i} e_{ij} \right) / n_i \\
&= (n_i^2 \mu^2 + n_i \mu \sum_{j=1}^{j=n_i} trt_i + n_i \mu \sum_{j=1}^{j=n_i} e_{ij} + n_i \mu \sum_{j=1}^{j=n_i} trt_i + \sum_{j=1}^{j=n_i} trt_i \sum_{j=1}^{j=n_i} trt_i + \sum_{j=1}^{j=n_i} trt_i \sum_{j=1}^{j=n_i} e_{ij} + \\
&\quad n_i \mu \sum_{j=1}^{j=n_i} e_{ij} + \sum_{j=1}^{j=n_i} trt_i \sum_{j=1}^{j=n_i} e_{ij} + \sum_{j=1}^{j=n_i} e_{ij} \sum_{j=1}^{j=n_i} e_{ij}) / n_i \\
&= n_i \mu^2 + 2\mu \sum_{j=1}^{j=n_i} trt_i + n_i trt_i^2 + 2 \sum_{j=1}^{j=n_i} trt_i e_{ij} + 2\mu \sum_{j=1}^{j=n_i} e_{ij} + \sum_{j=1}^{j=n_i} e_{ij} \sum_{j=1}^{j=n_i} e_{ij}
\end{aligned}$$

Note that $\sum_{j=1}^{j=n_i} e_{ij} = 0$

Therefore the above expression simplifies to

$$n_i \mu^2 + 2n_i trt_i + n_i trt_i^2$$

$$\text{therefore } \hat{\sigma}_i^2 = \left(\sum_{j=1}^{j=n_i} Y_{ij}^2 - n_i \bar{Y}_{ij}^2 \right) / (n_i - 1)$$

$$= \left[(n_i \mu^2 + 2n_i \mu trt_i + n_i trt_i^2 + \sum_{j=1}^{n_i} e_{ij}^2) - (n_i \mu^2 + 2n_i \mu trt_i + n_i trt_i^2) \right] / (n_i - 1)$$

$$= \sum_{j=1}^{n_i} e_{ij}^2 / (n_i - 1)$$

Thus what we are doing in computing the variance in each group in a Completely Randomized Design is calculating the deviation from the group mean ($Y_{ij} - \hat{Y}_{ij}$), which is:

$$[\mu + trt_i + e_{ij}] - [\mu + trt_i]$$

i.e. e_{ij} . Thus in this simple case of a CRD using Y_{ij} is equivalent to using the residuals, e_{ij} .

However, in most other designs/models this will not be the case. For example, if we look at a Randomized Complete Block Design and ask whether the variances in each block are homogeneous we can see immediately that the observations (experimental units) in each block each received different treatment effects.

If we look the the linear model for a RCB we have:

$$Y_{ij} = \mu + trt_i + block_j + e_{ij}$$

If we look the the observations in each block, then if we subtract out the j^{th} block effect ($\mu + block_j$), we can see that the deviation from the j^{th} block mean represents ($trt_i + e_{ij}$). Therefore, if we use these deviations to compute a variance we will be computing a variance based on treatments and residual variation; not at all what we want, AND completely wrong!

19.3 Rationale for the use of residuals

HOWEVER, what if we compute, and use, the estimated residuals? Recall that in our derivation of Least Squares we set out to estimate the various parameters that would provide the 'best' fit; *i.e.* minimize the Sums of Squares of Errors, or in other words the sum of the squared deviations of the actual observations from the fitted values, $\sum(Y - \hat{Y})^2$.

What were the assumptions for our Least Squares method? They were that the errors be uncorrelated; that the errors have an expectation of Zero (*i.e.* $E(e) = 0$) and a finite variance. Note, that for the method of Least Squares we are not requiring homogeneity of variance, nor normality; these are only required subsequently for statistical tests of significance. So we can use Least Squares to compute the fitted values (\hat{Y}), which represent ($\hat{Y}_{ij} = \mu + trt_i + block_k$). Thus $(Y - \hat{Y})$ provides us with an estimate of the e_{ij} 's,

$$\begin{aligned} Y - \hat{Y} &= (\mu + trt + i + block_j + e_{ij}) - (\mu + trt_i + block_k) \\ &= \hat{e}_{ij} \end{aligned}$$

Then we can use these estimated residuals, grouped appropriately, to test for homogeneity of variance.

19.4 Exercises

1. Using the data from the 1-Way, CRD (Section 11.5, page 83) test for homogeneity of variance in each group (treatments). Do this twice, once using the Y's and once using \hat{e} . Note that you get exactly the same answers, for the reasons outlined above. N.B. Make use of SAS and the OUTPUT option of PROC GLM.

2. Using the data from the 2-Way, RCB (Section 20.4, page 166) test for homogeneity of variance in each block. Do this twice, once using the Y's and once using the \hat{e} 's. Note the substantial difference. Again, make use of SAS.

20 Multiway Classification - STD Ch. 9

More than 1 source of variation, or classification, a cross-classified model.

Suppose that we are interested in looking at the weight gain of sows after they have had their piglets weaned off them. We want to see how quickly the sows regain weight; we want to compare 6 diets, to see whether there are differences amongst the diets. We have 24 sows, so we can have 4 on each diet. We could randomly assign sows to the 6 diets, that would give us a One-Way ANOVA. However, we note that the sows are not all of the same parity, there are 6 1st parity sows, 6 2nd, 6 3rd and 6 4th. This is an identifiable, systematic factor, hence we should [MUST] account for it in our design and model. We therefore decide to take the 6 first parity sows and randomly allocate one sow to each of the 6 diets (treatments). We then take the 6 second parity sows and randomly assign them to the 6 diets, and likewise for the third parity sows and the fourth parity sows. The 24 sows are then randomly assigned to individual cages, so that they cannot interfere with one another; so that our assumption of independence of the errors (and observations) will therefore be reasonable.

The layout is shown below.

$D_6 P_4$	$D_1 P_2$	$D_2 P_4$	$D_1 P_4$	$D_3 P_4$	$D_5 P_1$
$D_1 P_3$	$D_3 P_3$	$D_1 P_1$	$D_5 P_4$	$D_4 P_4$	$D_6 P_3$
$D_5 P_3$	$D_3 P_2$	$D_4 P_3$	$D_4 P_2$	$D_2 P_3$	$D_5 P_2$
$D_3 P_1$	$D_6 P_2$	$D_2 P_2$	$D_4 P_1$	$D_6 P_1$	$D_2 P_1$

Let us call diet (with its 6 levels) Factor A, and Parity (with its 4 levels) Factor B.

20.1 Linear model

$$Y_{ij} = \mu + \text{Factor } A_i + \text{Factor } B_j + AB_{ij} + \text{sow}_{ij} + \epsilon_{ij}$$

NOTE: we have only 1 experimental unit (sow) per AB (Diet*Parity) combination; therefore, the A*B interaction and the sow effects are confounded with one another. We can consider the AB interaction to be of a 'higher order' than the sow within AB effect. When we have a higher order effect confounded with an effect of a lower order, then we should go with the simpler effect, *i.e.* sow. If you think that there is, or might be, an interaction between these two factors then this design is **COMPLETELY AND ABSOLUTELY WRONG** and you should not even start your experiment with this design; you **NEED** a factorial design, see section 28. Thus our statistical model must be simplified to:

$$Y_{ij} = \mu + \text{Factor } A_i + \text{Factor } B_j + \text{sow}_{ij} + \epsilon_{ij}$$

NOTE: we have only 1 measurement per experimental unit, they are confounded and thus we cannot separate sow_{ij} from ϵ_{ij} . We shall have to consider only an 'error' term for the measurement made on the sow_{ij} ; the sow 'nested' within the i^{th} Factor A and also 'nested' within the j^{th} Factor B (note how this dovetails with what we saw in the One-way ANOVA model). Thus our statistical model becomes

$$Y_{ij} = \mu + \text{Factor } A_i + \text{Factor } B_j + e_{ij}$$

The trial is carried out starting in October; however, during the middle of December the students taking one of the production agriculture courses decide to have a barbeque and the staff decide to "appropriate" two pigs to serve as the input to human nutrition! We loose the second parity sow from Diet 1 and the fourth parity sow from Diet 2. These were the two pens most convenient for the students to take the animals from; a chance random event unrelated to the Diets or Parity *per se*.

A suitable statistical model would therefore be:

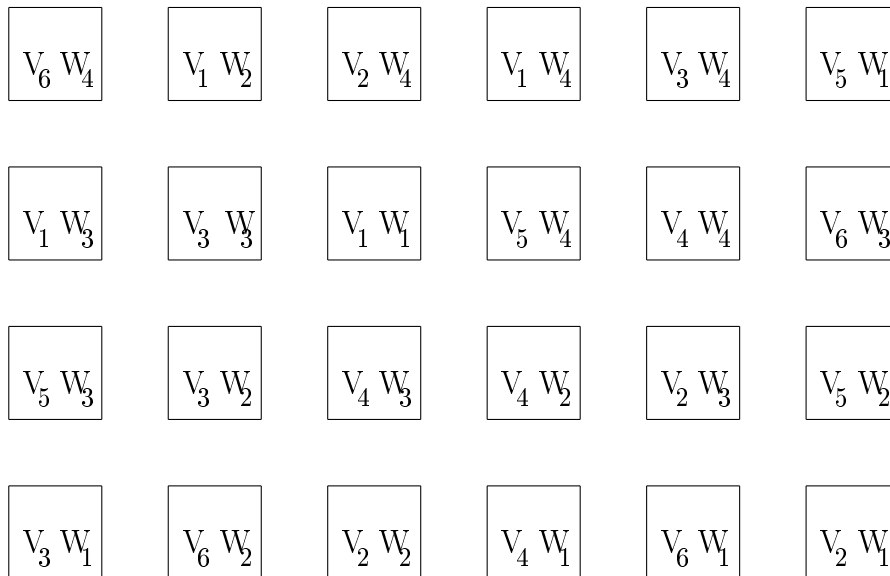
$$Y_{ij} = \mu + \text{Diet}_i + \text{Parity}_j + e_{ij}$$

20.2 Parameters of the Model

$$\mu, D_1, D_2, D_3, D_4, D_5, D_6, P_1, P_2, P_3, P_4, \sigma_e^2$$

Alternatively, consider an example relating to crops. Suppose that we have 6 varieties of maize that we want to compare and that we are going to have 4 plots for each variety, so that we will have 24 plots, each one 2 metres wide by 5 metres long. Our field where we are going to carry out this experiment is beside the autoroute and we mark out the 24 plots, separated one from another so that there is no interference between plots. So far this description is as per a Completely Randomised Design (One-way ANOVA). However, we decide that we also want to examine the effect of row width (the width between adjacent rows of maize in each plot). We have 4 row widths that we want to test, so we arrange that for the 4 plots of each variety, 1 plot will be planted at each of the row spacings (widths).

The layout in the field is shown below.



Unfortunately, part way through our experiment a large truck decides to leave the autoroute and become part of the "decor", ploughing through two of the plots; Variety 1 Row Width 2, and Variety 2 Row Width 4. We therefore lose these two observations!

A suitable statistical model would therefore be:

Table 19: Two-Way ANOVA, Data

Diet (Variety)		Parity		Row Width	
		1	2	3	4
Diet 1	Variety 1	3.4	.	4.1	7.0
Diet 2	Variety 2	2.3	1.9	7.1	.
Diet 3	Variety 3	3.4	4.0	3.1	5.5
Diet 4	Variety 4	5.8	6.6	6.4	8.0
Diet 5	Variety 5	5.3	4.9	7.1	6.9
Diet 6	Variety 6	5.4	7.3	6.7	8.7

$$Y_{ij} = \mu + \text{Variety}_i + \text{Width}_j + e_{ij}$$

20.3 Parameters

$$\mu, V_1, V_2, V_3, V_4, V_5, V_6, W_1, W_2, W_3, W_4, \sigma_e^2$$

20.4 Observations

This is a 2-way classification; the 2 factors being Diet, or Variety, (6 levels) and Parity, or Row Width, (4 levels).

20.5 Matrix Equations

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ \cdot \\ \cdot \\ Y_{22} \\ \cdot \\ \cdot \\ Y_{64} \end{bmatrix} = \begin{bmatrix} \mu & D_1 & D_2 & D_3 & D_4 & D_5 & D_6 & P_1 & P_2 & P_3 & P_4 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \\ D_6 \\ P_1 \\ P_2 \\ P_3 \\ P_4 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{21} \\ \cdot \\ \cdot \\ e_{22} \\ \cdot \\ \cdot \\ e_{64} \end{bmatrix}$$

Note that the ordering of the equations is immaterial, that is to say we could have Parity (Row Width) first and Diet (Variety) second, as:

$$Y_{ij} = \mu + Parity_j + Diet_i + e_{ij}$$

Try it and see!

$$Y = X b + e$$

Normal equations are:

$$X'X\tilde{b} = X'Y$$

20.6 Normal Equations

$$\begin{bmatrix}
 22 & 3 & 3 & 4 & 4 & 4 & 4 & 6 & 5 & 6 & 5 \\
 3 & 3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
 3 & 0 & 3 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
 4 & 0 & 0 & 4 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
 4 & 0 & 0 & 0 & 4 & 0 & 0 & 1 & 1 & 1 & 1 \\
 4 & 0 & 0 & 0 & 0 & 4 & 0 & 1 & 1 & 1 & 1 \\
 4 & 0 & 0 & 0 & 0 & 0 & 4 & 1 & 1 & 1 & 1 \\
 6 & 1 & 1 & 1 & 1 & 1 & 1 & 6 & 0 & 0 & 0 \\
 5 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 5 & 0 & 0 \\
 6 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 6 & 0 \\
 5 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 5
 \end{bmatrix}
 \begin{bmatrix}
 \tilde{\mu} \\
 \tilde{D}_1 \\
 \tilde{D}_2 \\
 \tilde{D}_3 \\
 \tilde{D}_4 \\
 \tilde{D}_5 \\
 \tilde{D}_6 \\
 \tilde{P}_1 \\
 \tilde{P}_2 \\
 \tilde{P}_3 \\
 \tilde{P}_4
 \end{bmatrix}
 =
 \begin{bmatrix}
 120.9 \\
 14.5 \\
 11.3 \\
 16 \\
 26.8 \\
 24.2 \\
 28.1 \\
 25.6 \\
 24.7 \\
 34.5 \\
 36.1
 \end{bmatrix}$$

20.7 A solution vector (GLM)

$$\tilde{b} = \begin{bmatrix}
 8.56808 \\
 -2.41540 \\
 -2.74397 \\
 -3.025 \\
 -0.325 \\
 -0.975 \\
 0.0 \\
 -2.72068 \\
 -2.21429 \\
 -1.23735 \\
 0.0
 \end{bmatrix}$$

$$\text{TSS} = 739.41$$

$$\text{SSR} = 722.114$$

$$\text{SSE} = 17.295$$

$$\text{CF} = 664.400$$

$$\text{SSR}_m = 57.714$$

20.8 Analysis using SAS/IML

USING SAS/PROC IML

```
/* Two-Way Analysis of Variance */
```

```
proc iml;  
reset print;
```

```
/*          Diets          Parities  
   mu    1 2 3 4 5 6    1 2 3 4  
*/  
x ={1    1 0 0 0 0 0    1 0 0 0,  
    1    1 0 0 0 0 0    0 0 0 1,  
    1    1 0 0 0 0 0    0 0 1 0,  
    1    0 1 0 0 0 0    1 0 0 0,  
    1    0 1 0 0 0 0    0 1 0 0,  
    1    0 1 0 0 0 0    0 0 1 0,  
    1    0 0 1 0 0 0    1 0 0 0,  
    1    0 0 1 0 0 0    0 1 0 0,  
    1    0 0 1 0 0 0    0 0 0 1,  
    1    0 0 1 0 0 0    0 0 1 0,  
    1    0 0 0 1 0 0    1 0 0 0,  
    1    0 0 0 1 0 0    0 1 0 0,  
    1    0 0 0 1 0 0    0 0 0 1,  
    1    0 0 0 1 0 0    0 0 1 0,  
    1    0 0 0 0 1 0    1 0 0 0,  
    1    0 0 0 0 1 0    0 1 0 0,  
    1    0 0 0 0 1 0    0 0 0 1,  
    1    0 0 0 0 1 0    0 0 1 0,  
    1    0 0 0 0 0 1    1 0 0 0,  
    1    0 0 0 0 0 1    0 1 0 0,  
    1    0 0 0 0 0 1    0 0 0 1,  
    1    0 0 0 0 0 1    0 0 1 0};
```

```
y={3.4,  
   7,  
   4.1,  
   2.3,  
   1.9,  
   7.1,  
   3.4,  
   4,
```

```

5.5,
3.1,
5.8,
6.6,
8,
6.4,
5.3,
4.9,
6.9,
7.1,
5.4,
7.3,
8.7,
6.7}];

xtx = x' * x;
xty = x' * y;

invxtx = ginv(xtx);
b = invxtx * xty;
tss = y' * y;
sumy = sum(y);
ssr = b' * xty;
ybar = sumy/nobs;
cf = nobs * ybar * ybar;
ssrm = ssr - cf;
dfd = 5;
dfp = 3;
rx = 1 + dfd + dfp;
dfe = nobs - rx;
sse = tss - ssr;
mse = sse/dfe;

/* Type III, Marginal Sums of Squares for Diets */
print /;
print " Type III Sums of Squares, Marginal, for Diets ";
kp= {0  1 -1 0 0 0 0  0 0 0 0,
      0  1 0 -1 0 0 0  0 0 0 0,
      0  1 0 0 -1 0 0  0 0 0 0,
      0  1 0 0 0 -1 0  0 0 0 0,
      0  1 0 0 0 0 -1  0 0 0 0};
k = kp';
df = nrow(kp);

```

```

kb = k' * b;
kxxk = k' * invvtx * k;
invkk = ginv(kxxk);
ssd = kb' * invkk * kb;
msd = ssd/df;
fd = msd/mse;
pr = 1 - probf(fd,df,dfe);

/* Type III, Marginal Sums of Squares for Parity */
print /;
print " Type III Sums of Squares, Marginal, for Parity ";
kp= {0  0 0 0 0 0 0  1 -1 0 0,
      0  0 0 0 0 0 0  1 0 -1 0,
      0  0 0 0 0 0 0  1 0 0 -1};
k = kp';
df = nrow(kp);
kb = k' * b;
kxxk = k' * invvtx * k;
invkk = ginv(kxxk);
ssp = kb' * invkk * kb;
msp = ssp/df;
fp = msp/mse;
pr = 1 - probf(fp,df,dfe);

/* Another Type III, Marginal Sums of Squares k' matrix
   for Parity */
print /;
print " Another Type III Sums of Squares, Marginal,
      for Parity ";
kp= {0  0 0 0 0 0 0  -1 0 0 1,
      0  0 0 0 0 0 0  0 -1 0 1,
      0  0 0 0 0 0 0  0 0 -1 1};
k = kp';
df = nrow(kp);
kb = k' * b;
kxxk = k' * invvtx * k;
invkk = ginv(kxxk);
ssp = kb' * invkk * kb;
msp = ssp/df;
fp = msp/mse;
pr = 1 - probf(fp,df,dfe);

```

```

/* Type III, Marginal Sums of Squares for Parity
   and Diet, i.e. SSRm */
print /;
print " Type III Sums of Squares, Marginal, for Parity and Diet ";
kp= {0  0 0 0 0 0 0  1 -1 0 0,
     0  0 0 0 0 0 0  1 0 -1 0,
     0  0 0 0 0 0 0  1 0 0 -1,
     0  1 -1 0 0 0 0  0 0 0 0,
     0  1 0 -1 0 0 0  0 0 0 0,
     0  1 0 0 -1 0 0  0 0 0 0,
     0  1 0 0 0 -1 0  0 0 0 0,
     0  1 0 0 0 0 -1  0 0 0 0};
k = kp';
df = nrow(kp);
kb = k' * b;
kxxk = k' * invvtx * k;
invkk = ginv(kxxk);
ssp = kb' * invkk * kb;
msp = ssp/df;
fp = msp/mse;
pr = 1 - probf(fp,df,dfe);

/* Estimates of fitted values, using the general k'b approach */

/* mu + d1 + p1 */
kp = {1  1 0 0 0 0 0  1 0 0 0};
k = kp';
kb = k' * b;
sv = k' * invvtx * k * mse;
se = sqrt(sv);

/* mu + d1 + p2 */
kp = {1  1 0 0 0 0 0  0 1 0 0};
k = kp';
kb = k' * b;
sv = k' * invvtx * k * mse;
se = sqrt(sv);

/* mu + d2 + p1 */
kp = {1  0 1 0 0 0 0  1 0 0 0};
k = kp';
kb = k' * b;

```

```

sv = k' * invxtx * k * mse;
se = sqrt(sv);

/* etc */

/* mu + d1 + average over parities */
kp = {4 4 0 0 0 0 0 1 1 1 1}/4;
k = kp';
kb = k' * b;
sv = k' * invxtx * k * mse;
se = sqrt(sv);

quit;

```

20.9 Analysis using SAS/GLM

USING SAS/PROC GLM

```

data twoway1; /* Two-way ANOVA */
input d p y;
cards;
1 1 3.4
1 3 7.0
1 4 4.1
2 1 2.3
2 2 1.9
2 4 7.1
3 1 3.4
3 2 4.0
3 3 5.5
3 4 3.1
4 1 5.8
4 2 6.6
4 3 8.0
4 4 6.4
5 1 5.3
5 2 4.9
5 3 6.9
5 4 7.1
6 1 5.4
6 2 7.3
6 3 8.7

```

```

6 4 6.7
;
proc glm data=twoway1;
classes d p;
model y = d p/xpx i solution;

/* Marginal, Type III, Sums of Squares for Diet */
contrast 'Diets, Type III' d 1 -1 0 0 0 0,
                        d 1 0 -1 0 0 0,
                        d 1 0 0 -1 0 0,
                        d 1 0 0 0 -1 0,
                        d 1 0 0 0 0 -1;

/* Marginal, Type III, Sums of Squares for Parities */
contrast 'Parities, Type III' p 1 -1 0 0,
                        p 1 0 -1 0,
                        p 1 0 0 -1;

/* Marginal, Type III, Sums of Squares for Diet and Parity */
contrast 'D and P, = SSRm' d 1 -1 0 0 0 0,
                        d 1 0 -1 0 0 0,
                        d 1 0 0 -1 0 0,
                        d 1 0 0 0 -1 0,
                        d 1 0 0 0 0 -1,
                        p 1 -1 0 0,
                        p 1 0 -1 0,
                        p 1 0 0 -1;

estimate 'Diet 1 - 2' d 1 -1 0 0 0 0;
estimate 'Diet 1 - 3' d 1 0 -1 0 0 0;
estimate 'Diet 1 - 4' d 1 0 0 -1 0 0;
estimate 'Diet 1 - 5' d 1 0 0 0 -1 0;
estimate 'Diet 1 - 6' d 1 0 0 0 0 -1;
estimate 'Diet 2 - 3' d 0 1 -1 0 0 0;
estimate 'Diet 2 - 4' d 0 1 0 -1 0 0;
estimate 'Diet 2 - 5' d 0 1 0 0 -1 0;
estimate 'Diet 2 - 6' d 0 1 0 0 0 -1;
estimate 'Diet 3 - 4' d 0 0 1 -1 0 0;
estimate 'Diet 3 - 5' d 0 0 1 0 -1 0;
estimate 'Diet 3 - 6' d 0 0 1 0 0 -1;
estimate 'Diet 4 - 5' d 0 0 0 1 -1 0;
estimate 'Diet 4 - 6' d 0 0 0 1 0 -1;
estimate 'Diet 5 - 6' d 0 0 0 0 1 -1;

```

```

estimate 'Parity 1 - 2' p 1 -1 0 0;
estimate 'Parity 1 - 3' p 1 0 -1 0;
estimate 'Parity 1 - 4' p 1 0 0 -1;

/* Estimate each fitted value */

estimate 'D1 P1' intercept 1 d 1 0 0 0 0 0 p 1 0 0 0;
estimate 'D1 P2' intercept 1 d 1 0 0 0 0 0 p 0 1 0 0;
estimate 'D1 P3' intercept 1 d 1 0 0 0 0 0 p 0 0 1 0;
estimate 'D1 P4' intercept 1 d 1 0 0 0 0 0 p 0 0 0 1;
estimate 'D2 P1' intercept 1 d 0 1 0 0 0 0 p 1 0 0 0;
estimate 'D2 P2' intercept 1 d 0 1 0 0 0 0 p 0 1 0 0;
estimate 'D2 P3' intercept 1 d 0 1 0 0 0 0 p 0 0 1 0;
estimate 'D2 P4' intercept 1 d 0 1 0 0 0 0 p 0 0 0 1;

/* Estimate sum of Parities on Diet 1 */

estimate ' mu + D1 + sum p ' intercept 4 d 4 0 0 0 0 0
p 1 1 1 1;

/* Estimate average of parities for Diet 1,
note avoid fractions */

estimate ' mu + d1 + av p' intercept 4 d 4 0 0 0 0 0
p 1 1 1 1 /divisor=4;

lsmeans d/pdiff stderr;
lsmeans d/pdiff stderr adjust=scheffe;
run;
quit;

```

20.10 Analysis of Variance

Tabulated F-values

$$F_{9,13,5\%} = 2.71$$

$$F_{9,13,1\%} = 4.19$$

$$F_{8,13,5\%} = 2.77$$

$$F_{8,13,1\%} = 4.30$$

Table 20: Two-Way ANOVA

ANOVA <i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F - ratio</i>	<i>E(MS)</i>
Total, TSS	$N = 22$	$Y'Y$ 739.41			
Model, SSR	$r(X)$ $= 9$	$\tilde{b}'X'Y$ 722.114	80.235	60.327**	
Mean, C.F.	1	$N\bar{y}^2$ 664.4	664.4	499.55**	
Model, after the mean, SSR_m $R(D, P Mean)$	$r(X) - 1$ $= 8$	$\tilde{b}'X'Y - N\bar{y}^2$ 57.714	7.214	5.42**	$\sigma_e^2 + Q(D, P)$
Error, Residual	$N - r(X)$ $22 - 9$ $= 13$	$Y'Y - \tilde{b}'X'Y$ 17.295	1.330		σ_e^2

The model accounts for a significant amount of variation.

The model, corrected for the mean, accounts for a significant amount of variation; *i.e.* over and above the mean there is a significant effect of Diet (Variety) and/or Parity (Row Width).

21 Hypotheses to be tested

Much as for the 1-Way ANOVA, the initial hypothesis will be that the Model (μ and D and P) does not explain variation in the dependent variable, *i.e.* that the terms in the Model (μ and D and P) are all equal to Zero, and the Alternative Hypothesis is that these terms (μ and D and P) are not all equal to Zero. Although we (again) do not have a full-rank model, we can write

$$H_o \quad \begin{bmatrix} \mu \\ D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \\ D_6 \\ P_1 \\ P_2 \\ P_3 \\ P_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and obviously H_A as \neq .

The next hypothesis is about the Mean; as before it is: $H_o: \bar{Y} = 0$, and $H_A: \bar{Y} \neq 0$

Continuing our subdivision of the source of variation, we have the Model over and above the Mean, which is the differences amongst Diets and the difference amongst Parities. This we shall then subdivide into differences amongst Diets (over and above the Mean

and Parities), and differences amongst Parities (over and above the Mean and Diets). We can specify, therefore, our Null Hypotesis (H_o) as being that the effects of the Dies are all equal, and that the effects of the Parities are all equal; our Alternative Hypothesis (H_A) will therefore be that the effects of the Diets are not all equal and/or that the effects of the Parities are not all equal. Thus:

$$H_o : D_1 = D_2 = D_3 = D_4 = D_5 = D_6$$

H_A : Diets are not all equal

The Null Hypothesis we can re-write as a series of comparisons:

6 Diets, 5 separate comparisons

i) $D_1 = D_2$

ii) $D_1 = D_3$

iii) $D_1 = D_4$

iv) $D_1 = D_5$

v) $D_1 = D_6$

which we can re-write as a series of comparisons with Null Hypotheses of Zero:

i) $D_1 - D_2 = 0$

ii) $D_1 - D_3 = 0$

iii) $D_1 - D_4 = 0$

iv) $D_1 - D_5 = 0$

v) $D_1 - D_6 = 0$

$$H_o \quad \begin{bmatrix} D_1 - D_2 \\ D_1 - D_3 \\ D_1 - D_4 \\ D_1 - D_5 \\ D_1 - D_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and

$$H_A \begin{bmatrix} D_1 - D_2 \\ D_1 - D_3 \\ D_1 - D_4 \\ D_1 - D_5 \\ D_1 - D_6 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Similarly we can look at Parities.

$$H_o : P_1 = P_2 = P_3 = P_4$$

H_A : Parities are not all equal

$$H_o \begin{bmatrix} P_1 - P_2 \\ P_1 - P_3 \\ P_1 - P_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$H_A \begin{bmatrix} P_1 - P_2 \\ P_1 - P_3 \\ P_1 - P_4 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Thus, the Model over and above the Mean, $R(\text{Diet}, \text{Parity}|\mu)$ is:

$$H_o \begin{bmatrix} D_1 - D_2 \\ D_1 - D_3 \\ D_1 - D_4 \\ D_1 - D_5 \\ D_1 - D_6 \\ P_1 - P_2 \\ P_1 - P_3 \\ P_1 - P_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$H_A \begin{bmatrix} D_1 - D_2 \\ D_1 - D_3 \\ D_1 - D_4 \\ D_1 - D_5 \\ D_1 - D_6 \\ P_1 - P_2 \\ P_1 - P_3 \\ P_1 - P_4 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

22 Partitioning Sums of Squares for the Model

Partitioning SSR_m into effects due to Diets and Parities. These are also called the '**CON-TRAST**' Sums of Squares, because they arise from contrasts, or comparisons, between levels of effects.

22.1 Sums of Squares for Factor 1

6 Diets, 5 separate comparisons

- i) $D_1 - D_2$
- ii) $D_1 - D_3$
- iii) $D_1 - D_4$
- iv) $D_1 - D_5$
- v) $D_1 - D_6$

Are these comparisons estimable? Well let us look at what we know! We know that fitted values are estimable, hence $\hat{Y} = X\tilde{b}$.

Thus the fitted value for the animal (experimental unit) receiving Diet 1, which was Parity 1 will be:

$$\begin{aligned} \hat{Y}_{11} &= (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0)\tilde{b} \\ &= \hat{Y}_{11} = \tilde{\mu} + \tilde{D}_1 + \tilde{P}_1 \end{aligned}$$

And the fitted value for the animal (experimental unit) receiving Diet 2, which was Parity 1 will be:

$$\begin{aligned}\hat{Y}_{21} &= (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0)\tilde{b} \\ &= \hat{Y}_{21} = \tilde{\mu} + \tilde{D}_2 + \tilde{P}_1\end{aligned}$$

These fitted values are both estimable, therefore a linear function of them (the difference) will also be estimable.

$$\begin{aligned}k'_{\hat{Y}_{11}} &= (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0) \\ -k'_{\hat{Y}_{21}} &= (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0) \\ &= k'_{1-2} = (0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)\end{aligned}$$

Similarly, the fitted value for the animal (experimental unit) receiving Diet 3, which was Parity 1 will be:

$$\begin{aligned}\hat{Y}_{31} &= (1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0)\tilde{b} \\ &= \hat{Y}_{31} = \tilde{\mu} + \tilde{D}_3 + \tilde{P}_1\end{aligned}$$

Again, these fitted values are both estimable, therefore a linear function of them (the difference) will also be estimable.

$$\begin{aligned}k'_{\hat{Y}_{11}} &= (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0) \\ -k'_{\hat{Y}_{31}} &= (1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0) \\ &= k'_{1-3} = (0 \quad 1 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)\end{aligned}$$

Etc, and putting these 5 comparisons together, we have

$$k' = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Marginal, Type III, Sums of Squares

$$\text{Sums of Squares} = (k'\tilde{b})'[k'(X'X)^{-1}k]^{-1}(k'\tilde{b})$$

$$\text{Type III Sums of Squares Diets} = 31.853$$

22.2 SAS CONTRAST statement for Factor 1

Note these CONTRAST statements come after the MODEL statement in PROC GLM.

```
/* Marginal, Type III, Sums of Squares for Diet */
contrast 'Diets, Type III' d 1 -1 0 0 0 0,
                        d 1 0 -1 0 0 0,
                        d 1 0 0 -1 0 0,
                        d 1 0 0 0 -1 0,
                        d 1 0 0 0 0 -1;

contrast 'Parities, Type III' p 1 -1 0 0,
                        p 1 0 -1 0,
                        p 1 0 0 -1;
```

22.3 Sums of Squares for Factor 2

$$k'_{Parities} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}$$

Similarly the Type III Sums of Squares Parities = 22.096

22.4 SAS CONTRAST statement for Factor 2

```
/* Marginal, Type III, Sums of Squares for Parities */
contrast 'Parities, Type III' p 1 -1 0 0,
                        p 1 0 -1 0,
                        p 1 0 0 -1;
```

Thus we can construct a more detailed ANOVA

Table 21: Two-Way ANOVA

<i>Source</i>	<i>Analysis of Variance</i> <i>df</i>	<i>SS</i>	<i>MS</i>	<i>F - ratio</i>	
Total, TSS	$N = 22$	$Y'Y$ 739.41			
Model, SSR	$r(X)$ $= 9$	$\tilde{b}'X'Y$ 722.114	80.235	60.327**	
Mean, C.F.	1	$N\bar{y}^2$ 664.4	664.4	499.55**	
Model, after the mean, SSR_m $R(D, P Mean)$	$r(X) - 1$ $= 8$	$\tilde{b}'X'Y - N\bar{y}^2$ 57.714	7.214	5.42**	$\sigma_e^2 + Q(D, P)$
D $R(D \mu, P)$	5	31.853	6.371	4.79*	$\sigma_e^2 + Q(D)$
P $R(P \mu, D)$	3	22.096	7.365	5.54*	$\sigma_e^2 + Q(P)$
Error, Residual	$N - r(X)$ $22 - 9$	$Y'Y - \tilde{b}'X'Y$ 17.295	1.330		

Table 22: E(MS)

Source	d.f.	E(MS)
Model, after the mean, SSR_m	$r(X) - 1$ $= 8$	$\sigma_e^2 + Q(D, P)$
D $R(D \mu, P)$	5	$\sigma_e^2 + \frac{1}{6} \sum_{i=1}^6 n_i (d_i - \bar{d})^2 = \sigma_e^2 + Q(D)$
P $R(P \mu, D)$	3	$\sigma_e^2 + \frac{1}{4} \sum_{j=1}^4 n_j (p_j - \bar{p})^2 = \sigma_e^2 + Q(P)$
Error,	$N - r(X)$	σ_e^2

Compute the differences between Diets and their standard errors, using our general approach of a k' matrix to generate contrasts and hence differences.

22.5 Expectations of Mean Squares

23 Gains in Efficiency STD 9.7

A randomised complete block design (RCB), as above, may be compared with that expected from a completely random design (CR).

$$\text{Estimate MSE (CR)} = \frac{f_b * MSB + (f_t + f_e) * MSE}{f_b + f_t + f_e}$$

MSB = block Mean Square

MSE = error Mean Square

f_b , f_t . & f_e are block, treatment and error d.f.

Where do these degrees of freedom come from and why do we use them? We are attempting to compute what the MSE would have been if we had had a CR design. If we had not included Block in our model then all the Sums of Squares for Block would be included in the Residual; f_b MSB. The Residual Mean Square that we have presently estimates σ_e^2 . The Residual has f_e degrees of freedom. The Mean Square for Treatments has Expectation of $\sigma_e^2 + Q(trt)$, and has f_t degrees of freedom. Thus we add $(f_e + f_t)$ MSE to the Sums of Squares of Block! This, divided by the total of these degrees of freedom, gives us an estimate of the MSE that we might have expected to get if we had carried out a CR design.

Note the importance of this concept, since we shall return to it when looking at the Relative Efficiency of other designs. This same general principle will allow us to examine the Relative Efficiency of other designs, even ones that we shall not explicitly cover in this course.

$$\text{Thus MSE(CR)} = \frac{3(7.365) + (5 + 13)1.33}{3 + 5 + 13} = 2.192$$

$$\begin{aligned} & \text{Then Relative Efficiency, RE (RCB to CRD)} \\ &= \frac{(f_1 + 1)(f_2 + 3)MSE(CR)}{(f_2 + 1)(f_1 + 3)MSE(RCB)} * 100 \end{aligned}$$

Where f_1 = the residual degrees of freedom for a RCB (model with the factor include) and f_2 = the residual degrees of freedom for a CRD (the model without the factor).

$$\frac{(13 + 1)(16 + 3)2.192}{(16 + 1)(13 + 3)1.33} * 100 = 161 \text{ percent}$$

in this example the RE was 161 %, consistent with differences between blocks

24 Expectation of Mean Squares

See Steel, Torrie and Dickey Ch. 9.9, Page 225

If the effects we consider are fixed effects then we will be interested in the differences between the various treatments or factors. However, if the effects that we are considering are classed as random effects then it is the variability in the population that we should be interested in.

The two-way model that we have been considering, the RCB design, would usually have treatment as a fixed effect; that being the purpose of the experiment in all likelihood. In the design as proposed, with the effects of block, we might well not be interested in the specific differences between blocks; particularly if they are fields or some other such effect. Why? Because other producers using the treatment(s) will not have the same fields. Thus our fields are a random sample of fields, and we should probably consider fields as a random effect. However, if the "block" effect was a specific effect, such as sex (male vs female!), then we should probably consider block to be a fixed effect. These cases are shown in Table 9.8 of STD (P226) for the cases of Mixed model no interaction no sampling, and Fixed model no interaction no sampling respectively.

Consider that we have r blocks and t treatments, and that their effects are β_j and α_i , respectively, expressed as deviations from their means.

The above models are probably the most common situations, but they are by no means the only possibilities. Another possibility is that both factors be random. We might have a random sample of fields (blocks) and we might have a random sample of treatments (see STD).

Table 23: Expectations of Mean Squares

Source of Variation	df	Block Effect	
		Random	Fixed
Block	r-1	$\sigma_e^2 + t\sigma_\beta^2$	$\sigma_e^2 + Q(\beta_j^2)$
Treatments	t-1	$\sigma_e^2 + Q(\alpha_i^2)$	$\sigma_e^2 + Q(\alpha_i^2)$
Residual	(t-1)(r-1)	σ_e^2	σ_e^2

Using the same data, for the balanced case, work out what the expectations of the various Mean Squares are for the situation of both factors being random effects and hence the various variance components.

25 Least Squares Means

In the scientific literature it is very common to see the term "Least Squares Means" or "LS Means". What are LS Means? How do we compute them? What are their advantages and disadvantages? Upon what assumptions are they dependent? As mentioned previously, the term least squares means is somewhat of a shorthand for a more correct statement - estimates of means calculated using the method of least squares.

Least Squares Means are based on "fitted values", so they are statistically estimable. Remember that fitted values represent our (unbiased) estimates of the corresponding linear function of the same real parameters. They are meant to estimate the means, using our method of least squares, and not to be affected by having unequal numbers of observations in each group; they should be better estimators than the simple averages. To illustrate this let us use the Two-Way ANOVA and the One-Way, Completely Randomized Design.

We need to look initially at the CRD example (Section 11.5). Suppose we ask "What is the clover production from each of the 6 treatments?" The answer to this is the fitted value estimates for:

$$(\tilde{\mu} + \tilde{tr}t_1) = 28.82 \pm 1.60$$

$$(\tilde{\mu} + \tilde{tr}t_2) = 23.90 \pm 1.79$$

$$(\tilde{\mu} + \tilde{tr}t_3) = 14.35 \pm 1.79$$

$$(\tilde{\mu} + \tilde{tr}t_4) = 19.92 \pm 1.60$$

$$(\tilde{\mu} + \tilde{tr}t_5) = 13.26 \pm 1.60$$

$$(\tilde{\mu} + \tilde{tr}t_6) = 18.70 \pm 1.60$$

These are the "Least Squares Means" for Treatment. $(\tilde{\mu} + \tilde{tr}t_6)$ is our unbiased estimate of $(\mu + trt_6)$.

Suppose that we ask a similar question about Diets from our Two-Way model. We do not have $(\tilde{\mu} + \tilde{diet}_1)$; our fitted values are $(\tilde{\mu} + \tilde{diet}_1 + \tilde{parity}_1)$, $(\tilde{\mu} + \tilde{diet}_1 + \tilde{parity}_2)$,

$(\tilde{\mu} + \tilde{diet}_1 + \tilde{parity}_3)$ and $(\tilde{\mu} + \tilde{diet}_1 + \tilde{parity}_4)$. Is this insurmountable? Suppose that we decide to define our various effects of Diet and Parity as being expressed as deviations about the mean, so that $\sum Diet_i = 0$ and $\sum Parity_j = 0$. Then if we take the 4 fitted values for Diet 1 and average them we have:

$$\begin{aligned} & (\tilde{\mu} + \tilde{diet}_1 + \tilde{parity}_1) \\ & + (\tilde{\mu} + \tilde{diet}_1 + \tilde{parity}_2) \\ & + (\tilde{\mu} + \tilde{diet}_1 + \tilde{parity}_3) \\ & + (\tilde{\mu} + \tilde{diet}_1 + \tilde{parity}_4) \\ & = 4\tilde{\mu} + 4\tilde{Diet}_1 + \sum_{j=1}^{j=4} \tilde{Parity}_j \end{aligned}$$

However, note that we have defined that the Parity effects will be expressed as deviations about the mean (μ), *i.e.* $\sum Parity_j = 0$. Thus, dividing by 4 to bring it back to a "per-unit" basis we have

$$\tilde{\mu} + \tilde{Diet}_1 + \frac{1}{4} \sum_{j=1}^{j=4} \tilde{Parity}_j$$

which we can consider to estimate

$$\mu + Diet_1 + \frac{1}{4} \sum_{j=1}^{j=4} Parity_j \quad \text{Note this is the ACTUAL, REAL model}$$

and since $\sum Parity_j = 0$ this therefore can be considered to estimate

$$\mu + Diet_1 + \frac{1}{4} 0$$

which equals $\mu + Diet_1$

It is important to note that we are NOT saying that the Parity effects are Zero, only that the sum of their effects is zero, and we are talking about the REAL parameters summing to Zero, by definition. Our solution vector has NOT changed, and we should note that the sum of the solutions for Parity does NOT equal Zero! (*i.e.* $\sum \tilde{P}_j \neq 0$).

25.1 LSMEANS using SAS/GLM

USING SAS/PROC GLM

```
proc glm data=twoway1;
classes d p;
model y = d p;
/* lsmeans computed explicitly */
estimate 'lsmean Diet 1' intercept 1 d 1 0 0 0 0 0 p .25 .25 .25 .25;
estimate 'lsmean Diet 2' intercept 1 d 0 1 0 0 0 0 p .25 .25 .25 .25;
estimate 'lsmean Diet 3' intercept 1 d 0 0 1 0 0 0 p .25 .25 .25 .25;
estimate 'lsmean Diet 4' intercept 1 d 0 0 0 1 0 0 p .25 .25 .25 .25;
estimate 'lsmean Diet 5' intercept 1 d 0 0 0 0 1 0 p .25 .25 .25 .25;
estimate 'lsmean Diet 6' intercept 1 d 0 0 0 0 0 1 p .25 .25 .25 .25;

/* Note, use /divisor= option to avoid fractions in the estimate
statement */
estimate 'lsmean Parity 1' intercept 6 d 1 1 1 1 1 1 p 6 0 0 0/divisor=6;
estimate 'lsmean Parity 2' intercept 6 d 1 1 1 1 1 1 p 0 6 0 0/divisor=6;
estimate 'lsmean Parity 3' intercept 6 d 1 1 1 1 1 1 p 0 0 6 0/divisor=6;
estimate 'lsmean Parity 4' intercept 6 d 1 1 1 1 1 1 p 0 0 0 6/divisor=6;
lsmeans d/pdiff stderr;
lsmeans d/pdiff stderr adjust=scheffe;
run;
quit;
```

26 Multiway Classification - Fixed effect and Random Effect, STD Ch. 9

We continue our multi-way model (two factors, but now we consider one of them to be a random effect rather than a fixed effect).

Suppose that we are interested looking at 4 varieties of maize, in real, field conditions. We recruit 12 farmers who agree to allow us to use 4 plots (fields) on each of their farms. On each farm we randomly allocate the 4 varieties of maize, one to each plot. At harvest we record the maize yield from each plot (field). For our analysis we shall need to account for the effects of variety AND farm. This two-factor model looks very similar to the preceding two-factor model (of diets and parities). HOWEVER, there is a difference; in this model we are not interested in these 12 farms, we consider them to be a reasonably random representative sample of Quebec farms, and we wish our results to be extrapolatable from these 12 farms to all farms in general; therefore we shall consider farms to be a random effect, rather than a fixed effect. IF we considered farm to be a fixed effect, then these results would only be applicable to these exact 12 farms and could not be extrapolated or considered applicable to yields from other farms. This, in all likelihood, goes exactly against the reason for the trial, which is to be able to generalise from these 12 samples farms to say that the results can be considered applicable to farms in general.

A suitable statistical model would therefore be:

$$Y_{ij} = \mu + \text{Variety}_i + \text{farm}_j + e_{ij}$$

$$\text{farm}_j \sim N(0, \sigma_f^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

26.1 Parameters

$$\mu, V_1, V_2, V_3, V_4, \sigma_f^2, \sigma_e^2$$

26.2 Observations

To analyse these data, where we have a fixed effect and a random effect, we should use a mixed model procedure (eg SAS PROC MIXED, or R lmerTest). PROC MIXED will give us an F-test of the fixed effect(s), and we can ask for lsmeans, and we can ESTIMATE differences, etc.

IF we want to know what are the effects of the farms (i.e. not jsut the variance amongst farms) then such a question is asking for the PREDICTION of a random effect. This is called BLUP: Best Linear Unbiased Prediction; we can obtain such predictions by adding the /SOLUTION option to the random statement (see SAS code below). In the below SAS code the ODS OUTPUT statement means that we also capture these output BLUPs to a new SAS dataset (solr1) for possible further manipulations etc.

26.3 Analysis using SAS/MIXED

USING SAS/PROC MIXED

```
data twoway2; /* Two-way RCBD */
input variety farm y;
cards;
1 1 51.9
2 1 53.7
3 1 54.9
4 1 55.6
1 2 51.6
2 2 50.9
3 2 52.2
4 2 53.8
1 3 44.0
```

Table 24: Two-Way RCBD, Data

Variety	Farm	Yield	Variety	Farm	Yield
1	1	51.9	1	7	52.1
2	1	53.7	2	7	53.4
3	1	54.9	3	7	55.5
4	1	55.6	4	7	56.1
1	2	51.6	1	8	52.6
2	2	50.9	2	8	52.4
3	2	52.2	3	8	53.6
4	2	53.8	4	8	55.8
1	3	44.0	1	9	49.3
2	3	46.0	2	9	49.4
3	3	47.3	3	9	52.1
4	3	48.4	4	9	51.6
1	4	52.6	1	10	51.1
2	4	52.8	2	10	52.5
3	4	54.0	3	10	52.9
4	4	56.3	4	10	57.1
1	5	56.5	1	11	58.6
2	5	57.5	2	11	57.6
3	5	58.2	3	11	58.9
4	5	61.7	4	11	61.0
1	6	48.5	1	12	45.1
2	6	50.8	2	12	46.7
3	6	51.6	3	12	47.9
4	6	52.0	4	12	48.0

```
2 3 46.0
3 3 47.3
4 3 48.4
1 4 52.6
2 4 52.8
3 4 54.0
4 4 56.3
1 5 56.5
2 5 57.5
3 5 58.2
4 5 61.7
1 6 48.5
2 6 50.8
3 6 51.6
4 6 52.0
1 7 52.1
2 7 53.4
3 7 55.5
4 7 56.1
1 8 52.6
2 8 52.4
3 8 53.6
4 8 55.8
1 9 49.3
2 9 49.4
3 9 52.1
4 9 51.6
1 10 51.1
2 10 52.5
3 10 52.9
4 10 57.1
1 11 58.6
2 11 57.6
3 11 58.9
4 11 61.0
1 12 45.1
2 12 46.7
3 12 47.9
4 12 48.0
```

```
;
```

```
ods output SolutionR=solr1;
proc mixed data=twoway2;
```

```

class variety farm;
model y = variety/ddfm=kr;
random farm/solution;
lsmeans variety;
estimate 'v1 - v2' variety 1 -1 0 0;
run;
quit;

```

26.4 Results

Model Information

Data Set	WORK.TWOWAY2
Dependent Variable	y
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Kenward–Roger
Degrees of Freedom Method	Kenward–Roger

Class Level Information

Class Level Information		
Class	Levels	Values
variety	4	1 2 3 4
farm	12	1 2 3 4 5 6 7 8 9 10 11 12

Dimensions

Dimensions	
Covariance Parameters	2
Columns in X	5
Columns in Z	12
Subjects	1
Max Obs Per Subject	48

[2]Number of Observations

Number of Observations	
Number of Observations Read	48
Number of Observations Used	48
Number of Observations Not Used	0

Iteration History

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	254.39543556	
1	1	162.98882144	0.00000000

Convergence Status

Convergence criteria met.

Covariance Parameter Estimates

Covariance Parameter Estimates	
Cov Parm	Estimate
farm	14.5449
Residual	0.6041

Fit Statistics

Fit Statistics	
-2 Res Log Likelihood	163.0
AIC (smaller is better)	167.0
AICC (smaller is better)	167.3
BIC (smaller is better)	168.0

Type 3 Tests of Fixed Effects

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
variety	3	33	49.78	<.0001

Estimates

Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr > t
v1 - v2	-0.8167	0.3173	33	-2.57	0.0147

Least Squares Means

Least Squares Means						
Effect	variety	Estimate	Standard Error	DF	t Value	Pr > t
variety	1	51.1583	1.1236	11.7	45.53	<.0001
variety	2	51.9750	1.1236	11.7	46.26	<.0001
variety	3	53.2583	1.1236	11.7	47.40	<.0001
variety	4	54.7833	1.1236	11.7	48.76	<.0001

IF we fit a model without the random effect of farm we find that the BIC value was 258.2 (as compared with the 168 when the random effect of farm was included). This difference is substantial, more than 8, indicating that we can consider the effect of farm to be significant.

27 Subsamples, or Nested Models

STD : Ch 7.6. P157

In many cases we have a model with subsamples. This arises when the experimental unit and the sampling unit are not the same. For example, imagine that we have 3 treatments that we wish to test on apple trees to see which treatment produces the heaviest apples. We have 12 apple trees and we randomly assign 4 trees to each treatment. We spray each tree with the appropriate treatment at the beginning of the growing season and then in the fall we randomly pick 6 apples from each tree and weigh them. We have 72 apples and hence 72 weights.

BUT, the experimental unit, to which the treatment was applied, was the tree and **NOT** the apple; apples are the subsampling unit. If one ignored this elementary fact and analysed the data one would in all likelihood come up with overly optimistic results; *i.e.* rubbish.

27.1 Linear model

$$Y_{ijk} = \mu + trt_i + tree_{ij} + e_{ijk}$$

27.2 Parameters for a Nested Model

The parameters of this model are: μ , each of the treatments (d_1, d_2 and d_3), the variance of the random effect of trees (nested within treatments) ($\sigma_{tree/trt}^2$) and the variance of the random effect of apples within trees (σ_e^2)

27.3 Hypotheses

Let us start with considering treatments. Our hypothesis will be very similar to that of previous fixed effects models; to test whether there are differences between the treatments (over and above the Mean). This we can describe (in words) in the form of a 'Null

Hypothesis' that the treatments are all equal, vs an 'Alternative Hypothesis' that the treatments are not all equal; *i.e.*

$$H_o : trt_1 = trt_2 = trt_3$$

H_A : treatments are not all equal

The Null Hypothesis we can re-write as a series of comparisons:

2 Treatments, 2 separate comparisons

i) $trt_1 = trt_2$

ii) $trt_1 = trt_3$

which we can re-write as a series of comparisons with Null Hypotheses of Zero:

i) $trt_1 - trt_2 = 0$

ii) $trt_1 - trt_3 = 0$

Which we can express statistically (as one hypothesis) as:

$$H_o \begin{bmatrix} trt_1 - trt_2 \\ trt_1 - trt_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and

$$H_A \begin{bmatrix} trt_1 - trt_2 \\ trt_1 - trt_3 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

If we turn now to the random effects of trees, we can consider that our Null Hypothesis will be that the variance amongst trees equals Zero.

$$H_o : \sigma_{tree/trt}^2 = 0$$

$$H_A : \sigma_{tree/trt}^2 > 0$$

Then, combining these two, our Null Hypothesis, for the Model over and above the Mean would be:

$$H_o \begin{bmatrix} trt_1 - trt_2 \\ trt_1 - trt_3 \\ \sigma_{tree/trt}^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$H_A \begin{bmatrix} trt_1 - trt_2 \\ trt_1 - trt_3 \\ \sigma_{tree/trt}^2 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

BUT, $\sigma_{tree/trt}^2 = 0$ means that there are no differences amongst trees, i.e. that they are all equal.

Thus, we could, by analogy, say that $\sigma_{tree/trt}^2 = 0$ is equal to a Null Hypothesis of:

H_o : Within treatments, there are no differences amongst trees, i.e. that, within treatments, trees are all equal.

$$\text{i.e. } tree_{11} = tree_{12} = tree_{13} = tree_{14}$$

$$\text{and } tree_{21} = tree_{22} = tree_{23} = tree_{24}$$

$$\text{and } tree_{31} = tree_{32} = tree_{33} = tree_{34}$$

which gives:

$$H_o \begin{bmatrix} tree_{11} - tree_{12} \\ tree_{11} - tree_{13} \\ tree_{11} - tree_{14} \\ tree_{21} - tree_{22} \\ tree_{21} - tree_{23} \\ tree_{21} - tree_{24} \\ tree_{31} - tree_{32} \\ tree_{31} - tree_{33} \\ tree_{31} - tree_{34} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad vs. \quad H_A \begin{bmatrix} tree_{11} - tree_{12} \\ tree_{11} - tree_{13} \\ tree_{11} - tree_{14} \\ tree_{21} - tree_{22} \\ tree_{21} - tree_{23} \\ tree_{21} - tree_{24} \\ tree_{31} - tree_{32} \\ tree_{31} - tree_{33} \\ tree_{31} - tree_{34} \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

We can therefore combine these to give a series of comparisons for the treatments and trees within treatments.

27.4 Matrix Equations

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ \vdots \\ Y_{116} \\ Y_{121} \\ \vdots \\ Y_{131} \\ \vdots \\ Y_{211} \\ \vdots \\ Y_{346} \end{bmatrix} = \begin{bmatrix} \mu & d_1 & d_2 & d_3 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & t_{11} & t_{12} \\ \mu & d_1 & d_2 & d_3 & t_{11} & t_{12} & t_{13} & t_{14} & t_{21} & t_{22} & t_{23} & t_{24} & t_{31} & t_{32} & t_{33} & t_{34} \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ trt_1 \\ trt_2 \\ trt_3 \\ tree_{11} \\ tree_{12} \\ tree_{13} \\ tree_{14} \\ tree_{21} \\ tree_{22} \\ tree_{23} \\ tree_{24} \\ tree_{31} \\ tree_{32} \\ tree_{33} \\ tree_{34} \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ \vdots \\ e_{116} \\ e_{121} \\ \vdots \\ e_{131} \\ \vdots \\ e_{211} \\ \vdots \\ e_{346} \end{bmatrix}$$

27.5 Normal Equations

$$\begin{bmatrix}
 72 & 24 & 24 & 24 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 \\
 24 & 24 & 0 & 0 & 6 & 6 & 6 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 24 & 0 & 24 & 0 & 0 & 0 & 0 & 0 & 6 & 6 & 6 & 6 & 0 & 0 & 0 \\
 24 & 0 & 0 & 24 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 6 & 6 \\
 6 & 6 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 6 & 6 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 6 & 6 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 6 & 6 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 6 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 \\
 6 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 \\
 6 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 \\
 6 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 \\
 6 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 \\
 6 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 \\
 6 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6
 \end{bmatrix}
 \begin{bmatrix}
 \tilde{\mu} \\
 \tilde{tr}t_1 \\
 \tilde{tr}t_2 \\
 \tilde{tr}t_3 \\
 \tilde{tree}_{11} \\
 \tilde{tree}_{12} \\
 \tilde{tree}_{13} \\
 \tilde{tree}_{14} \\
 \tilde{tree}_{21} \\
 \tilde{tree}_{22} \\
 \tilde{tree}_{23} \\
 \tilde{tree}_{24} \\
 \tilde{tree}_{31} \\
 \tilde{tree}_{32} \\
 \tilde{tree}_{33} \\
 \tilde{tree}_{34}
 \end{bmatrix}
 =
 \begin{bmatrix}
 25295.961 \\
 8003.348 \\
 8312.891 \\
 8979.722 \\
 1966.623 \\
 2005.769 \\
 2043.389 \\
 1987.567 \\
 2080.749 \\
 2172.589 \\
 2002.67 \\
 2056.883 \\
 2283.779 \\
 2230.488 \\
 2230.632 \\
 2234.823
 \end{bmatrix}$$

27.6 Analysis of Variance

<i>Analysis of Variance</i>				
<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F - ratio</i>
Total, TSS	$N = 72$	$Y'Y$ 8916948.5047		
Model, SSR	$r(X)$ $= 12$	$\tilde{b}'X'Y$ 8911427.8	742618.98	
Mean, C.F.	1	$N\bar{y}^2$ 8887300.6	8887300.6	
Model, after the mean, SSR_m $R(trt, tree Mean)$	$r(X) - 1$ $= 11$	$\tilde{b}'X'Y - N\bar{y}^2$ 24127.213741	2193.38307	23.84**
trt $R(trt Mean)$	$(d - 1)$ $= 2$	20747.0367	10373.518	$\frac{10373.518}{375.575}$ $= 27.62$
tree within trt $R(tree \mu, trt)$	$d(p - 1)$ $= 9$	3380.1771	375.575	4.08*
Error, Residual	$N - r(X)$ $72 - 12$	$Y'Y - \tilde{b}'X'Y$ 5520.695	92.01158	

27.7 Expectations of Mean Squares

trt	$(d - 1)$	$\sigma_e^2 + k_3 \sigma_{tree/trt}^2 + \frac{1}{d-1} \sum_i n_i (trt_i - \bar{trt})^2$ $= \sigma_e^2 + k_3 \sigma_{tree/trt}^2 + Q(trt)$
tree within trt	$d(p - 1)$	$\sigma_e^2 + k_2 \sigma_{tree/trt}^2$
Error,	$N - r(X)$	σ_e^2
Residual	$72 - 12 = 60$	

27.8 Computing Sums of Squares

After computing the Sums of Squares for the model corrected for the mean, $R(trt, tree(trt) | \mu)$, we need to compute the Reduction Sums of Squares for treatment, over and above the mean, or after the mean, $R(trt | \mu)$.

Let us return to our **[favourite]** fitted values:

$$\hat{Y}_{111} = \tilde{\mu} + \tilde{tr}t_1 + \tilde{tree}_{11} \quad k' = (1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$\hat{Y}_{121} = \tilde{\mu} + \tilde{tr}t_1 + \tilde{tree}_{12} \quad k' = (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

Then the difference between these two fitted values is estimable:

$$\hat{Y}_{111} - \hat{Y}_{121} = (\tilde{\mu} + \tilde{tr}t_1 + \tilde{tree}_{11}) - (\tilde{\mu} + \tilde{tr}t_1 + \tilde{tree}_{12}) = \tilde{tree}_{11} - \tilde{tree}_{12}$$

$$\text{with } k' = (0 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

Compare treatment 1, tree **1** vs treatment 1, tree **3**:

$$\hat{Y}_{111} = \tilde{\mu} + \tilde{tr}t_1 + \tilde{tree}_{11} \quad k' = (1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$\hat{Y}_{131} = \tilde{\mu} + \tilde{tr}t_1 + \tilde{tree}_{13} \quad k' = (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

Then the difference between these two fitted values is estimable:

$$\hat{Y}_{111} - \hat{Y}_{131} = (\tilde{\mu} + \tilde{t}r t_1 + \tilde{t}r \tilde{e}e_{11}) - (\tilde{\mu} + \tilde{t}r t_1 + \tilde{t}r \tilde{e}e_{13}) = \tilde{t}r \tilde{e}e_{11} - \tilde{t}r \tilde{e}e_{13}$$

$$\text{with } k' = (0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

Compare treatment 1, tree **1** vs treatment 1, tree **4**:

$$\hat{Y}_{111} = \tilde{\mu} + \tilde{t}r t_1 + \tilde{t}r \tilde{e}e_{11} \quad k' = (1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$\hat{Y}_{141} = \tilde{\mu} + \tilde{t}r t_1 + \tilde{t}r \tilde{e}e_{14} \quad k' = (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

Then the difference between these two fitted values is estimable:

$$\hat{Y}_{111} - \hat{Y}_{141} = (\tilde{\mu} + \tilde{t}r t_1 + \tilde{t}r \tilde{e}e_{11}) - (\tilde{\mu} + \tilde{t}r t_1 + \tilde{t}r \tilde{e}e_{14}) = \tilde{t}r \tilde{e}e_{11} - \tilde{t}r \tilde{e}e_{14}$$

$$\text{with } k' = (0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

Putting these comparisons together:

$$k'_{1-2} = (0 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$k'_{1-3} = (0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$k'_{1-4} = (0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$k' = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

We can repeat the same exercise to make the comparisons amongst the trees on treatment 2:

$$k' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

We can repeat the same exercise to make the comparisons amongst the trees on treatment **3**:

$$k' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}$$

Putting these 3 k' matrices together we obtain a k' for the effect of trees nested within treatments, $SS_{tree(trt)}$:

$$k' = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}$$

NOTE: k' is a 9*16 matrix, which corresponds to our 9 degrees of freedom.

What about treatments? Consider our 4 fitted values for the 4 trees on treatment 1:

$$\hat{Y}_{111} = \tilde{\mu} + \tilde{tr}t_1 + \tilde{tree}_{11} \quad k' = (1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$\hat{Y}_{121} = \tilde{\mu} + \tilde{tr}t_1 + \tilde{tree}_{12} \quad k' = (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$\hat{Y}_{131} = \tilde{\mu} + \tilde{tr}t_1 + \tilde{tree}_{13} \quad k' = (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$\hat{Y}_{141} = \tilde{\mu} + \tilde{tr}t_1 + \tilde{tree}_{14} \quad k' = (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

The sum of these is estimable, we know:

$$\sum_{j=1}^{j=4} = 4\tilde{\mu} + 4\tilde{tr}t_1 + \tilde{tree}_{11} + \tilde{tree}_{12} + \tilde{tree}_{13} + \tilde{tree}_{14} = 4\tilde{\mu} + 4\tilde{tr}t_1 + \sum_{j=1}^{j=4} \tilde{tree}_{1j}$$

$$k' = (4 \quad 4 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

and their average:

$$\frac{1}{4} \sum_{j=1}^{j=4} = \tilde{\mu} + \tilde{tr}t_1 + \frac{1}{4} \tilde{tr}ee_{11} + \frac{1}{4} \tilde{tr}ee_{12} + \frac{1}{4} \tilde{tr}ee_{13} + \frac{1}{4} \tilde{tr}ee_{14} = \tilde{\mu} + \tilde{tr}t_1 + \frac{1}{4} \sum_{j=1}^{j=4} \tilde{tr}ee_{1j}$$

$$k' = (1 \quad 1 \quad 0 \quad 0 \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

Note: this k' corresponds to the LSmean for treatment 1, i.e. the average over the 4 fitted values for (treatment, tree).

What about treatments? Consider our 4 fitted values for the 4 trees on treatment 2:

$$\hat{Y}_{211} = \tilde{\mu} + \tilde{tr}t_2 + \tilde{tr}ee_{21} \quad k' = (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$\hat{Y}_{221} = \tilde{\mu} + \tilde{tr}t_2 + \tilde{tr}ee_{22} \quad k' = (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$\hat{Y}_{231} = \tilde{\mu} + \tilde{tr}t_2 + \tilde{tr}ee_{23} \quad k' = (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

$$\hat{Y}_{241} = \tilde{\mu} + \tilde{tr}t_2 + \tilde{tr}ee_{24} \quad k' = (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0)$$

The sum of these is estimable, we know:

$$\sum_{j=1}^{j=4} = 4\tilde{\mu} + 4\tilde{tr}t_2 + \tilde{tr}ee_{21} + \tilde{tr}ee_{22} + \tilde{tr}ee_{23} + \tilde{tr}ee_{24} = 4\tilde{\mu} + 4\tilde{tr}t_2 + \sum_{j=1}^{j=4} \tilde{tr}ee_{2j}$$

$$k' = (4 \quad 0 \quad 4 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0)$$

and their average:

$$\frac{1}{4} \sum_{j=1}^{j=4} = \tilde{\mu} + \tilde{tr}t_2 + \frac{1}{4} \tilde{tr}ee_{21} + \frac{1}{4} \tilde{tr}ee_{22} + \frac{1}{4} \tilde{tr}ee_{23} + \frac{1}{4} \tilde{tr}ee_{24} = \tilde{\mu} + \tilde{tr}t_2 + \frac{1}{4} \sum_{j=1}^{j=4} \tilde{tr}ee_{2j}$$

$$k' = (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad 0 \quad 0 \quad 0 \quad 0)$$

Note: this k' corresponds to the LSmean for treatment 2, i.e. the average over the 4 fitted values for (treatment, tree).

Thus the difference between treatment 1 (averaged over the trees on treatment 1) and treatment 2 (averaged over the trees on treatment 2) is estimable:

$$\begin{aligned}
k'_{trt\ 1} &= (1 \quad 1 \quad 0 \quad 0 \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0) \\
-k'_{trt\ 2} &= (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad 0 \quad 0 \quad 0 \quad 0) \\
&= k'_{1-2} = (0 \quad 1 \quad -1 \quad 0 \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad -\frac{1}{4} \quad -\frac{1}{4} \quad -\frac{1}{4} \quad -\frac{1}{4} \quad 0 \quad 0 \quad 0 \quad 0)
\end{aligned}$$

In an exactly analogous manner the difference between treatment 1 (averaged over the trees on treatment 1) and treatment 3 (averaged over the trees on treatment 3) is estimable:

$$\begin{aligned}
k'_{trt\ 1} &= (1 \quad 1 \quad 0 \quad 0 \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0) \\
-k'_{trt\ 3} &= (1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4}) \\
&= k'_{1-3} = (0 \quad 1 \quad 0 \quad -1 \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad 0 \quad 0 \quad 0 \quad 0 \quad -\frac{1}{4} \quad -\frac{1}{4} \quad -\frac{1}{4} \quad -\frac{1}{4})
\end{aligned}$$

We can then combine these 2 contrasts (trt1-2 and trt1-3),

$$k'_{trt} = \begin{pmatrix} 0 & 1 & -1 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \end{pmatrix}$$

NOTE: this k' matrix is a 2*16 matrix, it has 2 rows, corresponding to the 2 degrees of freedom amongst our 3 treatments.

Actually this is still a simplification, we really have effects which are fixed (treatments) and effects which are random (trees) and hence we have what is more correctly known as a 'mixed model'. We should use the PROC MIXED procedure of SAS.

27.9 Comparisons amongst treatment means

The appropriate Mean Square/Variance to use for the treatment contrasts is the Mean Square between trees within treatments. This is the same Mean Square as used to test the statistical significance of treatments!

Similarly, for our lsmeans and estimates for treatment effects (e.g. differences amongst treatments) the Mean Square (σ^2) to use in our formula for the sampling variance is: $MS_{tree/trt}$

However, in GLM the ESTIMATE statement **DOES NOT ALLOW** of an error other than MSE, so you have to back-calculate yourself! PROC MIXED (being a proper mixed model) gets it right **AUTOMAGICALLY**.

27.10 Analysis using SAS

USING SAS/PROC GLM

```
data subsamp1;
input trt tree apple wt;
cards;
  1  1  1  313.063
  1  1  2  329.132
  1  1  3  334.278
  1  1  4  330.088
  1  1  5  334.987
  1  1  6  325.075
  1  2  1  333.936
  1  2  2  326.155
  1  2  3  352.854
  1  2  4  350.791
  1  2  5  318.560
  1  2  6  323.473
  1  3  1  345.494
  1  3  2  349.296
  1  3  3  339.190
  1  3  4  338.942
  1  3  5  331.370
  1  3  6  339.097
```

1	4	1	340.840
1	4	2	336.798
1	4	3	313.810
1	4	4	333.880
1	4	5	343.068
1	4	6	319.171
2	5	1	349.271
2	5	2	336.695
2	5	3	352.797
2	5	4	348.486
2	5	5	352.077
2	5	6	341.423
2	6	1	356.880
2	6	2	356.256
2	6	3	364.950
2	6	4	360.570
2	6	5	362.104
2	6	6	371.829
2	7	1	324.161
2	7	2	340.130
2	7	3	334.580
2	7	4	342.813
2	7	5	327.415
2	7	6	333.571
2	8	1	338.742
2	8	2	340.348
2	8	3	362.837
2	8	4	340.782
2	8	5	348.730
2	8	6	325.444
3	9	1	387.868
3	9	2	372.807
3	9	3	380.505
3	9	4	391.804
3	9	5	388.935
3	9	6	361.860
3	10	1	377.948
3	10	2	380.033
3	10	3	361.913
3	10	4	363.098
3	10	5	365.375
3	10	6	382.121
3	11	1	363.583

```

3 11 2 387.727
3 11 3 373.021
3 11 4 362.931
3 11 5 378.928
3 11 6 364.442
3 12 1 374.851
3 12 2 361.291
3 12 3 377.389
3 12 4 366.722
3 12 5 374.187
3 12 6 380.383
;
proc glm;
classes trt tree;
model wt = trt tree(trt);
random tree(trt)/test;
contrast 'SS trt' trt 1 -1 0 tree(trt) .25 .25 .25 .25
-.25 -.25 -.25 -.25 0 0 0 0,
trt 1 0 -1 tree(trt) .25 .25 .25 .25
0 0 0 0 -.25 -.25 -.25 -.25/E=tree(trt);
lsmeans trt/stderr pdiff e=tree(trt) adjust=bon;
run;
quit;

```

USING SAS/PROC MIXED

```

proc mixed data=subsamp1;
classes trt tree;
model wt = trt;
random tree(trt);
lsmeans trt/adjust=scheffe;
/* NOTE, in proc mixed we ONLY specify the fixed effects parts
in the ESTIMATE statements. There is no need to include
coefficients for the random part, proc mixed 'knows'
about the random effects correctly and automagically!
*/
estimate 'trt 1 - 2' trt 1 -1 0;
estimate 'trt 1 - 3' trt 1 0 -1;
estimate 'trt 2 - 3' trt 0 1 -1;
run;
quit;

```

27.11 SAS output

The SAS System

The GLM Procedure

Class Level Information		
Class	Levels	Values
trt	3	1 2 3
tree	12	1 2 3 4 5 6 7 8 9 10 11 12

Number of observations	72
------------------------	----

The SAS System

The GLM Procedure

Dependent Variable: wt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	24127.21374	2193.38307	23.84	<.0001
Error	60	5520.69496	92.01158		
Corrected Total	71	29647.90869			

R-Square	Coeff Var	Root MSE	wt Mean
0.813791	2.730251	9.592267	351.3328

Source	DF	Type I SS	Mean Square	F Value	Pr > F
trt	2	20747.03666	10373.51833	112.74	<.0001
tree(trt)	9	3380.17708	375.57523	4.08	0.0004

The SAS System

The GLM Procedure

Source	Type I Expected Mean Square
trt	$\text{Var}(\text{Error}) + 6 \text{Var}(\text{tree}(\text{trt})) + Q(\text{trt})$
tree(trt)	$\text{Var}(\text{Error}) + 6 \text{Var}(\text{tree}(\text{trt}))$

The SAS System

The GLM Procedure

Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: wt

Source	DF	Type I SS	Mean Square	F Value	Pr > F
trt	2	20747	10374	27.62	0.0001
Error: MS(tree(trt))	9	3380.177080	375.575231		

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tree(trt)	9	3380.177080	375.575231	4.08	0.0004
Error: MS(Error)	60	5520.694957	92.011583		

The SAS System

Least Squares Means

Adjustment for Multiple Comparisons: Bonferroni

Standard Errors and Probabilities Calculated Using the Type I MS for tree(trt) as an Error Term

trt	wt LSMEAN	Standard Error	Pr > t	LSMEAN Number
1	333.472833	3.955878	<.0001	1
2	346.370458	3.955878	<.0001	2
3	374.155083	3.955878	<.0001	3

Pr > t for H0: LSMean(i)=LSMean(j)Dep. Variable: wt			
i/j	1	2	3
1		0.1398	0.0001
2	0.1398		0.0023
3	0.0001	0.0023	

The SAS System

Dependent Variable: wt

Tests of Hypotheses Using the Type I MS for tree(trt) as an Error Term					
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
SS trt	2	20747.03666	10373.51833	27.62	0.0001

The SAS System

The Mixed Procedure

Model Information	
Data Set	WORK.SUBSAMP1
Dependent Variable	wt
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information		
Class	Levels	Values
trt	3	1 2 3
tree	12	1 2 3 4 5 6 7 8 9 10 11 12

Dimensions	
Covariance Parameters	2
Columns in X	4
Columns in Z	12
Subjects	1
Max Obs Per Subject	72
Observations Used	72
Observations Not Used	0
Total Observations	72

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	540.67374301	
1	1	530.01867630	0.00000000

Convergence criteria met.

Covariance Parameter Estimates	
Cov Parm	Estimate
tree(trt)	47.2606
Residual	92.0116

Fit Statistics	
-2 Res Log Likelihood	530.0
AIC (smaller is better)	534.0
AICC (smaller is better)	534.2
BIC (smaller is better)	535.0

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
trt	2	9	27.62	0.0001

Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr > t
trt 1 - 2	-12.8976	5.5945	9	-2.31	0.0466
trt 1 - 3	-40.6823	5.5945	9	-7.27	<.0001
trt 2 - 3	-27.7846	5.5945	9	-4.97	0.0008

Least Squares Means						
Effect	trt	Estimate	Standard Error	DF	t Value	Pr > t
trt	1	333.47	3.9559	9	84.30	<.0001
trt	2	346.37	3.9559	9	87.56	<.0001
trt	3	374.16	3.9559	9	94.58	<.0001

Differences of Least Squares Means									
Effect	trt	_ trt	Estimate	Standard Error	DF	t Value	Pr > t	Adjustment	Adj P
trt	1	2	-12.8976	5.5945	9	-2.31	0.0466	Scheffe	0.1239
trt	1	3	-40.6823	5.5945	9	-7.27	<.0001	Scheffe	0.0002
trt	2	3	-27.7846	5.5945	9	-4.97	0.0008	Scheffe	0.0026

27.12 Using Group Means

One could consider using the mean of the 6 apples as the observation and hence having 12 observations, 4 per diet, and hence a linear model:

$$Y_{ij} = \mu + trt_i + e_{ij}$$

This model would be adequate to test whether there were differences between the treatments.

Why then should we go to the bother of a subsampling model for our analysis?

Why not simply average the 6 individual weights and be done with it?

Well there are several reasons.

1. We may not have the same number of subsamples (apples) on each tree, hence using the mean would cause the variances to be non-homogeneous

2. We may want to know the variability between apples and between trees, so that we will be able to plan the optimal allocation for subsequent trials, see STD, Ch 7.9.

3. We will need this information when we design experiments where there is subsampling; to determine how many experimental units we need.

4. The variation from experimental unit to experimental unit, vs the variation amongst sub-samples may well be biologically quite interesting in its own right, over and above any differences amongst treatments!

27.13 Expectation of Mean Squares

Steel, Torrie and Dickey, Ch. 7.6, page 157, and Table 7.9, page 163

	Treatment fixed, tree random	
Source of Variation	df	Expected Mean Square

Treatment	a-1	$\sigma_e^2 + c\sigma_{tree}^2 + bc \sum \alpha^2 / (a - 1)$
Tree	a(b-1)	$\sigma_e^2 + c\sigma_{tree}^2$
Residual	ab(c-1)	σ_e^2

1) If σ_e^2 is large relative to σ_{tree}^2 then several samples per tree is beneficial.

2) If σ_e^2 is small relative to σ_{tree}^2 the use only a few samples and have more experimental units per treatment, i.e. more trees per treatment.

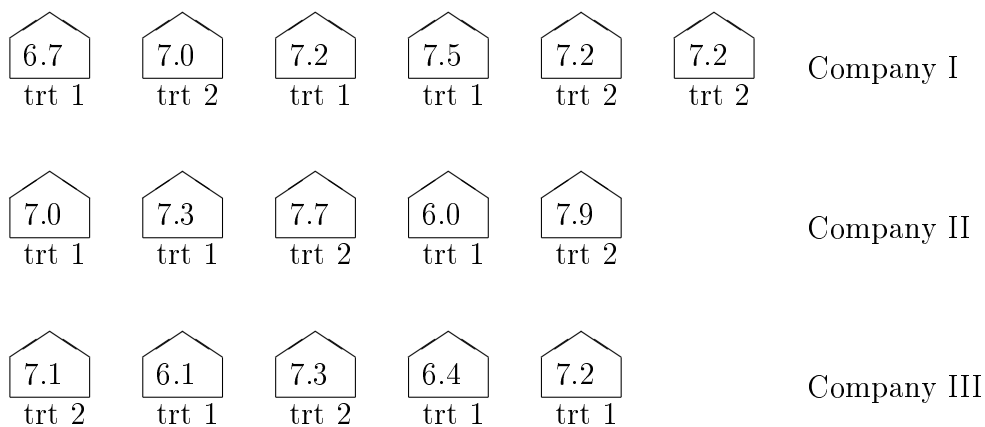
28 Factorial Experiments

STD. Ch.15. P352

When we have two or more factors, with two or more levels each, and we want to see if the effects of the factors are simply additive, or whether there is an 'interaction'. An 'interaction', between 2 factors (A and B for example) means that the effect of the difference between the levels of A (for example $A_1 - A_2$) depends upon the level of B, *i.e.* that the effects A and B are not independent of one another. In other words they interact with each other; hence the concept of 'interaction'.

Suppose that we are looking at the storage ability of potato stores built by 3 different companies and at 2 different chemical treatments used for treating potatoes as they are put into storage to help them keep longer before sprouting. So we have 2 factors to look at: Company and Treatment. We think that there may be an 'interaction' between the effects of the different companies' stores and the different treatments so we arrange to have several buildings from each Company on each Treatment. This type of design is what is known as a "Factorial Experiment" and allows us to examine for interaction effects. In this case, with Company having 3 levels and Treatment having 2 levels we say that it is a 3x2 Factorial; there are 6 combinations.

We have 6 buildings of Company I, 5 buildings of Company II and 5 buildings stores of Company III. The 2 treatments (1 and 2) will be randomly split amongst the 6 Company I buildings, for Company II and III we have 3 buildings on treatment 1 and 2 buildings on treatment 2.



28.1 Observations

		Trt	
		1	2
Store Company	1	6.7, 7.2 7.5	7.0, 7.2 7.2
	2	7.0, 7.3 6.0	7.7, 7.9
	3	6.1, 6.4 7.2	7.1, 7.3

28.2 SAS code for Data Step

```

data fac1;
/* Assume s = Store, t = Treatment */
input s t y;
cards;
1 1 6.7
1 1 7.2

```



```

1 1 7.5
1 2 7.0
1 2 7.2
1 2 7.2
2 1 7.0
2 1 7.3
2 1 6.0
2 2 7.7
2 2 7.9
3 1 6.1
3 1 6.4
3 1 7.2
3 2 7.1
3 2 7.3
;

```

28.3 Linear model

We can consider that there is an effect of the actual building (the experimental unit =, nested within store*trt) as well as the sampling error.

$$Y_{ijk} = \mu + store_i + trt_j + store * trt_{ij} + building_{ijk} + \epsilon_{ijk}$$

However, as we have previously noted, since we have only 1 measurement per building (experimental unit) we cannot separate the building effect from the sampling effect; they are 'confounded'. Thus we shall 'lump' them together into the error. The error can be considered to be the experimental unit error term and is the variation amongst buildings within store*trt.

$$Y_{ijk} = \mu + store_i + trt_j + store * trt_{ij} + e_{ijk}$$

28.4 Parameters of the Model

$\mu, s_1, s_2, s_3, t_1, t_2, s_1t_1, s_1t_2, s_2t_1, s_2t_2, s_3t_1, s_3t_2, \sigma_e^2$

28.5 Linear model in Matrix Notation

$$Y = Xb + e$$

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ \cdot \\ \cdot \\ Y_{121} \\ \cdot \\ \cdot \\ Y_{211} \\ \cdot \\ \cdot \\ Y_{221} \\ \cdot \\ \cdot \\ Y_{322} \end{bmatrix} = \begin{bmatrix} \mu & s_1 & s_2 & s_3 & t_1 & t_2 & s_1t_1 & s_1t_2 & s_2t_1 & s_2t_2 & s_3t_1 & s_3t_2 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \cdot & & & & & & & & & & & \\ \cdot & & & & & & & & & & & \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ \cdot & & & & & & & & & & & \\ \cdot & & & & & & & & & & & \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \cdot & & & & & & & & & & & \\ \cdot & & & & & & & & & & & \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ \cdot & & & & & & & & & & & \\ \cdot & & & & & & & & & & & \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} \mu \\ s_1 \\ s_2 \\ s_3 \\ t_1 \\ t_2 \\ s_1t_1 \\ s_1t_2 \\ s_2t_1 \\ s_2t_2 \\ s_3t_1 \\ s_3t_2 \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ \cdot \\ \cdot \\ e_{121} \\ \cdot \\ \cdot \\ e_{211} \\ \cdot \\ \cdot \\ e_{221} \\ \cdot \\ \cdot \\ e_{322} \end{bmatrix}$$

28.6 Normal Equations

$$X'X\tilde{b} = X'Y$$

$$\begin{bmatrix} 16 & 6 & 5 & 5 & 9 & 7 & 3 & 3 & 3 & 2 & 3 & 2 \\ 6 & 6 & 0 & 0 & 3 & 3 & 3 & 3 & 0 & 0 & 0 & 0 \\ 5 & 0 & 5 & 0 & 3 & 2 & 0 & 0 & 3 & 2 & 0 & 0 \\ 5 & 0 & 0 & 5 & 3 & 2 & 0 & 0 & 0 & 0 & 3 & 2 \\ 9 & 3 & 3 & 3 & 9 & 0 & 3 & 0 & 3 & 0 & 3 & 0 \\ 7 & 3 & 2 & 2 & 0 & 7 & 0 & 3 & 0 & 2 & 0 & 2 \\ 3 & 3 & 0 & 0 & 3 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 3 & 3 & 0 & 0 & 0 & 3 & 0 & 3 & 0 & 0 & 0 & 0 \\ 3 & 0 & 3 & 0 & 3 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 2 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 \\ 3 & 0 & 0 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 2 & 0 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 22 \end{bmatrix} \begin{bmatrix} \tilde{\mu} \\ \tilde{s}_1 \\ \tilde{s}_2 \\ \tilde{s}_3 \\ \tilde{t}_1 \\ \tilde{t}_2 \\ \tilde{s}_1\tilde{t}_1 \\ \tilde{s}_1\tilde{t}_2 \\ \tilde{s}_2\tilde{t}_1 \\ \tilde{s}_2\tilde{t}_2 \\ \tilde{s}_3\tilde{t}_1 \\ \tilde{s}_3\tilde{t}_2 \end{bmatrix} = \begin{bmatrix} 112.8 \\ 42.8 \\ 35.9 \\ 34.1 \\ 61.4 \\ 51.4 \\ 21.4 \\ 21.4 \\ 20.3 \\ 15.6 \\ 19.7 \\ 14.4 \end{bmatrix}$$

28.7 Solution to the linear model

$$\tilde{b} = (X'X)^{-1}X'Y$$

$$\begin{bmatrix} \tilde{\mu} \\ \tilde{s}_1 \\ \tilde{s}_2 \\ \tilde{s}_3 \\ \tilde{t}_1 \\ \tilde{t}_2 \\ s_1\tilde{t}_1 \\ s_1\tilde{t}_2 \\ s_2\tilde{t}_1 \\ s_2\tilde{t}_2 \\ s_3\tilde{t}_1 \\ s_3\tilde{t}_2 \end{bmatrix} = \begin{bmatrix} 7.2 \\ -0.66666 \\ 0.60 \\ 0.00 \\ -0.63333 \\ 0.00 \\ 0.63333 \\ 0.00 \\ -0.400 \\ 0.00 \\ 0.00 \\ 0.00 \end{bmatrix}$$

28.7.1 SAS code for Factorial model

```
/* Assume s = Store, t = Treatment */  
proc glm;  
classes s t;  
model y = s t s*t;  
run;
```

28.8 Basic Analysis of Variance

ANOVA					
<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F - ratio</i>	<i>E(MS)</i>
Total, TSS	$N = 16$	$Y'Y$ 799.36			
Model, SSR	$r(X)$ $= 6$	$\tilde{b}'X'Y$ 797.995	132.99	974**	
Mean, C.F.	1	$N\bar{y}^2$ 795.24	795.24	5825.9**	
Model, after the mean, SSR_m $R(t, s, t * s Mean)$	$r(X) - 1$ $= 5$	$\tilde{b}'X'Y - N\bar{y}^2$ 2.755	0.551	4.037	$\sigma_e^2 + Q(s, t, st)$
Error, SSE	$N - r(X)$ 10	$Y'Y - \tilde{b}'X'Y$ 1.365	0.1365	σ_e^2	

28.9 Hypotheses of Interest

- 1) H_o the levels of Store are all equal, *i.e.* $\text{effect}_{s_1} = \text{effect}_{s_2} = \text{effect}_{s_3}$
vs H_A the levels of Store are not all equal

- 2) H_o the levels of Treatment are all equal, *i.e.* $\text{Trt}_A = \text{Trt}_B$
vs H_A the levels of Treatment are not all equal; $\text{Trt}_A \neq \text{Trt}_B$

- 3) H_o the interaction effects are all equal
i.e. $a_1b_1 = a_1b_2 = a_2b_1 = a_2b_2$
vs H_A the interaction effects are not all equal

28.10 Derivation of Testable Hypotheses

What exactly is estimable and testable, how and why? As always we can, and should, start from the fact that the 'fitted values' are estimable and that anything and everything that is estimable can be and must be able to be written as a linear function of the 'fitted values' and the appropriate k' matrix. Thus $\hat{Y} = X\tilde{b}$.

$$\begin{aligned} \text{So } \hat{Y}_{11} &= \tilde{\mu} + \tilde{s}_1 + \tilde{t}_1 + s_1\tilde{t}_1 \\ k'_{11} &= (1 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0) \end{aligned}$$

$$\begin{aligned} \hat{Y}_{12} &= \tilde{\mu} + \tilde{s}_1 + \tilde{t}_2 + s_1\tilde{t}_2 \\ k'_{12} &= (1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0) \end{aligned}$$

$$\begin{aligned} \hat{Y}_{21} &= \tilde{\mu} + \tilde{s}_2 + \tilde{t}_1 + s_2\tilde{t}_1 \\ k'_{21} &= (1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0) \end{aligned}$$

$$\begin{aligned} \hat{Y}_{22} &= \tilde{\mu} + \tilde{s}_2 + \tilde{t}_2 + s_2\tilde{t}_2 \\ k'_{22} &= (1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0) \end{aligned}$$

$$\begin{aligned} \hat{Y}_{31} &= \tilde{\mu} + \tilde{s}_3 + \tilde{t}_1 + s_3\tilde{t}_1 \\ k'_{31} &= (1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0) \end{aligned}$$

$$\begin{aligned} \hat{Y}_{32} &= \tilde{\mu} + \tilde{s}_3 + \tilde{t}_2 + s_3\tilde{t}_2 \\ k'_{32} &= (1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1) \end{aligned}$$

28.10.1 SAS code for Fitted values

```
/* Assume s = Store, t = Treatment,
   Note how each statement goes to another line, terminated
   by a semi-colon
*/
estimate 'mu + s1 + t1 + s1t1' intercept 1 s 1 0 0 t 1 0
s*t 1 0 0 0 0 0;
estimate 'mu + s1 + t2 + s1t2' intercept 1 s 1 0 0 t 0 1
s*t 0 1 0 0 0 0;
```

```

estimate 'mu + s2 + t1 + s2t1' intercept 1 s 0 1 0 t 1 0
      s*t 0 0 1 0 0 0;
estimate 'mu + s2 + t2 + s2t2' intercept 1 s 0 1 0 t 0 1
      s*t 0 0 0 1 0 0;
estimate 'mu + s3 + t1 + s3t1' intercept 1 s 0 0 1 t 1 0
      s*t 0 0 0 0 1 0;
estimate 'mu + s3 + t2 + s3t2' intercept 1 s 0 0 1 t 0 1
      s*t 0 0 0 0 0 1;

```

28.10.2 Factor Store

Let us look at comparison(s) between levels of Factor Store, to see just what is estimable and what the suitable contrasts are.

We can see, from the section above, that \hat{Y}_{11} and \hat{Y}_{12} are both estimable, hence a linear function of them (their sum) is also estimable, *i.e.* $\hat{Y}_{11} + \hat{Y}_{12}$ and estimates:

$$\begin{aligned} \hat{Y}_{11} + \hat{Y}_{12} &= 2\tilde{\mu} + 2\tilde{s}_1 + \tilde{t}_1 + \tilde{t}_2 + s_1\tilde{t}_1 + s_1\tilde{t}_2 \\ k' &= (2 \quad 2 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0) \end{aligned}$$

Averaging, to bring it back to a 'per-unit' basis we get

$$(\hat{Y}_{11} + \hat{Y}_{12})/2 = \tilde{\mu} + \tilde{s}_1 + \frac{1}{2}\tilde{t}_1 + \frac{1}{2}\tilde{t}_2 + \frac{1}{2}s_1\tilde{t}_1 + \frac{1}{2}s_1\tilde{t}_2$$

$$\text{and } k'_1 = (1 \quad 1 \quad 0 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0 \quad 0 \quad 0)$$

Similarly for $\hat{Y}_{21} + \hat{Y}_{22}$

$$\hat{Y}_{21} + \hat{Y}_{22} = 2\tilde{\mu} + 2\tilde{s}_2 + \tilde{t}_1 + \tilde{t}_2 + s_2\tilde{t}_1 + s_2\tilde{t}_2$$

again, averaging to bring us back to a 'per-unit' basis we get

$$(\hat{Y}_{21} + \hat{Y}_{22})/2 = \tilde{\mu} + \tilde{s}_2 + \frac{1}{2}\tilde{t}_1 + \frac{1}{2}\tilde{t}_2 + \frac{1}{2}s_2\tilde{t}_1 + \frac{1}{2}s_2\tilde{t}_2$$

$$\text{and } k'_2 = (1 \quad 0 \quad 1 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0)$$

Similarly for $\hat{Y}_{31} + \hat{Y}_{32}$

$$\hat{Y}_{31} + \hat{Y}_{32} = 2\tilde{\mu} + 2\tilde{s}_3 + \tilde{t}_1 + \tilde{t}_2 + s_3\tilde{t}_1 + s_3\tilde{t}_2$$

again, averaging to bring us back to a 'per-unit' basis we get

$$(\hat{Y}_{31} + \hat{Y}_{32})/2 = \tilde{\mu} + \tilde{s}_2 + \frac{1}{2}\tilde{t}_1 + \frac{1}{2}\tilde{t}_2 + \frac{1}{2}s_3\tilde{t}_1 + \frac{1}{2}s_3\tilde{t}_2$$

$$\text{and } k'_3 = (1 \quad 0 \quad 0 \quad 1 \quad \frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0 \quad 0 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2})$$

k'_1 , k'_2 and k'_3 provide linear functions of the fitted values, therefore their differences will also be linear functions of the fitted values and hence estimable.

$$k'_1 = (1 \quad 1 \quad 0 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0 \quad 0 \quad 0)$$

$$-k'_2 = (1 \quad 0 \quad 1 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0)$$

$$k'_{1-2} = (0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2} \quad -\frac{1}{2} \quad -\frac{1}{2} \quad 0 \quad 0)$$

This estimates

$$(\tilde{s}_1 - \tilde{s}_2 + \frac{1}{2}s_1\tilde{t}_1 + \frac{1}{2}s_1\tilde{t}_2 - \frac{1}{2}s_2\tilde{t}_1 - \frac{1}{2}s_2\tilde{t}_2)$$

Likewise

$$k'_1 = (1 \quad 1 \quad 0 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0 \quad 0 \quad 0)$$

$$-k'_3 = (1 \quad 0 \quad 0 \quad 1 \quad \frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0 \quad 0 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2})$$

$$k'_{1-3} = (0 \quad 1 \quad 0 \quad -1 \quad 0 \quad 0 \quad \frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0 \quad -\frac{1}{2} \quad -\frac{1}{2})$$

This estimates

$$(\tilde{s}_1 - \tilde{s}_3 + \frac{1}{2}s_1\tilde{t}_1 + \frac{1}{2}s_1\tilde{t}_2 - \frac{1}{2}s_3\tilde{t}_1 - \frac{1}{2}s_3\tilde{t}_2)$$

This is the best that we can do to test the hypothesis that the stores are equal; *i.e.*:

$$s_1 = s_2$$

$$s_1 = s_3$$

It illustrates the very important point that with an interaction component implicating Factor*Store we cannot completely remove the interaction effects! Hence, if the Null Hypothesis for the interaction cannot be accepted, then the Sums of Squares for Stores will not be free of some effects of the interaction. Thus we shall not know if the Sums of Squares for Stores is really due to the main effect of Store and/or due to the effects of the interaction.

Combining the 2 k' matrices (k'_{1-2} and k'_{1-3}) as 2 rows we get

$$k' = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & .5 & .5 & -.5 & -.5 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & .5 & .5 & 0 & 0 & -.5 & -.5 \end{pmatrix}$$

Using this k' matrix we can compute the Sums of Squares for Factor Store. Note that since we have 3 levels of Store there are 2 degrees of freedom between levels of Store. If we had more than 3 levels of A we would have (a-1) linearly independent comparisons (a being the number of levels of this factor) and proceed accordingly.

28.10.3 SAS CONTRAST code for Store

```
/* Marginal, Type III, Sums of Squares for Stores */
contrast 'Stores' s 1 -1 0 s*t .5 .5 -.5 -.5 0 0,
                s 1 0 -1 s*t .5 .5 0 0 -.5 -.5;
```

28.10.4 Factor Treatment

Let us look at comparison(s) between levels of Factor Treatment, to see just what is estimable and what the suitable contrasts are.

We can see, from the above section on fitted values, that \hat{Y}_{11} , \hat{Y}_{21} and \hat{Y}_{31} are all estimable, hence a linear function of them (their sum) is also estimable, *i.e.* $\hat{Y}_{11} + \hat{Y}_{21} + \hat{Y}_{31}$ and estimates:

$$\begin{aligned} \hat{Y}_{11} + \hat{Y}_{21} + \hat{Y}_{31} &= 3\tilde{\mu} + \tilde{s}_1 + \tilde{s}_2 + \tilde{s}_3 + 3\tilde{t}_1 + s_1\tilde{t}_1 + s_2\tilde{t}_1 + s_3\tilde{t}_1 \\ k' &= (3 \quad 1 \quad 1 \quad 1 \quad 3 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0) \end{aligned}$$

Averaging, to bring it back to a 'per-unit' basis we get

$$(\hat{Y}_{11} + \hat{Y}_{21} + \hat{Y}_{31})/3 = \tilde{\mu} + \frac{1}{3}\tilde{s}_1 + \frac{1}{3}\tilde{s}_2 + \frac{1}{3}\tilde{s}_3 + \tilde{t}_1 + \frac{1}{3}s_1\tilde{t}_1 + \frac{1}{3}s_2\tilde{t}_1 + \frac{1}{3}s_3\tilde{t}_1$$

$$\text{and } k'_1 = (1 \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3} \quad 1 \quad 0 \quad \frac{1}{3} \quad 0 \quad \frac{1}{3} \quad 0 \quad \frac{1}{3} \quad 0)$$

Similarly, \hat{Y}_{12} , \hat{Y}_{22} and \hat{Y}_{32} are all estimable, hence a linear function of them (their sum) is also estimable, *i.e.* $\hat{Y}_{12} + \hat{Y}_{22} + \hat{Y}_{32}$ and estimates:

$$\begin{aligned} \hat{Y}_{12} + \hat{Y}_{22} + \hat{Y}_{32} &= 3\tilde{\mu} + \tilde{s}_1 + \tilde{s}_2 + \tilde{s}_3 + 3\tilde{t}_2 + s_1\tilde{t}_2 + s_2\tilde{t}_2 + s_3\tilde{t}_2 \\ k' &= (3 \quad 1 \quad 1 \quad 1 \quad 0 \quad 3 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1) \end{aligned}$$

Averaging, to bring it back to a 'per-unit' basis we get

$$(\hat{Y}_{12} + \hat{Y}_{22} + \hat{Y}_{32})/3 = \tilde{\mu} + \frac{1}{3}\tilde{s}_1 + \frac{1}{3}\tilde{s}_2 + \frac{1}{3}\tilde{s}_3 + \tilde{t}_2 + \frac{1}{3}s_1\tilde{t}_2 + \frac{1}{3}s_2\tilde{t}_2 + \frac{1}{3}s_3\tilde{t}_2$$

$$\text{and } k'_2 = (1 \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3} \quad 0 \quad 1 \quad 0 \quad \frac{1}{3} \quad 0 \quad \frac{1}{3} \quad 0 \quad \frac{1}{3})$$

k'_1 and k'_2 provide linear functions of the fitted values, therefore their difference will also be a linear function of the fitted values and hence estimable.

$$\begin{aligned} k'_1 &= (1 \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3} \quad 1 \quad 0 \quad \frac{1}{3} \quad 0 \quad \frac{1}{3} \quad 0 \quad \frac{1}{3} \quad 0) \\ -k'_2 &= (1 \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3} \quad 0 \quad 1 \quad 0 \quad \frac{1}{3} \quad 0 \quad \frac{1}{3} \quad 0 \quad \frac{1}{3}) \end{aligned}$$

$$k'_{1-2} = \left(0 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad \frac{1}{3} \quad -\frac{1}{3} \quad \frac{1}{3} \quad -\frac{1}{3} \quad \frac{1}{3} \quad -\frac{1}{3} \right)$$

With this k' matrix we can compute the Sums of Squares for Treatments. One problem is that in the above we have written $\frac{1}{3}$; however, for SAS we have to write the coefficients as decimals, and $\frac{1}{3}$ cannot be expressed exactly as a decimal. A solution is to scale this up to the Lowest Common Multiple, *i.e.* multiple all the coefficients by 3.

28.10.5 SAS CONTRAST code for Treatments

```
/* Marginal, Type III, Sums of Squares for Treatments */
contrast 'Treatments' t 3 -3 s*t 1 -1 1 -1 1 -1;
```

28.10.6 Store*Treatment Interaction

What about the interaction? The interaction measures whether the differences between t_1 and t_2 are the same at each of the 3 different levels of Store. Graphically we can represent these as:

$\hat{Y}_{11} = \tilde{\mu} + \tilde{s}_1 + \tilde{t}_1 + s_1\tilde{t}_1$	$\hat{Y}_{12} = \tilde{\mu} + \tilde{s}_1 + \tilde{t}_2 + s_1\tilde{t}_2$
$\hat{Y}_{21} = \tilde{\mu} + \tilde{s}_2 + \tilde{t}_1 + s_2\tilde{t}_1$	$\hat{Y}_{22} = \tilde{\mu} + \tilde{s}_2 + \tilde{t}_2 + s_2\tilde{t}_2$
$\hat{Y}_{31} = \tilde{\mu} + \tilde{s}_3 + \tilde{t}_1 + s_3\tilde{t}_1$	$\hat{Y}_{32} = \tilde{\mu} + \tilde{s}_3 + \tilde{t}_2 + s_3\tilde{t}_2$

28.10.7 Degrees of freedom for the A*B Interaction

How many degrees of freedom are there for the interaction? In general, if we have a levels of Factor A and b levels of Factor B, so that we have a*b different combinations (*i.e.* no combinations are missing) then we will have (a-1)*(b-1) degrees of freedom. This corresponds to the (a-1)*(b-1) linearly independent different contrasts, or comparisons, between differences. If we have any missing cells (combinations) then the degrees of freedom will be further reduced by the number of missing combinations. So in the above example we have a 3*2 factorial, which means that we have (3-1)*(2-1) = 2 degrees of freedom for the interaction. If we had no observations for Store 1, Treatment 1, then the degrees of freedom for the Interaction would be reduced by 1, to leave us with 1 d.f. for the Interaction!

28.10.8 Sums of Squares for the A*B Interaction

We have defined that the interaction measures the differences of the differences; *i.e.* we are looking to test whether the main effects are simply additive in action (no interaction), or whether their action is not simply additive in action (an interaction). So we need to look at the differences between treatments at each level of Store Company and see whether those differences are the same or whether they are sufficiently different from one another to conclude that there is an interaction. That is to say, if the differences amongst the differences are small then they would be consistent with the Null Hypothesis (H_0), that the interaction effects are all equal, *i.e.* what we commonly mean when we say that there was no interaction; whereas if the differences amongst the differences are larger than could reasonably be expected to arise by chance when H_0 is true, then we should reject H_0 and accept the Alternative Hypothesis (H_A), that there is an interaction, *i.e.* that the interaction effects are not all equal. Thus $(\hat{Y}_{11} - \hat{Y}_{12})$ represents the difference between (Store₁, Treatment₁ and Store₁-by-Treatment₁) and (Store₁, Treatment₂ and Store₁-by-Treatment₂), *i.e.*

$$\begin{aligned}\hat{Y}_{11} - \hat{Y}_{12} &= (\tilde{\mu} + \tilde{s}_1 + \tilde{t}_1 + s_1\tilde{t}_1) \\ &\quad - (\tilde{\mu} + \tilde{s}_1 + \tilde{t}_2 + s_1\tilde{t}_2) \\ &= (\tilde{t}_1 - \tilde{t}_2 + s_1\tilde{t}_1 - s_1\tilde{t}_2)\end{aligned}$$

If we similarly look at $(\hat{Y}_{21} - \hat{Y}_{22})$ we can see that this represents the difference between (Store₂, Treatment₁ and Store₂-by-Treatment₁) and (Store₂, Treatment₂ and Store₂-by-Treatment₂), *i.e.*

$$\begin{aligned}\hat{Y}_{21} - \hat{Y}_{22} &= (\tilde{\mu} + \tilde{s}_2 + \tilde{t}_1 + s_2\tilde{t}_1) \\ &\quad - (\tilde{\mu} + \tilde{s}_2 + \tilde{t}_2 + s_2\tilde{t}_2) \\ &= (\tilde{t}_1 - \tilde{t}_2 + s_2\tilde{t}_1 - s_2\tilde{t}_2)\end{aligned}$$

Then the difference between these 2 differences is one of the 2 comparisons we need for our Interaction CONTRAST. Thus

$$(\tilde{t}_1 - \tilde{t}_2 + s_1\tilde{t}_1 - s_1\tilde{t}_2)$$

$$\begin{aligned}
& - (\tilde{t}_1 - \tilde{t}_2 + s_2\tilde{t}_1 - s_2\tilde{t}_2) \\
& = (s_1\tilde{t}_1 - s_1\tilde{t}_2 - s_2\tilde{t}_1 + s_2\tilde{t}_2)
\end{aligned}$$

Thus a suitable k' matrix would be

$$k' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \end{pmatrix}$$

Which could be written as a SAS CONTRAST statement as:

```
/* Marginal, Type III, Sums of Squares for 1st Interaction */
contrast 'S1T1-S1T2-S2T1+S2T2' s*t 1 -1 -1 1 0 0;
```

Similarly, if we look at $(\hat{Y}_{31} - \hat{Y}_{32})$ we can see that this represents the difference between (Store₃, Treatment₁ and Store₃-by-Treatment₁) and (Store₃, Treatment₂ and Store₃-by-Treatment₂), *i. e.*

$$\begin{aligned}
\hat{Y}_{31} - \hat{Y}_{32} & = (\tilde{\mu} + \tilde{s}_3 + \tilde{t}_1 + s_3\tilde{t}_1) \\
& - (\tilde{\mu} + \tilde{s}_3 + \tilde{t}_2 + s_3\tilde{t}_2) \\
& = (\tilde{t}_1 - \tilde{t}_2 + s_3\tilde{t}_1 - s_3\tilde{t}_2)
\end{aligned}$$

Then the difference between this difference and our first difference is the second of the 2 comparisons we need for our Interaction CONTRAST. Thus

$$\begin{aligned}
& (\tilde{t}_1 - \tilde{t}_2 + s_1\tilde{t}_1 - s_1\tilde{t}_2) \\
& - (\tilde{t}_1 - \tilde{t}_2 + s_3\tilde{t}_1 - s_3\tilde{t}_2) \\
& = (s_1\tilde{t}_1 - s_1\tilde{t}_2 - s_3\tilde{t}_1 + s_3\tilde{t}_2)
\end{aligned}$$

Thus a suitable k' matrix would be

$$k' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \end{pmatrix}$$

Which could be written as a SAS CONTRAST statement as:

```
/* Marginal, Type III, Sums of Squares for 2nd Interaction */  
contrast 'S1T1-S1T2-S3T1+S3T2' s*t 1 -1 0 0 -1 1;
```

Then we can combine these 2 single degree of freedom contrasts together into a 2 degrees of freedom contrast:

28.10.9 SAS CONTRAST code for Interaction

```
/* Marginal, Type III, Sums of Squares for Interactions */  
contrast 'Interaction' s*t 1 -1 -1 1 0 0,  
s*t 1 -1 0 0 -1 1;
```

28.11 Analysis of Variance, partitioned

<i>Analysis of Variance</i>				
<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F - ratio</i>
Total, TSS	$N = 16$	$Y'Y$ 799.36		
Model, SSR	$r(X)$ $= 6$	$\tilde{b}'X'Y$ 797.995	132.99	974**
Mean, C.F.	1	$N\bar{y}^2$ 795.24	795.24	5825.9**
Model, after the mean, SSR_m $R(t, s, t * s Mean)$	$r(X) - 1$ $= 5$	$\tilde{b}'X'Y - N\bar{y}^2$ 2.755	0.551	4.037
$R(t \mu, s)$	1	1.6858	1.6858	12.35**
$R(s \mu, t)$	2	0.8098	0.4049	2.97 ^{n.s.s.}
$R(t * s \mu, t, s)$	2	0.9119	0.4560	3.347 ^{n.s.s.}
Error, SSE	$N - r(X)$ 10	$Y'Y - \tilde{b}'X'Y$ 1.365	0.1365	

WHAT DOES AN INTERACTION MEAN? STD, Page 352-358

That the differences between Treatments 1 and 2 are not the same at the different company stores. Also that the differences between company stores differ for Treatment 1 or 2.

The following sentences refer to the case where the 2 factors (A and B) are both considered as FIXED effects, i.e. we are interested in these specified levels of A and B and we are not and can not generalize to any other levels of A or B!

If the interaction is statistically significant then the main effects have little meaning or sense.

i.e. examine trt * store means and differences, "Simple Effects".

5 d.f., 5 comparisons, must be linearly independent.

If the interaction is n.s.s. then we can look at the main effects.

28.12 Expectation of Mean Squares

What are the expectations of the Mean Squares? What are/is the appropriate error term and which effects can we test?

These will depend upon whether the various factors are fixed and/or random. This is one of the major reasons why it is important to understand whether a factor is a random effect or not (and whether the sampling was hence random with respect to that factor!!!!).

See STD Ch 15.5 for a discussion of the computation of the expectations of the Mean Squares.

Both factors (A and B) fixed		
Source of Variation	df	Expected Mean Square
A	a-1	$\sigma_e^2 + b \sum \alpha_i^2 / (a - 1)$
B	b-1	$\sigma_e^2 + a \sum \beta_j^2 / (b - 1)$
A*B	(a-1)(b-1)	$\sigma_e^2 + r \sum (\alpha\beta)_{ij}^2 / (a - 1)(b - 1)$
Residual	ab(c-1)	σ_e^2

In the above case (both A and B are fixed effects), then all the terms (A, B, and A*B) are tested against the MSE.

Both factors (A and B) random		
Source of Variation	df	Expected Mean Square
A	a-1	$\sigma_e^2 + c\sigma_{\alpha\beta}^2 + bc\sigma_\alpha^2$
B	b-1	$\sigma_e^2 + c\sigma_{\alpha\beta}^2 + ac\sigma_\beta^2$
A*B	(a-1)(b-1)	$\sigma_e^2 + c\sigma_{\alpha\beta}^2$
Residual	ab(c-1)	σ_e^2

In the above case then MS A*B is tested against the MSE, and both MS A and MS B will be tested against the MS A*B.

Mixed. A fixed, B random => AB = random		
Source of Variation	df	Expected Mean Square
A	a-1	$\sigma_e^2 + c\sigma_{\alpha\beta}^2 + bc \sum \alpha^2 / (a - 1)$
B	b-1	$\sigma_e^2 + c\sigma_{\alpha\beta}^2 + ac\sigma_{\beta}^2$
A*B	(a-1)(b-1)	$\sigma_e^2 + c\sigma_{\alpha\beta}^2$
Residual	ab(c-1)	σ_e^2

In the above case then MS A*B is tested against the MSE, and both MS A and MS B will be tested against the MS A*B. Note further, that any tests relating to Factor A (fixed effects), lsmeans, differences amongst the levels of A, etc, will use the MS A*B as the 'Error' Mean Square.

Note the vast difference this makes in terms of which effects can be tested, and against which mean squares. This is another reason why the use of PROC MIXED in SAS is most strongly recommended. IF the 2 effects are fixed then we first test the interaction. IFF the interaction is not statistically significant is it meaningful to test the main effects. IF at least one of the effects is a random effect, then the interaction will be random, in which case it becomes the 'error' term for testing the main effects of A and B; in this case then whether or not the A*B interaction is significant or not it will still, and always, be appropriate and meaningful to test the A and B main effects.

```
/* Example: Assume A = fixed, B = random */
proc mixed;
class A B;
model y = A/ddfm=kr;
random B A*B;
lsmeans A/pdiff adjust=bon;
run;
```

28.12.1 SAS code for LSMeans

```
/* Assume s = Store, t = Treatment */
lsmeans s t/stderr;
```

```
lsmeans s*t/stderr pdiff adjust=scheffe;  
run;
```

Using the original definition of Least Squares Means given in the context of a Two-Way ANOVA extend this to a Factorial model and try and derive the above Least Squares Means, so that you can take the solution vector (\tilde{b}) and construct a k' vector to compute the LSMeans. Verify that you get the SAME answers as SAS!

29 Latin Square

See Steel, Torrie and Dickey, Ch 9.10, P227

Consider an example with crops. We have 5 treatments to test and we think/know that there is systematic variability across both rows and columns; so we want to try and control for this. We could lay out the 5 treatments in the following manner. Arrange treatments in block in 2 ways by rows and columns

Yield = 4.50 trt ₄	Yield = 2.05 trt ₁	Yield = 4.31 trt ₃	Yield = 3.85 trt ₂	Yield = 4.74 trt ₅
Yield = 4.00 trt ₂	Yield = 4.75 trt ₄	Yield = 2.30 trt ₁	Yield = 4.73 trt ₅	Yield = 4.74 trt ₃
Yield = 2.79 trt ₁	Yield = 4.82 trt ₃	Yield = 4.84 trt ₅	Yield = 5.01 trt ₄	Yield = 4.14 trt ₂
Yield = 4.50 trt ₅	Yield = 3.90 trt ₂	Yield = 4.98 trt ₄	Yield = 4.61 trt ₃	Yield = 2.60 trt ₁
Yield = 4.54 trt ₃	Yield = 4.54 trt ₅	Yield = 4.00 trt ₂	Yield = 2.34 trt ₁	Yield = 4.94 trt ₄

Often used in field experiments with crops.

Another example in animals, consider that we want to test 5 diets to look at their effects on intake. We know that there is substantial variability between animals in intake so we want to use each diet on each animal, so that the differences between animals will not contribute to between diet differences. BUT, there are also time effects (period) which cannot be ignored. So we use the following Latin Square layout:

Period 1	Period 2	Period 3	Period 4	Period 5	
Intake = 4.50 Animal ₄	Intake = 2.05 Animal ₁	Intake = 4.31 Animal ₃	Intake = 3.85 Animal ₂	Intake = 4.74 Animal ₅	Trt 5
Intake = 4.00 Animal ₂	Intake = 4.75 Animal ₄	Intake = 2.30 Animal ₁	Intake = 4.73 Animal ₅	Intake = 4.74 Animal ₃	Trt 4
Intake = 2.69 Animal ₁	Intake = 4.82 Animal ₃	Intake = 4.84 Animal ₅	Intake = 5.01 Animal ₄	Intake = 4.14 Animal ₂	Trt 3
Intake = 4.50 Animal ₅	Intake = 3.90 Animal ₂	Intake = 4.98 Animal ₄	Intake = 4.61 Animal ₃	Intake = 2.70 Animal ₁	Trt 2
Intake = 4.54 Animal ₃	Intake = 4.54 Animal ₅	Intake = 4.00 Animal ₂	Intake = 2.34 Animal ₁	Intake = 4.94 Animal ₄	Trt 1

Latin square does not permit of any interaction amongst rows, columns or treatments!!

If you think that there is an interaction DO NOT USE. Also if variances change!

Disadvantages of Latin Square: Numbers of rows, columns and treatments must be equal.

29.1 Linear Model

Linear model for crop example

$$Y_{ij(k)} = \mu + r_i + c_j + trt_{(k)} + plot(r \ c \ trt)_{ij(k)} + \epsilon_{ij(k)}$$

Linear model for animal example

$$Y_{ij(k)} = \mu + period_i + trt_j + animal_{(k)} + exp(period \ trt \ animal)_{ij(k)} + \epsilon_{ij(k)}$$

For the crop/soil example we must note that the plot which is nested within row, within column and within treatment, *i.e.* $plot(r \ c \ trt)$, cannot be dissociated from the random error ϵ , since there is only one measurement per row, column, treatment combination. Likewise, for the animal example, we have only 1 measurement for each period, treatment, animal combination, thus the $exp(erimental \ unit)$ within period, animal and treatment is confounded with ϵ , and hence again, we shall combine these 2 terms together into an 'error' term (e). However, we should remember that plot is in fact nested within treatment, and nested within row, and nested within column. Note also, that if there was an interaction between row and column (for example) we would note that there is only 1 experimental unit for each row*column combination and that hence the row*column interaction is confounded with the experimental unit (plot). This illustrates why we have noted that a Latin squares does not allow of an interaction between the 3 effects.

Thus, our linear model for the crop example will be:

$$Y_{ij(k)} = \mu + r_i + c_j + trt_{(k)} + e_{ij(k)}$$

and our linear model for the animal example will be:

$$Y_{ij(k)} = \mu + period_i + trt_j + animal_{(k)} + e_{ij(k)}$$

Let us analyse the animal example. We can set up our linear model as $Y = Xb + e$, as

usual, with columns of X for the mean, for the period, treatment and animal effects, thus:

$$\begin{bmatrix} Y_{11(4)} \\ Y_{21(1)} \\ Y_{31(3)} \\ Y_{41(2)} \\ Y_{51(5)} \\ Y_{12(2)} \\ Y_{22(4)} \\ Y_{32(1)} \\ Y_{42(5)} \\ Y_{52(3)} \\ Y_{13(1)} \\ Y_{23(3)} \\ Y_{33(5)} \\ Y_{43(4)} \\ Y_{53(2)} \\ Y_{14(5)} \\ Y_{24(2)} \\ Y_{34(4)} \\ Y_{44(3)} \\ Y_{54(1)} \\ Y_{15(3)} \\ Y_{25(5)} \\ Y_{35(2)} \\ Y_{45(1)} \\ Y_{55(4)} \end{bmatrix} = \begin{bmatrix} \mu & p_1 & p_2 & p_3 & p_4 & p_5 & trt_1 & trt_2 & trt_3 & trt_4 & trt_5 & a_1 & a_2 & a_3 & a_4 & a_5 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 5 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ period_1 \\ period_2 \\ period_3 \\ period_4 \\ period_5 \\ trt_1 \\ trt_2 \\ trt_3 \\ trt_4 \\ trt_5 \\ animal_1 \\ animal_2 \\ animal_3 \\ animal_4 \\ animal_5 \end{bmatrix} + \begin{bmatrix} e_{11(4)} \\ e_{21(1)} \\ e_{31(3)} \\ e_{41(2)} \\ e_{51(5)} \\ e_{12(2)} \\ e_{22(4)} \\ e_{32(1)} \\ e_{42(5)} \\ e_{52(3)} \\ e_{13(1)} \\ e_{23(3)} \\ e_{33(5)} \\ e_{43(4)} \\ e_{53(2)} \\ e_{14(5)} \\ e_{24(2)} \\ e_{34(4)} \\ e_{44(3)} \\ e_{54(1)} \\ e_{15(3)} \\ e_{25(5)} \\ e_{35(2)} \\ e_{45(1)} \\ e_{55(4)} \end{bmatrix}$$

Note that we have 3 dependencies: $period_5$, trt_5 , and $animal_5$; because the 5 columns corresponding to period sum to the first column (μ) and hence there are only 4 linearly independent columns amongst the 5 columns for period; and the 5 columns corresponding to treatment sum to the first column (μ) and hence there are only 4 linearly independent columns amongst the 5 columns for treatment; and the 5 columns corresponding to animal sum to the first column (μ) and hence there are only 4 linearly independent columns amongst the 5 columns for animals.

$$Y = Xb + e$$

$$X'X\tilde{b} = X'Y$$

$$\tilde{b} = (X'X)^- X'Y$$

Note: Not too many d.f. error

solution vector, \tilde{b}

$$\begin{pmatrix} \mu \\ \text{period} \ 1 \\ \ 2 \\ \ 3 \\ \ 4 \\ \ 5 \\ \text{trt} \ 1 \\ \ 2 \\ \ 3 \\ \ 4 \\ \ 5 \\ \text{animal} \ 1 \\ \ 2 \\ \ 3 \\ \ 4 \\ \ 5 \end{pmatrix}$$

If we look at the fitted values for each observation we have:

$$\begin{bmatrix} \mu + a_4 + r_1 + C_1 & \mu + a_1 + r_1 + C_2 & \mu + a_3 + r_1 + C_3 & \mu + a_2 + r_1 + C_4 & \mu + a_5 + r_1 + C_5 \\ \mu + a_2 + r_2 + C_1 & \mu + a_4 + r_2 + C_2 & \mu + a_1 + r_2 + C_3 & \mu + a_5 + r_2 + C_4 & \mu + a_3 + r_2 + C_5 \\ \mu + a_1 + r_3 + C_1 & \mu + a_3 + r_3 + C_2 & \mu + a_5 + r_3 + C_3 & \mu + a_4 + r_3 + C_4 & \mu + a_2 + r_3 + C_5 \\ \mu + a_5 + r_4 + C_1 & \mu + a_2 + r_4 + C_2 & \mu + a_4 + r_4 + C_3 & \mu + a_3 + r_4 + C_4 & \mu + a_1 + r_4 + C_5 \\ \mu + a_3 + r_5 + C_1 & \mu + a_5 + r_5 + C_2 & \mu + a_2 + r_5 + C_3 & \mu + a_1 + r_5 + C_4 & \mu + a_4 + r_5 + C_5 \end{bmatrix}$$

Let us sum the observations for animal 1

$$5\mu + 5a_1 + \sum_{i=1}^{i=5} \text{period}_i + \sum_{j=1}^{j=5} \text{trt}_j$$

Let us sum the observations for animal 2

$$5\mu + 5a_2 + \sum_{i=1}^{i=5} \text{period}_i + \sum_{j=1}^{j=5} \text{trt}_j$$

Divide by 5, to bring back to a per unit basis

$$\left(\mu + a_1 + \frac{1}{5} \sum_{i=1}^{i=5} period_i + \frac{1}{5} \sum_{j=1}^{j=5} trt_j \right) - \left(\mu + a_2 + \frac{1}{5} \sum_{i=1}^{i=5} period_i + \frac{1}{5} \sum_{j=1}^{j=5} trt_j \right)$$

Note that the mean, row and column coefficients cancel out one another, so that we are left with $a_1 - a_2$. $\Rightarrow a_1 - a_2$ is estimable

Partition $RDSS_m$

into	$SS_p, R(p \mu, trt, a)$	4
	$SS_{trt}, R(trt \mu, p, a)$	4
	$SS_a, R(a \mu, p, trt)$	4

Construct the k' matrices for the contrasts for each of the effects (row, column, treatment; or animal, treatment, period) for the Latin square. Construct the appropriate CONTRAST statements for SAS GLM and use them to compute the Sums of Squares for each of the effects; verifying that you obtain the same Sums of Squares as those given in the ANOVA table in these notes and in the ANOVA output from SAS.

Let us do the same thing for treatments as we did for animals, i.e. start from the various fitted values.

$$\hat{Y}_{113} = \mu + \tilde{p}_1 + \tilde{trt}_1 + \tilde{a}_3$$

$$\hat{Y}_{215} = \mu + \tilde{p}_2 + \tilde{trt}_1 + \tilde{a}_5$$

$$\hat{Y}_{312} = \mu + \tilde{p}_3 + \tilde{trt}_1 + \tilde{a}_2$$

$$\hat{Y}_{411} = \mu + \tilde{p}_4 + \tilde{trt}_1 + \tilde{a}_1$$

$$\hat{Y}_{514} = \mu + \tilde{p}_5 + \tilde{trt}_1 + \tilde{a}_4$$

Let us sum the observations for treatment 1

$$5\mu + \sum_{i=1}^{i=5} period_i + 5trt_1 + \sum_{k=1}^{k=5} a_k$$

Dividing by 5 we get

$$\mu + \frac{1}{5} \sum_{i=1}^{i=5} period_i + trt_1 + \frac{1}{5} \sum_{k=1}^{k=5} a_k$$

We should note in passing, that this corresponds to the Least Squares Mean for treatment 1, the corresponding k' could be used to explicitly estimate the lsmean, in IML, or usign the ESTIMATE statement in PROC GLM (and also PROC MIXED).

We can do the same for treatment 2, and then the difference ($trt_1 - trt_2$) will estimate the difference between the 2 treatments, completely free of any effects of the mean, of period and of animal; this is what we mean when we say 'over and above ..'.

29.2 Analysis of Variance

<i>Analysis of Variance</i>				
<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F - ratio</i>
Total, TSS	$N = 25$	$Y'Y$ 441.05		
Model, SSR	$r(X)$ $= 13$	$\tilde{b}'X'Y$ 440.873	33.913	2288.5**
Mean, C.F.	1	$N\bar{y}^2$ 420.414	420.414	28369**
Model, after the mean, SSR_m $R(p, trt, a Mean)$	$r(X) - 1$ $= 12$	$\tilde{b}'X'Y - N\bar{y}^2$ 20.459	1.705	115.04*
p	4	0.170	0.0425	$2.87^{n.s.s.}$
trt	4	0.432	0.1079	7.28*
a	4	19.857	4.964	334.98**
Error, SSE	$N - r(X)$ 12	$Y'Y - \hat{b}'X'Y$ 0.1778	0.0148	

So we can estimate treatment differences as before.

29.3 Fixed or Random ?

See STD Ch. 9.14, Page 239

Because of the particular (balanced) nature of the Latin Square whether the factors are fixed or random is not usually a complicating factor in our tests of significance; either way we compute the various degrees of freedom, Sums of Squares *etc* in the same manner. Note however, that whilst the estimates of the differences between treatments and the standard errors of such estimates are correct from GLM, the standard errors of the Least squares means (so-called sem's) ARE INCORRECT. If one or more of the effects in the model are random one should use PROC MIXED.

29.4 Analysis using SAS/GLM and MIXED

```
data latin1;
input a p t y;
cards;
4 1 5 4.50
1 2 5 2.05
3 3 5 4.31
2 4 5 3.85
5 5 5 4.74
2 1 4 4.00
4 2 4 4.75
1 3 4 2.30
5 4 4 4.73
3 5 4 4.74
1 1 3 2.69
3 2 3 4.82
5 3 3 4.84
4 4 3 5.01
2 5 3 4.14
5 1 2 4.50
2 2 2 3.90
4 3 2 4.98
3 4 2 4.61
1 5 2 2.70
3 1 1 4.54
5 2 1 4.54
2 3 1 4.00
1 4 1 2.34
4 5 1 4.94
;

proc glm data=latin1;
classes a p t;
model y = a p t;
estimate 't1-t2' t 1 -1 0 0 0;
estimate 't1-t3' t 1 0 -1 0 0;
estimate 't1-t4' t 1 0 0 -1 0;
estimate 't1-t5' t 1 0 0 0 -1;
lsmeans t/stderr pdiff adjust=bon;
run;

proc mixed data=latin1;
```

```

classes a p t;
model y = p t;
random a;
estimate 't1-t2' t 1 -1 0 0 0;
estimate 't1-t3' t 1 0 -1 0 0;
estimate 't1-t4' t 1 0 0 -1 0;
estimate 't1-t5' t 1 0 0 0 -1;
lsmeans t/pdiff adjust=scheffe;
run;
quit;

```

29.5 Gains in efficiency

Gains in efficiency relative to a randomized complete block design, STD, Ch 9.13, P237.

First of all compute what the MSE (RCB) would have been:

$$MSE (RCB) = \frac{f_c MSC + (f_t + f_e) MSE}{f_c + f_t + f_e}$$

Then compute the relative efficiency RE (LS to RCB) =

$$\frac{(f_1 + 1)(f_2 + 3)MSE(RCB)}{f_2 + 1)(f_1 + 3)MSE(LS)}$$

where f_1 = d.f. error in L.S. and f_2 = d.f. error in RCB

N.B. in a randomized complete block design we would have more degrees of freedom for the error and therefore more powerful test.

Example of a Latin Square

e.g. from P224 Ch. 9.11 Latin Square, wheat yields, 4 variates

	1	2	3	4
1	C = 10.5	D = 77	B = 12.0	A = 13.2

2	B = 11.1	A = 12.0	C = 10.3	D = 7.5
3	D = 5.8	C = 12.2	A = 11.2	B = 13.7
4	A = 11.6	B = 12.3	D = 5.9	C = 10.2

Reverse, à la animals

		Period			
		1	2	3	4
		Animal			
Treatment	A	4 = 11.6	2 = 12.0	3 = 11.2	1 = 13.2
	B	2 = 11.1	4 = 12.3	1 = 12.0	3 = 13.7
	C	1 = 10.5	3 = 12.2	2 = 10.3	4 = 10.2
	D	3 = 5.8	1 = 7.7	4 = 5.9	2 = 7.5

Rows, Columns and Treatments are independent of one another.

Type I and III Sums of Squares are equal, because it is a "balanced" design.

Unbalanced: imagine that we have a missing cell

Latin Square No. 2, 1 observation missing

Row 4, Column 1, Treatment Animal 1 GONE (Value 11.6)

Now in $(X'X)^-$ rows, Columns & Animals (Treatment) are no longer independent

Type I & III SS are different, the Type III marginal sums of squares do not add to $RDSS_m!$

30 Covariance

STD : Ch 17 P429

Analysis of Covariance is the combination of Regression parameters and Fixed, Classification effects in the same Analysis of Variance model. It is exceedingly simple; simply combine regression and classification factors in the same model.

For example, consider the case that we have 4 diets and we are interested in testing their effects on the growth rate of young heifers. We could randomly split the group of heifers into 4 groups and feed them each one of the diets, measuring their weight at the start and at the end of the trial period. Our analysis would be of the weight gain during the period of treatment. This analysis would be quite valid. However, we might recognise that the heifers were not all of exactly the same weight at the start of the experiment and we might consider that those which were larger/heavier at the start would have a propensity to grow faster, and hence be heavier. If we ignore this factor in our analysis then the random allocation to treatments will ensure that our analysis is unbiased, but it will have a slightly larger sampling variance (and error variance) than if we can/could correct for these differences between the animals. We can correct for the effect of the weight at the start of the trial period by including initial weight as a covariate (regression parameter) in our model.

Then the linear model would be

$$Y_{ij} = \mu + trt_i + b_1 X_{ij} + e_{ij}$$

where Y_{ij} = the weight gain for the j^{th} heifer on the i^{th} treatment
 μ = the effect of the overall mean
 trt_i = the effect of the i^{th} treatment
 b_1 = the regression of weight gain on initial weight
 X_{ij} = the initial weight for the j^{th} heifer on the i^{th} treatment
 e_{ij} = the random residual error associated with the j^{th} heifer
on the i^{th} treatment

Another example would be that from STD, Ch 17.4, Page 435. An experiment was conducted to compare 11 varieties of lima beans for ascorbic acid content, using a Randomised Complete Block design, with 5 blocks, and adding stage of maturity as a covariate.

30.1 Linear model

$$Y_{ij} = \mu + var_i + blk_j + b_1 X_{ij} + e_{ij}$$

where Y_{ij} = the ascorbic acid content for the i^{th} variety of lima bean
in the j^{th} block

μ = the effect of the overall mean

var_i = the effect of the i^{th} variety

blk_j = the effect of the j^{th} block

b_1 = the regression of ascorbic acid content on stage of maturity

X_{ij} = the stage of maturity for the i^{th} variety on the j^{th} block

e_{ij} = the random residual error associated with the i^{th} variety
on the j^{th} block

30.2 Matrix Equations

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{15} \\ Y_{21} \\ \vdots \\ Y_{25} \\ \vdots \\ Y_{111} \\ \vdots \\ Y_{115} \end{bmatrix} = \begin{bmatrix} \mu & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 & v_9 & v_{10} & v_{11} & blk_1 & blk_2 & blk_3 & blk_4 & blk_5 & X \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 34.0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 33.4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 36.1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 39.6 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 20.6 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 30.8 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 43.8 \end{bmatrix} = \begin{bmatrix} \mu \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \\ v_8 \\ v_9 \\ v_{10} \\ v_{11} \\ blk_1 \\ blk_2 \\ blk_3 \\ blk_4 \\ blk_5 \\ b_1 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{15} \\ e_{21} \\ \vdots \\ e_{25} \\ \vdots \\ e_{111} \\ \vdots \\ e_{115} \end{bmatrix}$$

30.3 Analysis of Variance

<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F - ratio</i>
Total, TSS	$N = 55$	$Y'Y$ 496801.1		
Model, SSR	$r(X)$ $= 16$	$\tilde{b}'X'Y$ 494598.23	30912.4	547.278***
Mean, C.F.	1	$N\bar{y}^2$ 434872.15	434872.15	7699.05***
Model, after the mean, SSR_m $R(var, blk, mature Mean)$	$r(X) - 1$ $= 15$	$\tilde{b}'X'Y - N\bar{y}^2$ 59726.077	3981.7	70.49***
var $R(var \mu, blk, mature)$	$(v - 1)$ $= 10$	7454.9	745.5	13.20***
blk $R(blk \mu, var, mature)$	$(b - 1)$ $= 4$	755.2	188.8	3.34*
mature $R(mature \mu, var, blk)$	1 $= 1$	3742.3	3742.3	66.25***
Error, Residual	$N - r(X)$ $55 - 16$	$Y'Y - \hat{b}'X'Y$ 2202.87	56.484	

30.4 Analysis using SAS

USING SAS/IML

```
proc iml;
reset print;
/* Analysis of Covariance, from STD, P435 */
```

```

/* mu varieties          block    covariate */
x = {1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 34.0,
     1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 33.4,
     1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 34.7,
     1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 38.9,
     1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 36.1,
     1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 39.6,
     1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 39.8,
     1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 51.2,
     1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 52.0,
     1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 56.2,
     1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 31.7,
     1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 30.1,
     1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 33.8,
     1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 39.6,
     1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 47.8,
     1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 37.7,
     1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 38.2,
     1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 40.3,
     1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 39.4,
     1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 41.3,
     1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 24.9,
     1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 24.0,
     1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 24.9,
     1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 23.5,
     1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 25.1,
     1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 30.3,
     1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 29.1,
     1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 31.7,
     1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 28.3,
     1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 34.2,
     1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 32.7,
     1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 33.8,
     1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 34.8,
     1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 35.4,
     1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 37.8,
     1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 34.5,
     1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 31.5,
     1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 31.1,
     1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 36.1,
     1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 38.5,
     1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 31.4,
     1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 30.5,

```

```

1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 34.6,
1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 30.9,
1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 36.8,
1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 21.2,
1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 25.3,
1 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 23.5,
1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 24.8,
1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 24.6,
1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 30.8,
1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 26.4,
1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 33.2,
1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 33.5,
1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 43.8};

```

```

y = {93.0,
     94.8,
     91.7,
     80.8,
     80.2,
     47.3,
     51.5,
     33.3,
     27.2,
     20.6,
     81.4,
    109.0,
     71.6,
     57.5,
     30.1,
     66.9,
     74.1,
     64.7,
     69.3,
     63.2,
    119.5,
    128.5,
    125.6,
    129.0,
    126.2,
    106.6,
    111.4,
     99.0,
    126.1,

```

```
95.6,  
106.1,  
107.2,  
97.5,  
86.0,  
88.8,  
61.6,  
83.4,  
93.9,  
69.0,  
46.9,  
80.5,  
106.5,  
76.7,  
91.8,  
68.2,  
149.2,  
151.6,  
170.1,  
155.2,  
146.1,  
78.7,  
116.9,  
71.8,  
70.3,  
40.9}];
```

```
xtx = x' * x;  
xty = x' * y;
```

```
invxtx = ginv(xtx);  
bhat = invxtx * xty;  
tss = y' * y;  
sumy = sum(y);  
nobs = nrow(x);  
cf = sumy * sumy/nobs;  
ssr = bhat' * xty;  
ssrm = ssr - cf;  
rx = 16;  
sse = tss - ssr;  
msr = ssr/rx;  
msrm = ssrm/(rx-1);  
mse = sse/(nobs-rx);
```

```

fm = cf/mse;
fr = msr/mse;
frm = msrm/mse;

/* k' for varieties */
kp = { 0 1 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
        0 1 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
        0 1 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
        0 1 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
        0 1 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0,
        0 1 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0,
        0 1 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0,
        0 1 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0,
        0 1 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0,
        0 1 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0};

k = kp';
kb = k' * bhat;
kgk = k' * invvtx * k;
invkgk = inv(kgk);
ssv = kb' * invkgk * kb;
df = nrow(kp);
msv = ssv/df;
fv = msv/mse;

/* k' for blocks */
kp = { 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 -1 0 0 0 0,
        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 -1 0 0 0,
        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 -1 0 0,
        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 -1 0};

k = kp';
kb = k' * bhat;
kgk = k' * invvtx * k;
invkgk = inv(kgk);
ssb = kb' * invkgk * kb;
df = nrow(kp);
msb = ssb/df;
fb = msb/mse;

/* k' for covariate */
kp = { 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1};
k = kp';
kb = k' * bhat;
kgk = k' * invvtx * k;

```

```

invkgk = inv(kgk);
ssc = kb' * invkgk * kb;
df = nrow(kp);
msc = ssc/df;
fc = msc/mse;

sv = k' * invvtx * k * mse;
se = sqrt(sv);

/* k' for SSRm, Varieties + Block + Covariate */
kp = { 0 1 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
       0 1 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
       0 1 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
       0 1 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
       0 1 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
       0 1 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
       0 1 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0,
       0 1 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0,
       0 1 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0,
       0 1 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0,
       0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 -1 0 0 0 0,
       0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 -1 0 0 0,
       0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 -1 0 0,
       0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 -1 0,
       0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1};

/* k' for a fitted value, variety 1, block 1,
covariate value = 34.0 */
kp = {1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 34.0};
k = kp';
kb = k' * bhat;
sv = k' * invvtx * k * mse;
se = sqrt(sv);

/* k' for an estimated value, variety 1, block 1,
covariate value = average
maturity (ie 33.9873) */
kp = {1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 33.9873};
k = kp';
kb = k' * bhat;
sv = k' * invvtx * k * mse;
se = sqrt(sv);

```

```

/* k' for an estimated value, variety 1, average over blocks,
   covariate value = average maturity (ie 33.9873),
   should correspond to the LSMean for variety 1 */
kp = {1 1 0 0 0 0 0 0 0 0 0 0 .2 .2 .2 .2 .2 33.9873};
k = kp';
kb = k' * bhat;
sv = k' * invvtx * k * mse;
se = sqrt(sv);

quit;

```

USING SAS/PROC GLM

```

data ancova;
input var block mature ascorb;
cards;
1 1 34.0 93.0
1 2 33.4 94.8
1 3 34.7 91.7
1 4 38.9 80.8
1 5 36.1 80.2
2 1 39.6 47.3
2 2 39.8 51.5
2 3 51.2 33.3
2 4 52.0 27.2
2 5 56.2 20.6
3 1 31.7 81.4
3 2 30.1 109.0
3 3 33.8 71.6
3 4 39.6 57.5
3 5 47.8 30.1
4 1 37.7 66.9
4 2 38.2 74.1
4 3 40.3 64.7
4 4 39.4 69.3
4 5 41.3 63.2
5 1 24.9 119.5
5 2 24.0 128.5
5 3 24.9 125.6
5 4 23.5 129.0
5 5 25.1 126.2
6 1 30.3 106.6

```

```

6 2 29.1 111.4
6 3 31.7 99.0
6 4 28.3 126.1
6 5 34.2 95.6
7 1 32.7 106.1
7 2 33.8 107.2
7 3 34.8 97.5
7 4 35.4 86.0
7 5 37.8 88.8
8 1 34.5 61.6
8 2 31.5 83.4
8 3 31.1 93.9
8 4 36.1 69.0
8 5 38.5 46.9
9 1 31.4 80.5
9 2 30.5 106.5
9 3 34.6 76.7
9 4 30.9 91.8
9 5 36.8 68.2
10 1 21.2 149.2
10 2 25.3 151.6
10 3 23.5 170.1
10 4 24.8 155.2
10 5 24.6 146.1
11 1 30.8 78.7
11 2 26.4 116.9
11 3 33.2 71.8
11 4 33.5 70.3
11 5 43.8 40.9
;

```

```

proc means data=ancova;
var ascorb mature;
run;

```

```

proc glm data=ancova;
class var block;
model ascorb = var block mature/XPX I SOLUTION;
estimate ' b mature' mature 1;
/* fitted value for observation 1: var 1, block 1, maturity = 34 */
estimate 'v1, block1, mature=93' intercept 1
  var 1 0 0 0 0 0 0 0 0 0 0 block 1 0 0 0 0 mature 34.0;
/* estimated value for var1, block 1, at average maturity = 33.9873 */

```



```

estimate 'v1, block1, mature=av (33.9873)' intercept 1
  var 1 0 0 0 0 0 0 0 0 0 0 0 block 1 0 0 0 0 mature 33.9873;
/* estimated value for var1, average over blocks,
  at average maturity = 33.9873 */
estimate 'Own calc of lsmean for var 1' intercept 1
  var 1 0 0 0 0 0 0 0 0 0 0 block .2 .2 .2 .2 .2 mature 33.9873;
/* estimated value for var2, average over blocks,
  at average maturity = 33.9873 */
estimate 'Own calc of lsmean for var 2' intercept 1
  var 0 1 0 0 0 0 0 0 0 0 0 block .2 .2 .2 .2 .2 mature 33.9873;
/* estimated value for var3, average over blocks,
  at average maturity = 33.9873 */
estimate 'Own calc of lsmean for var 3' intercept 1
  var 0 0 1 0 0 0 0 0 0 0 0 block .2 .2 .2 .2 .2 mature 33.9873;
/* lsmeans calculated for average maturity */
lsmeans var block/stderr pdiff adjust=scheffe;
/* lsmeans calculated, AT a maturity of 34.0 */
lsmeans var block/stderr pdiff adjust=scheffe at mature=34.0;
/* estimated value for var1, average over blocks,
  at maturity = 34.0 */
estimate 'v1, average over blcoks, mature=34.0' intercept 1
  var 1 0 0 0 0 0 0 0 0 0 0 block .2 .2 .2 .2 .2 mature 34.0;
run;
quit;

```

30.5 Least Squares Means

What will be the Least Squares Means for such a model including a covariate? SAS sets the covariate (X) to the average value.

What would happen if we have a covariate fitted as a linear and as a quadratic effect?

31 Split Plot

STD Ch 16 P400; SAS System for Mixed Models Ch 2

We have 4 blocks and 4 treatments

A_4	A_1	A_2	A_3	Block 1
A_2	A_1	A_4	A_3	Block 2
A_2	A_1	A_4	A_3	Block 3
A_3	A_4	A_2	A_1	Block 4

The basic arrangement, so far, is that of a Randomised Complete Block (RCB) experiment, the four treatments (Factor A, 4 levels) once and only once in each block. We then further sub-divide each A into 4 and apply b_1 , b_2 , b_3 , or b_4 randomly assigned. Thus we now have

b ₁	b ₄	b ₁	b ₃	b ₃	b ₄	b ₄	b ₃	Block 1
A ₄		A ₁		A ₂		A ₃		
b ₃	b ₂	b ₂	b ₄	b ₁	b ₂	b ₂	b ₁	
b ₄	b ₃	b ₄	b ₁	b ₂	b ₃	b ₄	b ₂	Block 2
A ₂		A ₁		A ₄		A ₃		
b ₁	b ₂	b ₂	b ₃	b ₁	b ₄	b ₃	b ₁	
b ₄	b ₁	b ₁	b ₃	b ₂	b ₃	b ₁	b ₄	Block 3
A ₂		A ₁		A ₄		A ₃		
b ₃	b ₂	b ₄	b ₂	b ₄	b ₁	b ₂	b ₃	
b ₃	b ₁	b ₂	b ₃	b ₂	b ₄	b ₄	b ₃	Block 4
A ₃		A ₄		A ₂		A ₁		
b ₂	b ₄	b ₁	b ₄	b ₁	b ₃	b ₂	b ₁	

Uses

1. Use when we have to apply larger amounts (for factor A).
2. When we later add an additional factor (b).
3. If larger differences are to be expected among the a's.
4. If greater precision is required for some factors (b).

31.1 Linear Model

What is the linear model for this analysis? Probably the easiest way to build up the model is, in my opinion, to follow the experimental design. What did we start with? Well, we started with a Randomized Complete Block (RCB) experiment. Thus, for this type of design the model would be:

$$Y_{ij} = \mu + block_i + A_j + e_{ij}$$

What is this error/residual term e_{ij} ? Well, after STD Table 9.8, this error is composed of the interaction between Block and Factor A + the 'real' residual variation from plot to plot (which plots are nested within block, and also within Factor A, and also within the block*A interaction). However, with only 1 measurement per plot we CANNOT separate the interaction from the residual, hence it is all 'lumped' together into the 'Error' term, but we could write the model as:

$$Y_{ij} = \mu + block_i + A_j + block * A_{ij} + plot(block * A)_{ij} + e_{ij}$$

although we would not be able to separate the $block * A_{ij}$ and $plot(block * A)_{ij}$ and e_{ij} effects. However, the addition of the sub-plots and the Factor B effect means that we do end up with more than 1 measurement per plot; thus we can extend the model by adding the B_k effects and the interaction between Factor A and Factor B, thus:

$$Y_{ijk} = \mu + block_i + A_j + block * A_{ij} + B_k + A * B_{jk} + e_{ijk}$$

Note that we still have only 1 plot (experimental unit for A and block) for each block*A combination. Therefore we shall not be able to separate the block*A interaction from the plot variability. Thus what is labelled as block*A is synonymous with plot(block A) and this will be the appropriate error term to use to test the significance of block and Factor A.

Thus it would be most sensible, convenient and least confusing to write

$$Y_{ijk} = \mu + block_i + A_j + plot(block * A)_{ij} + B_k + A * B_{jk} + e_{ijk}$$

However, that is not the way that it is normally written, so one just has to learn to make the connection, and hence work out that the $block * A_{ij}$ term is in fact the correct error term for the main-plot effect (A) and the block effect!

Analysis

	d.f.	M.S.	F ratio
Mean	1		
Blocks	3	MS_{Block}	$MS_{Block}/MS_{A*Block}$
Factor A	3	MS_A	$MS_A/MS_{A*Block}$
Error a, \equiv block * A \equiv plot(block*A)	9	$MS_{A*Block}$	$MS_{A*Block}/MSE$
Factor B	3	MS_B	MS_B/MSE
A * B	9	MS_{A*B}	MS_{A*B}/MSE
Residual, \equiv Error b	36		
Total	64 obs		

31.2 SAS/PROC GLM Model

```
proc glm;
classes blocks A B plot;
model Y = blocks A plot(blocks*A) B A*B;
test h=blocks A e=plots(blocks*A);
random blocks plots(blocks*A);
estimate 'A1 vs A2' A 1 -1 0 0
  block*A .25 -.25 0 0 .25 -.25 0 0 .25 -.25 0 0 .25 -.25 0 0
  A*B .25 .25 .25 .25 -.25 -.25 -.25 -.25 0 0 0 0 0 0 0 0;
lsmeans 'Trt' A/stderr;
run;
```

31.3 SAS/PROC MIXED Model

```
proc mixed;
classes blocks A B plot;
model Y = A B A*B;
random blocks plot(blocks*A);
estimate 'A1 vs A2' A 1 -1 0 0;
```

```
lsmeans 'Trt' A;
run;
```

Residual for Blocks & A	= Error _a M.S. (Block * A)
Residual for B	= MSE
Residual for A*B	= MSE

Ch 16 Table 16.2 P403

	measured as	s.e. of difference
Two A means	$a_j - a_{j'}$	$\sqrt{\frac{2E_a}{rb}}$
Two B means	$b_k - b_{k'}$	$\sqrt{\frac{2E_b}{ra}}$
Two B means at same level of A	$a_j b_k - a_j b_{k'}$	$\sqrt{\frac{2E_b}{r}}$
Two A means at		
1. Same level of B	$a_i b_j - a_k b_j$	$\sqrt{\frac{2[(b-1)E_b + E_a]}{rb}}$
2. Different levels of B	$a_i b_j - a_k b_l$	$\sqrt{\frac{2[(b-1)E_b + E_a]}{rb}}$

Note that the determination of the standard error of the difference is a non-trivial affair, and in almost all cases does not correspond to anything that SAS/GLM can directly and automatically compute. This is not surprising since GLM is a fixed effects model, and we have a model with both fixed and random effects, *i.e.* a mixed model, thus we should use PROC MIXED. What difference does it make? Using PROC GLM we can obtain an ANOVA, which with judicious use of the correct Mean Squares for each F-test can give us valid F-tests. We will also, for the balanced case (no missing observations), obtain Least squares means which are correct; BUT their standard errors will be incorrect!!! This is one

of the major problems to using a fixed effects model to analyse mixed model data, and is one of the reasons why PROC MIXED is to be preferred.

Example from STD, P 406 Table 16.3

Seed lot A	Blocks	Treatment		B	
		Check	Ceresan M	Panogen	Agrox
Vicland 1	1	42.9	53.8	49.5	44.4
	2	41.6	58.5	53.8	41.8
	3	28.9	43.9	40.7	28.3
	4	30.8	46.3	39.4	34.7
Vicland 2	1	53.3	57.6	59.8	64.1
	2	69.6	69.6	65.8	57.4
	3	45.5	42.4	41.4	44.1
	4	35.1	51.9	45.4	51.6
Clinton	1	62.3	63.4	64.5	63.6
	2	58.5	50.4	46.1	56.1
	3	44.6	45.0	62.6	52.7
	4	50.3	46.7	50.3	51.8
Branch	1	75.4	70.3	68.8	71.7
	2	65.6	67.3	65.3	69.4
	3	54.0	57.6	45.6	56.6
	4	52.7	58.5	51.0	47.4

Linear Model

$$Y_{ijk} = \mu + seed_i + block_j + seed * block_{ij} + trt_k + seed * trt_{ik} + e_{ijk}$$

Vicland 1

$$\sum = 16\mu + 16s_1 + \sum_{j=1}^{j=4} 4b_j + \sum_{j=1}^{j=4} 4sb_{1j} + \sum_{k=1}^{k=4} 4t_k + \sum_{k=1}^{k=4} 4st_{1k}$$

average

$$= \mu + s_1 + \frac{1}{16} \sum_{j=1}^{j=4} 4b_j + \frac{1}{16} \sum_{j=1}^{j=4} 4sb_{1j} + \frac{1}{16} \sum_{k=1}^{k=4} 4t_k + \frac{1}{16} \sum_{k=1}^{k=4} 4st_{1k}$$

$$= \mu + s_1 + \frac{1}{4} \sum_{j=1}^{j=4} b_j + \frac{1}{4} \sum_{j=1}^{j=4} sb_{1j} + \frac{1}{4} \sum_{k=1}^{k=4} t_k + \frac{1}{4} \sum_{k=1}^{k=4} st_{1k}$$

Similarly for Vicland 2

$$\sum = 16\mu + 16s_2 + \sum_{j=1}^{j=4} 4b_j + \sum_{j=1}^{j=4} 4sb_{2j} + \sum_{k=1}^{k=4} 4t_k + \sum_{k=1}^{k=4} 4st_{2k}$$

average

$$\begin{aligned} &= \mu + s_2 + \frac{1}{16} \sum_{j=1}^{j=4} 4b_j + \frac{1}{16} \sum_{j=1}^{j=4} 4sb_{2j} + \frac{1}{16} \sum_{k=1}^{k=4} 4t_k + \frac{1}{16} \sum_{k=1}^{k=4} 4st_{2k} \\ &= \mu + s_2 + \frac{1}{4} \sum_{j=1}^{j=4} b_j + \frac{1}{4} \sum_{j=1}^{j=4} sb_{2j} + \frac{1}{4} \sum_{k=1}^{k=4} t_k + \frac{1}{4} \sum_{k=1}^{k=4} st_{2k} \end{aligned}$$

Thus the contrast between Vicland 1 and Vicland 2 is

$$= s_1 - s_2 + \frac{1}{4} \sum_{j=1}^{j=4} sb_{1j} - \frac{1}{4} \sum_{j=1}^{j=4} sb_{2j} + \frac{1}{4} \sum_{k=1}^{k=4} st_{1k} - \frac{1}{4} \sum_{k=1}^{k=4} st_{2k}$$

31.4 Analysis of Variance

<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F - ratio</i>
.	.			
Model, after the mean, SSR_m $R(block, seed, trt Mean)$	$r(X) - 1$ $= 27$	$\tilde{b}'X'Y - N\bar{y}^2$ 7066.192	261.711	12.89*
block $R(block \mu, seed)$	$(b - 1)$ $= 3$	2842.9	947.6	$\frac{947.6}{68.70}$ $= 13.795^{**}$
seed $R(seed \mu, block)$	$(s - 1)$ $= 3$	2848.0	949.3	$\frac{949.3}{68.70}$ $= 13.818^{**}$
block*seed $R(block * seed \mu, block, seed)$ $R(plot(block * seed) \mu, block, seed)$	$(b - 1)(s - 1)$ $= 9$	618.3	68.70	3.38**
trt $R(trt \mu, block, seed, block * seed)$	$(t - 1)$ $= 3$	170.5	56.85	2.80 ^{n.s.s.}
trt*seed $R(trt * seed \mu, block,$ $seed, block * seed, trt)$	$(t - 1)(s - 1)$ $= 9$	586.4	65.16	3.21**
Error, Residual	$N - r(X)$ 64 - 28	$Y'Y - \hat{b}'X'Y$ 731.203	20.311	

31.5 Analysis using SAS

```

USING SAS/PROC GLM
data split;
  input block seed $ trt $ plot yield;
  cards;
1 vic1 check    1 42.9

```

1 vic1 ceresan 1 53.8
1 vic1 panogen 1 49.5
1 vic1 agrox 1 44.4
2 vic1 check 2 41.6
2 vic1 ceresan 2 58.5
2 vic1 panogen 2 53.8
2 vic1 agrox 2 41.8
3 vic1 check 3 28.9
3 vic1 ceresan 3 43.9
3 vic1 panogen 3 40.7
3 vic1 agrox 3 28.3
4 vic1 check 4 30.8
4 vic1 ceresan 4 46.3
4 vic1 panogen 4 39.4
4 vic1 agrox 4 34.7
1 vic2 check 5 53.3
1 vic2 ceresan 5 57.6
1 vic2 panogen 5 59.8
1 vic2 agrox 5 64.1
2 vic2 check 6 69.6
2 vic2 ceresan 6 69.6
2 vic2 panogen 6 65.8
2 vic2 agrox 6 57.4
3 vic2 check 7 45.4
3 vic2 ceresan 7 42.4
3 vic2 panogen 7 41.4
3 vic2 agrox 7 44.1
4 vic2 check 8 35.1
4 vic2 ceresan 8 51.9
4 vic2 panogen 8 45.4
4 vic2 agrox 8 51.6
1 clinton check 9 62.3
1 clinton ceresan 9 63.4
1 clinton panogen 9 64.5
1 clinton agrox 9 63.6
2 clinton check 10 58.5
2 clinton ceresan 10 50.4
2 clinton panogen 10 46.1
2 clinton agrox 10 56.1
3 clinton check 11 44.6
3 clinton ceresan 11 45.0
3 clinton panogen 11 62.6
3 clinton agrox 11 52.7

```

4 clinton check    12 50.3
4 clinton ceresan 12 46.7
4 clinton panogen 12 50.3
4 clinton agrox   12 51.8
1 branch check    13 75.4
1 branch ceresan  13 70.3
1 branch panogen  13 68.8
1 branch agrox    13 71.6
2 branch check    14 65.6
2 branch ceresan  14 67.3
2 branch panogen  14 65.3
2 branch agrox    14 69.4
3 branch check    15 54.0
3 branch ceresan  15 57.6
3 branch panogen  15 45.6
3 branch agrox    15 56.6
4 branch check    16 52.7
4 branch ceresan  16 58.5
4 branch panogen  16 51.0
4 branch agrox    16 47.4

```

```
;
```

```

proc glm;
  classes block seed trt;
  model yield = block seed block*seed trt trt*seed;
  random block*seed;
  test h=block seed e=block*seed;
  lsmeans block seed/pdiff stderr e=block*seed;
  lsmeans trt/pdiff stderr;
  estimate 'vic1 - vic2' seed 1 -1 0 0;
run;
quit;

```

```

proc mixed;
  classes block seed trt;
  model yield = seed trt trt*seed;
  random block block*seed;
  lsmeans seed;
  lsmeans trt;
  lsmeans trt*seed;
  estimate 'vic1 - vic2' seed 1 -1 0 0;
run;
quit;

```

```

proc glm;
  classes block seed trt plot;
  model yield = block seed plot(block*seed) trt trt*seed;
  random plot(block*seed)/test;
  lsmeans block seed/pdiff stderr e=plot(block*seed);
  lsmeans trt/pdiff stderr;
  estimate 'vic1 - vic2' seed 1 -1 0 0;
run;
quit;

proc mixed;
  classes block seed trt plot;
  model yield = seed trt trt*seed;
  random block plot(block*seed);
  lsmeans seed;
  lsmeans trt;
  lsmeans trt*seed;
  estimate 'vic1 - vic2' seed 1 -1 0 0;
run;
quit;

```

31.6 Expectation of Mean Squares

See STD Ch. 16.6, Page 422

If the effects we consider are fixed effects then we will be interested in the differences between the various treatments. However, if the effects that we are considering are classed as random effects then it is the variability in the population that we should be interested in.

Some of the same considerations as were discussed in factorial designs also pertain to Split Plot designs, namely the necessity of deciding which factor(s) are fixed and which are random, and hence which Mean Squares should be tested against which.

If we look at the Least squares means produced by GLM and MIXED we see that they

are the same. Note, this is only true for a completely balanced case, if we had unequal numbers of observations and/or missing values then the GLM Least squares means would not be correct. What about the standard errors? The table below shows the standard errors of the various LSMeans produced by GLM and MIXED from the above-listed SAS statements.

Effect	LSMean	Standard Errors	
		GLM	MIXED
Seeds			
Branch	61.069	2.072	4.246
Clinton	54.306	2.072	4.246
Vic1	42.456	2.072	4.246
Vic2	53.406	2.072	4.246
Trt			
Agrox	52.225	1.127	3.970
Ceresan	55.200	1.127	3.970
etc			
vic1 - vc2	6.763	1.593	2.930

Note the LARGE differences! Using GLM you get underestimates of the standard errors of both the least squares means of the treatments and varieties as well as differences; all things being considered a thoroughly undesirable situation.

32 Cross Over Design

32.1 General comments

In experiments where the animal/person/experimental unit remains on the treatment from the start of the experiment until the end we can call this a continuous trial. Completely Randomised Design (One-Way ANOVA), Two-Way ANOVA (Randomised Complete Block Designs) and Factorial Models are all examples of continuous trials. In a cross-over (also called a change-over trial), however, each animal will receive consecutively two or more experimental treatments during the course of the experiment; this has similarities with the Latin Square design. The period of comparison (C.P.) is therefore divided into a number of sub-periods, which are sometimes referred to as C.P.₁, C.P.₂, etc. We could think of the cross-over design as being a 2-by-2 Latin Square replicated several times contemporaneously.

In a continuous trial, particularly with animals, it is common to place animals on a standard diet/treatment, prior to their random allocation to the experimental treatments. For example, one might have a standardisation period (S.P.) prior to the experiment; this might be the preceding lactation if one was carrying out a whole (complete) lactation study with dairy cattle, or it might be the weight gain in the month preceding the start of the trial in a feeding trial. We take account of, or exploit, the high repeatability of lactation milk yield from one lactation to another, or the relatively high correlation between successive weights on a growth trial; all these with the objective of reducing the experimental error, by covariance adjustment for the measures taken during the standardisation period. Since, in the change-over design, two or more treatments are contrasted on the same experimental unit (e.g. animal, cow) the between-experimental unit (between cow) variation does not enter into the experimental error. Thus, the covariance feature is not needed, and the standardisation period (S.P.) plays a minor role, if any. However, in view of the value of standardising experimental conditions it would seem eminently desirable to routinely employ a short standardisation period, although such data will not (and cannot) be used in the analysis. The basic cross-over design and analysis presented here assumes that there are no carry-over effects, or equivalently, that they are removed by any 'washout' period between the treatment periods, or that the length of time on the treatments is sufficient to remove such residual effects. For a more advanced consideration of cross-over designs (which include this simple two-factor crossover as well as Latin squares) where carry-over

effects may be present see Ratkowsky et al; Cross-over experiments, Design, Analysis and Application.

32.2 Description

The basic cross-over or simple reversal trial can be defined as one in which two treatments (A and B) are studied, and each animal (cow, experimental unit) receives both treatments in either of the sequences A, B or B, A. Thus, the basic pattern of the design is simply:

Basic Pattern

Comparison period	Sequence Group	
	1	2
1	A	B
2	B	A

where the letters in the table represent the treatments. The two periods should be the same length (of time). The experimental units (animals, cow, people) available for the experiment should be allocated to the two sequence groups at random. Usually the same number of animals should be allocated to both groups, since this provides the maximum information per experimental unit, and equivalently the smallest sampling variances. If an odd number of experimental units (animals) are available, however, the number of animals allocated to one sequence can exceed by one the number allocated to the other sequence. There is no need to discard animals (experimental units) just to obtain equal numbers in the two sequence groups. Higher precision will be obtained than by leaving out the odd animal, although it should be recognised that the information (in the statistical sense) per unit is not quite maximum.

The cross-over design exploits the fact that in each time period we have both treatments; hence comparisons between treatments are free of period effects. We effectively remove the period effect from the comparison of treatments. Likewise, each animal receives both

treatments, so the comparison of treatments is within animal, thereby removing between-animal variation from the treatment differences.

32.3 Linear Model

Linear model for dairy cow example

$$Y_{ijhk} = \mu + seq_i + cow_{ij} + per_k + trt_h + e_{ijhk}$$

where Y_{ijhk} = the performance during the k^{th} period of the j^{th} cow in the i^{th} group ($i = 1, 2; j = 1, 2, \dots, n_i$; on the h^{th} treatment, $k = 1, 2$)

μ = the overall mean effect

seq_i = the effect of the i^{th} sequence group ($i = 1, 2$)

cow_{ij} = the effect of the j^{th} cow on the i^{th} sequence

($j = 1, 2, \dots, n_i$), $cow_{ij} \sim N(0, \sigma_{cow}^2)$

per_k = the effect of the k^{th} period ($k = 1, 2$)

trt_h = the effect of the h^{th} treatment ($h = 1, 2$; being a function of i and k)

e_{ijhk} = the random error, $e_{ijhk} \sim N(0, \sigma_e^2)$

32.4 Parameters of the model

Parameters of the model are the mean (μ), the effect of the sequence group (seq_i), the variance amongst animals (experimental units) (σ_{cow}^2), the effect of periods (per_k), the effect of the treatment (trt_h), and the random residual variation (σ_e^2). We are considering that periods are a fixed effect; it is possible to consider periods as a random effect.

32.5 Matrix Equations

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ \vdots \\ Y_{1n_1 2} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ \vdots \\ Y_{2n_2 1} \\ Y_{2n_2 2} \end{bmatrix} = \begin{bmatrix} \mu & seq_1 & seq_2 & a_{11} & \cdot & a_{1n_1} & a_{21} & a_{22} & \cdot & a_{2n_2} & p_1 & p_2 & trt_1 & trt_2 \\ 1 & 1 & 0 & 1 & \cdot & 0 & 0 & 0 & \cdot & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & \cdot & 0 & 0 & 0 & \cdot & 0 & 0 & 1 & 0 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & 0 & 0 & \cdot & 1 & 0 & 0 & \cdot & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & \cdot & 1 & 0 & 0 & \cdot & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & \cdot & 0 & 1 & 0 & \cdot & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & \cdot & 0 & 0 & 1 & \cdot & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & \cdot & 0 & 0 & 1 & \cdot & 0 & 0 & 1 & 0 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & 1 & 0 & \cdot & 0 & 0 & 0 & \cdot & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & \cdot & 0 & 0 & 0 & \cdot & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ seq_1 \\ seq_2 \\ animal_{11} \\ animal_{12} \\ \cdot \\ animal_{1n_1} \\ animal_{21} \\ animal_{22} \\ \cdot \\ animal_{2n_2} \\ period_1 \\ period_2 \\ trt_1 \\ trt_2 \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ e_{121} \\ \cdot \\ e_{1n_1 2} \\ e_{211} \\ e_{212} \\ e_{221} \\ \cdot \\ e_{2n_2} \end{bmatrix}$$

$$Y = Xb + e$$

$$X'X\tilde{b} = X'Y$$

$$\tilde{b} = (X'X)^{-1}X'Y$$

solution vector, \tilde{b}

$$= \begin{pmatrix} \mu \\ seq_1 \\ seq_2 \\ animal_{11} \\ animal_{12} \\ \cdot \\ animal_{1n_1} \\ animal_{21} \\ animal_{22} \\ \cdot \\ animal_{2n_2} \\ period_1 \\ period_2 \\ trt_1 \\ trt_2 \end{pmatrix}$$

32.6 Example data set

Period	Trt	Data				
		Sequence	group 1			
		Cow 1	Cow 2	Cow 3	Cow 4	
1	1	29.9	54.0	41.6	28.5	
2	2	27.8	49.7	38.4	26.5	
		Sequence	group 2			
		Cow 5	Cow 6	Cow 7	Cow 8	Cow 9
1	2	22.2	55.5	43.5	33.2	18.2
2	1	21.4	49.1	41.3	34.3	17.1

32.7 Derivation of CONTRASTS

Treatments: Consider the fitted values

$$\begin{array}{r} \hat{Y}_{111} = \tilde{\mu} + \tilde{s}eq_1 + \tilde{c}ow_{11} + \tilde{p}er_1 + \tilde{t}rt_1 \\ - \\ \hat{Y}_{112} = \tilde{\mu} + \tilde{s}eq_1 + \tilde{c}ow_{11} + \tilde{p}er_2 + \tilde{t}rt_2 \\ \hline \hat{Y}_{111} - \hat{Y}_{112} \quad = (\tilde{p}er_1 - \tilde{p}er_2) + (\tilde{t}rt_1 - \tilde{t}rt_2) \end{array}$$

$$\begin{array}{r} \hat{Y}_{211} = \tilde{\mu} + \tilde{s}eq_2 + \tilde{c}ow_{21} + \tilde{p}er_1 + \tilde{t}rt_2 \\ - \\ \hat{Y}_{212} = \tilde{\mu} + \tilde{s}eq_2 + \tilde{c}ow_{21} + \tilde{p}er_2 + \tilde{t}rt_1 \\ \hline \hat{Y}_{211} - \hat{Y}_{212} \quad = (\tilde{p}er_1 - \tilde{p}er_2) + (\tilde{t}rt_2 - \tilde{t}rt_1) \end{array}$$

$$\begin{array}{r} \text{Then } (\hat{Y}_{111} - \hat{Y}_{112} - (\hat{Y}_{211} - \hat{Y}_{212})) \\ = (\tilde{p}er_1 - \tilde{p}er_2) + (\tilde{t}rt_1 - \tilde{t}rt_2) \\ - (\tilde{p}er_1 - \tilde{p}er_2) + (\tilde{t}rt_2 - \tilde{t}rt_1) \\ \hline = 2(\tilde{t}rt_1 - \tilde{t}rt_2) \end{array}$$

Thus we can see that $\frac{1}{2}[(\hat{Y}_{111} - \hat{Y}_{112} - (\hat{Y}_{211} - \hat{Y}_{212})]$ provides us with a CONTRAST between the two treatments free of BOTH period effects and animal effects.

32.8 Analysis using SAS/MIXED

```
data cross;
input per trt seq cow my;
cards;
1 1 1 1 29.9
2 2 1 1 27.8
1 1 1 2 54.0
2 2 1 2 49.7
1 1 1 3 41.6
2 2 1 3 38.4
1 1 1 4 28.5
2 2 1 4 26.5
1 2 2 5 22.2
2 1 2 5 21.4
1 2 2 6 55.5
2 1 2 6 49.1
1 2 2 7 43.5
2 1 2 7 41.3
1 2 2 8 33.2
2 1 2 8 34.3
1 2 2 9 18.2
2 1 2 9 17.1
;

proc mixed;
classes per trt seq cow;
model my = seq trt per/dfm=kr;
random cow(seq);
lsmeans trt;
estimate 'trt 1-2' trt 1 -1;
run;

proc mixed;
classes per trt seq cow;
model my = seq trt per/dfm=kr;
run;
```

32.9 Parameter Estimates And Significance

Covariance parameters	
Cow(Seq)	171.19
Residual	2.4777

Model Fitting Information	
Observations	18
-2Res. Log Likelihood	94.4
Akaike's Information Criterion	98.4
Schwarz's Bayesian Criterion	98.8

Tests of Fixed Effects				
Source	NDF	DDF	Type III F	Pr > F
Sequence	1	7	0.16	0.7054
Trt	1	7	0.47	0.5165
per	1	7	10.25	0.0150

trt_A - trt_B 0.510 ± 0.746

lsmeans

Trt A 35.57 ± 4.42

Trt B 35.06 ± 4.42

Sequence 1 37.050 ± 6.57

Sequence 2 33.580 ± 5.87

Note, that since sequence 2 has one more experimental unit (cow) than sequence 1 it arrives at having a smaller sampling variance and standard error for the Least squares mean. The standard errors for the two treatments are equal, due to the balance of the design. If these data had been analysed using SAS PROC/GLM we would have obtained essentially the same estimates of the Least squares means for the treatments, but the

standard errors of these Least squares means would have been a factor of 10 times too small!

32.10 Output from PROC MIXED, including animal effect

First Analysis, including animal effect

The Mixed Procedure

Model Information	
Data Set	WORK.CROSS
Dependent Variable	my
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Class Level Information		
Class	Levels	Values
per	2	1 2
trt	2	1 2
seq	2	1 2
cow	9	1 2 3 4 5 6 7 8 9

Dimensions	
Covariance Parameters	2
Columns in X	7
Columns in Z	9
Subjects	1
Max Obs Per Subject	18
Observations Used	18
Observations Not Used	0
Total Observations	18

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	119.30777444	
1	1	94.36117735	0.00000000

Convergence criteria met.

Covariance Parameter Estimates	
Cov Parm	Estimate
cow(seq)	171.19
Residual	2.4777

Fit Statistics	
-2 Res Log Likelihood	94.4
AIC (smaller is better)	98.4
AICC (smaller is better)	99.5
BIC (smaller is better)	98.8

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
seq	1	7	0.16	0.7054
trt	1	7	0.47	0.5165
per	1	7	10.25	0.0150

Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr > t
trt 1-2	0.5100	0.7466	7	0.68	0.5165

Least Squares Means							
Effect	trt	seq	Estimate	Standard Error	DF	t Value	Pr > t
trt	1		35.5700	4.4201	7.1	8.05	<.0001
trt	2		35.0600	4.4201	7.1	7.93	<.0001
seq		1	37.0500	6.5655	7	5.64	0.0008
seq		2	33.5800	5.8724	7	5.72	0.0007

32.11 Output from PROC MIXED, omitting animal effect

Second Analysis, no animal effect

The Mixed Procedure

Model Information	
Data Set	WORK.CROSS
Dependent Variable	my
Covariance Structure	Diagonal
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Residual

Class Level Information		
Class	Levels	Values
per	2	1 2
trt	2	1 2
seq	2	1 2
cow	9	1 2 3 4 5 6 7 8 9

Dimensions	
Covariance Parameters	1
Columns in X	7
Columns in Z	0
Subjects	1
Max Obs Per Subject	18
Observations Used	18
Observations Not Used	0
Total Observations	18

Covariance Parameter Estimates	
Cov Parm	Estimate
Residual	173.66

Fit Statistics	
-2 Res Log Likelihood	119.3
AIC (smaller is better)	121.3
AICC (smaller is better)	121.6
BIC (smaller is better)	121.9

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
seq	1	14	0.31	0.5876
trt	1	14	0.01	0.9361
per	1	14	0.15	0.7079

32.12 Interpretation and comparison of Model 1 and 2

We have to fit the two models, with and without the random effect of animal (cow nested within sequence) to test whether there is a statistically significant effect of the random effect (cow). How do we do this, test the significance? As we know from previous work, we can use the log likelihoods of the 2 models to test this: $-2(\text{Difference})$ has a χ^2 distribution.

Log Likelihood comparison	
(-2Res. Log Likelihood, model 2)	119.3
- (-2Res. Log Likelihood, model 1)	94.4
Difference	24.9

From our statistical tables we can see that, for 1 d.f. (since we have the one parameter σ_{cow}^2) the tabulated value (for a 5% probability level) is 3.84. Since our calculated χ^2 exceeds the tabulated value we can safely reject the Null Hypothesis (that $\sigma_{cow}^2=0$) and accept the Alternative Hypothesis, that σ_{cow}^2 is $\neq 0$; our best estimate of σ_{cow}^2 is 171.19. Thus our use of the cross-over design and inclusion of the random effect of animal in our model has been very effective.

33 Repeated measurements

33.1 Background

- What are repeated measures and how should we analyse them?
- When we have more than 1 measure on each experimental unit;

- hence the importance of clearly understanding what is the experimental unit!

Analysis of repeated measures has some similarities to nested models; there is more than one measurement on each experimental unit.

However, with nested models, we are assuming that the subsamples (conditional upon coming from the given experimental unit) are independent of one another.

For repeated measures, we are usually assuming that this may well not be true, e.g. that we may have several measurements over time on an experimental unit, and that the further apart they are in time, the less correlated they will be.

Easiest to start by examining the design, layout and model where there is no repeated measurement,

i.e. if there is only 1 observation per experimental unit,

what is the experimental unit, what would be your model?

then think of the repeated measurements as being a sub-level within the experimental unit (subject/person/animal/whatever).

Examples

- 10 people, 5 randomly assigned to each of 2 diets (treatments)

- person is the experimental unit

- suppose we recorded their weight at the start and then 8 weeks later at the end of the experimental period.

- we can (should) compute ΔWt (Weight change) as: Final Weight - Initial Weight

Then we have a simple One-way ANOVA, Completely Randomized Design,

$$\Delta Wt_{ij} = Y_{ij} = \mu + Diet_i + e_{ij}$$

- suppose however, that we are interested in **HOW** the weight changes over the 8 weeks; so we decide to weight the people each week.

Now we have 9 weekly weight measurements; repeated measurements on the experimental units (the people/subjects). Our initial design is still the same, the treatment was applied to the experimental unit (person).

- we need to account for the effect of time (the 9 weekly measurements of weight), the possibility that the effect of diet may have different effects over time (i.e. a diet by time interaction), and the fact that the residual errors may be correlated, since they are measurements in the same experimental unit, and hence cannot be presumed to be necessarily independent.

33.2 Linear model

Thus we can extend our model:

$$Y_{ijk} = \mu + diet_i + person_{ij} + time_k + diet_i * time_k + e_{ijk}$$

Note, at this stage we have not specified the distribution of the e_{ijk} 's, the variance-covariance matrix of the random effects errors; we have to; our model is incomplete without!

Plant growth

Consider that we have an experiment with 3 treatments, 12 plots / treatment.

We measure the average height of the crop growth in each plot 2 times per week (Monday and Thursday).

We want to see if growth is different for different treatments.

Experimental unit? The plot, because it is the plot that receives the treatment.

If we start by just considering the model if we had simply recorded the height at the end (or the final height - initial height), *i.e.* only 1 observation per plot, then we would have, as per the previous example, a simple CRD:

$$Y_{ij} = \mu + trt_i + e_{ij}$$

Adding in the fact that we have repeated measurements, and that there is an effect of time, and that there may be a treatment*time interaction, we obtain:

$$Y_{ijk} = \mu + trt_i + plot_{ij} + day_k + trt_i * day_k + e_{ijk}$$

For both models:

- what is the appropriate variance-covariance structure of the e_{ijk} 's?

1) Quite independent of one another?

2) All errors equally correlated, regardless of how far apart they are in time?

3) Errors further apart are likely less correlated?

Note, it is essential for repeated measures analyses using SAS proc mixed that the data be sorted according to the experimental unit within which we have the repeated measures

aand by the repeated measures factor within the experimental unit. What does this mean in practice? Well it is advisable to ALWAYS sort your data just before running proc mixed, and the proc sort procedure of SAS is the easiest and safest way to ensure this:

```
proc sort data=SASdatasetname nodupkey;
  by diet person time;
run;
```

```
proc mixed data = SASdatasetname lognote;
class diet person time;
model y = diet time diet*time / dfm = kr;
random person(diet);
repeated time / type = ? subject = person(diet);
lsmeans diet*time / pdiff adjust = scheffe slice = time;
run;
```

We have the repeated measures on each person, so the person is the experimental unit. Since each person is on only one diet, it therefore follows that person is nested within diet, but cross classified with time. This we are sorting the data by diet (treatment) and person within diet (experimental unit within treatment)and then by time within each person.

What should type=? be? the possibilities are:

- 1) type = vc
- 2) type = cs
- 3) type = ar(1)
- 4) type = sp(pow)(time)

NOTE we are assuming that σ_e^2 is the same across time;

that we have homogeneous variances.

If we do not then some possibilities are using `cs` and `arh(1)` heterogenous variance options (see the SAS on-line documentation).

33.3 Specifying the covariance structure

In terms of modelling the variance-covariance structure it is a good idea to have some ideas about the form of the relationships;

this can be from a theoretical basis and/or by plotting the data.

The Mark 1 ocular estimation device is still unparalleled as an analytical tool!

So, start by graphing things! Graph all the observations; graph them by experimental units, use line graphs, *etc.*

33.4 Common covariance structures

There are 3 main covariance structures that are common for repeated measures:

1. Equal covariance amongst all observations on the same experimental unit; this is called Compound Symmetry (CS).

2. The covariance amongst observations declines the further apart they are in time; proportional to the 'time' distance apart that they are, $= \sigma^2 \rho^w$, where w = number of time units apart that any two observations are; this is called Auto-Regressive(1), AR(1). Note, it is assumed that the time intervals are all equal.

3. If the time measurements are not all equal then we can generalise AR(1) to a spatial power model, SP(POW)(time).

Note that in all of this we are assuming that the variances are homogeneous over time,

if they are not then we **NEED** to account for this. By this we mean that perhaps the residual variance increases with time; for example as animals age and get bigger we often find that the variance of weight increases, therefore if we were looking at bodyweight over time (as the animal grows) we probably do need to allow for the variance to increase, *i.e.* the variances would not be homogeneous, they would be heterogeneous.

Note also, that the above are not all the possibilities, there are many, depending upon the type of models and possible covariance structures that are appropriate to the particular research field (see the SAS PROC MIXED procedure and on-line documentation as well as the SAS System for Mixed Models books, and others for more examples).

Let's consider an example.

I collaborated with a researcher at INRA in France looking at the effect of male canary songs on the production of eggs from females (from the birds who listened to the songs). This was a study about evolutionary biology.

There were 24 females, 12 were exposed to 'superior' male songs, 12 were exposed to 'inferior' male songs. This looks like a CRD, with 2 treatments, and 12 birds per treatment. So far so good.

What was the response, the dependent variable?

We measured the egg weight and testosterone content of the eggs laid by each bird.

Each bird laid 1 egg per day, 5 to 7 eggs per bird.

We have to consider that the eggs were laid over the course of a week, the eggs laid closer together in time might be more similar.

Also, there may be a change in the mean with day, so we have to consider the time(day) factor,

and we have to consider that there might be a different response (to time/days) between the 2 treatments.

Before we have the repeated measurements, we can consider this as a CRD:

$$Y_{ij} = \mu + trt_i + bird_{ij} + \epsilon_{ij}$$

Note, just as for our basic CRD the $bird_{ij}$ and ϵ_{ij} terms are confounded, since (at this stage) there is only 1 measurement per bird; note that the subscripts ij are the same for both $bird$ and ϵ .

If we add the day (of lay) as our time effect we get

$$Y_{ijk} = \mu + trt_i + bird_{ij} + day_k + trt_i * day_k + e_{ijk}$$

The covariance structures that we looked at were the Compound Symmetry (equal covariance between eggs regardless of how many days apart they were laid), and the Autoregressive model (AR(1)), where the correlation between eggs declines the further apart in time the eggs are laid.

We were also interested to see whether there were statistically significant differences amongst birds.

The initial hypothesis was that we expected the AR(1) model to best model the covariance amongst eggs, and we thought that there might be differences amongst birds, *i.e.* $\sigma_{bird}^2 > 0$.

```
proc sort data=SASdatasetname nodupkey;
  by trt bird day;
run;

/* First model */
proc mixed data = SASdatasetname lognote;
class trt bird day;
model y = trt day trt*day / dfm = kr;
random bird(trt);
repeated day / type = cs subject = bird(trt);
run;
```

```

/* Second model */
proc mixed data = SASdatasetname lognote;
class trt bird day;
model y = trt day trt*day / dfm = kr;
random bird(trt);
repeated day / type = ar(1) subject = bird(trt);
run;

/* Third model, note no random effect, to compare
   with vs without the effect of bird nested within trt */
proc mixed data = SASdatasetname lognote;
class trt bird day;
model y = trt day trt*day / dfm = kr;
repeated day / type = ar(1) subject = bird(trt);
run;

```

33.5 Results

Table 25: Covariance structure

Covariance structure	BIC value	
CS	138.7	
AR(1)	109.1	<- Best
sp(pow)	111.8	

Table 26: Random effects

Trait	Correlation	Female χ^2
Egg Wt	0.60	30.0
Testosterone Conc	0.23	6.3

33.6 References

SAS System for Mixed Models. 2nd edition Littell, Milliken, Stroup and Wolfinger.

Littell, R. C., Henry, P. R., and Ammerman, C. B. 1998. Statistical Analysis of Repeated measures Data using SAS Procedures. *Journal of Animal Science*, v76:P1216-1231.

Wang, Z., and Goonewardene, L. A. 2004. The use of MIXED models in the analysis of animal experiments with repeated measures. *Canadian Journal of Animal Science*, v84:P1-11.

SAS on-line documentation for PROC MIXED, specifically about the RANDOM and REPEATED statements

Gil, D., Leboucher, G., Lacroix, A., Cue, R. I., and Kreutzer, M. 2004.

Female canaries produce eggs with greater amounts of testosterone when exposed to preferred male song.

Hormones and Behaviour, 45:64-70.