

Quantitative Methods in Biosciences (M3402-420)

Hans-Peter Piepho

Institute for Crop Science
Bioinformatics Unit

Fruwirthstrasse 23

piepho@uni-hohenheim.de

Hohenheim, June 2012

1: Basic Statistics

1. Simple random sampling in finite populations	3
1.1 Estimating the population mean and variance	3
1.2 Computing a confidence interval for the mean	5
1.3 The population total	9
1.4 SAS code for computing confidence limits for the mean and the total	11
1.5 Sample size	13
Appendix	
*A.1.1 Some theory to explain the variance of a mean formula	14
*A.1.2 Some theory to explain the degrees of freedom (n–1)	16
2. Stratified sampling in finite populations	18
2.1 The model	23
2.2 SAS hints	27
2.3 Optimal allocation	29
2.4 How to find n	31
2.5 Other sampling methods	32
*Appendix	33
3. Regression and correlation	35
3.1 Histogram	37
3.2 Correlation	39
3.2.1 Test of correlation	44
3.2.2 Confidence interval for correlation	??
3.3 Linear regression	45
3.4 Residuals	47
3.5 The normal distribution	51
3.6 Quantile plots	54
3.7 Answering the original research question	56
4. Linear models	62
4.1 Comparing two groups (one-way ANOVA)	64
4.2 Comparing more than two groups (one-way ANOVA)	72
4.3 Linear regression	74
4.4 Simultaneously assessing the effect of one categorical and one quantitative factor	79
4.4.1 Do slopes differ among groups?	80
4.4.2 Test of main effect for a qualitative factor, controlling for a quantitative factor	83
4.4.3 Test of main effect for quantitative factor, controlling for a qualitative factor	84
4.4.4 What has the analysis shown for the baby data (intermediate summary)?	86
4.4.5 Which factor is more important?	87
4.4.6 A further illustration of the models considered in Section 4.4	88
4.5 A general method for comparing nested models	89
4.6 Looking at a sequence of models	90
4.7 Multiple linear regression	96
4.7.1 Multicollinearity	97

Part 2: Biometrics

5. Designed experiments - one treatment factor	105
5.1 Randomization	106
5.2 Blocking	109
5.3 Replication and balance	113
5.4 Pseudo-replication and true replication	114
5.5 Incomplete blocks	115
5.6 Statistical analysis of experiments with one qualitative treatment factor	116
5.6.1 Completely randomized design	116
5.6.2 Randomized complete block design	122
5.6.3 Latin square	137
5.7 Treatment structure - contrasts	139
5.8 Statistical analysis of experiments with one quantitative treatment factor	144
5.8.1 How many x-levels?	164
5.9 Analysis of covariance	165
5.9.1 Models	168
6. Factorial experiments	179
6.1 Interaction	179
6.2 Mean comparisons	180
6.3 Linear model	181
6.3.1 Analysis of variance	183
6.3.2 Mean comparisons	187
6.3.3 Unbalanced data	192
6.3.4 More on comparison of means	196
6.4 Split-plot designs	201
6.4.1 Linear model	203
6.4.2 Analysis of variance	205
6.4.3 Mean comparisons	210
6.5 Factorial experiments with quantitative factors	216
7. Repeated measures	230
7.1 The goats data	230
7.2 The Sorghum data	241
7.3 An addendum to the cheese data	246
7.4 The colon data	250
7.4.1 Groups - the unpaired t-test	251
7.4.2 Regions - the paired t-test	252
7.4.3 Unpaired t-test for groups and one-way ANOVA	253
7.4.4 Paired t-test for regions and ANOVA for block design	255
7.4.5 A t-test for interaction	258
7.4.6 Comparing marginal means for groups by an unpaired t-test (assuming no interaction)	259
7.4.7 Comparing marginal means for regions by a paired t-test (assuming no interaction)	260
7.4.8 A linear model for two-way analysis	261
7.4.9 Two-way ANOVA	263

7.4.10 Mean comparisons (balanced data)	267
7.4.11 The t-test for interaction revisited	270
7.4.12 Analysis for three groups (unbalanced)	272
7.4.13 Analysis of replicate data (unbalanced)	274
7.4.14 Summary	278
8. Sample size and power for elementary procedures	281
8.1. Theoretical background	281
8.1.1 Confidence interval	282
8.1.2 t-test	282
8.2 Single population mean	284
8.2.1 Confidence interval	285
8.2.2 t-test	286
8.3 Two unpaired samples	287
8.3.1 t-test	287
8.3.2 Confidence interval	287
8.4. Two paired samples	287
8.4.1 t-test	289
8.4.2 Confidence interval	289
Appendices	291
A. Some linear model theory	291
B. Some mixed linear model theory	301

Quantitative Methods in Biosciences

Prof. Dr. H. P. Piepho (piepho@uni-hohenheim.de)

Universität Hohenheim

Institute 340

Bioinformatics Unit

Fruwirthstrasse 23

When? Fridays 8-12

Where? PC room 3 (HS 37, Fruwirthstrasse 49, "Kavaliershäuser")

Who? Participants of Master-Programme "AgriTropics" and everyone else who is interested, provided there are enough PCs.

Contents

The objective of this course is to equip you with a working knowledge of a set of important statistical techniques, including their implementation using standard software. I will introduce important statistical methods mainly by presenting real data first and explaining what were the objectives of the study. We will then discuss what methods are appropriate to address the problems posed, and what is the theory behind these methods. Occasionally, simple and artificial examples will be employed to exemplify a point.

For ease of reading the lecture notes, I have marked heavy theory parts with an asterisk (*). Most of the time, these parts have been moved to an Appendix, and they can usually be skipped on first reading without losing the main thread.

We will be using the SAS System as a tool to analyse data. The software is available in all four PC Labs on campus, including a comprehensive online-documentation. Also, small examples can be mastered using a pocket calculator.

Exam

The exam will be written after the course using a computer. All problems are to be solved using the statistical package SAS. Solutions will be written in a word document that is collected at the end of the exam. In addition, program files can be handed in with the word document containing the solutions. You will be allowed to carry these lecture notes, your own notes, and program files you have prepared during the lecture, books, and pocket calculator.

Recommended Literature

For Basics Statistics:

Mead R, Curnow RN, Hasted AM 2002 Statistical methods in agriculture and experimental biology. CRC Press, Boca Raton (ZB 2002/KW05+ZB 2002/KW06).

For Biometrics:

Dean A, Voss DT 1998 Design and analysis of experiments. Springer-Verlag, Berlin.

I strongly recommend that you read and consult these books during the course.

We are assuming that you have been exposed to a first course in statistics. The Mead et al. book can be used to refresh your memory on basic principles. You may also consult the ebook CAST, which is available on ILIAS or at http://cast.massey.ac.nz/collection_public.html. Please note that the module “Quantitative methods in biosciences” is NOT a first course in statistics. So if you have not had such a course, or if despite having attended such a course you find that you do not have the necessary background on basic concepts, it is your responsibility to fill any gaps by your own reading and study.

Note that a module at the University of Hohenheim comprises 56 hours “contact hours”. This is just the contact time. The workload of a module is 150-180 hours. Thus, for every hour of lecture, you are expected to do about two hours of reading and exercising at your own initiative.

A motto for this book

'When the lord created the world and people to live in it - an enterprise which, according to modern science, took a very long time - I could well imagine that He reasoned with Himself as follows: 'If I make everything predictable, these human beings, whom I have endowed with pretty good brains, will undoubtedly learn to predict everything, and they will thereupon have no motive to do anything at all, because they will recognise that the future is totally determined and cannot be influenced by any human action. On the other hand, if I make everything unpredictable, they will gradually discover that there is no rational basis for any decision whatsoever and, as in the first case, they will thereupon have no motive to do anything at all. Neither scheme would make sense. I must therefore create a mixture of the two. Let some things be predictable and let other things be unpredictable. They will then, among other things, have the very important task of finding out which is which.'

(E. F. Schumacher. 1973. Small is beautiful. A study of economics as if people mattered. Blond & Briggs, London, pp. 209-210)

One might say that statistics is one answer of mankind to this tiny little problem so nicely illustrated in Schumacher's book. This course will provide you with some tools to find out for your data what is predictable (pattern, structure) and what is unpredictable (noise).

1. Simple random sampling in finite populations

1.1 Estimating the population mean and variance

I am assuming that all of you have had a first course in statistics, in which you have been exposed to the ideas of random sampling. Most introductory courses consider sampling from infinite populations, and most of the time these infinite populations are hypothetical and do not literally exist. In contrast to this common practice, I believe that many basic statistical ideas can best be understood if looked at in the context of a finite population.

Example 1.1 (artificial): Consider a population of $N = 5$ plants. The leaf contents of a trace element are (ppm)

$$y_1 = 3 \quad y_2 = 5 \quad y_3 = 4 \quad y_4 = 2 \quad y_5 = 6$$

Assume that these values are unknown to the researcher. The researcher wants to know the average content of the trace element in the population of five plants. Assume he/she can afford to analyse just a sample of $n = 3$ plants. So he/she draws a random sample of size $n = 3$. What are the possible outcomes of the study, i.e., what are the possible sample means?

There are ten possible samples of size $n = 3$. The associated sample means

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

are as given in Table 1.1.

Table 1.1: Sample means and variances of all 10 possible samples of size $n = 3$ from population of $N = 5$ values. Observations in sample marked with an "x".

Possible sample no.						Sample		
	y_1	y_2	y_3	y_4	y_5	Sum	Mean	Variance
	3	5	4	2	6	$\sum_{i=1}^n y_i$	\bar{y}	s^2
1	x	x	x			12	4	1
2	x	x		x		10	10/3	7/3
3	x	x			x	14	14/3	7/3
4	x		x	x		9	3	1
5	x		x		x	13	13/3	7/3
6	x			x	x	11	11/3	13/3
7		x	x	x		11	11/3	7/3
8		x	x		x	15	5	1
9		x		x	x	13	13/3	13/3
10			x	x	x	12	4	4
Mean of sample statistics						4	2.5	2.0

The mean of the sample means is 4. This is equal to the mean of all contents in the population:

$$\mu = \frac{\sum_{i=1}^N y_i}{N} = \frac{3+5+4+2+6}{5} = 4$$

which is the **population parameter** the researcher desires to estimate. This shows that when we take a random sample and compute the sample mean, we will, on average, get the right answer: The estimator "sample mean" is an unbiased estimator of the population mean. For this unbiasedness to hold true in practice, we have to make sure that in our sampling procedure, each of the 10 possible samples is equally likely. In other words, each **member** or **element** of the population (**population unit**) has to have the same chance of entering our sample; the probability of selection is the same for each element. This scheme is called **simple random sampling**.

A measure of spread in the population is the variance defined as

$$\sigma^2 = \frac{\sum_{i=1}^N (y_i - \mu)^2}{N}$$

The population variance is unknown in most real applications. In the present example, we are in the exceptional position of being able to compute it:

$$\sigma^2 = \frac{(3-4)^2 + (5-4)^2 + (4-4)^2 + (2-4)^2 + (6-4)^2}{5} = 2$$

In a sample, the (usually unknown) population variance σ^2 is commonly estimated by

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$

(see Appendix A.1.1 for an explanation of the magical divisor $n - 1$). A convenient alternative computational formula for s^2 is

$$s^2 = \frac{1}{n-1} \left[\left(\sum_{i=1}^n y_i^2 \right) - n(\bar{y})^2 \right]$$

i.e., s^2 can be easily computed from the sum of squared observations $\sum_{i=1}^n y_i^2$ and the sample mean \bar{y} .

Example 1.1 (cont'd): Assume a sample of size $n = 3$ from the population of $N = 5$ leaves yielded the following result (line 1 in Table 1):

$y_1 = 3$	$y_2 = 5$	$y_3 = 4$
-----------	-----------	-----------

y_i	y_i^2	
3	9	
4	16	
5	25	
Sum: 12	50	
$\sum_{i=1}^n y_i$	$\sum_{i=1}^n y_i^2$	$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = 12/3 = 4$ $s^2 = \frac{1}{n-1} \left[\left(\sum_{i=1}^n y_i^2 \right) - n(\bar{y})^2 \right] = \frac{1}{2} (50 - 3 * 4^2) = 1$

The sample variances have been computed for all possible samples of size $n = 3$ (Table 1.1). Note that the mean of the sample variances s^2 is not exactly equal to the true variance σ^2 . It can be shown that to remove the bias, we need to multiply s^2 by a factor of $N/(N-1)$, i.e. an unbiased estimate of the population variance σ^2 is given by

$$\frac{N-1}{N} s^2 \quad (1.1)$$

(see Appendix A.1.1). In Table 1.1, the mean of all possible sample variances, multiplied by the factor $(N-1)/N$, does, in fact, equal the true population variance, demonstrating the unbiasedness of this estimator. Note that for $N \rightarrow \infty$, (1.1) tends to s^2 , the usual sample variance commonly used in an infinite population set up.

At this point it is useful to introduce and summarize some terminology used in the context of survey sampling:

- **Population units** make up the population that we want to know more about. In the leaf example, there are five population units, corresponding to five plants.
- **Population size**, usually denoted by N , is the total number of units in the population. For very large populations, often the exact size of the population is not known. In the leaf example, we have $N = 5$ plants, which is a very small population size.
- **Unit characteristic** is a particular piece of information about each member of the population. In the leaf example, the unit characteristic that interests us is the trace element content in ppm.
- **Population parameter** is a summary of the characteristic for all units in the population, such as the average value of the characteristic or the total. In the leaf example, we are interested in the mean trace element content across the five plants.

1.2 Computing a confidence interval for the mean

Example 1.2: A sample of 125 farms in Hertfordshire was taken to assess the number of wheat growers and the wheat growing area in the county. The number of farms in the county is known to be $N = 2496$ (data stored in wheat1.dat). Below is a list of wheat areas (in acres) reported for the 125 farms.

16 0 0 0 0 33 0 92 0 0 0 0 0 0 0 29 107 0 0 65 0 58 67 0 58 45 20 44 0 0 0 0 0
0 0 82 0 0 0 0 11 0 0 59 0 0 0 75 33 0 102 0 0 0 6 0 0 0 62 28 71 0 0 80 265 112 0 50
0 27 12 0 0 24 0 24 28 75 0 0 0 0 0 80 60 0 102 0 0 5 0 0 20 0 0 0 0 0 0 14 0 0 0 72
20 0 0 0 0 0 0 24 0 0 0 6 3 6 0 29 0 0

The objective is to estimate the mean wheat growing area per farm, based on the simple random sample. The estimate is to be assigned a confidence interval, i.e. an interval that is guaranteed to contain the population mean with a prespecified probability of 95%.

To assess the accuracy of the sample mean, we first need to compute its variance. As opposed to samples from an infinite population, for which sample means have variance σ^2/n , the variance in finite samples is

$$\text{var}(\bar{y}) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$$

(see Appendix 1.1). The factor $(N-n)/(N-1)$ is the so-called **finite population correction (fpc)**. This correction may be intuitively understood as follows: When the sample size equals $n = N$, we have assessed the whole population, i.e. the sample mean equals the population mean. Thus, the variance of the sample mean should be zero. Indeed, the fpc is zero in this case, and so is the variance of the mean. To estimate the variance of a mean, we essentially plug in sample variance, s^2 (more specifically, the unbiased estimator $s^2(N-1)/N$), for σ^2 . The square root of the variance of a mean is called the **standard error (s.e.)**:

$$s.e.(\bar{y}) = \sqrt{\text{var}(\bar{y})} = \sqrt{\frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)}$$

It is perhaps surprising that we can estimate the variance (and s.e.) of a sample mean, even though there is only one sample mean. The reason is that there is a mathematical relationship between the variance σ^2 of observations y_i and the variance of a mean \bar{y} , explicit in the above equation, which we can exploit.

When the sample size is large, the probability distribution of the sample mean is well approximated by a normal distribution. This follows from the Central Limit Theorem.

Central Limit Theorem: If Y_1, Y_2, \dots, Y_n are independent and identically distributed with mean μ and variance σ^2 then, for large n ,

$$Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

follows a standard normal distribution.

For finite populations, observations in the sample are not strictly independent, because there is a covariance between pairs of observations. But the covariance is negligible for large n . Thus, for large samples, the sample mean is approximately normal with variance σ^2/n . This striking fact is true independently of the distribution of the values attached to the population units, the so-called **parent distribution**. According to the Central Limit Theorem, the sample mean will

be approximately normal irrespective of the form of the underlying parent distribution. This powerful result can be used to construct a confidence interval for the population mean.

Based on the approximate normality of the sample mean (which follows from the Central Limit Theorem), an approximate $(1-\alpha)100\%$ -confidence interval for the sample mean when n is large ($n > 30$) is given by

$$\bar{y} \pm z_{1-\alpha/2} \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N}\right)}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ -percentile of the standard normal distribution:

α	$z_{1-\alpha/2}$
0.01	2.58
0.05	1.96
0.10	1.64

To be more accurate, specifically when $n < 30$, we can compute

$$\bar{y} \pm t_{1-\alpha/2;n-1} \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N}\right)}$$

where $t_{1-\alpha/2;n-1}$ is the critical value of a t -distribution with significance level α and $n-1$ degrees of freedom (d.f.).

Use of the t -distribution instead of the standard normal accounts for the fact that the variance σ^2 is unknown and needs to be replaced by its estimate s^2 . Some pertinent values for $t_{1-\alpha/2;n-1}$ at $\alpha = 0.05$ are as follows:

DF=(n-1)	t
3	3.18
4	2.78
5	2.57
6	2.45
7	2.36
8	2.31
9	2.26
10	2.23
11	2.20
12	2.18
13	2.16
14	2.14
15	2.13
16	2.12
17	2.11
18	2.10
19	2.09

20	2.09
21	2.08
22	2.07
23	2.07
24	2.06
30	2.04
50	2.00
100	1.98
124	1.98
∞	1.96 (infinite sample size)

We see that for larger n the critical values approach the critical value of the standard normal.

For the mean wheat area (**Example 1.2**) we find for $\alpha = 5\%$:

$$\bar{y} = 18.41$$

$$s^2 = 1330$$

$$n-1 = 124$$

$$t_{0.975;124} = 1.98$$

$$\left[\bar{y} \pm t_{1-\alpha/2;n-1} \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N} \right)} \right] = \left[18.41 \pm 1.98 \sqrt{\frac{1330}{125} \left(1 - \frac{125}{2496} \right)} \right] = [12.1; 24.7]$$

With confidence probability 95%, the interval (12.1; 24.7) contains the population mean of the wheat acreage per farm.

Using the standard normal distribution, we find

$$z_{0.975} = 1.96$$

$$\left[\bar{y} \pm z_{1-\alpha/2} \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N} \right)} \right] = \left[18.41 \pm 1.96 \sqrt{\frac{1330}{125} \left(1 - \frac{125}{2496} \right)} \right] = [12.2; 24.6]$$

This interval is slightly more narrow than the one based on the t-distribution.

It is important to note that a confidence interval is a random variable, just as the sample mean or sample variance. Thus it will vary from sample to sample. The idea of a confidence interval is to construct limits around the sample value, which will cover the population mean with pre-defined probability. In repeated samples, an $(1-\alpha)100\%$ -confidence interval will cover the population mean $(1-\alpha)100\%$ of the time. Of course, in reality we will be computing just one interval. To understand the implications of the interval, however, it is instructive to imagine what will happen under repeated sampling.

Exercise 1.1 (Example 1.2): For the wheat data the fraction (percentage) of farms growing wheat in Herfordshire was estimated by defining a dummy variable y with $y=1$ for farms growing wheat and $y=0$ for farms not growing wheat. The sample mean and variance were

$$\bar{y} = 0.36 \text{ and } s^2 = 0.23226$$

Convince yourself that the mean of $0.36 = 36\%$ is an estimate of the fraction of farms growing wheat. The sample size is $n = 125$, the total number of farms is $N = 2496$. Compute a 95% confidence interval for the fraction of farms growing wheat.

1.3 The population total

The total in the population is defined as

$$\tau = \sum_{i=1}^N y_i = N\mu$$

To derive an estimator of the total, consider a single observation from the population. On average, this observation will represent $1/N$ -th of the total. We say that its expected value equals $(1/N)\tau$.

$$E(y_i) = (1/N)\tau = \mu$$

The sample sum will therefore have the following expected value:

$$E\left(\sum_{i=1}^n y_i\right) = (n/N)\tau$$

To obtain an unbiased estimator of τ , it is necessary to multiply the sample sum by the inverse of (n/N) , yielding

$$T = \sum_{i=1}^n \frac{N}{n} y_i = N\bar{y}$$

The equation for T can be looked at in different ways. One way is to note that the sum in the sample is multiplied by a **raising factor** $g = N/n$ to estimate the population total (Yates, 1981):

$$T = \sum_{i=1}^n gy_i$$

The raising factor g is sometimes referred to as **weight**. The weights will be important when we use SAS to compute estimates and confidence limits for the population mean and total.

An alternative, perhaps more intuitively appealing derivation of the estimator for the population total is as follows: We can expect that the sample mean approximately equals the population mean given by τ/N . Equating the two means and solving for τ yields the estimator T .

$$\frac{\sum_{i=1}^n y_i}{n} \approx \frac{\tau}{N} \Rightarrow \frac{\sum_{i=1}^n y_i}{n} = \frac{T}{N} \Leftrightarrow T = \frac{N \sum_{i=1}^n y_i}{n} = \sum_{i=1}^n gy_i$$

The variance of the total T is related to the variance of a mean by

$$\text{var}(T) = \text{var}(N\bar{y}) = N^2 \text{ var}(\bar{y})$$

Thus, the standard error is

$$s.e.(T) = N \times s.e.(\bar{y})$$

To compute a confidence interval for the population total, the standard error needs to be replaced by its estimate in much the same way as for the population mean.

A $(1-\alpha)100\%$ -confidence interval for the population total is given by

$$T \pm t_{1-\alpha/2; n-1} N \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N}\right)}$$

where

$$T = \sum_{i=1}^n \frac{N}{n} y_i = N\bar{y}$$

For the total wheat area (**Example 1.2**) we find:

$$T = \sum_{i=1}^n \frac{N}{n} y_i = N\bar{y} = 18.41 * 2496 = 45946$$

$$\left[T \pm t_{1-\alpha/2; n-1} N \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N}\right)} \right] = \left[45946 \pm 1.98 * 2496 \sqrt{\frac{1330}{125} \left(1 - \frac{125}{2496}\right)} \right] = [30234; 61658]$$

The confidence interval for the total acreage is pretty wide, so the sample is too small to make precise statement regarding the population total.

Remark: The ratio n/N is also referred to as the **sampling fraction**, because $n/N100\%$ of the population appear in the sample. Under simple random sampling as used in this section, n/N is the **selection probability**, i.e. the probability for a member or element of the population to end up in a sample of size n . The probability of being selected is the same for each member of the population. As noted above, the inverse $g = N/n$ is the weight factor needed to estimate the population total. The ratio n/N is also called the **sampling rate**, i.e. the rate of the population that is sampled.

It should be stressed that there are other sampling plans in which the probability of selection is not constant for each element of the population. For example the probability may be proportional to some covariate. For a valid statistical analysis it is important that the probabilities be known. It is not important that the selection probabilities be equal.

Knowledge of the selection probabilities is the prerequisite for obtaining valid population estimates. The probabilities enter the estimates as **weights**. Under simple random sampling, the selection probability is the same for each population unit, and so the weights for all observations in the sample are the same, and the resulting estimators are particularly simple.

Exercise 1.2 (Example 1.2): For the wheat data the fraction (percentage) farms growing wheat in Herfordshire was estimated by defining a dummy variable y with $y=1$ for farms growing wheat and $y=0$ for farms not growing wheat. The sample mean and variance were

$$\bar{y} = 0.36 \text{ and } s^2 = 0.23226$$

From this, the number of farms in the county may be estimated. The sample size is $n = 125$, the total number of farms is $N = 2496$. Estimate the total number of wheat-growing farms and compute a 95% confidence interval.

1.4 SAS code for computing confidence limits for the mean and the total

To do the computations for Example 1.2 in SAS, you may use the following code:

```
data;
input area;
g=2496/125;
if area=0 then y=0; else y=1;
datalines;
16
0
0
<more data>
29
0
0
;
proc surveymeans sum mean total=2496 clm clsum;
var area y;
weight g;
run;
```

The IF/ELSE clause in the datastep defines a 0-1 valued variable y used to estimate the fraction of farms growing wheat. The WEIGHT statement submits the weight factor $g = N/n$, which is needed in computing the sample estimate of the total. The TOTAL option specifies the value of N . The options CLM and CLSUM invoke computation of a confidence interval for the mean and the total, respectively.

Exercise 1.3 (Example 1.2): Use SAS to compute confidence limits for the population mean and the population total of the wheat growing area.

Example 1.3: Consider this class as a finite population and let us determine the average age of all members in this population. We can then interview a random sample of students and ask them for their age. The average in the sample is expected to come close to the population mean. Repeated sampling will yield a set of means. To see what the properties are of sample means and how they vary, we will do the following class-room experiment:

- (1) We compile a list of participants, including their age. The mean age is our population mean. The list of all elements in a population is known as the **sampling frame**.
- (2) Each of you independently generates a random sample of $n = 5$ participants and asks them their age.

(3) Everyone of you computes the sample mean and a 95% confidence interval (note: $t_{0.975,4} = 2.36$), either by hand or using a computer.

An interesting question is for how many of the samples, the confidence interval contains the true mean. Theory says that on average 95% of the intervals should cover the population mean, while 5% do not cover the population mean. Thus the error probability equals the pre-specified value of $\alpha\%$. Our class-room experiment will produce just a limited number of samples, so the fraction of intervals covering the true mean may depart noticeably from the expected 95%. At any rate, the experiment will show that a confidence interval is a random statistic, just as the sample mean or sample variance, and that the interval is not guaranteed to cover the population value.

Here is how to draw a random sample of size n from a list of names using random numbers:

- (1) For each name ("element"/"unit" in the population) generate a uniform random number (each value between 0 and 1 is equally likely).
- (2) Sort the list by the random number
- (3) Select the first n elements as your sample

The following SAS code will do the job (using a slightly more sophisticated algorithm than the one described above), without you actually seeing the random numbers.

```
data namelist;
input name$13. firstnam$13. ;
datalines;
Makengele Michael
Aalandia Erika
Musavaya Katinka
Wamatu Jane
Aloo Frederick
SchmidtMoreno Lucia
Hongjie Zhou
Mergenthaler Marcus
Kirfu Gebreyel
Garoma Lemma
Cordero MariaVida
Hoang TheHungTha
Tinoco Roberto
Peylo Birgit
Legesse Alemu
Melkamu Jate
Lorrata Iroh
Dimasai Ossama
Pflanz Wilhelm
Ninako Dominic
Tesfamariam Tsehaye
;

proc print data=namelist;

proc surveyselect out=sample data=namelist method=srs sampsize=5;

proc print data=sample;
run;
```

Note: The "\$13." Following the variable labels NAME and FIRSTNAME tell SAS to read up to 13 letters for the name. Without this addition, names will be truncated after 8 letters.

Alternatively, you can generate a uniform random number in a datastep, sort the data using PROC SORT, and print the first 5 observations in the sorted dataset (this option allows you to see what SAS actually does in selecting the sample).

```

data namelist;
set namelist;
r=ranuni(1);

proc sort data=namelist out=namelist;
by r;

data sample;
set namelist;
if _N_<=5;

proc print data=sample;
run;

```

Exercise 1.4 (Example 1.3): Draw a sample of size $n = 5$ from the list of participants (see preliminary list in participants.dat). Compute the confidence limit for the mean age in your sample by hand. For comparison or alternatively, use SAS to do the job. We will collect all confidence intervals computed by participants in the course to study the coverage of the true mean age in the population.

1.5 Sample size

An important question regards the appropriate sample size for a survey or experiment. To answer this question, we need to specify how accurate an estimate we desire of a corresponding population parameter. Perhaps the easiest way to do this is to specify the width of a confidence interval one is willing to accept.

Half the width (HW) of a confidence interval for the mean is

$$HW = t_{1-\alpha/2;n-1} \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N}\right)}$$

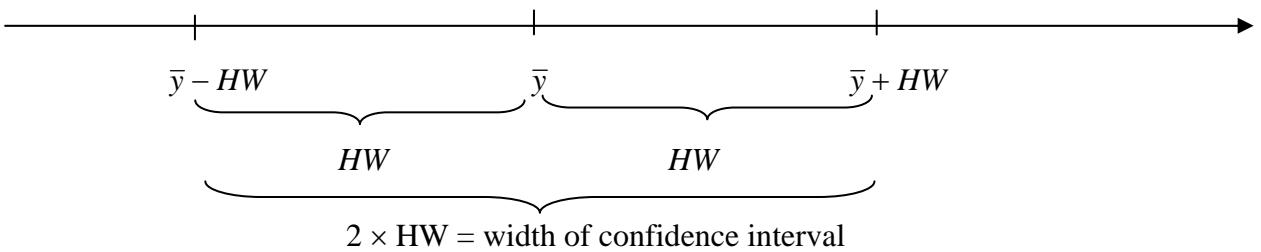


Fig. 1.1: Graphical display of confidence interval and its half width (HW)

To use this formula, it is helpful to make a few simplifying assumptions:

- (1) The sample size n will be small relative to the population size N .

- (2) The sample size will be so large that the sample variance s^2 is practically equal to the population variance σ^2 .
- (3) The sample size n is so large, that the critical t-value is practically identical to that of the standard normal distribution.

Under these assumptions, the half width of the confidence interval equals for $\alpha = 5\%$:

$$HW = 1.96 \sqrt{\frac{\sigma^2}{n}}$$

This equation can be rearranged to yield

$$n = \frac{1.96^2 \sigma^2}{(HW)^2}$$

Obviously, to use this equation for planning sample size, we need to know the population variance σ^2 . If this is not known or cannot be reasonably estimated *a priori*, there is no basis for planning the sample size.

Example 1.4: A simple random sample of farms is to be selected in a region to assess the average farm size in hectares (ha). From similar surveys in the past the variance is predicted to equal approximately 4.5 ha^2 . Average farm size is expected to be in the area of 10 ha to 20 ha. The researcher wants to assess average farm size and make sure that the half width of a 95% confidence interval equals 0.5 ha.

$$n = \frac{1.96^2 \times 4.5}{0.5^2} = 69.15 \approx 69$$

Thus, to achieve the desired accuracy, we need a sample of 69 farms.

Exercise 1.5 (Example 1.4): How does the required sample size change, if we want to achieve a half width of 0.1 ha? What happens if the variance equals 2 ha^2 ?

Appendix

*A.1.1 Some theory to explain the variance of a mean formula

Simple random sampling imposes a probability distribution for our sample, which we will now study. For the following theory it is helpful to consider the order in which elements are sampled. Specifically, the simple random sampling scheme ensures that the probability for each element to be the first element selected equals $1/N$. To state this more formally, we introduce the random index $I(1)$. This index may take values between 1 and N , depending on which element of the population is the first to enter the sample. For example, when observation number 4 enters first, we have $I(1) = 4$. We may state the following:

$$P[I(1)=4] = \frac{1}{N} \quad (\text{A1})$$

This says that the probability for the forth element in the population to enter the sample as first observation is $1/N$. This probability is the same for each element, and the probabilities sum to unity across all elements as needs to be required, since with probability one we will select one of the N elements. Thus, we have

$$P[I(1) = i] = \frac{1}{N} \quad \text{for } i = 1, \dots, N$$

An analogous statement can be made with respect to the subsequent observations in the sample, $I(2), \dots, I(n)$, i.e.

$$P[I(j) = i] = \frac{1}{N} \quad \text{for } j = 1, \dots, n \text{ and } i = 1, \dots, N.$$

Now let us look at the expected value of the first observation $y_{I(1)}$. Note that $y_{I(1)}$ is a random variable since the index $I(1)$ is random with probability function given by (A1). The expected value of the first observation is defined as

$$\begin{aligned} E(y_{I(1)}) &= \sum_{i=1}^N y_i P[I(1) = i] \\ &= \sum_{i=1}^N y_i \frac{1}{N} \\ &= \mu \end{aligned}$$

In loose language we may say that averaged over many samples, the mean of the first observation entering the sample will equal the population mean, because each observation has the same chance of entering the sample first. The same statement can be made regarding $y_{I(2)}, \dots, y_{I(n)}$, i.e.

$$E(y_{I(j)}) = \mu \quad \text{for } j = 1, \dots, n$$

Now we are ready to look at the sample mean and its expectation. We have

$$\begin{aligned} E(\bar{y}) &= E\left(\frac{y_{I(1)} + y_{I(2)} + \dots + y_{I(n)}}{n}\right) \\ &= \frac{1}{n}[E(y_{I(1)}) + E(y_{I(2)}) + \dots + E(y_{I(n)})] \\ &= \frac{1}{n}[nE(y_{I(1)})] \\ &= \mu \end{aligned}$$

This shows that the sample mean is, in fact, an unbiased estimator of the population mean. We had seen a numerical example of this in Example 1.1.

Next, we look at the variance of $y_{I(1)}$, i.e. the variance of observations in the sample. The variance is defined as the expected value of the squared deviation from the population mean:

$$\begin{aligned}
\text{var}(y_{I(1)}) &= E[(y_{I(1)} - \mu)^2] \\
&= \sum_{i=1}^N P(I(1) = i)(y_i - \mu)^2 \\
&= \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2 \\
&= \sigma^2
\end{aligned}$$

So we have shown that the variance of observations in the sample equals the population variance σ^2 .

Now let us turn to the variance of a mean. This is given by

$$\begin{aligned}
\text{var}(\bar{y}) &= \frac{1}{n^2} \text{var}\left(\sum_{j=1}^n y_{I(j)}\right) \\
&= \frac{1}{n^2} \sum_{j=1}^n \text{var}(y_{I(j)}) + \frac{1}{n^2} \sum_{j \neq k} \text{cov}(y_{I(j)}, y_{I(k)}) \\
&= \frac{1}{n} \sigma^2 + \frac{n-1}{n} \text{cov}(y_{I(1)}, y_{I(2)})
\end{aligned}$$

where $\text{cov}(y_{I(j)}, y_{I(k)}) = E[(y_{I(j)} - \mu)(y_{I(k)} - \mu)]$ is the covariance, a measure of how strongly $y_{I(j)}$ and $y_{I(k)}$ "co-vary". The last equality follows from the fact that all pairs $(y_{I(j)}, y_{I(k)})$ are identically distributed and that there are $n(n-1)$ pairings of observations in the sample. The covariance is not equal to zero here due to the fact that we are sampling from a finite population. To see this, consider a population with $N = 2$ and $y_1 = 2$ and $y_2 = 4$, and assume that the sample size $n = 2$, i.e. we sample the whole population. The covariance must be negative for if the first observation in the sample is the smaller one, the second will be the larger one, and vice versa.

A smart way to find the covariance for any N is to note that if we sample the whole population ($n = N$), the variance of the sample mean is zero:

$$\begin{aligned}
\text{var}(\bar{y} | n = N) &= \frac{1}{N} \sigma^2 + \frac{N-1}{N} \text{cov}(y_{I(1)}, y_{I(2)}) = 0 \\
\Leftrightarrow \text{cov}(y_{I(1)}, y_{I(2)}) &= -\frac{\sigma^2}{N-1}
\end{aligned}$$

Using this result, we find

$$\text{var}(\bar{y}) = \frac{1}{n} \sigma^2 - \frac{n-1}{n} \frac{\sigma^2}{N-1} = \frac{\sigma^2}{n} \frac{N-n}{N-1}$$

*A.1.2 Some theory to explain the degrees of freedom (n-1)

For those of you interested in theory, this section will explain the "magical" fact that the divisor in the sample variance s^2 needs to be $n - 1$ rather than n to achieve unbiasedness in

large (infinite) populations. We will use the fact that the variance of a random variable z is defined as follows:

$$\text{var}(z) = E[(z - E(z))^2] = E[z^2 - 2zE(z) + (E(z))^2] = E(z^2) - [E(z)]^2$$

Thus,

$$E(z^2) = \text{var}(z) + [E(z)]^2$$

Using this result, the expected value s^2 is

$$\begin{aligned} E(s^2) &= E\left(\frac{\sum_{j=1}^n (y_{I(j)} - \bar{y})^2}{n-1}\right) \\ &= E\left(\frac{\sum_{j=1}^n (y_{I(j)})^2 - n\bar{y}^2}{n-1}\right) \\ &= \frac{1}{n-1} E\left(\sum_{j=1}^n (y_{I(j)})^2\right) - \frac{n}{n-1} E(\bar{y}^2) \\ &= \frac{n}{n-1} (\sigma^2 + \mu^2) - \frac{n}{n-1} (\text{var}(\bar{y}) + \mu^2) \\ &= \frac{n}{n-1} \sigma^2 - \frac{n}{n-1} \frac{\sigma^2}{n} \frac{N-n}{N-1} \\ &= \frac{n}{n-1} \sigma^2 \left(1 - \frac{N-n}{n(N-1)}\right) \\ &= \frac{n}{n-1} \sigma^2 \left(\frac{n(N-1)-N+n}{n(N-1)}\right) \\ &= \frac{n}{n-1} \sigma^2 \left(\frac{nN-n-N+n}{n(N-1)}\right) \\ &= \frac{n}{n-1} \sigma^2 \left(\frac{N(n-1)}{n(N-1)}\right) \\ &= \frac{N}{N-1} \sigma^2 \end{aligned}$$

Use $E(z^2) = \text{var}(z) + [E(z)]^2$ on $y_{I(j)}$

μ^2 cancels out

This shows, why we have to multiply s^2 by $(N-1)/N$ to obtain an unbiased estimate of σ^2 in finite populations. It also shows that in infinite populations, where $(N-1)/N$ equals unity, s^2 is, in fact, an unbiased estimator. Finally, it shows that division of the sum of squares by n rather than $n-1$ would introduce a bias in the estimation by σ^2 proportional to $(n-1)/n$.

2. Stratified sampling in finite populations

In some cases, the population to be sampled can be stratified into more or less homogeneous **strata** (groups). If so, **stratification** should be incorporated into the sampling plan, because this has the potential of increasing accuracy, as will be explained using an artificial example.

Example 2.1 (artificial): A population of rice fields in a defined area may be stratified into irrigated and rainfed. Possibly the yields on irrigated fields tend to be higher than in rainfed fields. In this case it is preferable to take a sample of fixed size from each of the two strata rather than a simple random sample from the whole population (for the latter, the sample size per stratum would be random). Assume for simplicity that yield on all irrigated fields is 3 t/ha, while on all rainfed fields the yield is 1 t/ha. Also, assume that all fields have equal size and are square in shape. Finally, the total irrigated and rainfed areas are equal in size. Thus, the population mean is

$$\mu = (1 + 3)/2 = 2$$

Of course, this is a highly idealistic example, but it demonstrates the advantages of stratification. In the figure below, black squares are irrigated, while white squares are rainfed. The mean rice yield is to be assessed from a random sample. Assume that the researcher does not know that yields are the same in each **stratum**, but he correctly guesses that yields are more homogeneous within strata than between strata.

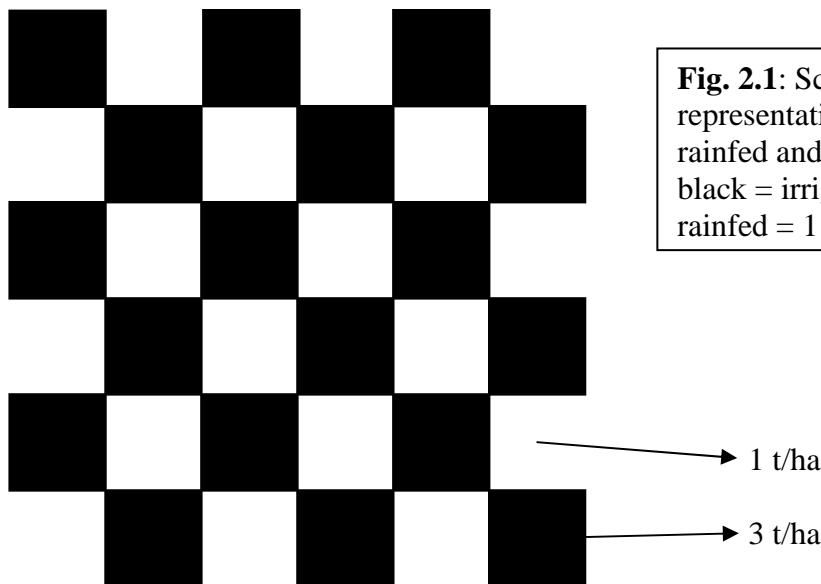


Fig. 2.1: Schematic representation of a population of rainfed and irrigated rice fields.
black = irrigated = 3 t/ha, white = rainfed = 1 t/ha.

If a simple random sample of size 2 is taken, there are four possible samples with regard to the resulting sample mean (Table 2.1).

Table 2.1: Four possible outcomes of a simple random sample of size $n = 2$ for the rice example.

First field Second field Sample mean

irrigated	irrigated	(3+3)/2=3
irrigated	rainfed	(3+1)/2=2
rainfed	irrigated	(1+3)/2=2
rainfed	rainfed	(1+1)/2=1

50% of the samples will have a mean equal to the population mean ($\mu = 2$), while for the other 50% sample mean and population mean will differ.

If instead of taking a simple random sample, the researcher decides to randomly select one irrigated and one rainfed field (this is a stratified sample), there is only one possible outcome: The mean will equal $(3+1)/2 = 2$ with probability one! Thus, the sample mean has a variance of zero, and we will be guaranteed to obtain the population mean!

In this simplistic example, the within-stratum variance equals zero, so the mean of a stratified sample has zero variance, too. By contrast, the mean from the simple random sample has considerable variance. In most practical situations, within-stratum variance will not equal zero. The example should make it clear, however, that the smaller the within stratum variance the more accurate will be the sample estimate. Thus, if you conduct a survey, it is worthwhile to check if there are variables according to which the population can be stratified so that within-stratum variability is smaller than between-stratum variability.

Often, stratification is done according to administrative boundaries. This type of stratification may not be optimal in terms of accuracy, but it is often the most practical way to stratify a sample in large surveys. Often, administrative units vary according to important covariates, so some gain in accuracy can usually be expected.

Example 2.1: Minnesota Radon levels (Nolan and Speed, 2000). Radon is a radioactive gas with a very short half life, yet it is considered a serious risk to the general public. It has long been known to cause lung cancer. In 1987, the US Environmental Protection Agency (EPA) conducted a survey in Minnesota to assess the fraction of households for which the radon levels exceeded a critical action limit of 4 pCi/l. For a sample of 1003 households, radon concentrations were measured for two days. The table below gives an excerpt of the data collected (counties one and two).

County	1	1	1	1	2	2	2	2	2	
Radon (pCi/l)	1.0	2.2	2.2	2.9	2.4	0.5	4.2	1.8	2.5	5.4

Survey design: The houses included on the list to be surveyed (sampling frame) were all those with permanent foundations, at least one floor at or below ground level, owner occupied, and with a listed phone number. The real population of interest is all occupied residencies. These restrictions resulted from the difficulties in gaining permission to conduct the survey in rental units and finding houses without listed phone numbers, and from the fact that houses entirely above ground tend to have very low radon concentrations.

Houses were selected county by county for the sample. Within a county, each house had an equal chance of being included in the survey. The county population and a radon index were used to determine how many houses to choose from each county. Table 2.1 contains, for each county, the total number of households in the county and the number of houses sampled from the county.

To select the houses to be surveyed, telephone numbers were randomly chosen from a directory of listed telephone numbers. For each county, the list of randomly selected phone numbers was 5 times the desired number of households to be contacted in the county. The phone numbers were arranged in lists of 50, and the contacts for each county were made in waves of 50, until the desired number of participants was obtained.

Table 2.1: Structure of stratified sample for Minnesota Radon levels. Population sizes are given in hundreds (multiply by 100 to get the appropriate figure; for county 1, e.g., $N = 5400$).

Size of			Size of		
County	Sample	Population	County	Sample	Population
1	4	54	45	8	97
2	57	719	46	13	111
3	4	110	47	5	77
4	7	115	48	3	72
5	4	95	49	10	103
6	3	29	50	14	149
7	14	186	51	1	39
8	4	102	52	4	95
9	11	105	53	3	78
10	6	141	54	3	33
11	5	84	55	26	361
12	5	56	56	11	199
13	6	100	57	4	57
14	15	172	58	6	75
15	4	31	59	4	41
16	2	18	60	4	125
17	4	52	61	2	46
18	12	172	62	42	1809
19	69	794	63	0	18
20	3	55	64	5	67
21	11	114	65	3	74
22	6	73	66	11	159
23	2	79	67	3	38
24	10	134	68	14	47
25	15	146	69	122	81
26	0	27	70	14	165
27	119	3925	71	9	111
28	6	64	72	4	55
29	5	57	73	27	360
30	4	90	74	10	110
31	12	166	75	2	37
32	7	48	76	4	47
33	4	49	77	4	94
34	4	141	78	5	17
35	3	23	79	7	74
36	9	62	80	5	50
37	2	38	81	4	69
38	10	44	82	50	424
39	5	15	83	3	47
40	6	87	84	1	28
41	4	29	85	13	161
42	10	90	86	14	216
43	1	16	87	3	46
44	9	45			

The phone caller determined whether the candidate was eligible and willing to participate. If this was the case, the candidate was mailed a packet of materials containing a charcoal canister for measuring radon levels, an instruction sheet, a questionnaire, literature on radon, and a postage-paid return envelope. Eligible candidates who were unwilling to participate were mailed information on radon, and a second phone contact was made later to see if they had changed their minds about participating in the study.

The original survey design was to use sampling rates proportional to county populations, with the proportion determined by one of three factors according to whether the county was considered a potentially high-radon, medium-radon, or low-radon area. In reality, the sampling rates were far more varied. Nonetheless, for each county, a simple random sample of willing households was obtained.

A goal of the analysis was to provide an estimate of the number/fraction of houses in the state that exceed the EPA recommended action level of 4 pCi/l (see Table 2.2). In addition, it is of interest to estimate the average radon level.

Table 2.2: EPA Action Guidelines.

Radon concentration	Recommended urgency of reduction efforts
200 or above	Action to reduce levels far below 200 pCi/l as possible is recommended within several weeks after measuring these levels.
20 to 200	Action to reduce levels as far below 20 pCi/l as possible is recommended within several months.
4 to 20	Action to reduce levels to under 4 pCi/l is recommended within a few years, and sooner if levels at the upper end of this range.
Less than 4	While these levels are at or below the EPA guideline, some homeowners might wish to attempt further reductions.

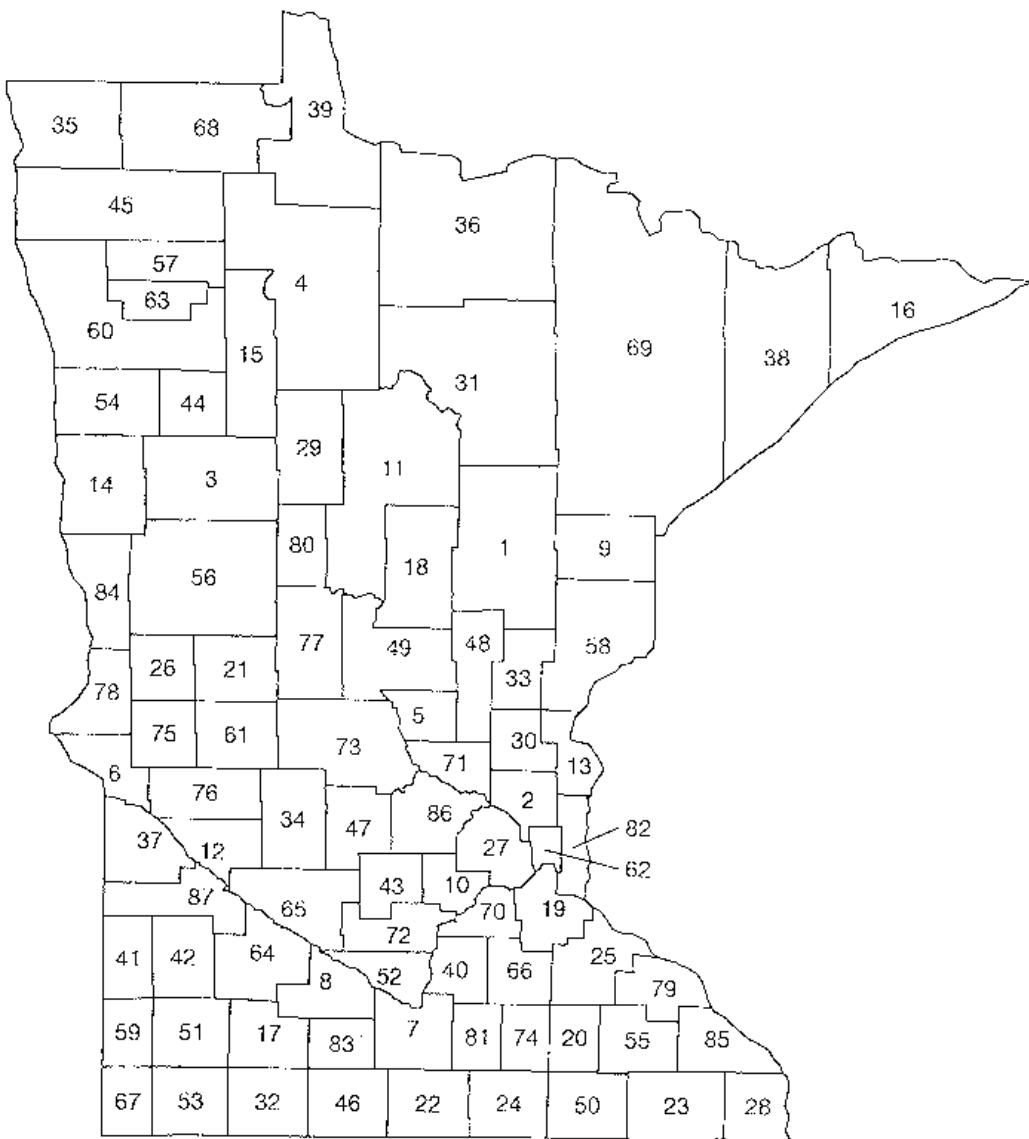


Fig. 2.1. County map of Minnesota. See Table 2.1 to identify counties.

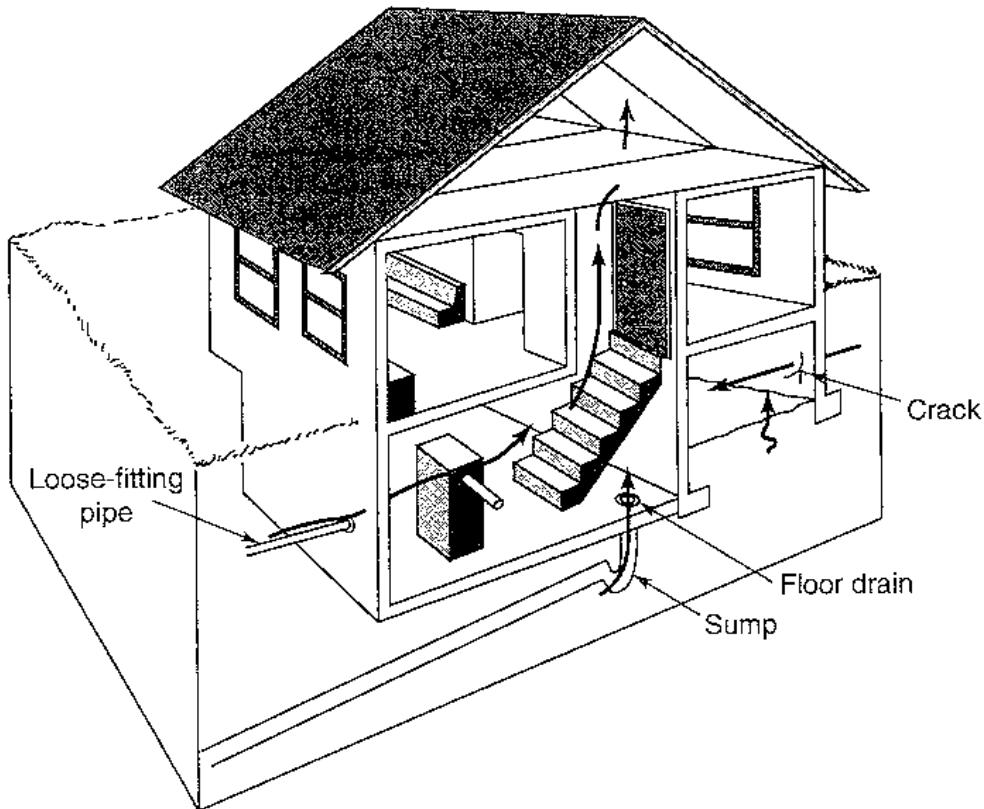


Fig. 2.2: Radon can enter a house through many paths (Nolan and Speed, 2000).

Before presenting a line of attack to this problem, it should be stressed that the data can be analysed in different ways, so what we will be discussing here is just one option (Nolan and Speed, 2000).

2.1 The model

It is helpful to start looking at the data from county level first and then to think about how county estimates can be integrated to come up with a reasonable estimate at state level. At the county level, we are dealing with a simple random sample. To estimate the fraction of houses exceeding the critical limit, define a variable y , which takes the value $y = 1$, when the critical radon level of 4 pCi/l is exceeded, and $y = 0$ otherwise. The mean of this variable will equal the fraction of households for which the critical radon level is exceeded. When estimating the mean radon level, y is the radon level measured for an individual house (using the charcoal canister provided to the household). It is helpful to index the variable y both by county and by household within county to reflect the structure of the sample. Thus,

$$y_{ij} = \text{value of unit characteristic for the } i\text{-th house in the } j\text{-th county.}$$

The unit characteristic here is either the radon level of the indicator variable taking values 0 or 1, depending on whether or not the threshold of 4 pCi/l is exceeded.

The mean of the unit characteristic in the j -th county is given by

$$\mu_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{ij}$$

where N_j is the number of houses in the j -th county. The mean for the whole state is

$$\mu = \frac{1}{N} \sum_{j=1}^J \sum_{i=1}^{N_j} y_{ij}$$

where N is the total population size in the state and J is the number of counties. Note that

$$N = N_1 + N_2 + \dots + N_J$$

Also, notice that

$$\mu = \frac{N_1}{N} \mu_1 + \dots + \frac{N_J}{N} \mu_J$$

This last equation gives us a clue as to how the mean of the unit characteristic can be estimated for the state based on the county-wise mean estimates: We just plug in the county-wise sample means as estimates for μ_j ($j = 1, \dots, J$):

$$\bar{y} = \frac{N_1}{N} \bar{y}_1 + \dots + \frac{N_J}{N} \bar{y}_J$$

Note that this is a weighted average of sample means per stratum.

The total sample size n is split among the J counties so that

$$n = n_1 + \dots + n_J$$

where n_j is the sample size in the j -th county. The mean can be re-expressed as

$$\bar{y} = \frac{1}{N} \left(\frac{N_1}{n_1} \sum_{i=1}^{n_1} y_{i1} + \dots + \frac{N_J}{n_J} \sum_{i=1}^{n_J} y_{iJ} \right) \quad (2.1)$$

This equation shows, that each observation y_{ij} receives a weight proportional to

$$g_j = N_j/n_j,$$

i.e. \bar{y} is a weighted mean of observations y_{ij} in the sample, with weights proportional to g_j . The role of weights is the same as with simple random samples (see Section 1). It will be useful to consider this second equation for the mean, which explicitly involved weights g_j , when using the SAS procedure SURVEYMEANS.

In order to construct a confidence interval, we need to first compute the variance of \bar{y} . To do this, note that the sample means from different counties are independent, since independent simple random samples were taken in each county. Thus,

$$\text{var}(\bar{y}) = \left(\frac{N_1}{N} \right)^2 \text{var}(\bar{y}_1) + \dots + \left(\frac{N_J}{N} \right)^2 \text{var}(\bar{y}_J)$$

We already know the equation for the variance of a mean in a simple random sample from Section 1. Thus, for the j -th county we have

$$\text{var}(\bar{y}_j) = \frac{\sigma_j^2}{n_j} \left(\frac{N_j - n_j}{N_j - 1} \right)$$

where n_j is the size of the sample in the j -th county and σ_j^2 the within-stratum variance for the j -th stratum (county). The question of how to determine the optimal n_j , given n , as well as the optimal n , is postponed into the next section.

Plugging this variance into our variance formula we find

$$\text{var}(\bar{y}) = \frac{\sigma_1^2}{n_1} \left(\frac{N_1}{N} \right)^2 \frac{N_1 - n_1}{N_1 - 1} + \dots + \frac{\sigma_J^2}{n_J} \left(\frac{N_J}{N} \right)^2 \frac{N_J - n_J}{N_J - 1}$$

To apply this formula, we need to estimate σ_j^2 . Using the unbiased estimator known from simple random sampling (Section 1), we can compute the estimated variance, from which a confidence interval can be obtained.

Estimated variance of a mean in a stratified sample:

$$\text{est. var}(\bar{y}) = \frac{s_1^2}{n_1} \left(\frac{N_1}{N} \right)^2 \left(1 - \frac{n_1}{N_1} \right) + \dots + \frac{s_J^2}{n_J} \left(\frac{N_J}{N} \right)^2 \left(1 - \frac{n_J}{N_J} \right) \quad (2.2)$$

where s_j^2 is the sample variance in the j -th county.

Estimated standard error:

$$\text{est.s.e.}(\bar{y}) = \sqrt{\text{est. var}(\bar{y})}$$

An approximate $(1-\alpha)100\%$ confidence interval is given by:

$$\bar{y} \pm z_{1-\alpha/2} \text{est.s.e.}(\bar{y})$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ -percentile of the standard normal distribution:

α	$z_{1-\alpha/2}$
0.01	2.58
0.05	1.96
0.10	1.64

Note: SAS uses the t-distribution instead of the standard normal, but it does not compute appropriate degrees of freedom. When the sample size is not small, the difference between the critical t and the critical z -value will be negligible.

When the stratum sample size n_j is large relative to the stratum size N_j , the factor $(1 - n_j/N_j)$ is close to unity, and the variance can be estimated as

$$\text{est.var}(\bar{y}) = \frac{s_1^2}{n_1} \left(\frac{N_1}{N} \right)^2 + \dots + \frac{s_J^2}{n_J} \left(\frac{N_J}{N} \right)^2 \quad (2.3)$$

Example 2.2: For the moment we treat Minnesota as a state with only three counties: Hennepin, Ramsey, and St. Louis (i.e. counties #27, #62, and #69). Their population sizes (in hundreds of houses) are 3925, 1809 and 81. The total population of this mini-state is 5815 hundred houses. The sample sizes in the three counties are 119, 42, and 122, respectively. The sample means for radon levels in pCi/l in the three counties are

$$\bar{y}_1 = 4.64 \quad \bar{y}_2 = 4.54 \quad \bar{y}_3 = 3.06$$

The sample standard deviations are

$$s_1 = 3.4 \quad s_2 = 4.9 \quad s_3 = 3.6$$

With

$$N_1 = 3925 \times 10^2$$

$$N_2 = 1809 \times 10^2$$

$$N_3 = 81 \times 10^2$$

$$N = 5815 \times 10^2$$

the population mean of the radon level (unit characteristic) is estimated as

$$\bar{y} = \frac{N_1}{N} \bar{y}_1 + \frac{N_2}{N} \bar{y}_2 + \frac{N_3}{N} \bar{y}_3 = \frac{3925}{5815} \times 4.64 + \frac{1809}{5815} \times 4.54 + \frac{81}{5815} \times 3.06 = 4.59$$

With

$$n_1 = 119$$

$$n_2 = 42$$

$$n_3 = 122$$

the variance is estimated to be

$$\begin{aligned} \text{est.var}(\bar{y}) &= \frac{s_1^2}{n_1} \left(\frac{N_1}{N} \right)^2 + \frac{s_2^2}{n_2} \left(\frac{N_2}{N} \right)^2 + \frac{s_3^2}{n_3} \left(\frac{N_3}{N} \right)^2 \\ &= \frac{3.4^2}{119} \left(\frac{3925}{5815} \right)^2 + \frac{4.9^2}{42} \left(\frac{1809}{5815} \right)^2 + \frac{3.6^2}{122} \left(\frac{81}{5815} \right)^2 = 0.10 \end{aligned}$$

and the estimated standard error is

$$est.s.e.(\bar{y}) = \sqrt{est.var(\bar{y})} = \sqrt{0.10} = 0.316$$

Thus, 95% confidence limits are given by

$$(4.59 \pm 1.96 \times 0.316) = (4.59 \pm 0.62) = (3.97; 5.21)$$

The mean radon level in the population of counties #27, #62 and #69 is contained in the interval from 3.97 to 5.21 with confidence probability 95%.

Exercise 2.1 (Example 2.2): For the three counties of Example 2.2, a dummy variable y was defined to equal 0 if the radon level is below 4 pCi/l and 1 otherwise. The data were analysed as follows:

$$\bar{y}_1 = 0.496 \quad \bar{y}_2 = 0.381 \quad \bar{y}_3 = 0.188$$

The sample standard deviations are

$$s_1 = 0.502 \quad s_2 = 0.492 \quad s_3 = 0.393$$

With

$$N_1 = 3925 \times 10^2$$

$$N_2 = 1809 \times 10^2$$

$$N_3 = 81 \times 10^2$$

$$N = 5815 \times 10^2$$

and

$$n_1 = 119$$

$$n_2 = 42$$

$$n_3 = 122$$

estimate the fraction of households in the population of counties #27, #62 and #69 exceeding the critical level of 4 pCi/l. You may use the approximate equation (2.3), since $n_j \ll N_j$. Do this using a pocket calculator. In addition, you may use SAS to check your results (see SAS hints below).

Exercise 2.2 (Example 2.1): Estimate the fraction of houses in Minnesota exceeding the critical radon level of 4 pCi/l. Use the data stored in the datasets **radon.dat** and **minnesota counties.dat**. The data set **radon.dat** contains the variables county, indexing the county, and the variable radon, which represents the radon level. Each line in the data set corresponds to a house in the sample. The dataset **minnesota counties.dat** contains the sample size and population size per stratum (county). Do the computations using the SAS procedure SURVEYMEANS. Computations are the same as in Exercise 2.1, but now all 87 counties are included. Also, PROC SURVEYMEANS can be prompted to use the exact formula (2.2).

2.2 SAS hints

Open the dataset radon.dat into the program editor and write a datastep to read the data. Generate the indicator variable y using an IF statement as follows:

```
data radon;
input county radon;
if radon<4 then y=0; else y=1;
datalines;
1      1.0
1      2.2
1      2.2
<more data>
87     3.7
87     2.9
87     3.7
;
```

To compute the variance of the sample mean with a finite population correction (fpc ; see p.6), i.e., accounting for the factor $(1 - n_j/N_j)$ in the variance formula, we need to provide the sampling fractions per stratum (state) in a separate data set, which also contains the county ID. We may read the data from counties.dat, which contains the variables SAMPLE (sample size per stratum = n) and POPULATION (population size per stratum = N_j , in hundreds). To read the data set, simply open it into the program editor and write a data step. In the data step, multiply POPULATION by 100 to obtain the population sizes of the strata (N_j). Define a variable _RATE_ equal to the sampling rate n/N . The SURVEYMEANS procedure will recognize the _RATE_ variable as the one needed to compute the fpc , if the dataset is submitted to the procedure via the RATE= option. Also, in the datastep we need to define a weighting variable g as in simple random sampling to compute a weighted mean according to (2.1):

```
data counties;
input county sample population;
population = population*100;
_rate_=sample/population;
g = population/sample;
datalines;
1      4      54
2      57     719
3      4      110
<more data>
85     13     161
86     14     216
87     3      46
;
```

The weighting variable g generated in the dataset labeled COUNTIES needs to be made available in the dataset RADON containing the variable to be analysed (y). This may be done using the MERGE command within a datastep as follows:

```
data radon2;
merge radon counties;
by county;
```

Finally, analyse the augmented dataset RADON2 and the datasep COUNTIES as follows:

```

proc surveymeans data=radon2 rate=counties;
stratum county;
weight g;
var y;
run;

```

If the RATE=COUNTIES option is dropped, the fpc is not computed.

Each time you generate a new SAS dataset, it is useful to print and check the result by calling the PRINT procedure, e.g.

```

proc print data=radon;
run;

```

If in Exercise 2.1 you want to compare your result to that obtained with SAS, you can proceed as in Exercise 2.2 and use an IF statement select counties 27, 62 and 69 in the data step that merges the datasets RADON and COUNTIES as follows:

```

data radon2;
merge radon counties;
by county;
if county in (27, 62, 69);

```

2.3 Optimal allocation

An important question in stratified sampling regards the optimal sample size per stratum. This optimum can be computed for a given total sample size. The optimal total sample size will need to be determined along the same lines as for simple random samples (see Example 1.4). For a given total sample size n , we need to observe the constraint

$$n = n_1 + n_2 + \dots + n_J$$

The optimal allocation of the total sample size to the different strata will depend on the variances within the different strata (σ_j^2).

If we make the simplifying assumption that sample sizes per stratum (n_j) will be small relative to the stratum size, i.e. the number of elements per stratum (N_j), and that the sample size will be so large that sample variances can be replaced by population variances, the variance of the mean estimator is

$$\text{var}(\bar{y}) = \frac{\sigma_1^2}{n_1} \left(\frac{N_1}{N} \right)^2 + \dots + \frac{\sigma_J^2}{n_J} \left(\frac{N_J}{N} \right)^2 \quad (2.3)$$

This approximate equation is more convenient, because it is easier to manipulate algebraically. The task then is to minimise the variance subject to the constraint

$$n = n_1 + n_2 + \dots + n_J$$

We will now show how to find the optimal allocation for $J = 2$. Computations are similar, but slightly more complex for $J > 2$ (see Appendix). Observing the constraint $n_1 + n_2 = n$, the variance may be expressed as a function of n_1 :

$$\text{var}(\bar{y}) = \frac{\sigma_1^2}{n_1} \left(\frac{N_1}{N} \right)^2 + \frac{\sigma_2^2}{n-n_1} \left(\frac{N_2}{N} \right)^2$$

To find the optimum, we compute the first derivative with respect to n_1 and set this equal to zero.

$$\frac{\partial \text{var}(\bar{y})}{\partial n_1} = -\frac{\sigma_1^2}{n_1^2} \left(\frac{N_1}{N} \right)^2 + \frac{\sigma_2^2}{(n-n_1)^2} \left(\frac{N_2}{N} \right)^2 = 0$$

Some algebraic rearrangement of this equation yields

$$n_1 = n \frac{\sigma_1 \frac{N_1}{N}}{\sigma_1 \frac{N_1}{N} + \sigma_2 \frac{N_2}{N}}$$

It can be shown (see Appendix) that for $J \geq 2$ the **optimal allocation** is

$$n_j = n \frac{\sigma_j \frac{N_j}{N}}{\sum_{i=1}^J \sigma_i \frac{N_i}{N}} = n \frac{\sigma_j \frac{N_j}{N}}{\sigma_1 \frac{N_1}{N} + \sigma_2 \frac{N_2}{N} + \dots + \sigma_J \frac{N_J}{N}}$$

where

n = total sample size

N_j = number of population units in j -th stratum

σ_j = standard deviation in j -th stratum

Obviously, the optimal allocation depends on the stratum variances.

Under the simple assumption that the **stratum variances (standard deviations) are the same across all strata** ($\sigma_j = \sigma$ for every j , so that σ_j cancels out) the optimal allocation is

$$n_j = n \frac{N_j}{N}$$

Example 2.3: For the moment treat Minnesota as a state with only three counties: Hennepin, Ramsey, and St. Louis (i.e. counties #27, #62, and #69). Their population sizes (in hundreds of houses) are 3925, 1809 and 81. The total population of this mini-state is 5815 hundred houses. Assume the desired sample size is $n = 158$ and that the variances in the three counties are the same. With

$$N_1 = 3925 \times 10^2$$

$$N_2 = 1809 \times 10^2$$

$$N_3 = 81 \times 10^2$$

$$N = 5815 \times 10^2$$

$n = 158$

the optimal allocation is found to be

$$n_1 = n \frac{N_1}{N} = 158 \frac{3925}{5815} \approx 107$$

$$n_2 = n \frac{N_2}{N} = 158 \frac{1809}{5815} \approx 49$$

$$n_3 = n \frac{N_3}{N} = 158 \frac{81}{5815} \approx 2$$

Note that this allocation deviates notably from the allocation used in Example 2.2.

Exercise 2.3 (Example 2.3): For the three counties of Example 2.2, the standard deviations of the dummy variable y defined to equal $y=0$ if the radon level is below 4 pCi/l and $y=1$ otherwise, were:

$$s_1 = 0.502 \quad s_2 = 0.492 \quad s_3 = 0.393$$

With

$$N_1 = 3925 \times 10^2$$

$$N_2 = 1809 \times 10^2$$

$$N_3 = 81 \times 10^2$$

$$N = 5815 \times 10^2$$

find the optimal allocation, assuming a total sample size of $n = 283$, assuming the estimated standard deviations (s_j) equal the true standard deviations (σ_j). Compare your result to the allocation actually used ($n_1 = 119$, $n_2 = 42$, $n_3 = 122$, $n = 283$) [The optimal allocation could not be used in the actual survey because the standard deviations were not known *a priori*. Specifically, for binary data (0-1) as used here the sample variance will depend on the mean, and if this were known in advance, the survey would not need to be done in the first place!].

2.4 How to find n

In the preceding section we have considered the problem of finding the optimal allocation of a given total sample size n to strata. We may plug the optimal allocation formula for n_j into the formula for the variance of the overall mean \bar{y} . This will yield an equation that depends only on the unknown n . From there, we may use the method based on the half width of the confidence interval described for simple random sampling (Section 1). This will not be elaborated here detail. The procedure is only briefly sketched.

(1) Decide on a specific form of the optimal allocation, depending on the type of information available on, or the type of assumptions made with regard to, the stratum standard deviations σ_i .

(2) Use the relevant equation for the optimal allocation in (1) and plug it into the variance of the overall estimator, such as the simplified equation (2.3). This equation depends only on the overall sample size n .

(3) Get a priori information on the variance.

(4) Quantify the required precision for the overall estimator in terms of the half width (HW) of the 95% confidence interval and equate this to $2\sqrt{\text{var}(\bar{y})}$. Solve this equation for n .

(5) Compute sample sizes for the strata based on the formula chosen for the optimal allocation from (1).

Example 2.4:

(1) We decide to use the allocation $n_j = n \frac{N_j}{N}$. This assumes that $\sigma = \sigma_i$ for all strata. For this common variance, prior information is required.

$$(2) \text{ var}(\bar{y}) = \frac{\sigma^2}{n_1} \left(\frac{N_1}{N} \right)^2 + \dots + \frac{\sigma^2}{n_J} \left(\frac{N_J}{N} \right)^2 = \frac{\sigma^2}{n} \left(\frac{N_1}{N} + \frac{N_2}{N} + \dots + \frac{N_J}{N} \right) = \frac{\sigma^2}{n}$$

(3) Prior information suggests $\sigma = 10$

(4) We require $HW = 5 \Rightarrow$

$$5 = 2\sqrt{\text{var}(\bar{y})} \Leftrightarrow$$

$$25 = 4 \text{ var}(\bar{y}) = 4 \frac{\sigma^2}{n} \Leftrightarrow$$

$$n = 4 \frac{\sigma^2}{25}$$

Now plug in prior information about the variance:

$$n = 4 \frac{\sigma^2}{25} = 400 / 25 = 16$$

(5) Assume there are 3 strata with $N_1 = 1000$, $N_2 = 3000$, and $N_3 = 4000$. Then

$$n_1 = n \frac{N_1}{N} = 16 \frac{1000}{8000} = 2, \quad n_2 = n \frac{N_2}{N} = 16 \frac{3000}{8000} = 6, \quad n_3 = n \frac{N_3}{N} = 16 \frac{4000}{8000} = 8$$

2.5 Other sampling methods

In Sections 1 and 2 we have looked at two sampling methods. There is a wide variety of sampling methods as well as estimation methods for survey sampling, which cannot be covered here. For details you may take a look at a pertinent textbook, e.g. Yates (1981) or Särndal et al. (1993).

*Appendix

The variance of a mean is

$$\text{var}(\bar{y}) = V = \text{var}\left(\sum_{i=1}^J N_i \bar{y}_i\right) = \sum_{i=1}^J \frac{N_i^2 \sigma_i^2}{n_i} \quad (\text{A1})$$

σ_j^2 = variance of j -th stratum

J = number of strata

Need to observe constraint:

$$n = \sum_{i=1}^J n_i \quad (\text{A2})$$

To find the minimum of V subject to the constraint (A2), we may use the **method of Lagrange-multipliers** and minimize

$$F = V + \lambda \left(n - \sum_{i=1}^J n_i \right) \quad (\text{A3})$$

Note that F equals V plus a multiple of the constraint $\left(n - \sum_{i=1}^J n_i \right)$. We find for the derivatives

$$\frac{\partial F}{\partial n_j} = -\frac{\sigma_j^2 N_j^2}{n_j^2} - \lambda = 0 \quad \text{for all } i \text{ and} \quad (\text{A4})$$

$$\frac{\partial F}{\partial \lambda} = n - \sum_{i=1}^J n_i = 0 \quad (\text{A5})$$

Note that (A5) yields the constraint, which shows that minimisation of (A3) will make sure the constraint holds for the solution. We first try to find a solution for the Lagrange-multiplier λ . Rearranging (A4) yields:

$$\sigma_i N_i = n_i \sqrt{-\lambda} \quad \text{for every } i \quad (\text{A6})$$

Adding equations (A6) across strata and inserting (A5) we find:

$$\sum_{i=1}^J \sigma_i N_i = n \sqrt{-\lambda} \Leftrightarrow \lambda = -\frac{\left(\sum_{i=1}^J \sigma_i N_i \right)^2}{n^2} \quad (\text{A7})$$

Inserting (A4) for λ yields:

$$\frac{\sigma_j^2 N_j^2}{n_j^2} = \left(\sum_{i=1}^J \sigma_i N_i \right)^2 / n^2 \Leftrightarrow n_j = n \frac{\sigma_j N_j}{\sum_{i=1}^J \sigma_i N_i} \quad (\text{A8})$$

References

- Nolan D, Speed T 2000 StatLabs. Mathematical statistics through applications. Springer, Berlin (SK 850 N 787)
- Särndal CE, Swensson B, Wretman J 1993 Model assisted survey sampling. Springer, New York.
- Yates F 1981 Sampling methods for censuses and surveys. Charles Griffin, London.

3. Regression and correlation

Example 3.1 (Nolan and Speed, 2000):

The first dataset (crab1.dat): With the assistance of the California Department of Fish and Game and commercial fishers from northern California and southern Oregon, a group of researchers collected growth data on the adult female Dungeness crab (*Cancer magister*). For a sample of 472 crabs, the following variables were recorded:

```
presz = Premolt carapace size in millimeters  
postsz = Postmolt carapace size in millimeters  
incr = Increment = postmolt size - premolt size  
year = Year (81=1981, 82=1982, 92=1992, NA = not assessed)  
lf = Measurement site 0=field, 1=lab
```

The researchers wanted to study the relation between premolt and postmolt carapace size. Specifically, they were interested in deriving an equation for predicting premolt size from postmolt size. This equation can be used to predict premolt sizes for crabs, on which only postmolting measurements have been taken. Such predictions are needed to better understand the growth biology of the crab.

Biological background: The Dungeness crab has a broad, flattened hard shell, or **carapace**, that covers the back of the animal. The shell is an external skeleton that provides protection for the crab (Fig. 3.1). To accommodate growth, the crab **molts** periodically, casting off its shell and growing a new one. The molting process takes about four to five days. Immediately after the molting season, it is fairly easy to determine whether a crab has recently molted; its shell is clean, free of barnacles, and lighter in color.

Crabs mate in April and May when females molt. The male crabs molt later in the year in July and August. During the female crab's molting season male and female crabs enter shallow water; a male and female will embrace prior to the female's molting. When the female leaves her shell, the male deposits a sperm packet in the female. Once the female crab's shell has hardened, the male and female separate. The female stores the sperm for months as her eggs develop. In the fall, she extrudes her eggs and fertilizes them with the stored sperm. The juveniles molt every few months for about two years, until they reach maturity. This occurs when the capraces (shells) are 90 to 100 mm in width.

Ecological background: In US waters, nearly the entire adult male Dungeness crab population is fished each year. Female crabs are not fished in order to maintain the viability of the crab population. However, the question of fishing female crabs has been raised as a means of controlling large fluctuations in yearly catches of crabs. To support the change in the US fishing law, it has been noted that the fishing industry in Canada allows female crabs to be fished and does not suffer from such large fluctuations in catches.

Size restrictions on male crabs are set to ensure that they have at least one opportunity to mate before being fished. To help determine similar size restrictions for female crabs, more needs to be known about the female crab's growth.

The lack of growth marks on crab shells makes it difficult to determine the age of a crab. This is because crabs molt regularly, casting off their old shell and growing a new one. Adult female crabs molt in April and May, although they do not necessarily molt yearly. Biologists require size-specific information on molting to understand the female crab's growth pattern.

Of particular interest is the size of the increase in the width of the shell having observed only the size after the crab molted; for example, for a female crab with a postmolt shell that measures 150 mm across, the scientists want to provide a prediction for how much the shell changed in size. This information is needed in developing recommendations for size restrictions on fishing female crabs.

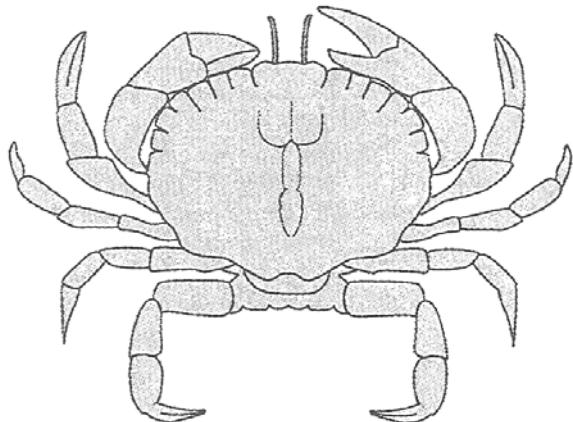


Fig. 3.1 Dungeness crab (*Cancer magister*).

Some additional information on the first dataset: Part of the data were obtained by capture-recapture in the years 1981, 1982 and 1992. 12,000 crabs were caught, measured, tagged with a unique identification number, and returned to the water. This was done before the molting season (January to March). Commercial fisheries brought tagged crabs they caught in their traps to the laboratory for second measurements. Commercial traps have netting designed to catch the larger male crabs; female crabs caught with these traps were typically larger than 155 mm.

The laboratory data were collected during the molting season for female crabs. Crabs that were in a prematuring embrace were caught and brought to the lab. The premolt carapace width was measured when the crab was first collected, and the postmolt measurements were made three or four days after the crab left its old shell to ensure that the new shell had time to harden.

Studies suggest that crabs in captivity have smaller molt increments than those in the wild. Although the crabs in this study were held in captivity for only a few days, a comparison of the crabs caught by these two collection methods is advisable. Also it needs to be kept in mind that the crabs were fished in the early 1980s and ten years later.

A second data set (crab2.dat) was collected in late May 1983, after the molting season. The carapace width was recorded as well as information on whether the crab had molted in the most recent molting season or not. The crabs were collected in traps designed to catch adult female crabs of all sizes, and so it is thought that the sample is **representative** of the adult female Dungeness crab population. This sample consists of 362 crabs.

```
size = Carapace size  
shell = 0=clean shell (crab did not molt), 1=fouled shell (crab molted)
```

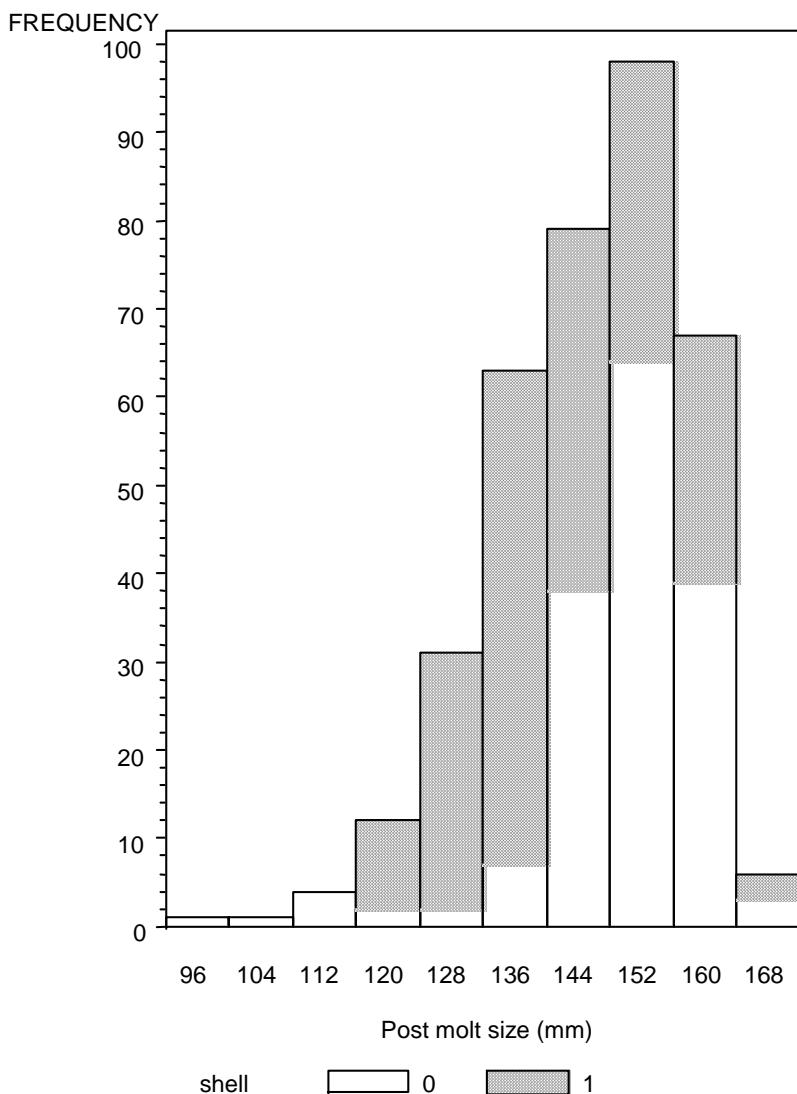


Fig. 3.2: Size distribution of 362 adult female Dungeness crabs shortly after the 1983 molting season. Shell = 0: crabs did not molt. Shell = 1: Crabs molted. Numbers on the abscissa are class means (**crab2.dat**).

The ultimate objective of our analysis is to obtain an estimate of the premolt (January to March) distribution of carapace sizes and to display this in a histogram as in Fig. 3.2. To derive useful catching size restrictions, the histogram needs to be representative of the crab population. Dataset 1, which also has pre-molt measurements, cannot be used to obtain the histogram. The reason is that part of the crabs were brought by commercial fisheries. Commercial traps have netting designed to catch the larger male crabs; female crabs caught with these traps were typically larger than 155 mm. Thus, crabs smaller than 155 mm are probably underrepresented. By contrast, dataset 2 has a representative sample, but no premolt data! However, we can use the results of dataset 1 to predict premolt sizes for dataset 2!

3.1 Histogram

A histogram displays the frequency of measurements by vertical bars. The data are classified into equally spaced classes and the frequency in each class is used to determine the height of

bars corresponding to each class. The resulting histogram gives a visual impression of the distribution of the data. From the histogram, we can deduce the fraction of observations falling below a certain threshold. In Figure 3.2, class borders are at 92, 100, 108, 116, 124, 132, 140, 148, 156 and 164 mm. The class width is 8 mm, and there are 10 classes.

The software used to generate the histogram in Fig. 3.2 (SAS PROC GCHART) uses a rule as the following to determine class width and the number of classes:

Number of classes (k)

$$k \geq (2n)^{1/3}$$

(Terrel GR, Scott DW 1985 Oversmoothed nonparametric density estimates. JASA 80, 209-214)

Class width (b):

$$b > \frac{V}{k}$$

where

$$V = y_{max} - y_{min}$$

is the **range** of observed values.

Example 3.1 For **crab2.dat** we find:

$$n = 362 \rightarrow k \geq (2*362)^{0.333} = 8.97 ; \text{ chose } k = 10$$

$$y_{min} = 95.6$$

$$y_{max} = 168$$

$$V = 72.6$$

$$b > 72.6/10 = 7.2 ; \text{ chose } b = 8$$

Exercise 3.1: Use PROC GCHART and PROC UNIVARIATE to make a histogram for the 1983 crab data in **crab2.dat**.

Exercise 3.2: Suppose you have a sample of size $n = 100$ and the range is 700. How many classes and what class width would you use to draw a histogram?

SAS hints:

```
data crabs;
input size shell;
datalines;
116.8      1
117.1      1
<more data>
166.6      0
168.0      0
;
```

```

proc univariate data=crabs;
histogram size;
run;

proc gchart data=crabs;
vbar size /subgroup=shell space=0;
run;

```

3.2 Correlation

We will now study the relationship between pre- and post-molting size in the first dataset (**crab1.dat**). To do so, we plot premolt size versus postmolt size. Fig. 3.3 shows that there is a close association of premolt and postmolt size.

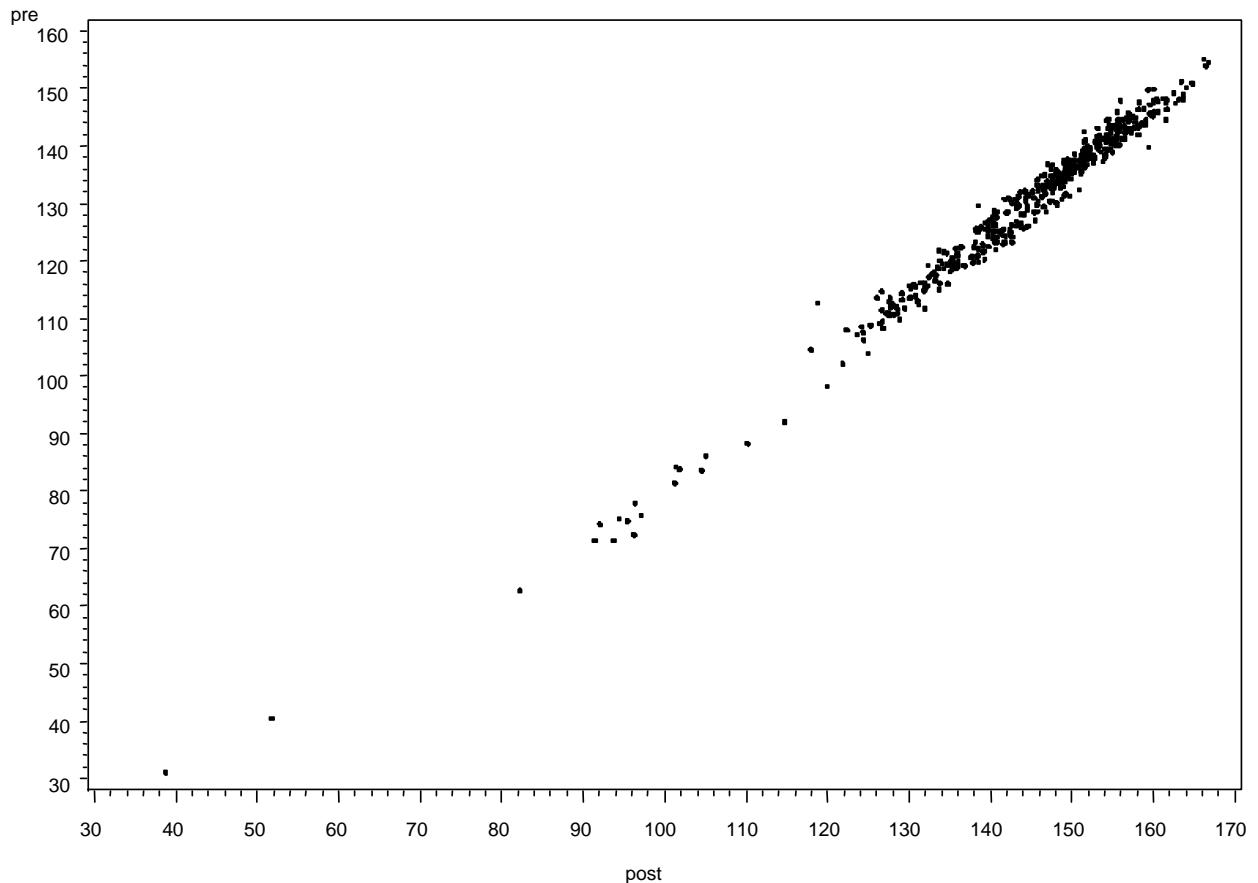


Fig. 3.3: Scatter plot of premolt and postmolt carapace width for 472 adult females Dungeness crabs (Nolan and Speed, 2000); dataset 1.

A useful measure to quantify the association between two quantitative measurements is the correlation. The correlation is a quantity scaled to fall between -1 and 1. A value of 0 indicates that the two variables are not correlated. A correlation of -1 or +1 means that all points of the scatter plot fall exactly on a line. For the crab data in Fig. 3.3, the correlation is 0.99, indicating that the correlation is very tight. The correlation coefficient is a dimensionless measure of linear association, i.e., if premolt site is converted to centimeters or inches, the value of r remains unchanged.

Positive correlation coefficients indicate that above average values in one variable, such as postmolt size, tend to be associated with above average values in the second variable, such as premolt size. It also indicates that below average values in the first variable are typically associated with below average values in the second. Conversely, negative correlation coefficients indicate that above average values in one variable tend to be associated with below average values in the second variable, and vice versa.

To compute the correlation coefficient, let $(x_1, y_1), \dots, (x_n, y_n)$ be the pairs of postmolt and premolt sizes for all laboratory crabs. Further let

$$\bar{x} = \text{average of postmolt size}$$

$$\bar{y} = \text{average of premolt size}$$

$$s_x = \text{standard deviation of postmolt size}$$

$$s_y = \text{standard deviation of premolt size}$$

The correlation is defined as follows:

$$r = \frac{1}{n-1} \sum_{i=1}^n \frac{x_i - \bar{x}}{s_x} \times \frac{y_i - \bar{y}}{s_y}$$

In our example, $\bar{x}=143.9$ mm, $s_x = 14.6$ mm, $\bar{y}=129.2$ mm, $s_y = 15.9$ mm, and $r = 0.99$. For computational purposes, the formulae below are preferable.

When the correlation is to be computed by hand, the following equations are helpful:

$$r = \frac{CP_{xy}}{\sqrt{SS_x SS_y}}$$

$$CP_{xy} = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n} \quad (\text{sum of cross-products})$$

$$SS_x = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n}$$

$$SS_y = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n}$$

(x_i, y_i) = i -th pair of values.

Example 3.1: For the crab data in **crab1.dat**, we find

$$CP_{xy} = 8884374.91 - \frac{67919.7 * 60988}{472} = 108343.84$$

$$SS_x = 9874444.09 - \frac{67919.7^2}{472} = 100957.55$$

$$SS_y = 7998915.88 - \frac{60988^2}{472} = 118542.69$$

$$r = \frac{108343.84}{\sqrt{100957.55 * 118542.69}} = 0.99$$

Exercise 3.3: Compute the correlation for premolt and postmolt size in **crab1.dat** using PROC CORR.

SAS hints

```
data crabs;
input
presz    postsz     inc      year$   lf;
datalines;
113.6    127.7    14.1     NA      0
118.1    133.2    15.1     NA      0
<more data>
134.1    148.3    14.2     92      1
114.4    129.2    14.8     92      1
;
proc corr data=crabs;                                (computes correlation)
var presz postsz;
run;

proc gplot;                                         (plots premolt size versus postmolt size)
plot presz*postsz;
run;
```

Exercise 3.4: Compute the correlation for the following artificial data using a pocket calculator:

x	3	5	8	10
y	8	8	5	1

Verify your result using PROC CORR.

The size increment

The correlation between premolt and postmolt size is very high ($r = 0.99$). The main reason is that postmolt size is made up of premolt size plus a small growth increment. That is, when a group of crabs molt, the big crabs stay big and the small crabs stay small, relatively speaking. This may be stated more formally by expressing the postmolt size as

$$x = y + z$$

where

x = postmolt size
 y = premolt size
 z = increment

Rearranging this, the increment, z , is

$$z = x - y$$

It is informative to correlate the increment z , with postmolt size x (Fig. 3.4).

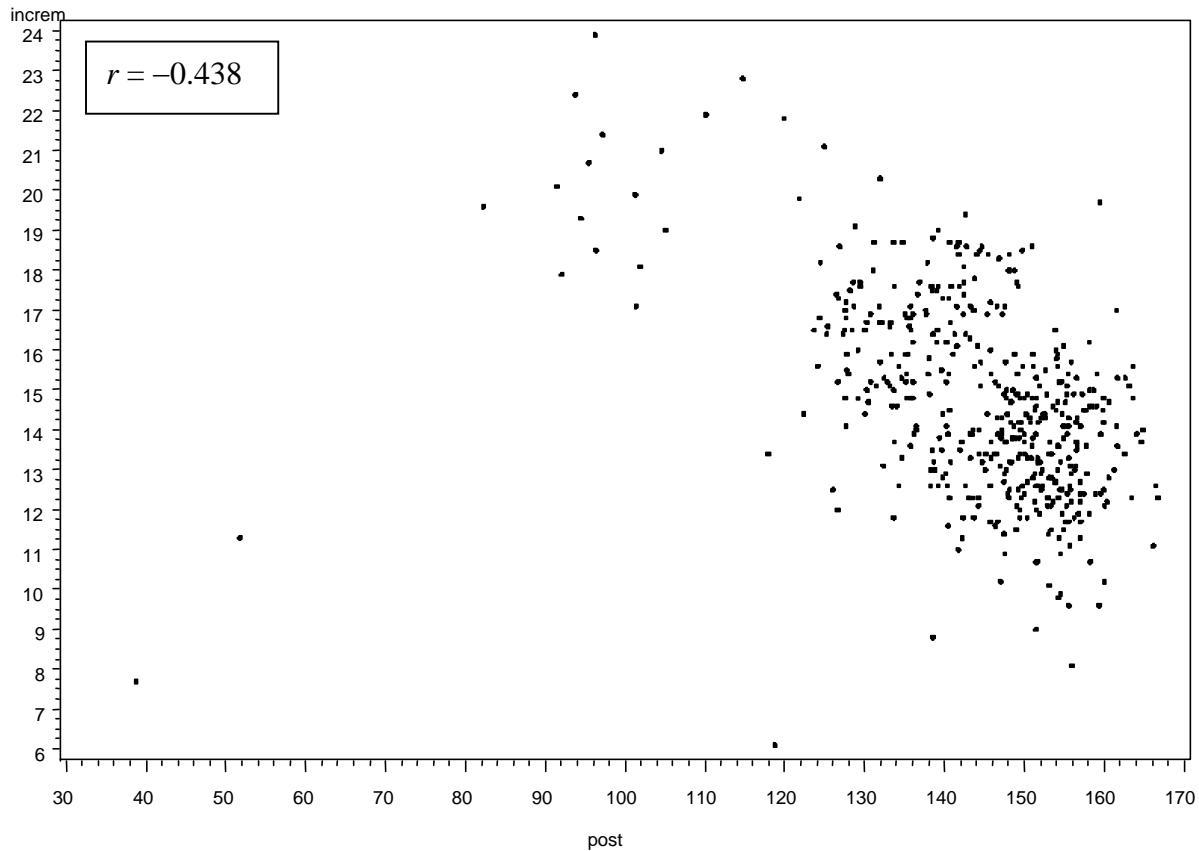


Fig. 3.4: Scatter plot of size increment and postmolt carapace width for 472 adult females Dungeness crabs (**crab1.dat**).

The plot shows that size increments are smaller for large crabs than for small ones, a feature that goes unnoticed when plotting premolt versus postmolt size. The association is not very tight, though. It is useful here to test whether the correlation is, in fact, statistically significant. This will be done in Section 3.2. Figure 3.4 also reveals that there are two rather atypical observations with a postmolt size of about 40 and 50. Such atypical observations are called **outliers**. These are atypical for two reasons: The post molt size is far below average and the observations seem to suggest that for low postmolt sizes, the size increment becomes larger as postmolt size increases. Overall, this would suggest a curvilinear relationship between increment and postmolt size. It must be noted, however, that this impression is based on only two observations. Therefore, care should be exercised when drawing conclusions from the scatter plot. The two small crabs are probably juvenile crabs. If we keep in mind that predictions are needed for adult crabs only, for which postmolt size is usually well above 90, we may safely ignore these two outlying observations. Generally, the question of how to deal with outliers is a difficult one, but here the answer is rather obvious for biological reasons. When dropping the two outliers, the scatter plot suggests a linear association (Fig. 3.5). In face of the two outliers, the correlation coefficient would not be a useful measure to summarize the scatter plot for the whole data. This is so because the correlation coefficient is

a measure of linear association. With the two outliers, the correlation is $r = -0.438$, without them it changes $r = -0.564$. This is a relatively dramatic change, which shows that the two outliers are rather influential, diluting the correlation.

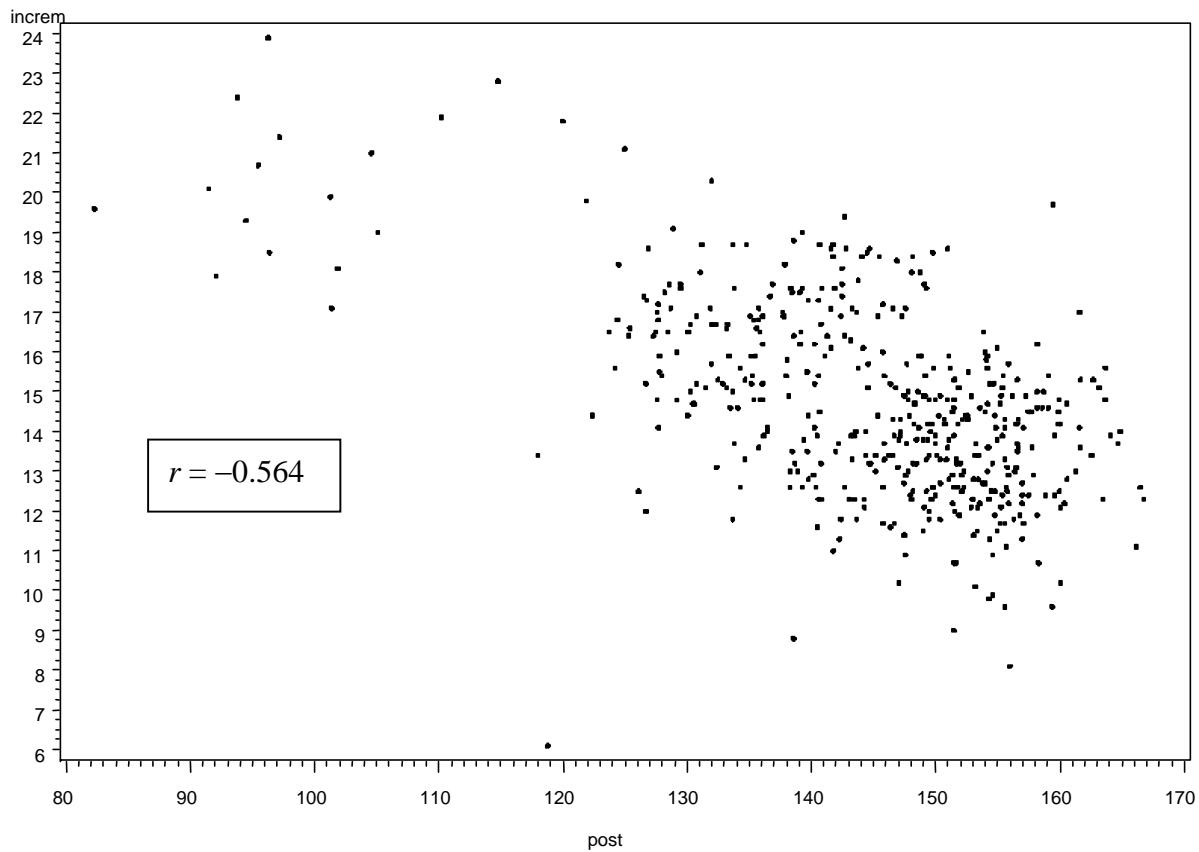


Fig. 3.5: Scatter plot of size increment and postmolt carapace width for 470 adult females Dungeness crabs with postmolt size >80 (**crab1.dat**).

Exercise 3.5: For the association of size increment (z) and postmolt size (x), we find:

$$\sum_{i=1}^n x_i = 67829$$

$$\sum_{i=1}^n z_i = 6913$$

$$\sum_{i=1}^n x_i^2 = 9870255.41$$

$$\sum_{i=1}^n z_i^2 = 104434.20$$

$$\sum_{i=1}^n x_i z_i = 989232.81$$

$$n = 470$$

Compute the sample correlation.

3.2.1 Test of correlation

The sample correlation r is just an estimate of the correlation in the population of crabs, ρ . Even if the population correlation equals zero, we will most likely obtain a sample correlation differing from zero (similarly, in survey sampling, the sample mean is not usually equal to the population mean; see Section 1). Thus, it is often useful to perform a test of the null hypothesis, that the population correlation is zero. This is particularly true, when the estimated correlation is rather small, so there is doubt whether there is any real association. The correlation of postmolt and premolt equals $r = 0.99$ ($n = 472$). The scatter plot in Fig. 3.3 leaves little doubt, that this correlation is "real". For the correlation of size increment and postmolt size, the picture is less clear (Fig. 3.5), though the scatter plot still is rather suggestive of a real association. For this regression, it may be useful to perform a test of the null hypothesis of no correlation ($H_0: \rho = 0$).

Test of correlation ρ

Question: Is there a real association between two variables X and Y ?

Assumption: The data follow a bivariate normal distribution

$H_0: \rho = 0$ (null hypothesis)

$H_A: \rho \neq 0$ (alternative hypothesis)

Computations:

$$(1) \text{ Compute } t_{obs} = \frac{|r|}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

(2) Read the critical t-value $t_{tab} = t_{1-\alpha/2; n-2}$ (see p. 7, Section 1), where α is the significance level. If n is large ($n > 100$), replace t_{tab} by $z_{tab} = z_{1-\alpha/2}$.

(3) Compare t_{obs} with t_{tab} :

If $t_{obs} \leq t_{Tab} \Rightarrow H_0 (\rho = 0)$ (no association)

If $t_{obs} > t_{Tab} \Rightarrow H_A (\rho \neq 0)$ (significant association)

Example 3.1: The correlation of postmolt size and size increment equals $r = -0.564$ ($n = 470$). We perform a t-test of the null hypothesis of no correlation ($H_0: \rho = 0$) at a significance level of 5%.

$r = -0.564, n = 470, \alpha = 5\%$

$$t_{obs} = \frac{|-0.564|}{\sqrt{1 - (-0.564)^2}} \sqrt{470 - 2} = 14.78$$

$t_{tab} \approx z_{tab} = 1.96 < t_{obs} = 14.78 \Rightarrow$ The correlation is significant.

If the test is performed using a statistical package, the output does not provide t_{tab} and t_{obs} . Instead it computes what is usually called a **p-value**.

The **p-value** is the probability of observing a value for the test statistic (t_{obs} in this case), which is at least as extreme as the one observed for the data, **providing the null hypothesis is true**. When this probability is smaller than some pre-specified limit α , it is concluded that the null hypothesis is not plausible - the null hypothesis is then rejected and the test is said to be significant.

Example 3.1: Under the null hypothesis of no real correlation, one would expect r , and thus t_{obs} , to be close to zero. For the crab data, the p-value of $r = -0.564$ ($t_{obs} = 14.78$) is smaller than $p = 0.0001$. Thus, under the null hypothesis, the probability of observing $t_{obs} = 14.78$, or larger, is smaller than $p = 0.0001$. This is smaller than $\alpha = 5\%$, so the null hypothesis is rejected, i.e. the test is significant at the 5% level of significance.

Exercise 3.6: Use PROC CORR to compute the correlations of postmolt size and size increment in **crabs1.dat** and perform a t-test of the null hypothesis that there is no correlation.

Exercise 3.7: Use a pocket calculator to test $H_0: \rho = 0$ for the correlation of postmolt size and premolt size ($r = 0.99, n = 472$).

Exercise 3.8: Use PROC CORR to see how the correlation changes among postmolt size and premolt size, when the two outliers identified in Fig. 3.4 are deleted.

SAS hints

To delete observations with postmolt size < 80 , you may invoke an IF statement in the dataset as follows:

```
data crabs;
input
presz    postsz     inc      year$   lf;
if postsz > 80;
datalines;
113.6    127.7    14.1    NA      0
118.1    133.2    15.1    NA      0
<more data>
134.1    148.3    14.2    92      1
114.4    129.2    14.8    92      1
;
```

3.2.2 Confidence interval for the correlation

It can be shown (A. Stuart & K. Ord: Kendall's advanced theory of statistics. 6th edition. Volume 1, § 16.33) that for large n the so-called Fisher's z-transformation

$$q = \frac{1}{2} \ln \left[\frac{1+r}{1-r} \right]$$

is approximately normal with mean

$$\theta = \frac{1}{2} \ln \left[\frac{1+\rho}{1-\rho} \right]$$

and variance

$$\sigma_q^2 = \frac{1}{n-3}$$

Thus, the random variable

$$z = \frac{q - \theta}{\sigma_q}$$

has an approximate standard normal distribution. Thus,

$$\theta_u = q - z_{1-\alpha/2} \sigma_q \text{ and } \theta_o = q + z_{1-\alpha/2} \sigma_q$$

provide approximate $(1-\alpha)100\%$ -confidence limits on θ , where $z_{1-\alpha/2}$ is the $(1-\alpha/2)100\%$ -quantile of the standard normal distribution. The corresponding limits for ρ are easily obtained by back-transformation of the limits θ_u and θ_o . The whole procedure is summarized below.

Confidence interval for correlation ρ

Compute:

$$q = \frac{1}{2} \log \left[\frac{1+r}{1-r} \right] \text{ and}$$

$$\theta_L = q - \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \text{ und } \theta_U = q + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}$$

where

$z_{1-\alpha/2}$ = $(1-\alpha/2)100\%$ -quantile of standard normal distribution

The $(1-\alpha)100\%$ -confidence limits are given by

$$\rho_L = \frac{e^{2\theta_L} - 1}{e^{2\theta_L} + 1} \text{ and } \rho_U = \frac{e^{2\theta_U} - 1}{e^{2\theta_U} + 1}$$

Assumption: Data follow a bivariate normal distribution.

Example: $r = 0.99037$, $n = 472$, $\alpha = 5\%$

$$q = \frac{1}{2} \log \left[\frac{1+0.990373}{1-0.99037} \right] = 2.6656$$

$$z_{1-\alpha/2} = 1.96$$

$$\theta_L = 2.6656 - \frac{1.96}{\sqrt{469}} = 2.6656 - 0.0915 = 2.5742 \text{ and } \theta_U = 2.6656 + 0.0915 = 2.7570$$

$$\rho_L = \frac{e^{2*2.5742} - 1}{e^{2*2.5742} + 1} = 0.9885 \text{ and } \rho_o = \frac{e^{2*2.7570} - 1}{e^{2*2.7570} + 1} = 0.9920$$

With 95% probability the interval from 0.9885 to 0.9930 covers the true correlation ρ .

SAS hints

```
proc corr data=crabs fisher(type=twosided biasadj=no);
var presz postsz;
run;
```

3.3 Linear regression

The scatter plot in Fig. 3.3 as well as the correlation of $r = 0.99$ suggest that premolt size (y) increases linearly with postmolt size (x). But this finding alone does not allow us to predict premolt size from postmolt size, our primary objective. Since the association is linear, it would seem reasonable to draw a line through the scatter of points and use this for prediction. Specifically, if the line is of the form

$$\text{premolt} = \alpha + \beta \times \text{postmolt}$$

the **prediction** for the premolt size of a crab with premolt size equal to 136 would be

$$\text{premolt} = \alpha + \beta \times 136$$

It is perhaps counterintuitive to speak of prediction here, since the predicted variable precedes the explanatory variable in time. However, from a statistical point of view, the term prediction applies whenever an unknown variable is computed (predicted) from a known variable, regardless of any time-related ordering among the two variables.

The values for α and β can be found by the **method of least squares**. This method finds the line that minimizes the sum of squared difference between the observed premolt size and the premolt size on the line:

$$\sum (\text{premolt} - \alpha - \beta \times \text{postmolt})^2$$

More formally, the least squares method minimizes the following sum of squares with respect to α and β :

$$\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

for pairs (x_i, y_i) of postmolt and premolt sizes for n crabs. The minimizing values of α and β , the **least squares estimates**, are denoted by $\hat{\alpha}$ and $\hat{\beta}$, and the resulting line, $\hat{y} = \hat{\alpha} + \hat{\beta}x$, is called the **regression line** of premolt size on postmolt size. For the data in **crab1.dat** we find, deleting the two outliers:

$$\text{premolt} = -29.68 + 1.10 \times \text{postmolt}$$

The regression line with the scatter plot is shown in Fig. 3.6.

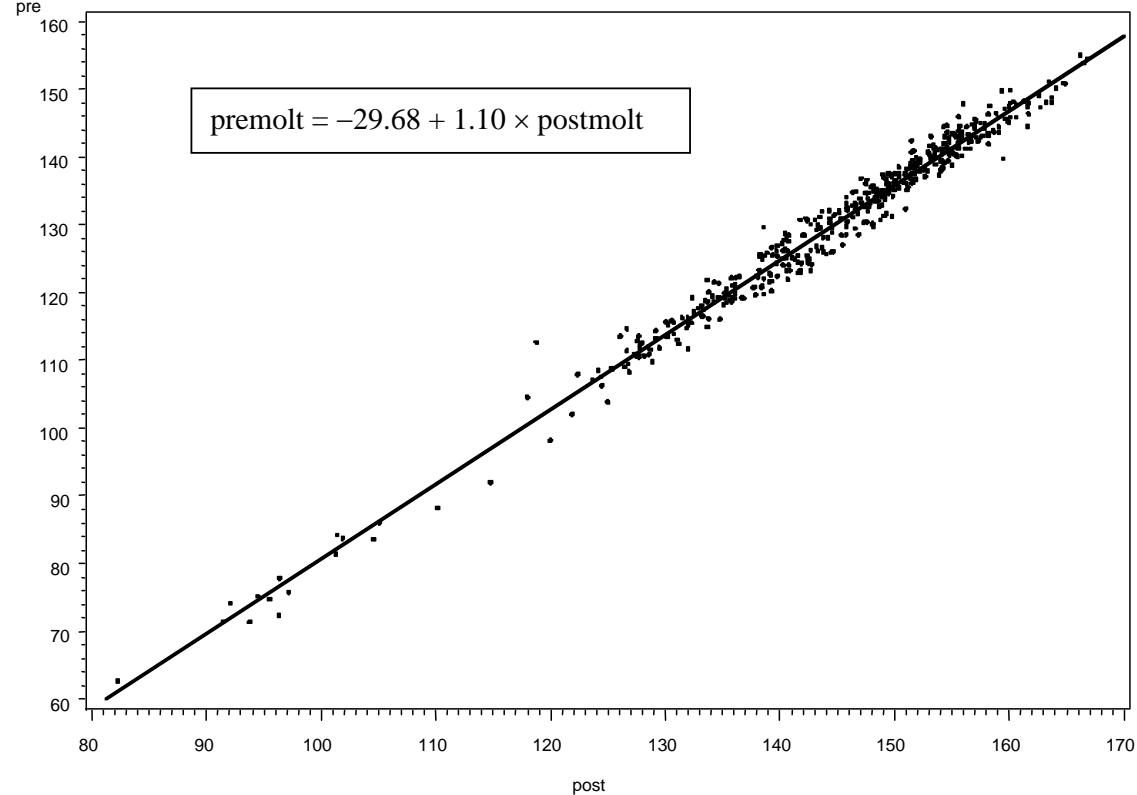


Fig. 3.6: Scatter plot of size increment and postmolt carapace width for 470 adult females Dungeness crabs with postmolt size >80 (**crab1.dat**).

Some more terminology: In the regression equation ($\text{premolt} = \alpha + \beta \times \text{postmolt}$), premolt is called the **response**, while postmolt is denoted as **explanatory variable**. This is because the response may (at least partly) be explained by the regression on the explanatory variable. Alternative terms for response (y) and explanatory variable (x) are as follows:

y	x
response	explanatory variable
predicted variable	predictor variable
dependent variable	independent variable

The least squares estimates of a linear regression can be computed as follows:

$$\hat{\beta} = r \frac{s_y}{s_x}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Exercise 3.9: For the data in **crab1.dat** we find, deleting the two outliers, $s_y = 14.67$, $s_x = 13.17$, $r = 0.99$, $\bar{x} = 144.3$ and $\bar{y} = 129.6$. Compute the least squares regression line. Check your results using PROC REG of the SAS System. Plot the data and the regression line using PROC GPLOT.

SAS hints

```
proc reg data=crabs;
model presz=postsz;           (This does the regression)
run;

symbol value=dot i=r1;        (This plots the regression)
proc gplot data=crabs;
plot presz*postsz;
run;
```

3.4 Residuals

The points in the scattergram do not fall exactly on the regression line. As a result, predictions are not perfect. This can be seen from the differences of observed and predicted premolt sizes. These differences are called **residuals**, and they are computed by

$$r_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$$

or

$$r_i = y_i - \hat{y}_i$$

where the regression line prediction is

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

The more variable the residuals, the less accurate are the predictions from the regression.

The (approximate) **standard deviation of the residuals** can be computed as

$$s_r = s_y \sqrt{1 - r^2}$$

The variance is given by

$$s_r^2 = (1 - r^2)s_y^2$$

For the crab data we find $s_r = 2.00$, whereas $s_y = 14.67$.

Note that the variance of the residuals, s_r^2 , will always be a fraction of the variance of premolt sizes, s_y^2 , since the factor $(1 - r^2)$ will always fall between 0 and 1. The ratio

$$\frac{s_r^2}{s_y^2} 100\%$$

tells us which percentage of the total variance in premolt size is **not** accounted for by the regression. Similarly, the fraction

$$\frac{s_y^2 - s_r^2}{s_y^2} 100\% = r^2 100\%$$

is the percentage of the total variance in premolt size, that is explained by the regression. This percentage is called the **coefficient of determination**.

The **coefficient of determination** for a linear regression of a response variable y on an explanatory variable x is computed by

$$CD = r^2 \quad (0 \leq CD \leq 1)$$

where r is the correlation coefficient. The CD is often expressed in %. The CD may be interpreted as the proportion/percentage of the total variance in the response that is explained by a regression on the explanatory variable.

Exercise 3.10: For the crab data (**crab1.dat**), assess the coefficient of determination ($r = 0.99$, $s_y = 14.67$). Compute the residual standard deviation (s_r) and variance (s_r^2).

Exercise 3.11: Generate residuals for the regression of premolt size on postmolt size for the crab data (**crab1.dat**) and draw a histogram.

SAS hints

If the regression is done with PROC GLM instead of PROC REG, residuals can be output into a dataset "RES" as follows:

```
proc glm data=crabs;
model presz=postsz;
output out=res residual=residual;
run;

proc print data=res;
run;
```

A histogram of the residuals in the dataset RES may be plotted as follows:

```
proc univariate data=res;
var residual;
histogram residual;
run;
```

3.5 The normal distribution

To assess model fit of a linear regression, one can study the distribution of the residuals. The histogram for the residuals in Fig. 3.7 is reasonably symmetric and unimodal. Often the empirical distribution of a variable (here the residuals) can be approximated by a normal distribution. The standard normal curve is depicted in Fig. 3.8. Its shape is rather similar to the histogram in Fig. 3.7. The normal curve is unimodal and symmetric around 0. The following can be stated about the standard normal distribution:

68% of the area under the curve is within 1 unit of its center

95% of the area under the curve is within 2 units of its center

99.7% of the area under the curve is within 3 units of its center

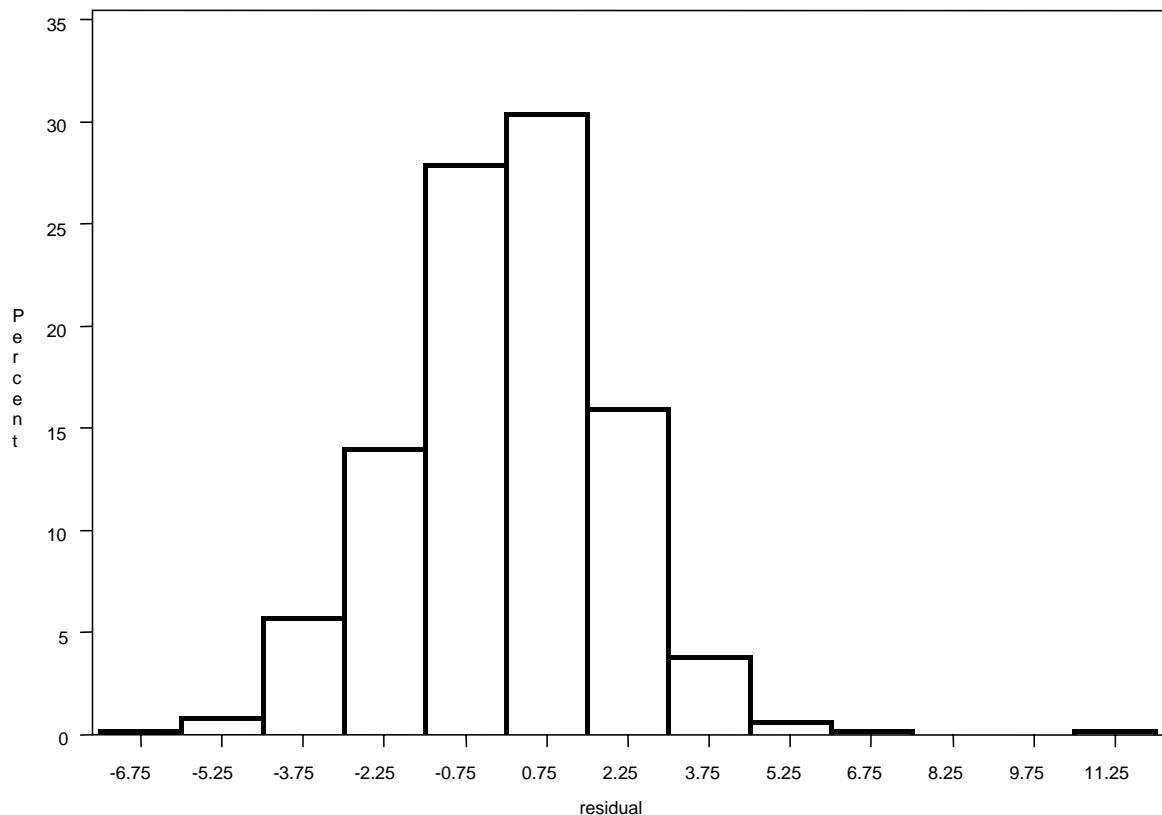


Fig. 3.7: Histogram of residuals r_i for regression of premolt size on postmolt size for crab data (**crab1.dat**).

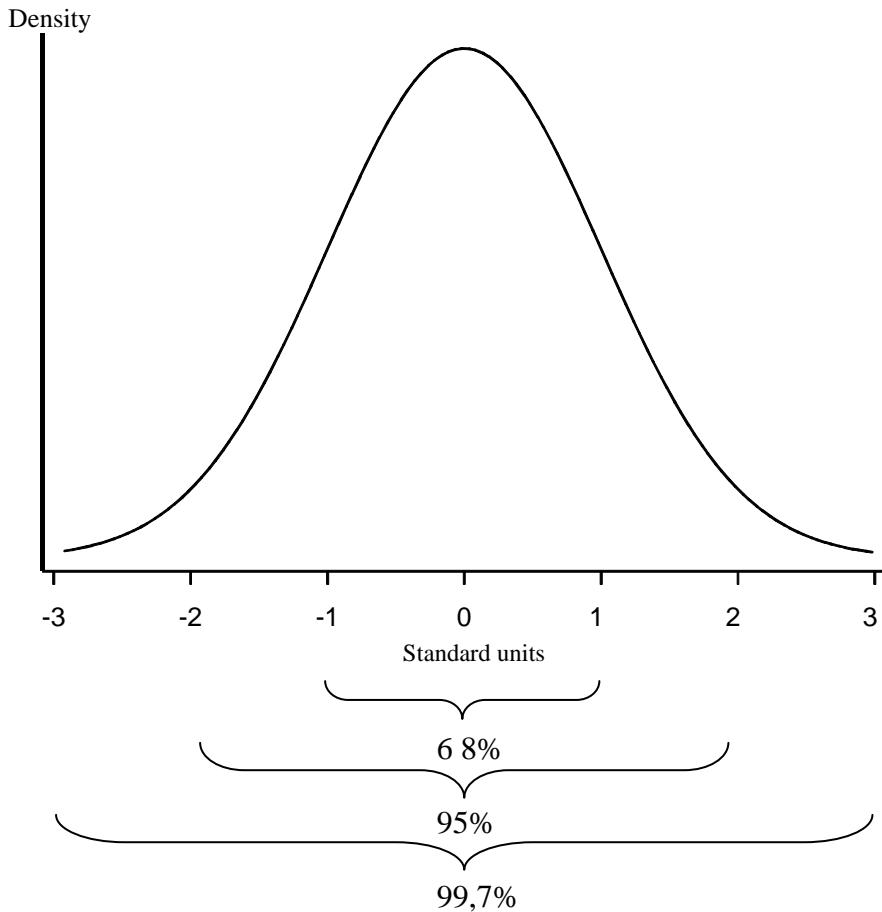


Fig. 3.8: The standard normal curve (density).

These areas and others are determined from the following analytic expression for the curve:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

Traditionally, $\Phi(z)$ represents the area under the normal curve to the left of z , namely,

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

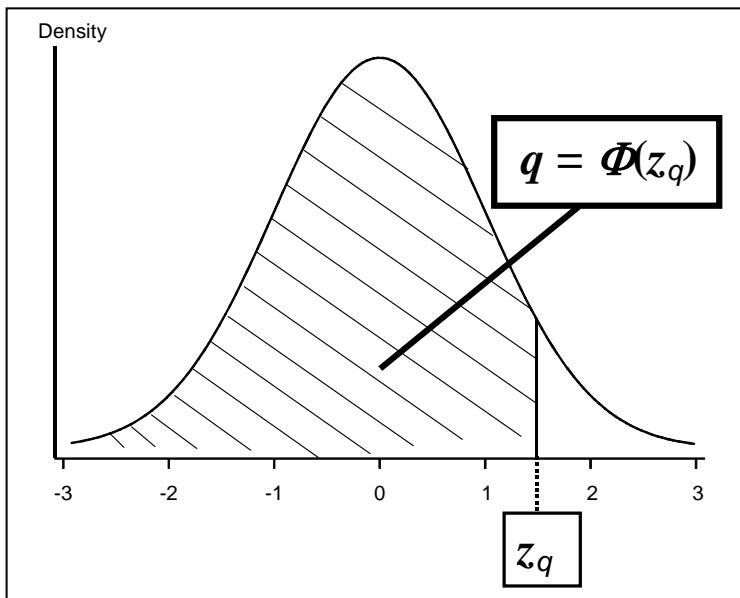
This area needs to be computed numerically, since the integral expression cannot be solved analytically. Tables 3.1 gives a few typical values for $\Phi(z)$ and z . For a given value of the probability $\Phi(z)$, z is referred to as a **quantile**. For example, if 2.5% of the area fall to the left of z , then z is the 2.5% quantile.

Tab. 3.1: Quantiles of the standard normal distribution. Values of z , so that $q = \Phi(z)$.

Probability quantile

$$q = \Phi(z_q) \quad z_q$$

0.600	0.25335
0.700	0.52440
0.800	0.84162
0.900	1.28155
0.950	1.64485
0.975	1.95996
0.990	2.32635
0.995	2.57583
0.84134	1.00
0.97725	2.00
0.99865	3.00



To compare the distribution of residuals to the standard normal distribution, it is necessary to standardize the residuals. The standardization is effected by subtracting the mean and dividing by the standard deviation. The mean of all residuals can be shown to equal 0 (you may check this algebraically!). Thus the standardized residuals are given by

$$z_i = \frac{r_i - 0}{s_r} = \frac{r_i}{s_r}$$

The histogram for the standardized residuals in Fig. 3.9 roughly resembles the standard normal distribution (Fig. 3.8). To gain further insight, one may compute the percentage of standardized residuals falling within limits ± 1 , ± 2 and ± 3 . These percentages (Table 3.2) are rather close to the percentages expected for the standard normal distribution. Thus, the residuals at least approximately follow a normal distribution.

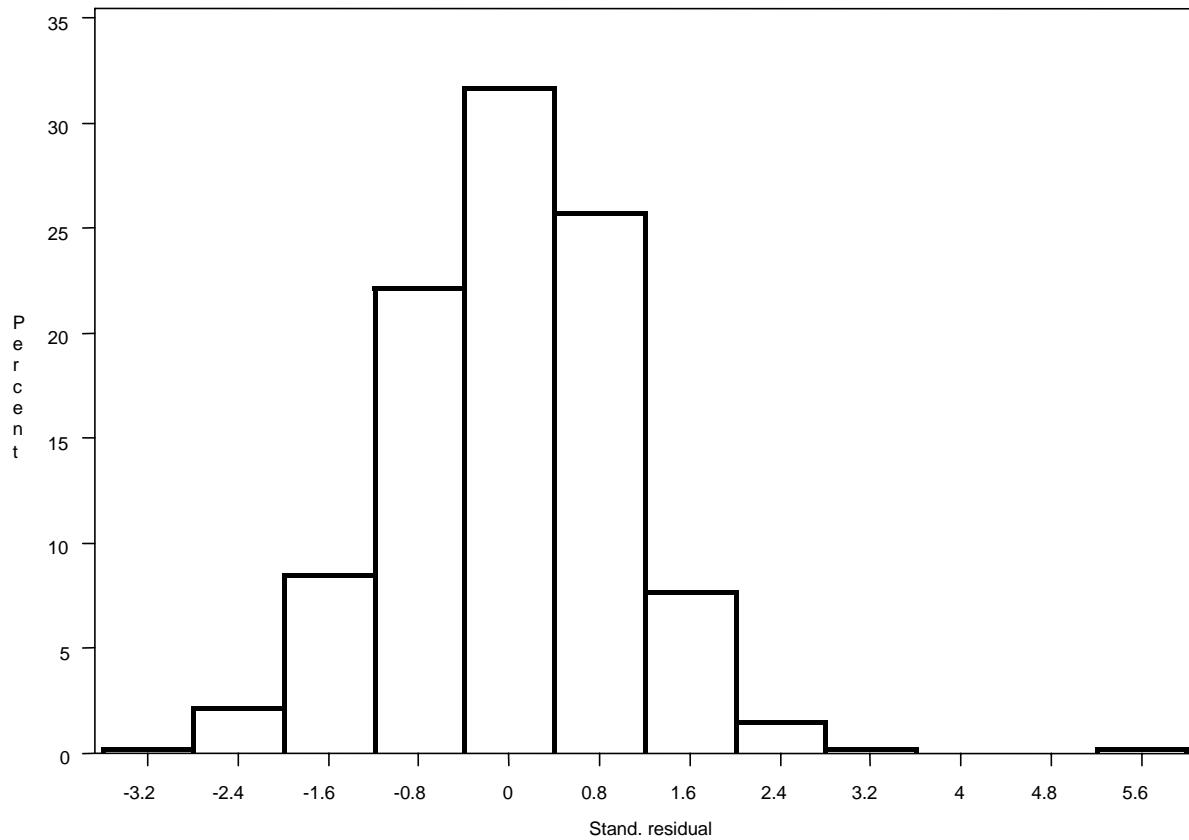


Fig. 3.9: Histogram of standardized residuals $z_i = r_i/s_r$ for regression of premolt size on postmolt size for crab data (**crab1.dat**).

Table 3.2: Expected and observed percentage of standardized residuals $z_i = r_i/s_r$ falling within specified limits.

Limits	Percentage of standardized residuals falling within limits	
	Observed	Expected
-3 to 3	99.4%	99.7%
-2 to 2	96%	95%
-1 to 1	71%	68%

3.6 Quantile plots

Comparison to a normal distribution can be carried further by looking at the quantiles. For the standard normal distribution, the q -th quantile is z_q , where

$$\Phi(z_q) = q ; \quad 0 < q < 1$$

For the residuals r_1, \dots, r_n , the sample quantiles are found by ordering the residuals from smallest to largest. We denote this ordering by $r_{(1)}, \dots, r_{(n)}$. Then $r_{(k)}$ is considered as the $k/(n+1)$ -th sample quantile. We divide by $n + 1$ rather than n to keep q less than 1.

The **normal-quantile plot**, also known as the **normal-probability plot** or **quantile-quantile plot (Q-Q-plot)**, provides a graphical means of comparing the data (residual) distribution to the normal. In our case, it graphs pairs $(z_{k/(n+1)}, r_{(k)})$. If the plotted points fall roughly on a line, then this indicates that the residuals have an approximate normal distribution. It is immaterial whether we plot ordered residuals $(r_{(i)})$ or ordered standardized residuals $(z_{(i)})$, since both differ just by a scaling factor. In either case, points should approximately fall on a line under normality. Fig. 3.10 shows the Q-Q-plot for residuals r_i . Except for one outlying observation, the points fall on a line, so the normality assumption is quite reasonable.

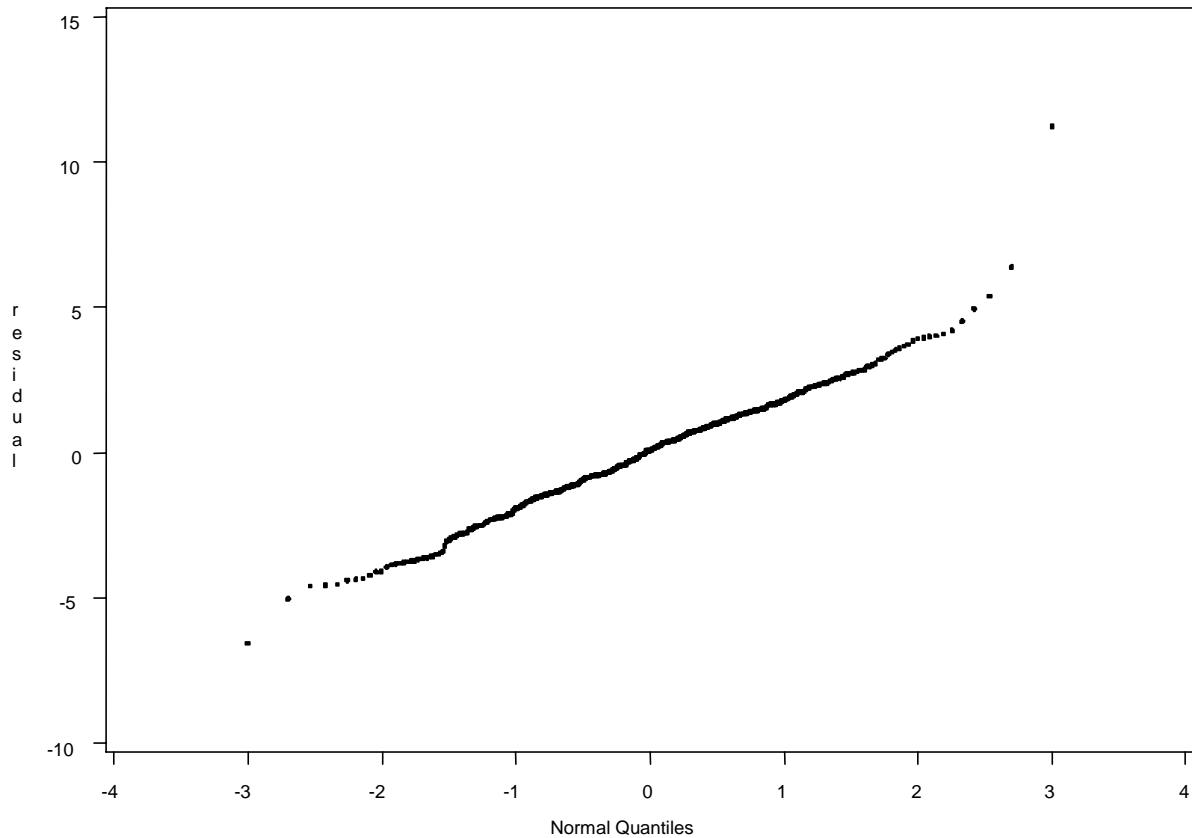


Fig. 3.10: Q-Q-plot of the residuals from the least squares fit of premolt size to postmolt size for the crab data (**crab1.dat**).

For the crab data, the standard deviation of the residuals equals $s_r = 2.00$ mm, so about 68% of the crabs' premolt sizes are within 2 mm of the regression line. This means that if we draw two lines parallel to the regression line, one 2 mm above the regression line and the other 2 mm below it, then roughly 68% of the points in the scatter plot would be expected to fall between these two lines. Similarly, if we draw the lines ± 4 mm off the regression line, 95% of the points would be expected to fall between the two lines. Such lines delineate what can be referred to as a **prediction band**. More exact calculations not reproduced here show that exact prediction bands are slightly bent away from the line, fanning out towards the ends. For the case at hand, the exact delimiting curves are very similar to a straight line. Fig. 3.11 shows 95% prediction band for the crab data. 21 of 370 points (4,5%) fall outside the band (not all of these can be detected easily by eye in Fig. 3.12, because some are very close to one of the lines bordering the band!), while 349 fall inside (95,5%). These percentages agree well with the expected values of 5% and 95%.

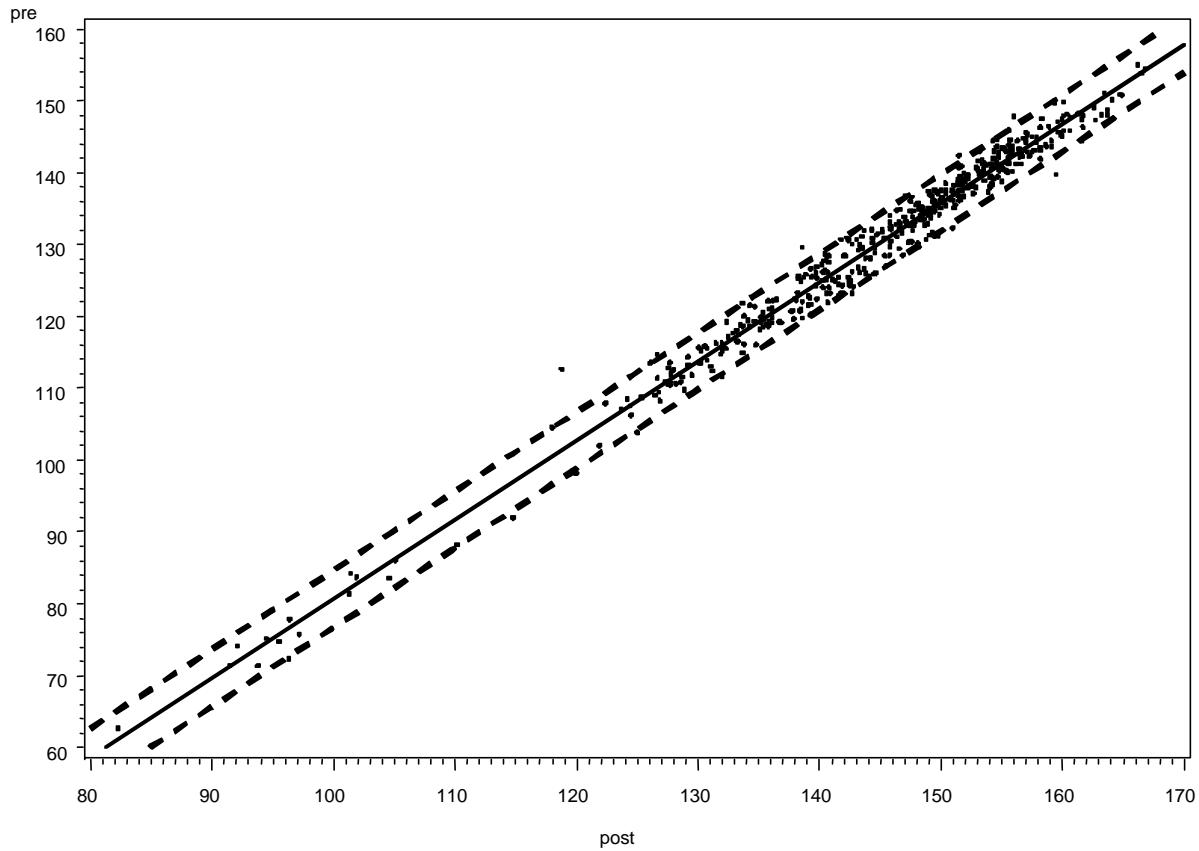


Fig. 3.11: 95% prediction band for the least squares fit of premolt size to postmolt size for the crab data (**crab1.dat**).

Exercise 3.12: Generate a Q-Q-plot for the residuals of the regression of premolt size on postmolt size for the crab data (**crab1.dat**). Produce 95% prediction bands.

SAS hint

QQ-plot

```
proc univariate data=res;
var residual;
qqplot residual/normal;
run;
```

Prediction bands

```
symbol i=rlcli value=dot;
proc gplot;
plot presz*postsz;
run;
```

Creates prediction bands

3.7 Answering the original research question

The original research objective was to predict the distribution of premolt sizes for the second crab data (**crab2.dat**) using the regression function estimated from the first data set (**crab1.dat**). This estimated distribution of premolt sizes can then be used to derive size restrictions for catching female Dungeness crabs at the Pacific coast of California and

Oregon. The estimated distribution alone does not give all the answers, but it is one useful building block for deriving effective size restrictions. For example, we may estimate what proportion of female crabs will remain, if all crabs larger than a threshold size (in terms of the carapace/shell diameter in mm) are caught.

It is tempting to use the first data set for estimating the distribution of premolt sizes. After all, premolt sizes are available only for this first data set, but not for the second. The problem is that the crabs in the first data set do not form a representative sample of crabs at the Pacific coast. This follows from the way in which the crabs were trapped. Recall, e.g., that part of the crabs were brought by commercial fisheries. Commercial traps have netting designed to catch the larger male crabs; female crabs caught with these traps were typically larger than 155 mm. Thus, crabs smaller than 155 mm are probably underrepresented. By contrast, the sampling for the second data set was such, that the postmolt size distribution well represents the population of crabs at the Pacific coast.

Our rather detailed regression analysis has shown that predictions of premolt size from postmolt size are quite accurate, though not perfect. Thus, it seems safe to use the regression for estimating the premolt size distribution. The estimated regression for the first data set (**crab1.dat**) is

$$\text{premolt} = -29.68 + 1.10 \times \text{postmolt}$$

The first 5 observations of postmolt size in **crab2.dat** for crabs that molted (shell = 1) are:

postmolt

116.8
117.1
118.4
119.6
120.1

Plugging the first observed postmolt size into the regression equation, we find the predicted value

$$\text{premolt} = -29.68 + 1.10 \times 116.8 = 98.8$$

For the first five values the predictions are:

postmolt	premolt
116.8	98.80
117.1	99.13
118.4	100.56
119.6	101.88
120.1	102.43

A large number of crabs did not molt in the current molting season (shell = 0). For these crabs, a reasonable assumption is that sizes did not change compared to the end of the premolting period, which precedes the time of measurement by just a few weeks. This assumption is reasonable because once the shells have hardened, which happens shortly after molting, they will not grow anymore.

Thus, to obtain a prediction of the premolting size distribution for the whole population, crabs are treated differently, depending on whether or not they molted in the current molting season:

Crabs that molted : predict premolt size by regression
 Crabs that did not molt: assume the size was the same before and after the molting season

The computations can be implemented in a SAS datastep using an IF statement as follows (using the data in **crab2.dat**):

```
data crabs2;
input size shell;
if shell=1 then pre=-29.68+1.1*size; else pre=size;
datalines;
116.8      1
117.1      1
<more data>
166.6      0
168.0      0
;
proc print; run;
```

Plotting the histogram:

```
proc gchart data=crabs2;
vbar pre /subgroup=shell space=0;
run;
```

The predicted histogram is shown in Fig. 3.13. Comparing this to the postmolt distribution in Fig. 3.2, we see that there is a major shift towards smaller sizes. The predicted distribution can be used to estimate the proportion of female crabs remaining if a given size restriction is imposed, i.e., if crabs are trapped only if their shell size is larger than some specified threshold t . To estimate this proportion, we may introduce a dummy variable d , which equals 1, if premolt $< t$ and which equals 0 otherwise. The mean of d equals the estimated proportion of crabs remaining, if crabs smaller t are not trapped and assuming that all crabs larger than t are caught. We may attach a confidence interval using the methods from Section 1. Here, the total population size N is unknown. Assuming the population size N is very large and large relative to the sample size n , we may set N equal to infinity. For example, the fraction remaining with a threshold of $t = 140$ mm is estimated as follows:

```
data crabs2;
input size shell;
if shell=1 then pre=-29.68+1.1*size; else pre=size;
t=140;
if pre < t then d = 1; else d = 0;
datalines;
116.8      1
117.1      1
<more data>
166.6      0
168.0      0
;

proc surveymeans data=crabs2 mean clm;
var d;
run;
```

We obtain an estimate fraction of 0.49, with 95 confidence limits given by 0.44 and 0.54. Thus, roughly half the female crabs will remain, if the catch size is restricted to be larger than $t = 140$ mm.

Exercise 3.13: Estimate the proportion of female crabs remaining for different thresholds t . What threshold should be used to make sure that 20% of the female crab population remains?

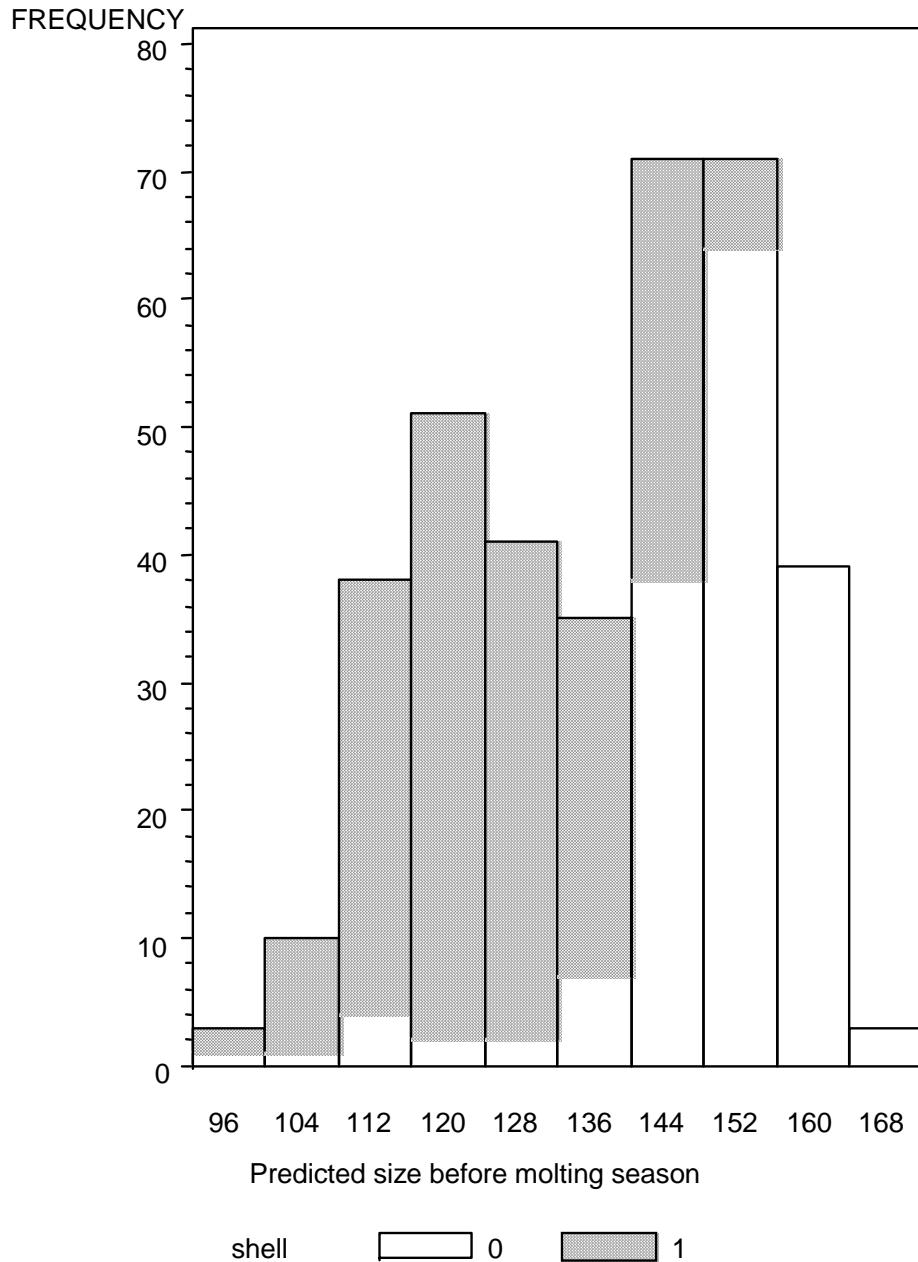


Fig. 3.13: Predicted size distribution of 362 adult female Dungeness crabs shortly before the 1983 molting season. Shell = 0: crabs did not molt. Shell = 1: Crabs molted. Numbers on the abscissa are class means (**crab2.dat**).

Final comment: Looking, as we did in this class, at the fraction of crabs falling above a threshold is only a first step in developing size restrictions for crab fishing. The complete procedure is described in the papers of Mohr and Hankin cited in Nolan and Speed (2000). Their procedure is more complex. They are aiming to set a size restriction that will ensure that the females have the opportunity to mature and mate for a few seasons. To do this they need

to figure out the age*shell distribution. Their first step is to determine the probability that a crab will molt given its shell size (premolt). From there they attempt to determine the age of crabs that are a particular size (Deb Nolan, personal communication, April 2002).

Exercise 3.14: In predicting premolt sizes, we have ignored prediction error. To account for prediction error, we may add a random number to the prediction from the regression. Specifically, we may add a simulated random variable, that follows a normal distribution with standard deviation equal to $s_r = 2.00$, the standard deviation of the residuals! In a SAS datastep, we can use the NORMAL function to generate standard normal deviates. Multiplying a standard normal deviate by s_r , we obtain a random normal deviate with zero mean and standard deviation s_r . Generate such normal deviates from within a SAS datastep as follows (example for $t = 140$):

```
data crabs2;
input size shell;
if shell=1 then pre=-29.68+1.1*size + 2*normal(-1); else pre=size;
t=140;
if pre < t then d=1; else d = 0;
datalines;
```

The negative argument to the NORMAL function prompts SAS to use the computer clock to find a seed (starting value) for the random number generator.

Repeatedly use this datastep to simulate the premolt size distribution. Each time, generate a histogram and compute the confidence interval for the proportion of crabs remaining with a size threshold $t = 140$. Compare the results of these simulations to the result obtained without adding a simulated normal deviate.

Exercise 3.15: In computing the regression for the first data set (**crab1.dat**), we pooled data from different years and from different sources (field, lab). Could there be a problem with this approach? How can we check if pooling is adequate, i.e. computation is a single regression line?

SAS hint

Plot regressions separately for the YEAR variable using the following code:

```
symbol value=dot i=r1;
proc gplot data=crabs;
plot presz*postsz=year;
run;
```

Exercise 3.16: Develop a prediction equation based on a regression of the size increment (postmolt - premolt) on postmolt size. Estimate this using PROC REG. Compare the resulting equation to that obtained by regressing premolt on postmolt size. Do you see a link between the two regressions? Hint: check algebraically, how the regression of increment on postmolt can be derived from the regression of premolt on postmolt size.

Exercise 3.17: The dataset **rain.dat** contains yields of wheat (dt/ha) from 26 consecutive years, together with the amount of rain from April to June (in mm) for each year. Compute the correlation between yield and rain. Develop an equation to predict yield from the rain from April to June. How accurate is the prediction? Does a linear model make sense? Consider

transforming the data. For example, plot the inverse of yield versus the inverse of rain and compare the coefficient of determination to the plot of the untransformed data. *SAS hint:* To compute the inverse of a variable Y in a datastep, use the following line inside the datastep:

```
INV_Y = 1/Y;
```

This generates a new variable INV_Y holding the inverse of Y.

Exercise 3.18: Pulp is a plant-based material for making paper (wook, cotton, etc.). Table 3.3 gives pulp prices and shipments recorded over a period of time (data taken from: Makridakis S, Wheelwright SC, Hyndman RJ 1998 Forecasting. Methods and applications. Third edition. Wiley, New York). Fit a regression to predict pulp shipments from the current world market price. The data are stored in **pulp.dat**.

Table 3.3: World pulp prices and shipments (pulp is a plant-based material for making paper).

Pulp shipments (millions metric tone)	World pulp price (dollars per ton)
10.44	792.32
11.40	868.00
11.08	801.09
11.70	715.87
12.74	723.36
14.01	748.32
15.11	765.37
15.26	755.32
15.55	749.41
16.81	713.54
18.21	685.18
19.42	677.31
20.18	644.59
21.40	619.71
23.63	645.83
24.96	641.95
26.58	611.97
27.57	587.82
30.38	518.01
33.07	513.24
33.81	577.41
33.19	569.10
35.15	516.75
27.45	612.18
13.96	831.04

4. Linear models

Example 4.1: (Backhaus et al. 2000 Multivariate Analysemethoden. Springer, Berlin). A market researcher of a company that produces margarine wants to investigate, which factors influence the margarine sales. Thus, she randomly selects a sample of 37 sales areas (geographical regions) and collects data on a number of variables, which may have an effect on sales and which are under the control of the company:

Sales	Quantity of margarine sold (boxes per area)
Price	Price of margarine (DM/box)
Advertise	Expenditure for advertisements (DM/area)
Visits	Number of visits by sales agents

The researcher wants to know, how sales are influenced by these variables. An equation is needed to predict sales from the explanatory variables. The data are stored in **margarine.dat**.

Table 4.0: Four sample observations for margarine data.

sales	price	advertisement	visits
2585	12.5	2000	109
1819	10	550	107
1647	9.95	1000	99
1496	11.5	800	70

One way of approaching the problem is by linear regression. One can do simple linear regressions in turn, using sales as the response and one of the other variables at a time as an explanatory variable. However, since there are several explanatory variables, one can regress the response (sales) on all three explanatory variables simultaneously:

$$\text{Sales} = \alpha + \beta_1 \times \text{price} + \beta_2 \times \text{advertise} + \beta_3 \times \text{visits}$$

This model is a multiple linear regression model, a special case of the **general linear model** (GLM). Simple linear regression considered in the previous chapter is a particularly simple case of a GLM. This chapter considers various models, which all fall into the class of GLMs.

Example 4.2 (Nolan and Speed, 2000, Chapter 10): The Child Care Health and Development Studies (CHDS) is a large study in the US to investigate factors affecting child health. In this example, we look at factors affecting the birth weight of children, using a subset of data covering 1236 male single births where the baby lived at least 28 days. The data comprise the following variables (**babies.dat**; Table 4.1):

Variable	Description
bwt	Birth weight in ounces (999 unknown)
gestation	Length of pregnancy in days (999 unknown)
parity	0= first born, 1 = not first born, 9=unknown
age	mother's age in years
height	mother's height in inches (999 unknown)
weight	Mother's prepregnancy weight in pounds (99 unknown)
smoke	Smoking status of mother 0=not now, 1=yes now, 9=unknown

It was found that the babies born to women who smoked during their pregnancy tended to weigh less than those born to women who did not smoke. But smokers may differ from nonsmokers in some essential ways that may affect the birth weight of the baby, no matter whether or not the mother smoked. The 1989 Surgeon General's Report contains three assertions:

- Mothers who smoke have increased rates of premature delivery (baby born too early).
- The newborns of smokers are smaller at every gestational age (length of pregnancy).
- Smoking seems to be a more significant determinant of birth weight than the mother's pregnancy height and weight or parity.

We will investigate each of these statements using various linear models.

Table 4.1: Sample observations and data description for the 1236 babies in the Child Health and Development Studies subset.

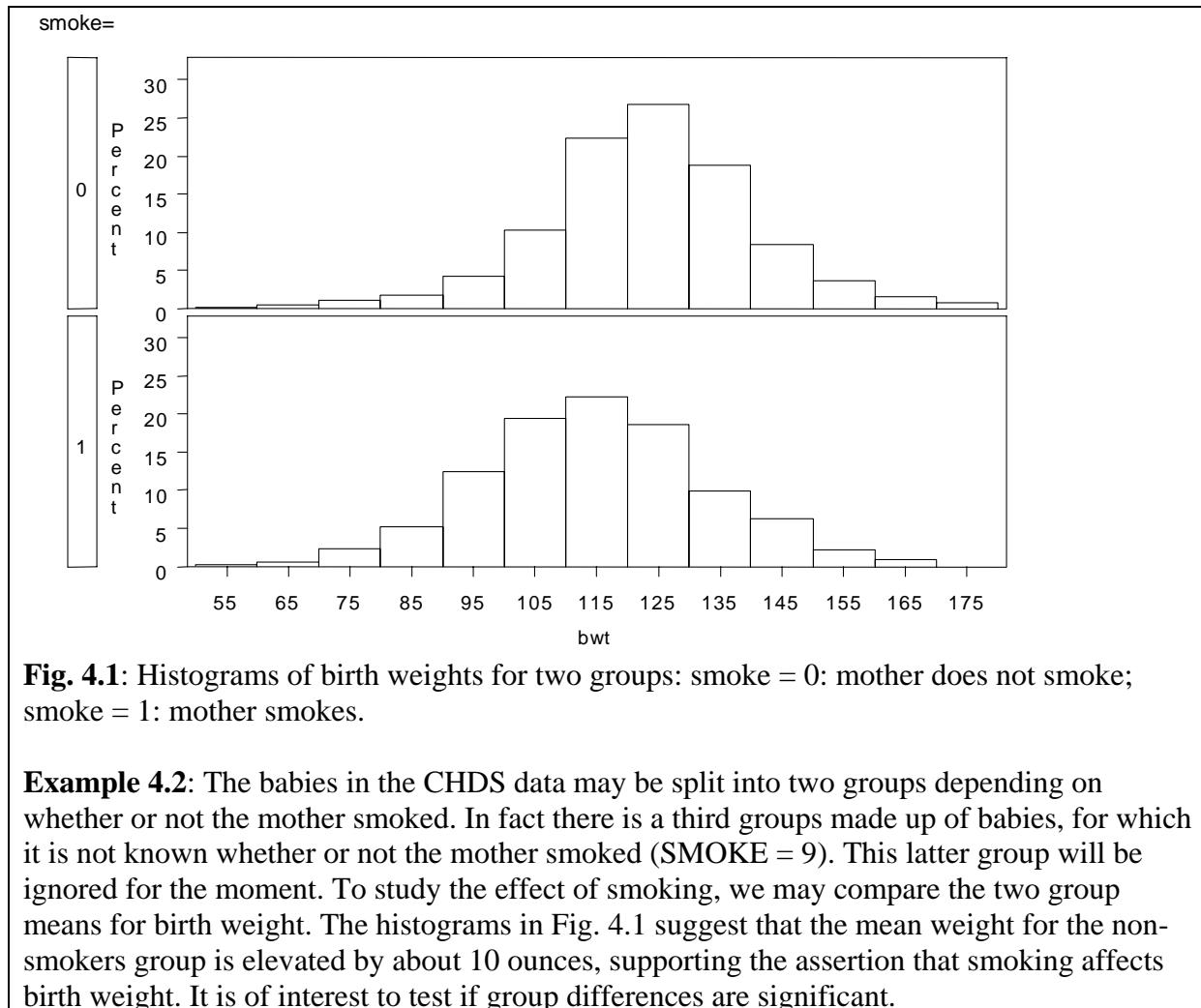
Birth weight	120	113	128	123	108	136	138
Gestation	284	282	279	NA	282	286	244
Parity	1	0	1	0	1	0	0
Age	27	33	28	36	23	25	33
Height	62	64	64	69	67	62	62
Weight	100	135	115	190	125	93	178
Smoking status	0	0	1	1	1	1	0

Example 4.3 (Mead et al., p. 52; **melons.dat**): An experiment was performed to compare four melon varieties. Each variety was tested on six field plots. The allocation of treatments (varieties) to experimental units (plots) was completely at random (**completely randomized design**; see Chapt. 5). The yields were as follows:

Variety	A	B	C	D
Yields	25.12	40.25	18.30	28.55
	17.25	35.25	22.60	28.05
	26.42	31.98	25.90	33.20
	16.08	36.52	15.05	31.68
	22.15	43.32	11.42	30.32
	15.92	37.10	23.68	27.58
Mean (\bar{y}):	20.49	37.40	19.49	29.90

The objective of the analysis is to compare the six variety means.

4.1 Comparing two groups (one-way ANOVA)



You may be familiar with the simple t-test for comparing two independent group means (see Mead et al., 1993 and end of this section), which is appropriate to this task. Here, we will cast the problem into that of a comparison of two linear models, one of which is a special case of the other. This approach is quite generally applicable in the analysis of linear models, as will become apparent throughout this chapter.

If the two groups have a common mean μ (expected value, population mean), the observed data may be expressed as

$$y_{ij} = \mu + e_{ij} \quad (4.1)$$

where

- y_{ij} = weight of j -th baby in i -th group
 - ($i = 1$ for babies from mothers who do not smoke; SMOKE = 0;
 - $i = 2$ for babies from mothers who smoke; SMOKE = 1)
- μ = common mean
- e_{ij} = random deviation from the expected value (μ), assumed to follow a normal distribution with mean zero and variance σ^2 .

If, on the other hand, the groups differ in their means, the model is

$$y_{ij} = \mu_i + e_{ij} \quad (4.2a)$$

where

$$\mu_i = i\text{-th group mean} \quad (i = 1, 2)$$

The group mean in the second model may be expressed as

$$\mu_i = \mu + \alpha_i$$

where

$$\alpha_i = \text{effect of } i\text{-th group}$$

With this, the model becomes

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad (4.2b)$$

This alternative formulation in (4.2b) shows that model (4.1), which asserts that both groups have the same mean, is a special case of (4.2a/4.2b), which holds when group means differ. Specifically, (4.1) is obtained from (4.2b) by dropping the group effect (α_i). We say that (4.1) is a **reduced model**, while (4.2b) is the corresponding **full model**, which contains the full set of parameters (μ, α_i). The reduced model is said to be **nested** within the full model, because the former is a special case of the latter.

Both models can be estimated by the method of least squares. Specifically, the reduced model is estimated by minimizing

$$SS(\mu) = \sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \mu)^2$$

where n_i is the number of observations for the i -th group. This yields the least squares estimator of μ ,

$$\hat{\mu} = \bar{y}_{\bullet\bullet} = \frac{\sum_{i=1}^2 \sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^2 n_i}$$

which is just the simple mean of all observations. The minimized SS is given by

$$SS(\hat{\mu}) = \sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu})^2 = \sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\bullet\bullet})^2$$

For the full model, we minimize

$$SS(\mu, \alpha_i) = \sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2$$

which yields

$$\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i = \bar{y}_{i\bullet} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i},$$

the simple group mean. The minimized SS is

$$SS(\hat{\mu}, \hat{\alpha}_i) = \sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\alpha}_i)^2 = \sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2$$

Imposing the restriction $\alpha_1 + \alpha_2 = 0$, the least squares solution for the effects is

$$\begin{aligned}\hat{\mu} &= \bar{y}_{\bullet\bullet} = \frac{\bar{y}_{1\bullet} + \bar{y}_{2\bullet}}{2} \\ \hat{\alpha}_i &= \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\end{aligned}$$

The restriction $\alpha_1 + \alpha_2 = 0$ is just a matter of convenience. Alternatively, we might set $\alpha_2 = 0$, which is the constraint used by SAS procedures such as GLM. Some such restriction is needed to be able to find a least squares solution for the parameters μ and α_i . To see this, note that μ_i can be expressed as the sum of μ and α_i in an infinite number of ways, while still yielding the same values for μ_i . Whichever restriction we impose on the parameters, it will always be true that the estimate of the group mean, μ_i , is equal to the simple sample mean $\bar{y}_{i\bullet}$. Also note that the group mean (μ_i) itself has a direct interpretation, while the parameters (μ , α_i) do not.

Example 4.3: Assume that the birth weight population means for the two groups of babies are

$$\begin{aligned}\mu_1 &= 107 \text{ ounces} \\ \mu_2 &= 95 \text{ ounces}\end{aligned}$$

These two means follow for, e.g.,

$$\mu = 100, \alpha_1 = 7, \alpha_2 = -5$$

or

$$\mu = 95, \alpha_1 = 12, \alpha_2 = 0$$

To resolve this indeterminacy or overparameterisation, some restriction on the parameters is needed. For example, with the restriction $\alpha_1 + \alpha_2 = 0$, we need find

$$\mu = 101, \alpha_1 = 6, \alpha_2 = -6$$

An important fact to remark about the full and reduced model is that the variation about the group means is smaller than or equal to the variation about the overall mean:

$$SS(\hat{\mu}) - SS(\hat{\mu}, \hat{\alpha}_i) \geq 0$$

Thus, there always is a **reduction in SS** for the full model compared to the reduced model. This is true regardless of whether or not the full model holds. This can be shown algebraically in many different ways, which is not done here. To just illustrate the point, consider Examples 4.4 through 4.6.

Example 4.4: (artificial)

Group	Observations	Group mean
1	1, 2, 3	$\hat{\mu}_1 = \hat{\mu} + \hat{\alpha}_1 = 2$
2	21, 22, 23	$\hat{\mu}_2 = \hat{\mu} + \hat{\alpha}_2 = 22$

$$\left. \begin{array}{l} \\ \end{array} \right\} \text{Overall mean: } \hat{\mu} = 12$$

$$\begin{aligned} SS(\hat{\mu}, \hat{\alpha}_i) &= (1-2)^2 + (2-2)^2 + (3-2)^2 + (21-22)^2 + (22-22)^2 + (23-22)^2 = 1+0+1+1+0+1=4 \\ SS(\hat{\mu}) &= (1-12)^2 + (2-12)^2 + (3-12)^2 + (21-12)^2 + (22-12)^2 + (23-12)^2 \\ &= 121 + 100 + 81 + 81 + 100 + 121 = 604 \end{aligned}$$

$$SS(\hat{\mu}) - SS(\hat{\mu}, \hat{\alpha}_i) = 600$$

The reduction in SS for the full model compared to the reduced model is 400. The reduction is so large, because most of the variation in the data occurs between groups (between group means), while the within-group variation is relatively small. So there are marked differences among the SS of the full and reduced models.

Example 4.5: (artificial)

Group	Observations	Group mean
1	1, 2, 3	2
2	3, 4, 5	4

$$\left. \begin{array}{l} \\ \end{array} \right\} \text{Overall mean} = 3$$

$$\begin{aligned} SS(\hat{\mu}, \hat{\alpha}_i) &= (1-2)^2 + (2-2)^2 + (3-2)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 = 1+0+1+1+0+1=4 \\ SS(\hat{\mu}) &= (1-3)^2 + (2-3)^2 + (3-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 = 4+1+0+0+1+4=10 \end{aligned}$$

$$SS(\hat{\mu}) - SS(\hat{\mu}, \hat{\alpha}_i) = 6$$

The reduction in SS for the full model compared to the reduced model is only 6. The reduction is small, because a considerable portion of the variation in the data occurs within groups, while the between-group variation is relatively small.

Example 4.6: (artificial)

Group	Observations	Group mean
1	1, 2, 3	2
2	0, 2, 4	2

$$\left. \begin{array}{c} \\ \end{array} \right\} \text{Overall mean} = 2$$

$$SS(\hat{\mu}, \hat{\alpha}_i) = (1-2)^2 + (2-2)^2 + (3-2)^2 + (0-2)^2 + (2-2)^2 + (4-2)^2 = 1+0+1+4+0+4=10$$

$$SS(\hat{\mu}) = (1-2)^2 + (2-2)^2 + (3-2)^2 + (0-2)^2 + (2-2)^2 + (4-2)^2 = 1+0+1+4+0+4=10$$

$$SS(\hat{\mu}) - SS(\hat{\mu}, \hat{\alpha}_i) = 0$$

The reduction in SS for the full model compared to the reduced model is zero, because the overall mean is equal to the two group means. In other words, the reduction is zero, because there is no between group variation (variation of group means).

Example 4.2: For birth weight in the baby data we find for the grouping variable SMOKE (1 = nonsmokers; 2 = smokers)

$$\bar{y}_{..} = 119.52 \text{ (overall mean)}, \quad SS(\hat{\mu}) = 405,928$$

$$\bar{y}_{1.} = 123.05 \text{ (nonsmokers)}, \quad SS(\hat{\mu}, \hat{\alpha}_i) = 382,529$$

$$\bar{y}_{2.} = 114.11 \text{ (smokers)}$$

$$SS(\hat{\mu}) - SS(\hat{\mu}, \hat{\alpha}_i) = 23,400$$

The reduction in SS is relatively small compared to the total variation about the overall mean. Babies from nonsmokers have an average weight of 9 ounces above that of babies from smokers.

The **reduction in SS** from the reduced to the full model may be ascribed to the addition of the group effect α_i to the reduced model, which just contains μ . This may be expressed by

$$RSS(\hat{\alpha}_i | \hat{\mu}) = SS(\hat{\mu}) - SS(\hat{\mu}, \hat{\alpha}_i)$$

In RSS , the effects listed after the horizontal bar are the effects common to the full and the reduced models. In plain words, $RSS(\hat{\alpha}_i | \hat{\mu})$ may be read as "**the reduction in SS due to fitting α_i , after having fitted the general mean μ .**"

It can be shown that the RSS is itself a sum-of-squares. In the present case, it is of the form

$$RSS(\hat{\alpha}_i | \hat{\mu}) = \sum_{i=1}^2 n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

This expression shows that the RSS becomes large when the group means differ grossly. When the RSS is large relative to the SS for the full model, the group effect is large in

magnitude. Conversely, if the RSS is of about the same order of magnitude as the SS for the full model, group differences are small.

The reduction in SS is associated with a **reduction in the degrees of freedom** (d.f.). The d.f. of an error sum of squares for a particular model equals the number of observations ($n_1 + n_2$), minus the number of (free) parameters. For the reduced model, one parameter needs to be estimated (μ), so the d.f. is $n_1 + n_2 - 1$. For the full model, there are two means to be estimated, so the d.f. is $n_1 + n_2 - 2$. The d.f. for RSS is just the reduction in d.f. from the SS of the full to the SS of the reduced models, i.e. the d.f. = 1.

The sums of squares may be compiled into an **analysis-of-variance** (ANOVA) table. Each sum of squares in the table has associated degrees of freedom (d.f.). In the ANOVA table, a sum of squares is divided by its d.f. to yield a **mean square**. The mean square computed from the SS of the full model serves as an estimate, s^2 , of the population error variance, σ^2 :

$$s^2 = \frac{SS(\hat{\mu}, \hat{\alpha}_i)}{(n_1 + n_2 - 2)}$$

The RSS is compared against this estimate (s^2). The ratio of the two mean squares yields the so-called **F-value**. A large F-value is indicative of group differences, as will be explained below.

Source	Degrees of freedom (d.f.)	Sum of squares (SS)	Mean square (MS)	F_{obs}
Groups	1	$RSS(\hat{\alpha}_i \hat{\mu})$	$RSS(\hat{\alpha}_i \hat{\mu})/1$	$\frac{RSS(\hat{\alpha}_i \hat{\mu})}{s^2}$
Error	$n_1 + n_2 - 2$	$SS(\hat{\mu}, \hat{\alpha}_i)$	$s^2 = \frac{SS(\hat{\mu}, \hat{\alpha}_i)}{(n_1 + n_2 - 2)}$	
Corrected	$n_1 + n_2 - 1$	$SS(\hat{\mu})$		

Example 4.2: For the baby data we find

Source	Degrees of Freedom (d.f.)	Sum of squares (SS)	Mean square (MS)	F_{obs}	p
Groups	1	23400	23400	74.87	<0.0001
Error	1224	382529	$s^2 = 313$		
Corrected	1225	405928			

The ANOVA table is used to test the null hypothesis of equal group means:

$$H_0: \alpha_1 = \alpha_2$$

It can be shown that the mean squares in the ANOVA have the following expectations:

Source	Degrees of freedom (d.f.)	Mean square (MS)	Expected MS E(MS)
Groups	1	$RSS(\hat{\alpha}_i \hat{\mu})$	$\sigma^2 + \sum_{i=1}^2 n_i (\alpha_i - \bar{\alpha}_*)^2$
Error	$n_1 + n_2 - 2$	$s^2 = \frac{SS(\hat{\mu}, \hat{\alpha}_i)}{(n_1 + n_2 - 2)}$	σ^2

Under H_0 , $\sum_{i=1}^2 n_i (\alpha_i - \bar{\alpha}_*)^2 = 0$, so that both MS have the same expectation, and their ratio, i.e., the F -value, is expected to be close to unity. By contrast, when the null is not true, i.e. there are group differences, the group MS is expected to be much larger than the error MS . Thus, H_0 is rejected when F_{exp} is large. To decide how large F_{exp} needs to be in order to reject H_0 , one needs to consider the distribution of F_{exp} under H_0 . Specifically, we may compute the probability of observing an F -value larger than or as large as the one actually observed, **providing H_0 is true**. This probability is commonly referred to as the **p-value**. In the example (Example 4.2),

$$\text{p-value} = P(F_{obs} \geq 74.87) < 0.0001$$

where $P(\cdot)$ indicates the probability. Thus, observing an F -value of 74.87 or larger, when H_0 is true, is very unlikely: the probability is smaller than 0.0001. So H_0 is not plausible and can be rejected.

When the p-value is small, this means that the null hypothesis H_0 is not plausible, so it is rejected. It is customary to reject H_0 when $p < 5\%$.

Thus, for the baby data, we reject the null hypothesis and conclude that the group differences are significant.

Alternatively, F_{obs} may be compared against a tabular F -value (F_{tab}) for an F -distribution with 1 d.f. in the numerator and $n_1 + n_2 - 2$ d.f. in the denominator. H_0 is rejected when $F_{obs} > F_{tab}$. To determine F_{tab} , we need to specify the significance level α of the test. The tabular F -value (F_{tab}) for a specified significance level α is that value of F for which

$$P(F_{obs} > F_{tab}) = \alpha,$$

providing H_0 is true. Thus, the significance level α is the probability of falsely rejecting the null hypothesis, providing H_0 is true. Critical values for $\alpha = 5\%$ may be read from Table 4.II. Note that values of $F_{obs} > F_{tab}$ will automatically have p-values $< \alpha$.

Example 4.2: F_{exp} has 1 numerator d.f. and 1224 denominator d.f. The denominator d.f. is not in the Table 4.II, so we use the value for d.f. = ∞ (infinity). For $\alpha = 5\%$ we find $F_{tab} = 3.84 < F_{obs} = 74.87$, so the null hypothesis is rejected. This result is the same as that found by looking at the p-value. [Generally, the test will yield the same result, no matter whether you use the p-value and reject when $p < \alpha$ or you reject when $F_{obs} > F_{tab}$ at $\alpha = 5\%$].

Exercise 4.1: Reproduce the above ANOVA for the baby data using the SAS procedure GLM.

SAS hints

The missing observations for BWT are coded 999. These need to be replaced by a dot in SAS, which is effected by a simple IF statement. Missing values for the other variables have to be coded by dots in a similar fashion.

```

data;
input
bwt gestation parity age height weight smoke;
if bwt=999 then bwt=.;
if weight=999 then weight=.;
if height=99 then height=.;
if gestation = 999 then gestation = .;
if smoke=9 then smoke=.;
datalines;

<data>

proc glm;
class smoke;
model bwt=smoke;
lsmeans smoke/pdiff;
means smoke;
run;

```

This code produces an ANOVA table and means. The option PDIFF on the LSMEANS statement yields a t-test for the comparison of means. The t-statistic is computed as

$$t_{\text{exp}} = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{s.e(\hat{\mu}_1 - \hat{\mu}_2)} = \frac{|\hat{\alpha}_1 - \hat{\alpha}_2|}{s.e(\hat{\alpha}_1 - \hat{\alpha}_2)}$$

where the standard error is computed using a general method not described here.

For the case at hand, the t-statistic for comparing two means has the following simple form:

$$t_{\text{obs}} = \frac{|\bar{y}_{1\bullet} - \bar{y}_{2\bullet}|}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s is the square root of the error MS of the analysis of variance. This is compared against a t-distribution with error d.f. $(n_1 + n_2 - 2)$. The tabular t-value (t_{tab}) for three different choices of α may be read from Table 4.I. The null hypothesis of equal means is rejected when $t_{\text{obs}} > t_{\text{tab}}$. In the case of two groups the t-test is identical to the F-test, since in this case $F = t^2$.

Exercise 4.2: For birth weight in the baby data we find for the grouping variable SMOKE (0 = nonsmokers; 1 = smokers)

$$s^2 = 313, \quad \bar{y}_{1\bullet} = 123.05 \quad n_1 = 742 \quad (\text{mother does not smoke})$$

$$\bar{y}_{2\bullet} = 114.11 \quad n_2 = 484 \quad (\text{mother smokes})$$

Compute the t -statistic and determine if this exceeds the tabular t -value (t_{tab}) at $\alpha = 5\%$. What can you conclude from the result? Compare this to the p-value computed by SAS for the comparison among means ($p < 0.0001$). Verify that $(t_{obs})^2 = F_{obs} = 74.87$ of the ANOVA. Also, verify that $(t_{tab})^2 = F_{tab} = 3.84$.

4.2 Comparing more than two groups (one-way ANOVA)

Example 4.3: The comparison of the yield means four melon varieties is a problem, in which more than two group means (varieties = groups here) are to be compared. As we will see, we can use basically the same methods as for comparing two groups.

To compare more than two groups, the ANOVA proceeds in the same way as in section 4.1:

$$SS(\mu) = \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \mu)^2 \quad (\text{reduced model})$$

$$SS(\mu, \alpha_i) = \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2 \quad (\text{full model})$$

where g is the number of groups (varieties) and n_i is the number of plots in the i -th group (for the i -th variety). The full model has g means ($\mu_i = \mu + \alpha_i$), while the reduced model has one common mean (μ), so the reduction in d.f., i.e. the d.f. of RSS, is $(g-1)$. The ANOVA-table is computed as follows:

Source	Degrees of freedom (d.f.)	Sum of squares (SS)	Mean square (MS)	F_{obs}
Groups (varieties)	$g - 1$	$RSS(\hat{\alpha}_i \hat{\mu})$	$\frac{RSS(\hat{\alpha}_i \hat{\mu})}{g - 1}$	$\frac{RSS(\hat{\alpha}_i \hat{\mu})/(g - 1)}{s^2}$
Error	$\sum_{i=1}^g (n_i - 1)$	$SS(\hat{\mu}, \hat{\alpha}_i)$	$s^2 = \frac{SS(\hat{\mu}, \hat{\alpha}_i)}{\sum_{i=1}^g (n_i - 1)}$	
Corrected total	$\left(\sum_{i=1}^g n_i \right) - 1$	$SS(\hat{\mu})$		

where

$$RSS(\hat{\alpha}_i | \hat{\mu}) = SS(\hat{\mu}) - SS(\hat{\mu}, \hat{\alpha}_i)$$

The test statistic F_{obs} is used to test

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_g$$

F_{obs} needs to be compared against an F-distribution with $g - 1$ numerator d.f. and $\sum_{i=1}^g (n_i - 1)$ denominator d.f. When H_0 is rejected, it is useful to look at pairwise comparisons of group means by t-tests.

Exercise 4.2: Do an analysis of variance to compare the four melon varieties of Example 4.3 (**melons.dat**). Verify using SAS PROC GLM that the ANOVA table reads:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1291.477146	430.492382	23.42	<.0001
Error	20	367.653150	18.382658		
Corrected Total	23	1659.130296			

Compare group means by a t-test.

SAS hints

You can use exactly the same GLM code as in the two-group case! Multiple t-tests are obtained by the LSMEANS SMOKE /PDIFF statement. The variety is coded using character-valued (alphanumeric) symbols (letters). Therefore, the variable name needs to be followed by a \$-sign in the input statement. In Chapter 5, Section 5.6, the example will be taken up again, introducing the familiar letter or lines display for balanced data.

```
data;
input
yield      variety$;
datalines;
25.12      v1
17.25      v1
26.42      v1
16.08      v1
22.15      v1
15.92      v1
40.25      v2
35.25      v2
31.98      v2
36.52      v2
43.32      v2
37.10      v2
18.30      v3
22.60      v3
25.90      v3
15.05      v3
11.42      v3
23.68      v3
28.55      v4
28.05      v4
33.20      v4
```

```

31.68      v4
30.32      v4
27.58      v4
;
proc glm;
class variety;
model yield=variety;
lsmeans variety/pdiff;
run;

```

4.3 Linear regression

Example 4.2: ANOVA is appropriate to study the effect on birth weight (BWT) of **categorical** explanatory variables such as the grouping variable SMOKE, which had just two categories (0 and 1). If the effect of a **quantitative** explanatory variable such as the mother's WEIGHT on BWT is to be studied, linear regression as considered in Chapter 3 is appropriate. A plot of BWT versus WEIGHT (Fig. 4.2) and the fitted line suggests that there is a slight upward trend, i.e. high birth weight tends to be associated with high weight of the mother. The slope equals 0.13, so an increase of the mother's weight by one pound is associated with an expected increase of the baby's birth weight by 0.13 ounces. This prediction is far from perfect due to the wide scatter around the regression line. Also, it is difficult to tell, if this trend is, in fact, significant. The scatter of point is hard to distinguish from random noise. A statistical test can help to clarify the situation.

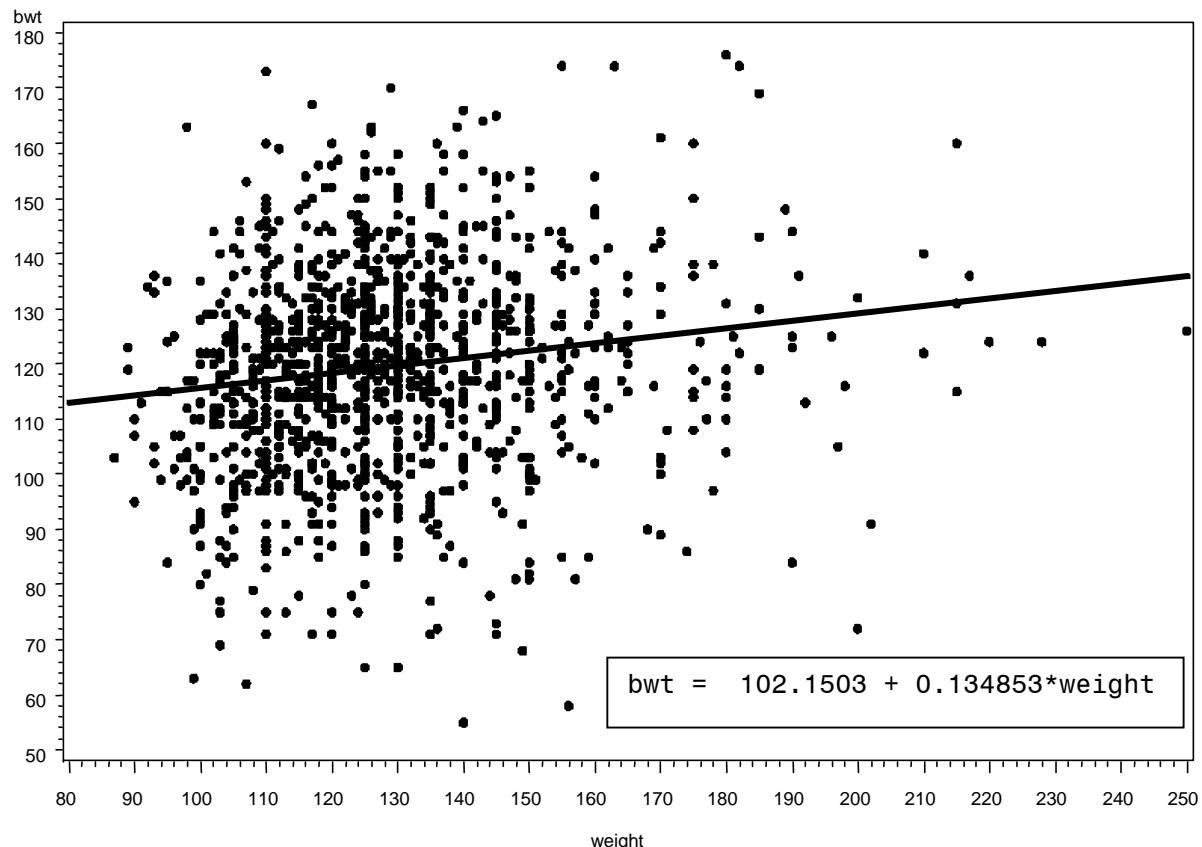


Fig. 4.2: Plot of BWT (ounces) vs. WEIGHT (pounds) for the baby data with fitted regression line.

In linear regression we are fitting the model

$$y_j = \alpha + \beta x_j + e_j$$

where

y_j = j -th observation of response (BWT)

x_j = j -th observation for explanatory variable (WEIGHT)

α = intercept

β = slope

e_j = random deviation ("error") of y_j from the true regression line $\alpha + \beta x_j$. Errors are assumed to follow a normal distribution with zero mean and variance σ^2 .

In the present case it is of interest to test

$$H_0: \beta = 0, \quad (\text{mother's weight is not associated with birth weight})$$

which implies that the model can be reduced to

$$y_j = \alpha + e_j$$

To test this null hypothesis, we can fit, by the method of least squares, a full model with the regression slope and the reduced model without the regression term. The associated error SS are:

$$SS(\alpha) = \sum_{j=1}^n (y_j - \alpha)^2 \quad (\text{reduced model})$$

$$SS(\alpha, \beta) = \sum_{j=1}^n (y_j - \alpha - \beta x_j)^2 \quad (\text{full model})$$

where n is the number of observations. There are explicit computational formulae for these SS, but these shall not concern us here, because in most applications the calculations will be done using a computer. The reduction in error SS between the two models will provide an SS that can be attributed to the regression coefficient β :

$$RSS(\beta | \alpha) = SS(\alpha) - SS(\alpha, \beta)$$

The full and the reduced model differ by one parameter, so the RSS has one d.f. The residual d.f. equals $n-2$, since the full model has two parameters. Again, the error variance is estimated from the SS of the full model:

$$s^2 = \frac{SS(\hat{\alpha}, \hat{\beta})}{n - 2}$$

The SS can be assembled into an ANOVA table as usual:

Source	Degrees of freedom (d.f.)	Sum of squares (SS)	Mean square (MS)	F_{obs}
Slope	1	$RSS(\hat{\beta} \hat{\alpha})$	$RSS(\hat{\beta} \hat{\alpha})/1$	$\frac{RSS(\hat{\beta} \hat{\alpha})/1}{s^2}$
Error	$n - 2$	$SS(\hat{\alpha}, \hat{\beta})$	$s^2 = \frac{SS(\hat{\alpha}, \hat{\beta})}{n - 2}$	
Corrected	$n - 1$	$SS(\hat{\alpha})$		

where

$$RSS(\hat{\beta} | \hat{\alpha}) = SS(\hat{\alpha}) - SS(\hat{\alpha}, \hat{\beta})$$

To test

$$H_0: \beta = 0$$

F_{obs} is compared to an F-distribution with 1 numerator d.f. and $n-2$ denominator d.f. Note that the error MS provides an estimate s^2 of the error variance σ^2 .

Example 4.2: For the regression of BWT on WEIGHT we find:

$$\begin{aligned} SS(\hat{\alpha}) &= 404,162 \\ SS(\hat{\alpha}, \hat{\beta}) &= 394,573 \\ RSS(\hat{\beta} | \hat{\alpha}) &= 404,162 - 394,573 = 9,590 \\ n &= 1200 \end{aligned}$$

Source	Degrees of Freedom (d.f.)	Sum of squares (SS)	Mean square (MS)	F_{obs}
Slope	1	9,590	9,590	29.12
Error	1198	394,573	$s^2 = 329.4$	
Corrected total	1199	404,162		

For $\alpha = 5\%$, we find $F_{tab} = 3.84 < F_{obs} = 29.12$, so the regression is significant.

While the regression is significant, the coefficient of determination (CD) is rather low. In Chapter 3 we saw how the CD for simple linear regression can be computed from the correlation coefficient r ($CD = r^2$). Alternatively, it may be computed from the SS , yielding the same result.

The coefficient of determination for a simple linear regression is computed as

$$CD = r^2 = \frac{RSS(\hat{\beta} | \hat{\alpha})}{SS(\hat{\alpha})}$$

The CD assesses the proportion/percentage of the total variation in the response explained by the regression. The SS for the reduced model, $SS(\hat{\alpha})$, assesses the variation around the overall mean. This is a measure of the total variation. The reduction in SS due to the slope, expressed in units of $SS(\hat{\alpha})$, is the CD . The CD is also denoted as r^2 or "R-square" in computer output.

Example 4.2: For the regression of birth weight (BWT) on the mother's weight (WEIGHT), we find:

$$SS(\hat{\alpha}) = 404,162$$

$$SS(\hat{\alpha}, \hat{\beta}) = 394,573$$

$$RSS(\hat{\beta} | \hat{\alpha}) = 9,590$$

$$CD = 9,590 / 404,162 = 0,0237 = 2,37\%$$

Thus, while the regression of birth weight (BWT) on mother's weight (WEIGHT) is highly significant ($p < 0.0001$), it explains only 2,37% of the total variation in birth weight. So while we have established that birth weight and mother's weight are associated, most of the variation in birth weight is due to other factors.

The regression coefficient β may also be tested by a t-test as follows:

$$t_{obs} = \frac{|\hat{\beta}|}{s.e.(\hat{\beta})}$$

It can be shown that

$$s.e.(\hat{\beta}) = \sqrt{\frac{s^2}{SS_x}} \quad , \text{ where } SS_x = \sum_{j=1}^n (x_j - \bar{x})^2$$

The null hypothesis

$$H_0: \beta = 0$$

is rejected at significance level α , when $t_{obs} > t_{tab}$, where t_{tab} is the critical value at significance level α of a t-distribution with $(n-2)$ d.f.

A $(1-\alpha) \times 100\%$ confidence interval for β is computed as

$$\hat{\beta} \pm t_{tab} s.e.(\hat{\beta})$$

where t_{tab} is the tabular t-value at given significance level α and $n-2$ d.f.

The intercept can be tested similarly. While both tests are part of the standard output of regression packages, the latter test is not usually of interest and it is not considered here for brevity.

Example 4.2: For the regression of BWT on WEIGHT we find

$$SS_x = 527,343$$

$$\hat{\beta} = 0.13485$$

$$s^2 = 329.4 \text{ (from the ANOVA table!)}$$

$$s.e(\hat{\beta}) = \sqrt{\frac{329.4}{527,343}} = 0.025$$

$$t_{obs} = \frac{0.13485}{0.025} = 5.40 > t_{tab} = 1.96 \Rightarrow \text{The regression is significant.}$$

A 95% confidence interval is given by

$$0.13485 \pm 1.96 \times 0.025 = (0.0858; 0.1839)$$

Exercise 4.3: Do a linear regression of BWT on WEIGHT using PROC GLM.

SAS hints

The analysis of variance is produced by

```
proc glm;
model bwt=weight/solution clparm;
run;
```

The SOLUTION option prints least squares estimates of the parameters, while CLPARM invokes computation of confidence intervals around parameter estimates. Note that the explanatory variable WEIGHT is **not** listed in a CLASS statement. This is because WEIGHT is a quantitative variable. By contrast, in the ANOVA described in Sections 4.1 and 4.2, the categorical explanatory variable SMOKE had to be listed in a CLASS statement. **The purpose of the CLASS statement is to make a distinction between categorical (qualitative) and quantitative explanatory variables: categorical explanatory variables need to be listed in the CLASS statement, while quantitative explanatory variables do not appear in the CLASS statement.**

The linear regression may be plotted by

```
symbol i=r1 value=dot;
proc gplot;
plot bwt*weight;
run;
```

The SYMBOL statement is used to invoke a least squares fit through the scatter (I=RL, or equivalently, INTERPOL = RL, where RL stands for "regression linear") and to represent plotted data by dots (VALUE=DOT).

Exercise 4.4: A regression of BWT on mother's height (HEIGHT) yielded the following result:

$$SS(\hat{\alpha}, \hat{\beta}) = 405,905$$

$$SS(\hat{\alpha}) = 389,908$$

$$n = 1214$$

$$\hat{\beta} = 1.4334$$

$$SS_x = 7785.229 \text{ (SS for heights)}$$

Do an ANOVA for this regression. In addition, perform a t-test for $H_0: \beta = 0$ and find a 95% confidence interval for β . Do all of this using a pocket calculator and check results using SAS.

Exercise 4.5: Investigate the assertion that mothers who smoke have increased rates of premature delivery, i.e. reduced gestation time (note: in **babies.dat**, time of gestation is stored in the variable GESTATION).

4.4 Simultaneously assessing the effect of one categorical and one quantitative factor

The CHDS study covers many potential explanatory variables, some of which are categorical (SMOKE), while others are quantitative (WEIGHT, HEIGHT). So far we have studied factors one at a time. If several important factors influence the response (BWT) simultaneously, it is useful to model them simultaneously.

Example 4.2: The data may be divided into two groups depending on whether or not the mother smokes (SMOKE=0 or 1). We may then look at the regressions of BWT on WEIGHT separately within each group. Fig. 4.3 shows both regressions in one single plot. Not unexpectedly, the regression line for babies of non-smokers is moved upwards relative to the smoking group. This agrees well with the ANOVA results, which yielded a birth weight mean of 123 for babies from non-smokers and a mean of 114 ounces for babies from smokers. Also, each group shows an upward trend for the regression on weight. The slope is slightly larger for the babies of nonsmokers (0.148) compared to babies of smokers (0.107). Thus, it appears that for smokers the effect of WEIGHT is stronger. One interesting question is whether the observed slope difference is, in fact, significant, i.e. whether the difference in slope estimates is just due to sampling errors or reflects real differences.

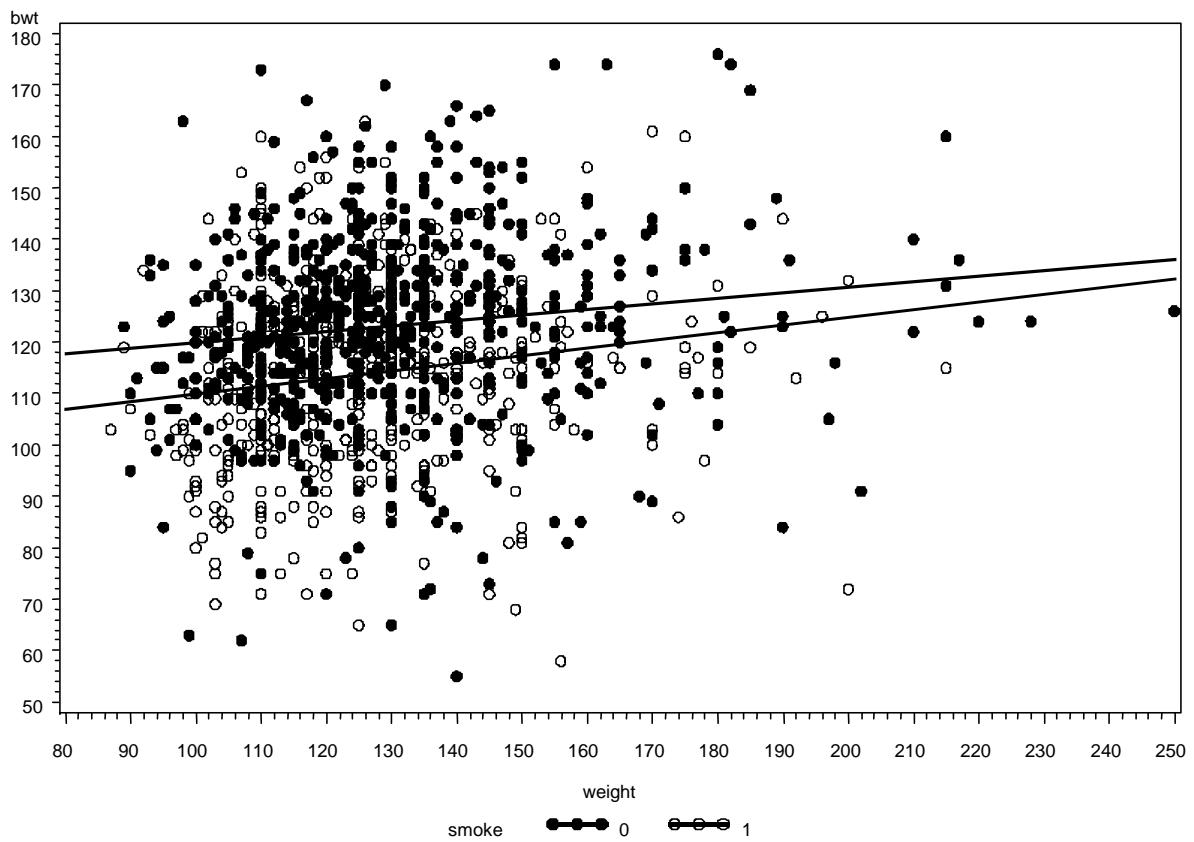


Fig. 4.3: Two linear regressions of birth weight (BWT) vs. mother's weight (WEIGHT).

SMOKE=0 (mother does not smoke): $BWT = 109.02 + 0.107 \times \text{WEIGHT}$

SMOKE=1 (mother smokes): $BWT = 95.13 + 0.148 \times \text{WEIGHT}$

4.4.1 Do slopes differ among groups?

If the difference in slopes is significant, then there is said to be a significant **interaction** among the factors SMOKE and WEIGHT, i.e. the effect of the mother's weight (WEIGHT) on birth weight (BWT) depends on whether or not the mother smokes (SMOKE). Conversely, the expected difference among groups depends on the mother's weight (WEIGHT). To find out whether or not such dependencies exist, we need to test for interaction. This may be done by fitting two nested models to all data simultaneously and comparing the fits by an F-test. The **full model** for the whole data is given by

$$y_{ij} = \alpha_i + \beta_i x_{ij} + e_{ij}$$

where

y_{ij} = j -th birth weight in the i -th group ($i = 1$ for SMOKE = 0; $i = 2$ for SMOKE = 1)

β_i = regression slope for i -th group

x_{ij} = WEIGHT corresponding to y_{ij}

e_{ij} = random deviation from regression line for y_{ij} , assumed to follow a normal distribution with zero mean and variance σ^2 (as usual)

If the slopes do not differ, a common slope may be assumed for both groups (while intercepts differ). The **reduced model** then reads

$$y_{ij} = \alpha_i + \beta x_{ij} + e_{ij},$$

where

$$\beta = \text{common slope}$$

The null hypothesis to be tested by comparing the two models is

$$H_0: \beta_1 = \beta_2 = \beta \quad (\text{slopes are the same in both groups})$$

Again, the full and reduced models are **nested**. To see this more clearly, it is useful to **reparameterize** the regression coefficient in the full model as follows:

$$\beta_i = \beta + \delta_i$$

where

$$\beta = \text{common slope}$$

$$\delta_i = \text{deviation from common slope in } i\text{-th group } (i = 1, 2), \text{ interaction term}$$

With this reparameterization, the full model reads

$$y_{ij} = \alpha_i + (\beta + \delta_i)x_{ij} + e_{ij}$$

$$= \alpha_i + \beta x_{ij} + \delta_i x_{ij} + e_{ij}$$

and the null hypothesis is translated as

$$H_0: \delta_1 = \delta_2 = 0$$

This formulation of the full model is helpful in that now the reduced model can be obtained from the full model by dropping the interaction term δ_i . Thus, we may use the ideas about reduction in SS we are already familiar with. The error SS for the full and reduced models are given by

$$SS(\alpha_i, \beta) = \sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \alpha_i - \beta x_{ij})^2$$

$$SS(\alpha_i, \beta, \delta_i) = \sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \alpha_i - \beta x_{ij} - \delta_i x_{ij})^2$$

And the reduction in SS due to δ_i is

$$RSS(\delta_i | \alpha_i, \beta) = SS(\alpha_i, \beta) - SS(\alpha_i, \beta, \delta_i)$$

A word on least squares estimation.

Example 4.7: Assume, the regression slopes are $\beta_1 = 4$ and $\beta_2 = 6$. Now there is an infinite number of ways in which β_1 and β_2 can be expressed as $\beta_i = \beta + \delta_i$ ($i = 1, 2$) to yield $\beta_1 = 4$ and $\beta_2 = 6$, e.g.

$$\beta = 6, \delta_1 = -2, \delta_2 = 0$$

or

$$\beta = 1200, \delta_1 = -1196, \delta_2 = -1194$$

Thus, there is no unique least squares solution. To resolve this indeterminacy, a restriction needs to be imposed, e.g. $\delta_1 + \delta_2 = 0$. SAS uses $\delta_2 = 0$. It is important and perhaps reassuring at this point to note that this restriction has no impact whatsoever on the resulting values for the slopes β_1 and β_2 . It is only the slopes that have a direct biological interpretation here (the slope tells us by how many units the response changes when x changes by one unit). The particular values for β , δ_1 and δ_2 have no such meaning.

Instead of presenting an ANOVA table at this stage, we will just look at the F-test to compare the full and the reduced model. To compute the F-statistic, we need to know the d.f. for the error SS of the full and reduced model. The full model has four parameters (two slopes and two regression coefficients), so the d.f. is $n-4$. The reduced model has a common slope, but two separate intercepts, so there are 3 parameters, and the d.f. is $n-3$. The d.f. for RSS is the difference of the two error d.f., i.e. the d.f. for RSS is 1.

The F-statistic for

$$H_0: \delta_1 = \delta_2 = \dots = \delta_g \quad (\text{no interaction} = \text{common slopes}) \quad (\text{same as } H_0: \beta_1 = \beta_2 = \dots = \beta_g)$$

$$F_{\text{exp}} = \frac{RSS(\hat{\delta}_i | \hat{\alpha}_i, \hat{\beta})/(g-1)}{s^2}$$

where

$$RSS(\hat{\delta}_i | \hat{\alpha}_i, \hat{\beta}) = SS(\hat{\alpha}_i, \hat{\beta}) - SS(\hat{\alpha}_i, \hat{\beta}, \hat{\delta}_i)$$

and

g is the number of groups ($g = 2$ in our example)

$$s^2 = \frac{SS(\hat{\alpha}_i, \hat{\beta}, \hat{\delta}_i)}{n - 2g}$$

This is compared against an F -distribution with $(g-1)$ numerator d.f. and $n-2g$ denominator d.f.

Example 4.2: For the regression of BWT on WEIGHT and SMOKE we find

$$SS(\hat{\alpha}_i, \hat{\beta}, \hat{\delta}_i) = 368,438$$

$$SS(\hat{\alpha}_i, \hat{\beta}) = 368,636$$

$$n = 1190$$

$$g = 2$$

$$RSS(\hat{\delta}_i | \hat{\alpha}_i, \hat{\beta}) = 198$$

$$F_{obs} = \frac{198/(2-1)}{368,438/1186} = 0.64 < F_{tab} = 3.84$$

\Rightarrow There is no significant interaction; it is concluded that the slopes are parallel.

4.4.2 Test of main effect for a qualitative factor, controlling for a quantitative factor

From the preceding analysis, we selected the reduced model

$$y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}$$

which implies parallel regression lines, i.e. **absence of interaction**. This model may be used to further test the effects of WEIGHT and SMOKE (α_i, β), the socalled **main effects**. We speak of main effects of two factors only when they do not interact. In the above model, α_i is a main effect for SMOKE, while β is a main effect for WEIGHT.

For example, the vertical distance of the regression lines may be interpreted as the differences between groups (SMOKE = 0 and SMOKE = 1). This distance relates to an evaluation of the two regressions at the same value of WEIGHT. If the weight at which each regression is evaluated is x_0 , then the predicted values are $\alpha_1 + \beta x_0$ and $\alpha_2 + \beta x_0$, the difference (vertical distance) being $\alpha_1 - \alpha_2$, regardless of the value of x_0 . In other words, the vertical distance of the regression lines, i.e. the group difference or difference in main effects, is $\alpha_1 - \alpha_2$ at any weight (WEIGHT; see Fig. 4.4). Clearly, under this model, the effect of smoking does not depend on WEIGHT (absence of interaction). To test

$$H_0: \alpha_1 = \alpha_2 \text{ (no group difference)}$$

we may compare two nested models. For this purpose, we replace the intercepts by

$$\alpha_i = \alpha + \gamma_i,$$

in much the same way as in analysis of variance for comparing groups. Thus, the **full** model reads

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + e_{ij}$$

The null hypothesis of common intercepts (no group differences) translates to

$$H_0: \gamma_1 = \gamma_2$$

The **reduced** model with a common intercept follows from the full model by dropping γ_i .

Assuming no interaction, the null hypothesis of **common intercepts (i.e. absence of differences among g groups)**, $H_0: \gamma_1 = \gamma_2 = \dots = \gamma_g$ (same as $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_g$), is tested by computing

$$SS(\hat{\alpha}, \hat{\beta}, \hat{\gamma}_i) \text{ on } n-g-1 \text{ d.f. and}$$

$$SS(\hat{\alpha}, \hat{\beta}) \text{ on } n-2 \text{ d.f. and}$$

$$F_{obs} = \frac{RSS(\hat{\gamma}_i | \hat{\alpha}, \hat{\beta}) / (g-1)}{s^2}$$

where

$$s^2 = SS(\hat{\mu}, \hat{\beta}, \hat{\gamma}_i) / (n-g-1)$$

F_{obs} is compared against an F-distribution with $(g-1)$ and $(n-g-1)$ d.f.

Example 4.2: The null hypothesis of common intercepts in the baby data implies that there is a common regression line, i.e. there is no effect on BWT due to smoking. To test this null hypothesis, we compute

$$SS(\hat{\alpha}, \hat{\gamma}_i, \hat{\beta}) = 368,363$$

$$SS(\hat{\alpha}, \hat{\beta}) = 389,567$$

$$RSS(\hat{\gamma}_i | \hat{\alpha}, \hat{\beta}) = 21,204$$

$$n = 1190, g = 2$$

$$s^2 = \frac{368,363}{1187} = 311$$

$$F_{obs} = \frac{21,204 / (2-1)}{311} = 68.3 > F_{tab} = 3.84$$

\Rightarrow The null hypothesis that smoking has no effect ($H_0: \alpha_1 = \alpha_2$ or equivalently $H_0: \gamma_1 = \gamma_2$), is rejected.

This test for a possible effect of smoking on birth weight controls for the effect of weight by inclusion of the regression term βx_{ij} . Thus, the test differs from a simple ANOVA. In fact, if the regression explains a substantial part of the total variation, the **power** (probability) to detect group differences may be substantially increased by a reduction in the residual variance, where the reduction is "explained" by the other factor.

4.4.3 Test of main effect for quantitative factor, controlling for a qualitative factor

Let us now turn to a test of the regression coefficient β for WEIGHT. Specifically, consider

$$H_0: \beta = 0 \quad (\text{no effect of WEIGHT on BWT})$$

In testing this, we will want to control for the other factor (SMOKE). This is effected by comparing the following two nested models:

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + e_{ij} \quad (\text{full model - two parallel lines}) \quad \text{and}$$

$$y_{ij} = \alpha + \gamma_i + e_{ij} \quad (\text{reduced model - two horizontal lines})$$

To test $H_0: \beta = 0$, compute

$$SS(\hat{\alpha}, \hat{\gamma}_i, \hat{\beta})$$

$$SS(\hat{\alpha}, \hat{\gamma}_i)$$

$$RSS(\hat{\beta} | \hat{\alpha}, \hat{\gamma}_i) = SS(\hat{\alpha}, \hat{\gamma}_i) - SS(\hat{\alpha}, \hat{\gamma}_i, \hat{\beta})$$

$$s^2 = \frac{SS(\hat{\alpha}, \hat{\gamma}_i, \hat{\beta})}{n-g-1}$$

$$F_{obs} = \frac{RSS(\hat{\beta} | \hat{\alpha}, \hat{\gamma}_i)}{s^2}$$

F_{obs} is compared against an F-distribution with 1 numerator d.f. and $n-g-1$ denominator d.f.

Example 4.2: The null hypothesis of a zero slope for the baby data implies that the mothers weight has no effect on birth weight. To test this, we compute

$$SS(\hat{\alpha}, \hat{\gamma}_i, \hat{\beta}) = 368,636$$

$$SS(\hat{\alpha}, \hat{\gamma}_i) = 376,342$$

$$RSS(\hat{\beta} | \hat{\alpha}, \hat{\gamma}_i) = 7,706$$

$$n = 1190, g = 2$$

$$s^2 = \frac{368,636}{1187} = 311$$

$$F_{obs} = \frac{7,706/(g-1)}{311} = 24.8 > F_{tab} = 3.84$$

\Rightarrow The null hypothesis that weight has no linear effect ($H_0: \beta = 0$), is rejected.

4.4.4 What has the analysis shown for the baby data (intermediate summary)?

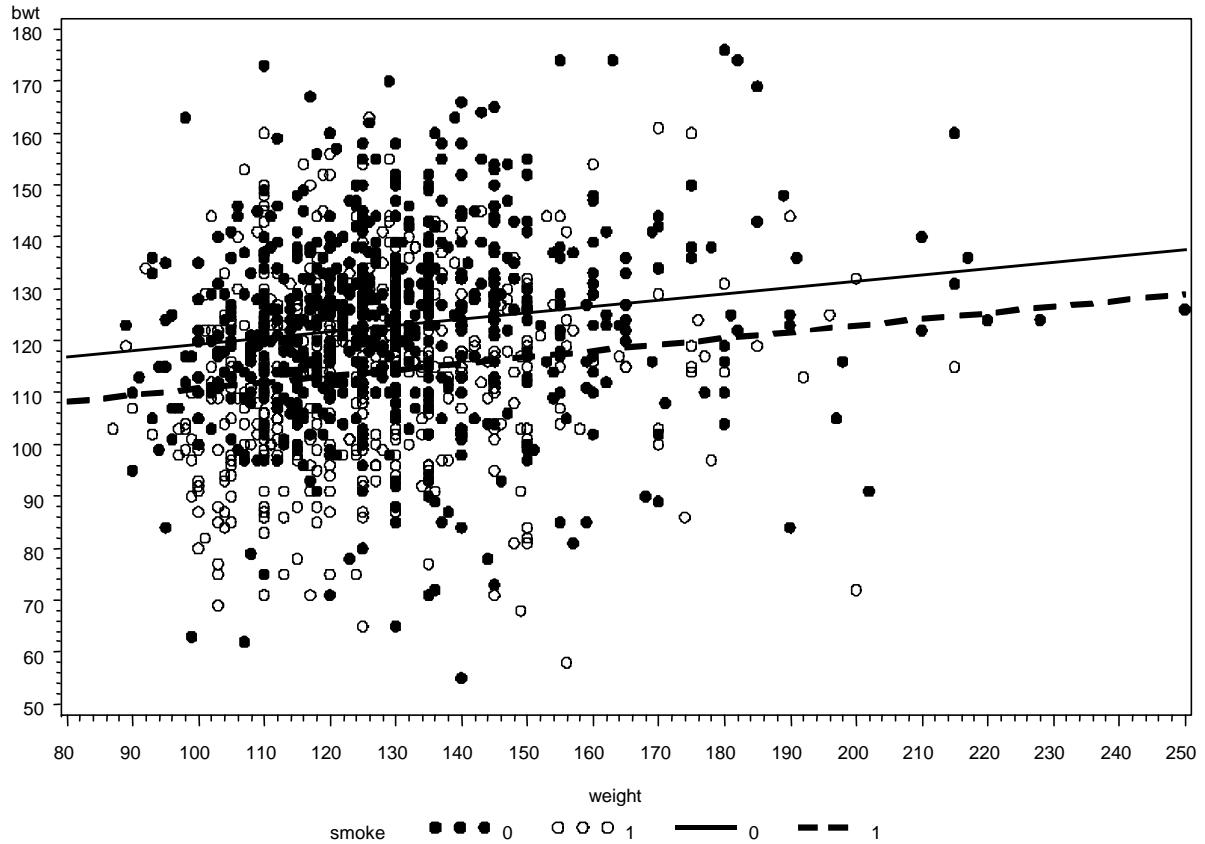


Fig. 4.4: Fitted regression model with common slope for both groups.

$$\text{SMOKE}=0 \text{ (mothers does not smoke): } \text{BWT} = 107.1 + 0.122 \times \text{WEIGHT}$$

$$\text{SMOKE}=1 \text{ (mother smokes): } \text{BWT} = 98.5 + 0.122 \times \text{WEIGHT}$$

Our analysis so far has revealed, that both smoking (SMOKE) and the mother's weight (WEIGHT) have an effect on birth weight. We also found that the effects of these factors are independent of one another, because the interaction was not significant. To conclude the analysis, we will therefore want to estimate the selected model

$$y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}$$

The least squares estimates are as follows:

$$\hat{\alpha}_1 = 107.1 \quad (\text{SMOKE}=0)$$

$$\hat{\alpha}_2 = 98.5 \quad (\text{SMOKE}=1) \quad \hat{\beta} = 0.122 \quad (\text{WEIGHT})$$

Thus, we obtain the following two prediction equations for the two groups:

SMOKE = 0 (mothers are nonsmokers):

$$\text{BWT} = 107.1 + 0.122 \times \text{WEIGHT}$$

SMOKE = 1 (mother is a smoker):

$$\text{BWT} = 98.5 + 0.122 \times \text{WEIGHT}$$

The intercept for the smokers group is lower, indicating that smoking tends to reduce the birth weight of the baby. In other words, **when the weight is held constant**, birth weight in the non-smokers group is expected to be above that in the smokers group by 8.6 ounces.

The common slope of 0.122 means that, **independently of whether or not the mother smokes**, an increase in the mother's weight by one pound is associated with an increase in the expected birth weight by 0.122 ounces. In other words, **when the group is fixed** (either SMOKE=0 or SMOKE=1), the slope of the regression is 0.122 for each.

Basically, this analysis confirms the findings from sections 4.1 and 4.2, and it adds the insight that there is no interaction between both factors. The fitted model is shown in Fig. 4.4.

4.4.5 Which factor is more important?

To assess the importance of a factor, we may look at the coefficient of determination for different models. In addition, the residual variance (s^2), or its square root, the standard deviation (s) may be inspected. The standard deviation is sometimes referred to as the **root mean squared error** (RMSE). The RMSE is part of the standard output of PROC GLM.

The coefficient of determination (CD or R^2) for any (full) linear model is computed from the RSS relative to a reduced model with a single intercept term.

$$CD = R^2 = \frac{RSS(\text{due to all model parameters except intercept})}{SS(\text{intercept-only model})}$$

The CD (R^2) is often expressed as a percentage. It assesses the percentage/proportion of the total variation in the response explained by the linear model.

Example 4.2: For the model including the factors WEIGHT and SMOKE,

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + e_{ij},$$

the SS is 368,636. The intercept-only model, $y_{ij} = \alpha + e_{ij}$, has SS equal to 399,356. Thus, the RSS is $399,356 - 368,636 = 30,720$. From this

$$CD = R^2 = \frac{30,720}{399,356} = 0,077 = 7.7\%$$

Thus, the model explains 7.7% of the total variation.

Factor in Model	s	Reduction in s (%)	s^2	Reduction in s^2 (%)	CD (%)
Intercept only	18.33	-	335,88	-	0%
SMOKE	17.80	2.9%	316,79	5,7%	5.8%
WEIGHT	18.11	1.1%	328,15	2,3%	2.4%
SMOKE, WEIGHT	17.62	3.9%	310,56	7,5%	7.7%

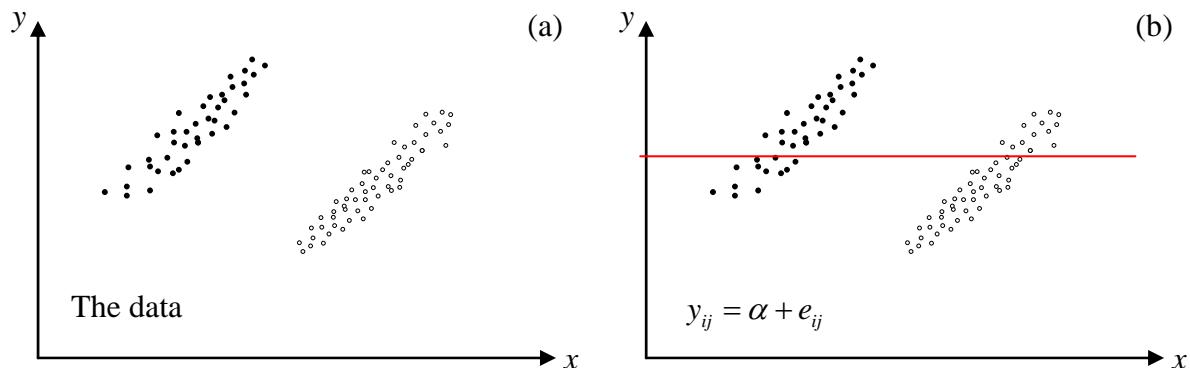
The factor SMOKE, if fitted alone, can explain 5.8% of the total variation, and it reduces the residual variance (s^2) from 335.88 to 316.79, i.e. by $(335.88 - 316.79)/335.88 = 5.7\%$. By contrast, the WEIGHT fitted alone explains less than 0.1% and reduces s^2 by 2.3%. When both factors are fitted together, the variance is reduced by 7.5%, which is roughly the same as the sum of 5.7% and 2.3%. From these figures it can be concluded that the factor SMOKE is more important than WEIGHT. The figures also show, however, that neither factor explains a large share of the total variation. There must be a number of other factors influencing birth weight.

A note of caution regarding the above interpretation is in order. The 5.7% reduction in s^2 by SMOKE and of 2.3% by WEIGHT add up to 8.0%, which is rather close to the reduction of 7.5% when both factors are fitted simultaneously, because the two factors are nearly uncorrelated: The correlation of the dummy variable for SMOKE (levels 0 and 1) and WEIGHT is -0.07 , which is rather close to zero. When two factors are highly correlated, the one factor can explain a large portion of what is explained by the other, and vice versa. In the extreme, when two factors are perfectly correlated, it is sufficient to have just one of them in the model; the added reduction in s when the other is added, will then be zero! This problem will be addressed in more detail in section 4.7 on multiple linear regression.

A second note of caution: the notion of "explained" variation does not imply a causal relationship. It just implies a statistical association. For example a regression of amount of damage caused by a fire on the number of firefighters that fought the fire may be highly significant. This does not show, however, that the presence of firefighters was the cause of the damage!

4.4.6 A further illustration of the models considered in Section 4.4.

Fig. 4.5 shows a hypothetical example that illustrates the different models considered for one qualitative and one quantitative factor. The different types of symbol (black dots, white circles) represent two groups. Panel (a) shows the raw data. Obviously, the points clearly fall into two groups. Panel (b) shows the fit of a single mean. The simple regression shown in panel (c) is clearly misleading, because the two groups which are ignored. Panel (d) shows the fit of two group means. This model ignores the influence of the quantitative variable x . Panel (e) shows parallel regressions, and this model fits quite well to this data. The full model in panel (f) allows separate slopes for the two groups, but slope differences are small. Visual inspection suggests that the parallel lines model in panel (e) is the best model. This would need to be substantiated by significance as explained in the preceding subsections.



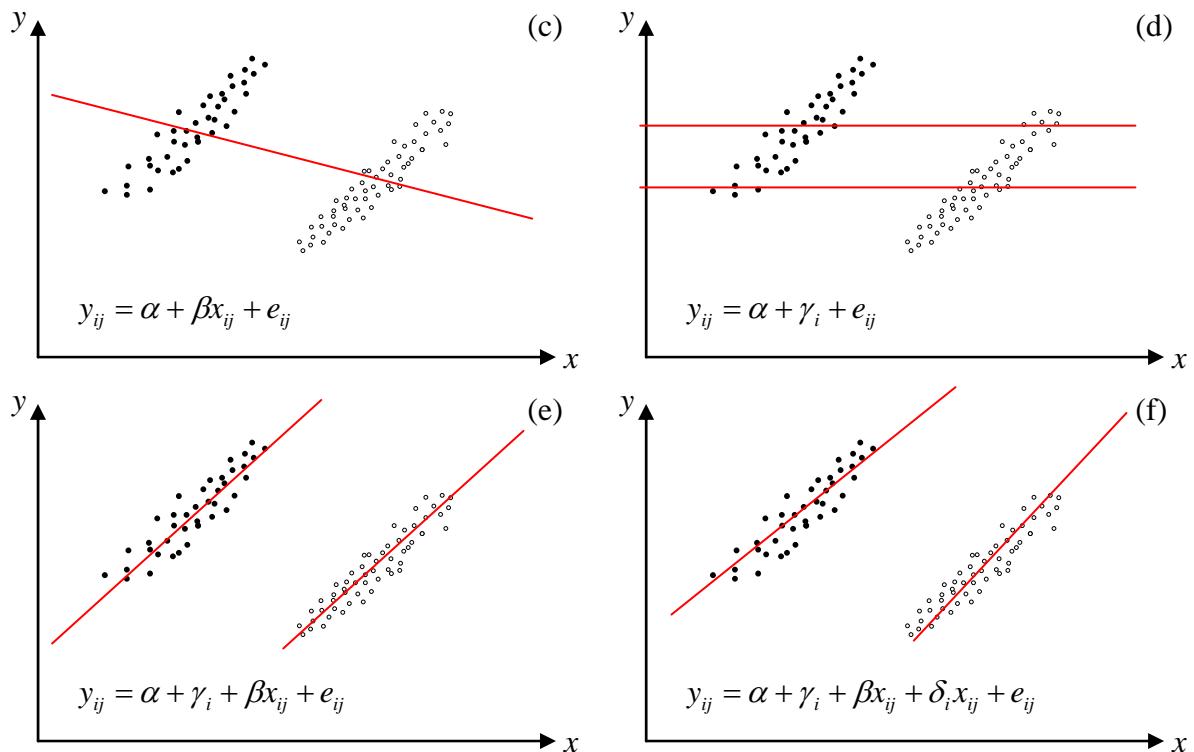


Fig. 4.5: Different models fitted to the same data.

4.5 A general method for comparing nested models

From the preceding sections, a general pattern emerges for comparing two nested linear models.

- (1) Fit the full and the reduced model by least squares and record the error SS for both models: $SS(\text{full})$ and $SS(\text{reduced})$
- (2) Determine the error d.f. for the full and reduced models: $df(\text{full})$ and $df(\text{reduced})$. The d.f. of a model equals the number of observations, minus the number of free parameters (some care is needed to find the d.f. when the model contains effects for categorical variables).
- (3) Estimate the error variance from the SS of the full model (s^2). Compare the RSS to s^2 via an F-statistic, i.e. compute

$$MS(\text{numerator}) = \frac{RSS(\text{reduced vs. full})}{Rdf(\text{reduced vs. full})}$$

where

$RSS(\text{reduced vs. full}) = SS(\text{reduced}) - SS(\text{full})$ is the reduction in SS and
 $Rdf(\text{reduced vs. full}) = df(\text{reduced}) - df(\text{full})$ is the reduction in d.f.,

$$s^2 = \frac{SS(\text{full})}{df(\text{full})}$$

$$F_{obs} = \frac{MS(numerator)}{s^2}$$

(4) Compare F_{exp} to an F-distribution with

$$df(numerator) = Rdf(reduced \ vs. \ full)$$

and

$$df(denominator) = df(s^2) = df(full)$$

(5) If F_{obs} is significant, reject the reduced model. Otherwise accept the reduced model.

4.6 Looking at a sequence of models

So far we have looked at pairs of full and reduced models to test different effects. The terms "full" and "reduced" models are relative, because a reduced model in one test may be a full model in another. For example, the model

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + e_{ij}$$

is a reduced model relative to the full model

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + \delta x_{ij} + e_{ij}$$

but it is itself a full model relative to the reduced model

$$y_{ij} = \alpha + \beta x_{ij} + e_{ij}$$

In fact, the intermediate model ($y_{ij} = \alpha + \gamma_i + \beta x_{ij} + e_{ij}$) may be seen as a member in a sequence of models starting from simple to more complex. For example, one may consider the sequence given in Table 4.2.

Table 4.2(a): A model sequence for the baby data (two groups: SMOKE=0 and SMOKE=1).

Term added	Model	d.f.	SS	RSS	Rdf
Intercept	$y_{ij} = \alpha + e_{ij}$	1189	399,356		
SMOKE	$y_{ij} = \alpha + \gamma_i + e_{ij}$	1188	376,342	23,014	1
WEIGHT	$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + e_{ij}$	1187	368,636	7,706	1
SMOKE*WEIGHT	$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + \delta x_{ij} + e_{ij}$	1186	368,438	198	1

For each model in the sequence, the immediately preceding model constitutes a reduced model, while the model that follows is a full model relative to the current one. Thus, at each step, we can compute the RSS and test whether it is significant by comparison to an estimate, s^2 , of the error variance, σ^2 . We saw earlier that the error variance s^2 is computed from the full model. Now for each step in the sequence, the current model is a full model relative to the preceding one, so we would be dealing with different estimates of s^2 . To simplify things, we can use a single estimate of s^2 , and naturally, this estimate should be based on the last model

in the sequence (the "fullest model"). Each RSS would then be compared against this one estimate of the error variance (s^2). The d.f. for error is, of course, the d.f. of the last model in the sequence. The d.f. for an RSS is given by the associated reduction in d.f. compared to the preceding model, i.e. by the Rdf . The F-statistic for each source of variation is computed with s^2 in the denominator. The full ANOVA table then reads as follows:

Source	Degrees of Freedom	Sum of squares	Mean square	F_{obs}	p-value
SMOKE	1	23,014	3,014	74.08	<0.0001
WEIGHT	1	7,706	7,706	24.81	<0.0001
SMOKE*WEIGHT	1	198	198	0.64	0.4250
Error	1186	368,438	$s^2 = 311$		
Corrected total	1189	399,356			

Now what does this ANOVA tell us about the baby data in addition to what we have learned in the preceding sections? Well, there is no real news here. It is just a matter of convenience, that the different analyses now appear compactly in a single ANOVA table. Incidentally, this table is in a format similar to that produced by many statistical packages, including SAS.

A few remarks on interpretation of the ANOVA table are in order. **The first test to look at in a sequential ANOVA table is always the one at the bottom!** Here, this is the test of the interaction term, denoted as SMOKE*WEIGHT. This term corresponds to δ_i , and so it tests the heterogeneity of slopes among the two groups (SMOKE=0 and SMOKE=1). The F-test is not significant, so we may conclude that slopes are parallel, i.e., there is a common slope.

If the interaction were significant, the interpretation of the ANOVA table would terminate, i.e. the tests for SMOKE and WEIGHT would be irrelevant. This is so because a significant interaction would tell us that there are two separate regression lines with a separate slope, and hence also with a separate intercept. No more than this can be learned from any further tests in the ANOVA table. In fact, the other tests appearing in the ANOVA table are misleading in the presence of significant interaction. For example, a test whether a "common" slope (WEIGHT) is significant is pointless, when it has been determined that slopes differ among groups.

In the present case, we conclude that lines run parallel, because the **interaction is not significant**. Because of the non-significant interaction, it makes sense to move up one row in the ANOVA table and look at the test for WEIGHT, i.e. the test of H_0 that the common slope β equals zero (no effect of WEIGHT). The sum-of-squares for WEIGHT corresponds to the RSS when adding the common slope to a model with effects α and γ_i , i.e. a model with a separate intercept term for each group (SMOKE=0 and SMOKE=1). Thus, the F-test for β controls for possible group differences in intercepts (factor SMOKE). This control is effected by first fitting SMOKE (γ_i) in the sequence, and then adding WEIGHT (β). The F-test for WEIGHT is highly significant, so the slope is different from zero.

Now what about the test for the other factor (SMOKE)? If we move to the top row, we have an F-test for SMOKE. But we need to be cautious here. The sum of squares for SMOKE is the RSS for adding γ_i to a reduced model that contains just a common intercept α term. The reduced model does not contain the regression on WEIGHT. Thus, the F-test for SMOKE in this sequence of models does not control for the other factor (WEIGHT)! The reason is

that we have fitted SMOKE before WEIGHT. Thus, the test for SMOKE is not a valid one, in case WEIGHT is significant. The reason is that apparent differences in SMOKE may be **confounded** with WEIGHT effects. To obtain a valid test, the order of fitting needs to be reversed. The resulting ANOVA is as follows:

Table 4.2(b): The other model sequence for the baby data (two groups: SMOKE=0 and SMOKE=1).

Term added	Model	d.f.	SS	RSS	Rdf
Intercept	$y_{ij} = \alpha + e_{ij}$	1189	399,356		
WEIGHT	$y_{ij} = \alpha + \beta x_{ij} + e_{ij}$	1188	369,840	9,516	1
SMOKE	$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + e_{ij}$	1187	368,636	21,204	1
SMOKE*WEIGHT	$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + \delta_k x_{ij} + e_{ij}$	1186	368,438	198	1

Source	Degrees of Freedom	Sum of squares	Mean square	F _{obs}	p-value
WEIGHT	1	9,516	9,516	30.63	<0.0001
SMOKE	1	21,204	21,204	68.23	<0.0001
SMOKE*WEIGHT	1	198	198	0.64	0.4250
Error	1186	368,438	311		
Corrected total	1189	399,356			

The F-test for SMOKE now controls for WEIGHT, because the latter is fitted first. Note that the sum-of-squares for WEIGHT and SMOKE have changed compared to the reverse order of fitting considered previously. This underlines the importance of the order in which terms are fitted (though in this case the F-values do not change dramatically).

It should be re-iterated that the tests for WEIGHT and SMOKE make sense here only because the interaction, SMOKE*WEIGHT was not significant. An alternative to the above analysis, in light of the non-significant interaction, would be to remove the interaction and consider only reduced sequences of models, comprising just the terms α (intercept), γ_i (SMOKE) and β (WEIGHT). As in section 4.4 we would then test WEIGHT (β) by comparing the following two models:

$$\begin{aligned} y_{ij} &= \alpha + \gamma_i + e_{ij} && (\text{SMOKE}) \\ y_{ij} &= \alpha + \gamma_i + \beta x_{ij} + e_{ij} && (\text{SMOKE, WEIGHT}) \end{aligned}$$

Conversely, to test SMOKE, the sequence would be

$$\begin{aligned} y_{ij} &= \alpha + \beta x_{ij} + e_{ij} && (\text{WEIGHT}) \\ y_{ij} &= \alpha + \gamma_i + \beta x_{ij} + e_{ij} && (\text{SMOKE, WEIGHT}) \end{aligned}$$

In either case, the term to be tested is entered last! **It is a general principle for testing terms in linear models that the correct test for a term is obtained by entering that term last in the model building sequence.** The two ANOVA tables corresponding to the two sequences are as follows:

Source	Degrees of Freedom	Sum of squares	Mean square	F_{obs}	p-value
SMOKE	1	23,014	23,014	74.11	<0.0001
WEIGHT	1	7,706	7,706	24.81	<0.0001
Error	1187	368,636	311		
Corrected total	1189	399,356			

Source	Degrees of Freedom	Sum of squares	Mean square	F_{obs}	p-value
WEIGHT	1	9,516	9,516	30.64	<0.0001
SMOKE	1	21,204	21,204	68.28	<0.0001
Error	1187	368,636	311		
Corrected total	1189	399,356			

In both tables, the error mean square is computed from the error SS for the model ($y_{ij} = \alpha + \gamma_i + \beta x_{ij} + e_{ij}$), which contains WEIGHT (β) and SMOKE (γ_i), but not the interaction WEIGHT*SMOKE (δ_i). Note that in both tables, the F-values are virtually the same as in the corresponding ANOVA tables which included the interaction term WEIGHT*SMOKE. The reason is that the interaction was non-significant, so the MS for the interaction estimates only error. Thus, only a minor change is to be expected when the interaction is dropped, which implies that the sum of squares for the interaction and for error in the full sequence (including interaction) is pooled.

In summary, it is immaterial whether or not we recompute the ANOVA table after finding a non-significant interaction. From a practical point of view it may be more convenient not to recompute. In terms of the interpretation this latter strategy implies that generally **a test in a sequential ANOVA table should be interpreted only if all tests in subsequent lines are not significant**.

The question of whether or not to reduce the model by omitting non-significant terms will, of course, appear in a different light, when it comes to estimation of the parameters of the final model. In our case, the final model will contain WEIGHT and SMOKE, but not the interaction WEIGHT*SMOKE, i.e. we will want to estimate the parameters of the model

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + e_{ij} \quad (\text{SMOKE, WEIGHT})$$

or, equivalently, of the model

$$y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}$$

(See end of section 4.4).

Exercise 4.6: To a joint analysis for the effect of WEIGHT and SMOKE on BWT using SAS PROC GLM. Reproduce the results given in this and the preceding two sections.

SAS hints

Fit the model

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + e_{ij} \quad (\text{SMOKE, WEIGHT})$$

by

```
proc glm;
  class smoke;
  model bwt=weight smoke/solution clparm;
run;
```

The order in which effects appear in the MODEL statement determines the model fitting sequence. The above code gives the correct test for SMOKE (assuming there is no interaction), because WEIGHT is fitted first. The correct test for WEIGHT is obtained by

```
proc glm;
  class smoke;
  model bwt=smoke weight/solution clparm;
run;
```

The model with interaction,

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + \delta x_{ij} + e_{ij}$$

is fitted by

```
proc glm;
  class smoke;
  model bwt=smoke weight smoke*weight/solution clparm;
run;
```

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + \delta x_{ij} + e_{ij}$$

Part of the output from the above code for $y_{ij} = \alpha + \gamma_i + \beta x_{ij} + \delta x_{ij} + e_{ij}$ is as follows:

Dependent Variable: bwt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	30918.1116	10306.0372	33.18	<.0001
Error	1186	368438.4061	310.6563		
Corrected Total	1189	399356.5176			

R-Square	Coeff Var	Root MSE	bwt Mean
0.077420	14.75732	17.62545	119.4353

Source	DF	Type I SS	Mean Square	F Value	Pr > F
smoke	1	23014.29693	23014.29693	74.08	<.0001
weight	1	7705.92387	7705.92387	24.81	<.0001
weight*smoke	1	197.89078	197.89078	0.64	0.4250

Source	DF	Type III SS	Mean Square	F Value	Pr > F
smoke	1	1365.789733	1365.789733	4.40	0.0362
weight	1	7761.640849	7761.640849	24.98	<.0001
weight*smoke	1	197.890782	197.890782	0.64	0.4250

SAS prints two different types of SS labeled Type I and Type III. The sequential SS discussed in this chapter correspond to Type I SS, and you may ignore the Type III SS. The ANOVA output starts with a table that has sources of variation labeled MODEL and ERROR. The F-test corresponds to a comparison of the full model that contains all effects listed in the MODEL statement, to a model that contains just the intercept term. For example, when the model

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + \delta_i x_{ij} + e_{ij}$$

is fitted, as in the example, the SS listed in the row labeled MODEL compares this model to the reduced model

$$y_{ij} = \alpha + e_{ij}$$

The associated RSS is

$$RSS(\hat{\gamma}_i, \hat{\beta}, \hat{\delta}_i | \hat{\alpha}) = SS(\hat{\alpha}) - SS(\hat{\alpha}, \hat{\gamma}_i, \hat{\beta}, \hat{\delta}_i)$$

If the test is significant, this only tells us that at least one of the effects in the model ($\gamma_i, \beta, \delta_i$) may be relevant, but it does not tell us which. Thus, this test is not usually of primary interest.

It should also be noted that the error SS is listed only once, i.e. in the first ANOVA table. It is not repeated in the tables that follow, which give the SS for the different effects in the model (Type I and Type III).

To estimate the selected model,

$$y_{ij} = \alpha + \gamma_i + \beta x_{ij} + \delta_i x_{ij} + e_{ij}$$

$$= \alpha_i + \beta x_{ij} + \delta_i x_{ij} + e_{ij}$$

it is convenient to not estimate α and γ_i , but the group specific intercept $\alpha_i = \alpha + \gamma_i$. This is effected by using the NOINT option on the MODEL statement as follows:

```
proc glm;
  class smoke;
  model bwt=smoke weight/noint solution clparm;
```

```
run;
```

To plot the separate regressions for the two groups simultaneously (different slopes), use

```
symbol i=r1 value=dot;
proc gplot;
plot bwt*weight=smoke;
run;
```

(Unfortunately, it is not so straightforward to plot parallel lines in one graph).

4.7 Multiple linear regression

So far we have studied the effect of one quantitative variable (WEIGHT) on birth weight (BWT). There are other factors, which may also have an effect on BWT, for example the mothers height (HEIGHT) or the duration of the gestation period (GESTATION). More than one quantitative factor can be fitted by a multiple linear regression model, e.g.

$$BWT = \alpha + \beta_1 \times \text{WEIGHT} + \beta_2 \times \text{HEIGHT} + \beta_3 \times \text{GESTATION}$$

More formally, this model may be written as

$$y_j = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j}$$

where

x_{1j} = mother's weight for j -th baby (WEIGHT)

x_{2j} = mother's height for j -th baby (HEIGHT)

x_{3j} = duration of gestation period for j -th baby (GESTATION)

Quantitative terms may be fitted sequentially in the same way as discussed throughout this chapter.

Exercise 4.7: Use multiple regression to investigate the effect of WEIGHT, HEIGHT, and GESTATION on BWT. Which of these variables is the most important (look at reduction in s and/or increase in CD/R^2)? Carefully look at different orders of fitting terms. Interpret the regression terms for the fitted model.

SAS hints

```
proc glm;
model bwt=height gestation weight/solution;
run;
```

Exercise 4.8: In addition to the variables considered in Exercise 4.7, add the qualitative variable SMOKE as well as the quantitative variable AGE to your analysis. Is smoking more important than the quantitative explanatory variables?

Exercise 4.9: Fit a multiple linear regression to the margarine data of Example 4.1 described at the beginning of this chapter (**margarine.dat**). Specifically, regress the sales (SALES) on the price of margarine (PRICE), expenditures for advertisement (ADVERTISEMENT) and number of visits (VISITS). Which of these three variables, which are under the control of the company, are the most important?

Exercise 4.10: For the baby data, check the Surgeon General's assertions that "The newborns of smokers are smaller (in terms of BWT) at every gestational age (GESTATION; length of pregnancy)." Hint: Test the interaction of SMOKE and GESTATION.

SAS hint:

To plot the regressions of BWT vs. GESTATION for the two groups (nonsmokers and smokers; SMOKE=0 and SMOKE=1), use

```
symbol i=r1 value=dot;
proc gplot;
plot bwt*gestation=smoke;
run;
```

4.7.1 Multicollinearity

Example 4.8: The following data were collected on the volume of black cherry trees for a sample of 31 black cherry trees in the Allegheny National Forest, Pennsylvania (modification from an original dataset available at <http://www.statsci.org/data/general/cherry.html>; see **trees.dat**):

VOLUME = volume of tree (cubic feet)
HEIGHT = height of tree (feet)
DIAM = diameter (inches) (at 54 inches above ground)
DIAM2 = diameter (inches) (at 30 inches above ground)

The data were collected in order to find an estimate for the volume of a tree (and therefore the timber yield), given its height and diameter. Here, we will look at a multiple regression to predict VOLUME from the two diameter measurements:

$$\text{VOLUME} = \alpha + \beta_1 \times \text{DIAM} + \beta_2 \times \text{DIAM2}$$

HEIGHT can be included as well (and usually is in these types of application), but for the moment this is not done here for simplicity (but see Exercise 4.11). One task in the statistical analysis is to test whether the regression coefficients β_1 and β_2 are significant.

To test the significance of the two terms (DIAM and DIAM2), two model fitting sequences are considered. The ANOVAs are as follows:

The test for DIAM, correcting for DIAM2:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Diam2	1	7532.412772	7532.412772	403.87	<.0001
Diam	1	51.450669	51.450669	2.76	0.1079
Error	28	522.220430	18.650730		

The test for DIAM2, correcting for DIAM:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Diam	1	7581.781332	7581.781332	406.51	<.0001
Diam2	1	2.082109	2.082109	0.11	0.7408
Error	28	522.220430	18.650730		

Both variables are not significant, if the analysis corrects for the other variable. However, if VOLUME is regressed on DIAM or DIAM2 alone, both are highly significant:

For DIAM fitted alone:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Diam	1	7581.781332	7581.781332	419.36	<.0001
Error	29	524.302539	18.079398		

For DIAM2 fitted alone:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Diam2	1	7532.412772	7532.412772	380.78	<.0001
Error	29	573.671099	19.781762		

This result is surprising at first sight, and the different analyses seem to suggest contradicting conclusions. How should we decide whether or not a regression on diameter is at all useful?

The counterintuitive results here are due to the high correlation among the two explanatory variables (predictor variables). The correlation is

$$r_{(\text{DIAM}, \text{DIAM2})} = 0.995$$

and it is highly significant ($p < 0.0001$). A high correlation of explanatory variables is referred to as **multicollinearity** in multiple regression. If one of two highly correlated variables is already in a regression model, addition of the other adds virtually no new information, and the reduction in *SS* is very small. From a practical point of view, regression with one of the two is just as good as regression on both.

Loosely speaking, when both regressors (predictors, explanatory variables) are in the model, the ANOVA cannot decide which one is the better predictor, because both are equally informative as predictors and both convey essentially the same type of information.

It is interesting to look at the standard errors of the parameter estimates. When fitting the full model with DIAM and DIAM2, we find:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-37.01550833	3.42469843	-10.81	<.0001
Diam	4.22090991	2.54131490	1.66	0.1079
Diam2	0.70699000	2.11596994	0.33	0.7408

$$R^2 = 0.94$$

With DIAM only, we find:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-36.94345912	3.36514495	-10.98	<.0001
Diam	5.06585642	0.24737695	20.48	<.0001

$$R^2 = 0.94$$

With DIAM2 only, we find:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-36.82529408	3.52503913	-10.45	<.0001
Diam2	4.20421886	0.21545211	19.51	<.0001

$$R^2 = 0.93$$

The standard errors in the full model (including both DIAM and DIAM2) are **inflated** by a factor of 10 compared to the simple regression models with only DIAM or DIAM2. The phenomenon is also known as **variance inflation** (remember: the variance of an estimate equals the squared standard error). Due to this inflation, none of the two predictors is significant in the full model. This inflation of standard errors is related to the non-significance of both terms in the full model. The two are just two sides of the same coin: multicollinearity. Note that the p-values for the t-tests are the same as for the F-tests of the same terms.

The R^2 values are about the same for all three models. Due to the high correlation among DIAM and DIAM2, one of the two is sufficient for prediction. The R^2 for DIAM alone is slightly larger than for DIAM2 alone, so we will prefer a regression on DIAM alone. The prediction equation one would want to use in practice is, therefore,

$$\text{VOLUME} = -36.94 + 5.07 \times \text{DIAM}$$

Exercise 4.11: Reproduce the analyses for the tree data (**trees.dat**) using PROC GLM. Does inclusion of HEIGHT improve the prediction of VOLUME.

4.8 Checking model assumptions

It was pointed out in Chapter 3, that normality of errors can be checked using quantile-quantile (Q-Q)-plots. These plot observed quantiles versus expected quantiles for the normal distribution. In addition, it is helpful to plot residuals against predicted values. This helps to see if variance increases with the mean (predicted value).

To obtain these plots, residuals must be computed first, preferable the so-called standardized or studentized residuals. If the MIXED procedure is used for fitting linear models, these can be very easily computed and plotted using the ODS statement. In addition, one needs to use the Option RESIDUAL with the model statement.

Example 4.3 (cont'd): For the melon data, the SAS code is as follows:

```
ods graphics on;
proc mixed;
class variety;
model yield=variety/residual;
lsmeans variety/pdiff;
run;
ods graphics off;
```

SAS - [Ausgabe - (Unbenannt)]

Datei Bearbeiten Ansicht Extras Lösungen Fenster Hilfe

Ergebnisse

Das S
Die Pro
Typ 3 Tests
Zähler
Effekt Freiheitsgrade F
variety 3

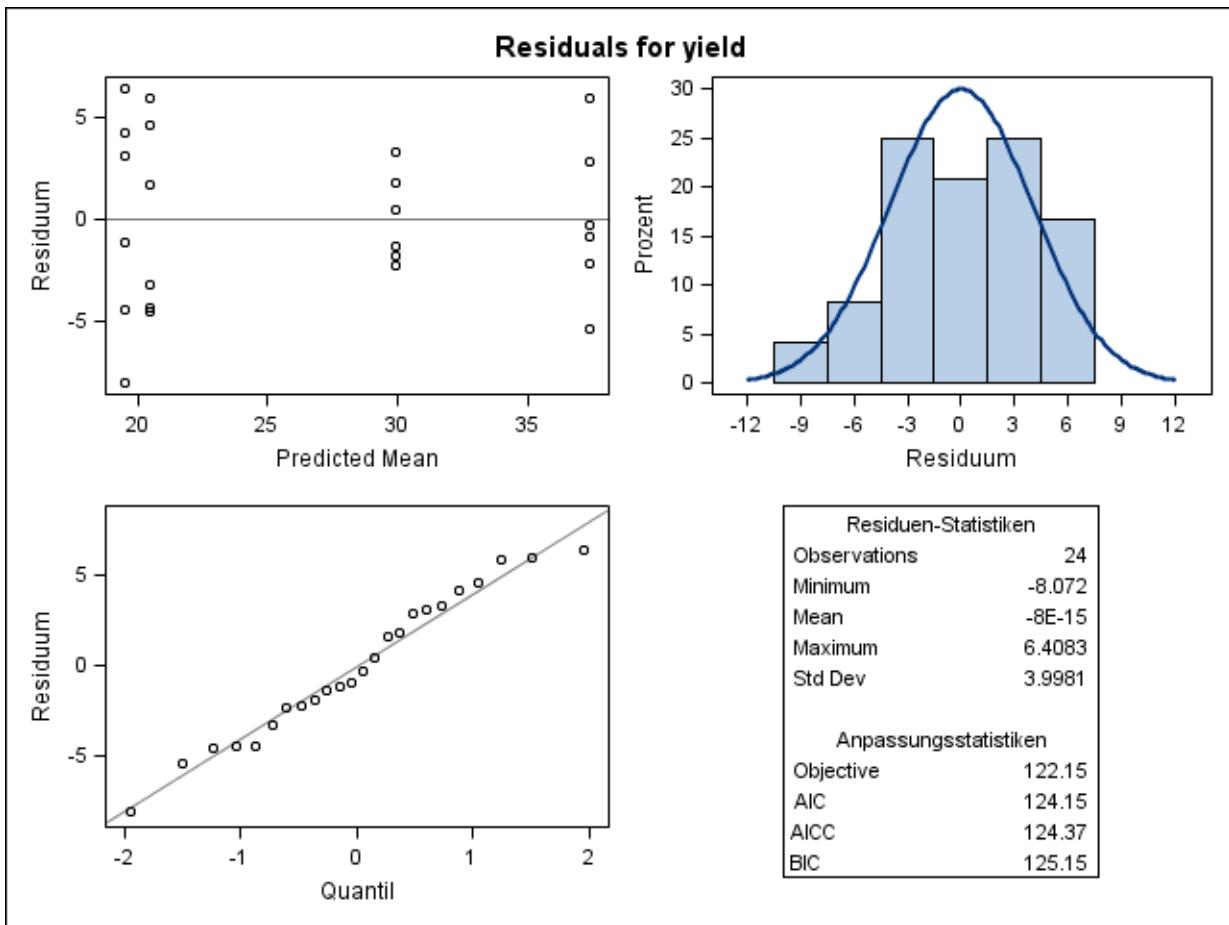
Kleinste-Quadrat
Effekt variety Schätzwert St
variety A 20.4900
variety B 37.4033
variety C 19.4917
variety D 29.8967

Differenzen Kleinste-Quadrat
Effekt variety _variety Schätzwert
variety A B -16.9133
variety A C 0.9983
variety A D -9.4067
variety B C 17.9117
variety B D 7.5067
variety C D -10.4050

Ausgabe - (Unbenannt) Log - (Unbenannt) me

Start 4 Fir... 2 Wi... 2 Mic... Eudora... Notes ...

Click here to get the plots!



The residual plots look okay. In particular, the residuals show no trend of increasing variance with increasing mean (top right plot), and the Q-Q-plot is nicely linear (lower left plot).

Tab. 4.I: Critical values t_{tab} for t -distribution (two-sided) with ν d.f.

$\nu \backslash \alpha$	0,05	0,01	0,001	$\nu \backslash \alpha$	0,05	0,01	0,001
1	12,706	63,657	636,619	51	2,008	2,676	3,492
2	4,303	9,925	31,599	52	2,007	2,674	3,488
3	3,182	5,841	12,924	53	2,006	2,672	3,484
4	2,776	4,604	8,610	54	2,005	2,670	3,480
5	2,571	4,032	6,869	55	2,004	2,668	3,476
6	2,447	3,707	5,959	56	2,003	2,667	3,473
7	2,365	3,499	5,408	57	2,002	2,665	3,470
8	2,306	3,355	5,041	58	2,002	2,663	3,466
9	2,262	3,250	4,781	59	2,001	2,662	3,463
10	2,228	3,169	4,587	60	2,000	2,660	3,460
11	2,201	3,106	4,437	61	2,000	2,659	3,457
12	2,179	3,055	4,318	62	1,999	2,657	3,454
13	2,160	3,012	4,221	63	1,998	2,656	3,452
14	2,145	2,977	4,140	64	1,998	2,655	3,449
15	2,131	2,947	4,073	65	1,997	2,654	3,447
16	2,120	2,921	4,015	66	1,997	2,652	3,444
17	2,110	2,898	3,965	67	1,996	2,651	3,442
18	2,101	2,878	3,922	68	1,995	2,650	3,439
19	2,093	2,861	3,883	69	1,995	2,649	3,437
20	2,086	2,845	3,850	70	1,994	2,648	3,435
21	2,080	2,831	3,819	71	1,994	2,647	3,433
22	2,074	2,819	3,792	72	1,993	2,646	3,431
23	2,069	2,807	3,768	73	1,993	2,645	3,429
24	2,064	2,797	3,745	74	1,993	2,644	3,427
25	2,060	2,787	3,725	75	1,992	2,643	3,425
26	2,056	2,779	3,707	76	1,992	2,642	3,423
27	2,052	2,771	3,690	77	1,991	2,641	3,421
28	2,048	2,763	3,674	78	1,991	2,640	3,420
29	2,045	2,756	3,659	79	1,990	2,640	3,418
30	2,042	2,750	3,646	80	1,990	2,639	3,416
31	2,040	2,744	3,633	81	1,990	2,638	3,415
32	2,037	2,738	3,622	82	1,989	2,637	3,413
33	2,035	2,733	3,611	83	1,989	2,636	3,412
34	2,032	2,728	3,601	84	1,989	2,636	3,410
35	2,030	2,724	3,591	85	1,988	2,635	3,409
36	2,028	2,719	3,582	86	1,988	2,634	3,407
37	2,026	2,715	3,574	87	1,988	2,634	3,406
38	2,024	2,712	3,566	88	1,987	2,633	3,405
39	2,023	2,708	3,558	89	1,987	2,632	3,403
40	2,021	2,704	3,551	90	1,987	2,632	3,402
41	2,020	2,701	3,544	91	1,986	2,631	3,401
42	2,018	2,698	3,538	92	1,986	2,630	3,399
43	2,017	2,695	3,532	93	1,986	2,630	3,398
44	2,015	2,692	3,526	94	1,986	2,629	3,397
45	2,014	2,690	3,520	95	1,985	2,629	3,396
46	2,013	2,687	3,515	96	1,985	2,628	3,395
47	2,012	2,685	3,510	97	1,985	2,627	3,394
48	2,011	2,682	3,505	98	1,984	2,627	3,393
49	2,010	2,680	3,500	99	1,984	2,626	3,392
50	2,009	2,678	3,496	∞	1,960	2,576	3,291

Tab. 4.II: Critical values F_{tab} for F -distribution with v_1 numerator d.f. and v_2 denominator d.f. for $\alpha = 5\%$.

v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	50	∞
v_2																
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.3	251.8	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.48	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.58	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.70	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.44	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.75	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.32	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.02	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.80	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.64	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.51	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.40	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.31	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.24	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.18	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.12	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.08	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.04	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.00	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	1.97	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	1.94	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.91	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.88	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.86	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.84	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.94	1.82	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.92	1.81	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.79	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.89	1.77	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.76	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.66	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.69	1.56	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.60	1.46	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.51	1.35	1.00

5. Designed experiments - one treatment factor

In the context of designed experiments, a number of key terms are frequently used. In particular, the terms **factor**, **variable** and **level** are often used in reference to treatments. A brief description of some of these terms is given here for quick reference.

A **factor** is generally some characteristic assumed to have an effect on the response. It is useful to distinguish treatment factors and blocking factors. **Blocking factors** pertain to the experimental units and to the grouping of such units, which is called blocking. Blocking may help increasing the efficiency of an experiment (section 5.2 and 5.5). Blocking factors are important in defining and establishing the experimental layout, but they are not themselves part of the research question. Examples: replicates, incomplete blocks, plots, animals, petri dishes, locations, barns, growth chambers, tables, tablets. **Treatment factors** are at the heart of the research question: we want to study treatment effects on the response. Examples: Type of feed in a feeding experiment and amount of fertilizer in a field experiment are treatment factors.

A factor generally has several **levels**. Example: When there are three blocks labelled 1, 2, 3, the factor block is said to have levels 1, 2 and 3. A treatment factor must have several levels that are systematically varied. Example: the factor fertilizer may have levels 0, 70 and 140 kg per ha. It is important that levels of the treatment factor must be randomly allocated to the experimental units. This process is known as randomization (sections 5.1 to 5.2). A particular level of a treatment factor is often referred to simply as treatment.

In order to represent a factor in analysis, we need to define a suitable **variable** to be entered in a dataset. The variable must have suitable levels. Example: To study the treatment factor “fertilizer” in an experiment with three different amounts of N-fertilizer (0, 70, 140), we may define a variable “fert” in an Excel spreadsheet with levels 0, 70 and 140.

A **treatment factor** may be either quantitative or qualitative. When the treatment factor is **qualitative**, we may perform analysis of variance, followed by mean comparison of means pertaining to the levels of the treatment factor (Sections 5.6 and 5.7). When the treatment factor is **quantitative**, it is usually preferable to perform a regression (section 5.8).

5.1 Randomization

The examples considered in this class so far were mainly drawn from **observational studies** and from surveys. In the biosciences, such observational studies undoubtedly have their role, and often the subject dictates this kind of study. Often, however, one may design an **experiment** to answer a research question. The main advantage of a designed experiment is that some of the explanatory variables, the so-called **treatments**, are under the control of the experimenter. Specifically, we may choose to keep certain factors constant throughout the experiment, while the levels of the **treatment factor** are varied systematically. Factors that cannot be kept constant and are not treatment factors, may be controlled by **randomization**, i.e., by random allocation of treatments to experimental units. This allows to reliably study the effect of treatment factors under controlled conditions.

Example 5.1: A researcher wants to study the effect of three different kinds of feed for pigs (barley-based, maize-based and potato-based). The study is to be performed on-farm. Consider the following two options:

(1) Observational study

The researcher selects a sample of farms and determines the types of feed used by farmers. In the sample, she selects those farms which use one of the three feeds of interest. In other words, the sample is subdivided into three groups depending on the type of feed used. Different traits, e.g. daily weight gain, are measured on the pigs from these farms. A one-way ANOVA is performed to detect group differences. Group differences are taken to be due to differences among the three types of feed.

(2) Designed experiment (on-farm)

The researcher prepares three types of feed. She finds a number of farms (experimental units) willing to participate in the experiment. The farms agree to use any of the three feeds. The researcher randomly allocates feeds to farms. In other words, three groups of farms are formed, one for each feed, and farms are randomly allocated to groups. This random allocation (**randomization**) is the main difference compared to the observational study. As in the observational study (1), different traits, e.g. daily weight gain, are measured on the pigs from these farms. A one-way ANOVA is performed to detect differences among the three types of feed.

The designed experiment has many advantages. Most importantly, significant differences among groups of farms are caused solely by differences in feed because of the **randomization**. By contrast, in the observational study it is quite possible that type of feed is correlated with other environmental factors. For example, farms feeding maize may tend to be located in regions with warmer climate than farms mainly feeding potatoes or barley. If differences among groups of farms are detected, it cannot be proved that these are due to the different feeds alone. In the worst case, the differences are entirely due to other factors. Such correlations do not exist in the designed experiment due to randomization.

Example 5.1 highlights the main advantage of designed experiments compared to observational studies: Randomized allocation of treatments to experimental units allows an unbiased assessment of treatment effects. Randomization, therefore, is perhaps the most important feature of a designed experiment.

Example 5.2: A researcher wants to compare five Sorghum cultivars (A, B, C, D, E) in a field experiment with four replications. A **replication** or **experimental unit** is made up of a **plot** of land, on which Sorghum is grown and harvested. Since there are five cultivars, 20 plots (experimental units) are needed. Hence, the experimental area is subdivided into twenty plots and the plots are numbered systematically from 1 to 20. The researcher considers the following two designs: systematic and randomized.

(1) Systematic

1 A	2 A	3 A	4 A
5 B	6 B	7 B	8 B
9 C	10 C	11 C	12 C
13 D	14 D	15 D	16 D
17 E	18 E	19 E	20 E

(2) Randomized

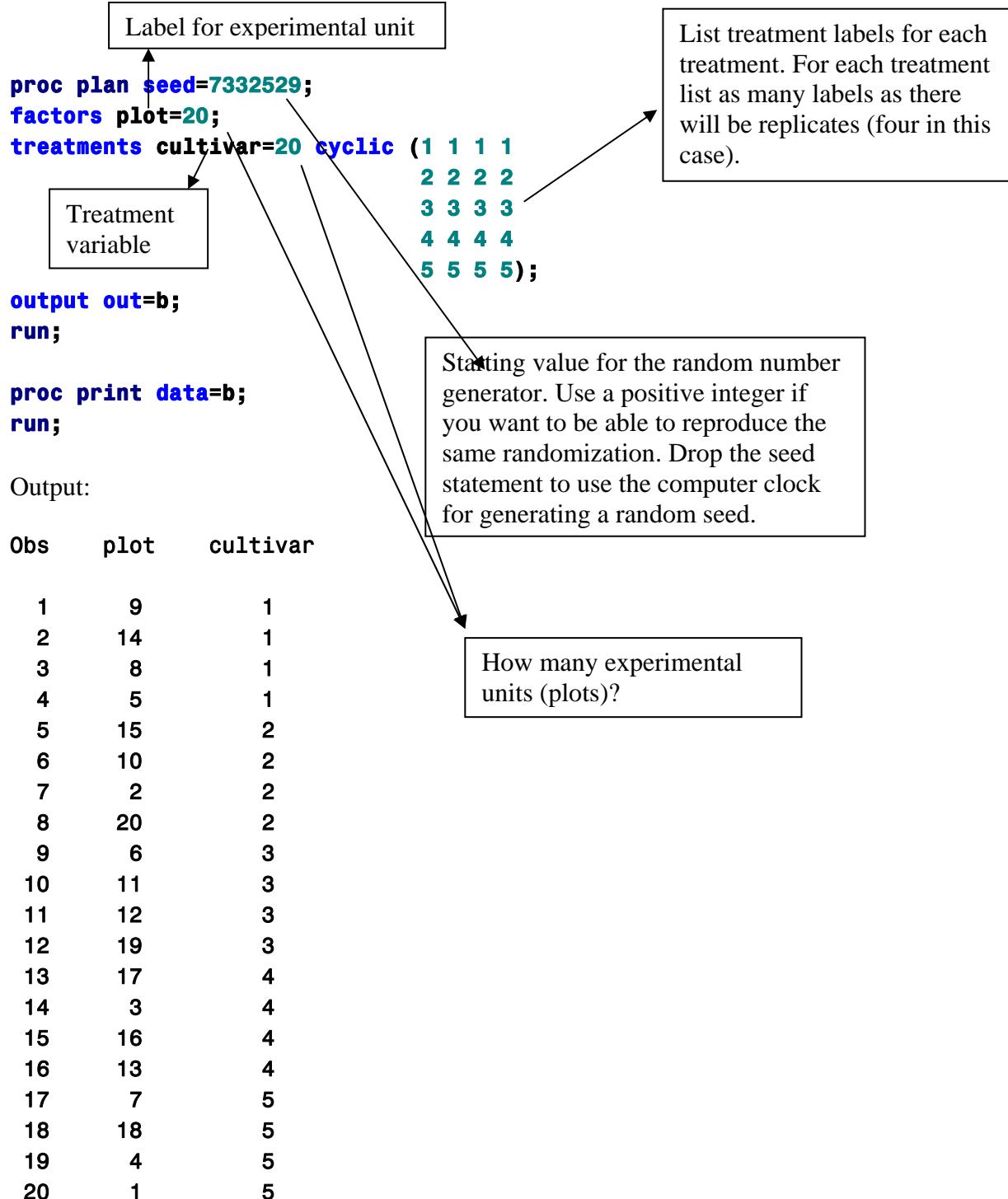
The five cultivars are randomly allocated to the 20 plots using random numbers (we will see below how this is achieved using SAS).

1 E	2 B	3 D	4 E
5 A	6 C	7 E	8 A
9 A	10 B	11 C	12 C
13 D	14 A	15 B	16 D
17 D	18 E	19 C	20 B

A disadvantage of the systematic design is that treatment effects may be confounded with fertility trends and other trends present in the experimental area. For example, if there is a trend of increasing fertility from top to bottom, cultivar A is at a disadvantage compared to

cultivar E. Differences in mean yield for cultivars A and E may, in fact, be due to soil differences alone, and not differences in the genotype. Randomization (second design) makes sure, that the distribution of cultivars among plots is fair/more even. It avoids biases that may occur in the systematic design.

The randomized design in Example 5.2 is called a **completely randomized design (CRD)**. The following code was used to generate the completely randomized design in Example 5.2:



The treatment labels here are 1 to 5. PROC PLAN does allow use of alpha-numeric labels A, B,, but implementation is more complicated. Thus, it is easier to use numeric coding 1, 2,

3, 4, 5, and then set 1 ≡ A, 2 ≡ B, 3 ≡ C, 4 ≡ D, and 5 ≡ E. For example, based on the output, we assign cultivar A (label 1) to plots 9, 14, 8 and 5, etc.

Example 5.1: Assume there are 36 participating farms, and we want to randomly allocate three feeds to the 36 farms. The following code will do the job:

```
proc plan seed=8935517;
factors farm=36;
treatments feed=36 cyclic (1 1 1 1 1 1 1 1 1 1 1 1
                           2 2 2 2 2 2 2 2 2 2 2 2
                           3 3 3 3 3 3 3 3 3 3 3 3);
output out=b;
run;

proc print data=b;
run;
```

It may be convenient to order the output by farms instead of by feeds. This is effected by using PROC SORT prior to PROC PRINT:

```
proc sort data=b;
by farm;
```

Exercise 5.1: Assume you want to test six different levels of nitrogen fertilizer in a greenhouse pot experiment with maize. Each fertilizer treatment is to be tested in 5 pots (experimental units). Generate a completely randomized design for this experiment using PROC PLAN.

Exercise 5.2: You want to perform an experiment with cows in which three different methods of husbandry are to be compared. You have a total of 24 cows. Randomly allocate cows to methods of husbandry using PROC PLAN.

5.2 Blocking

Complete randomization has the obvious disadvantage that all treatments may be allocated to the "good" experimental units purely by chance, in which case the experiment will yield a misleading result. For example, complete randomization might produce a design identical to the systematic design in Example 5.2. To avoid this, one can restrict randomization by **blocking**. In the simplest case, a block consists of as many experimental units as there are treatments, and each treatment occurs once in each block. Randomization is done separately within each block, i.e., treatments are randomly allocated to experimental units within blocks.

The randomized complete block design

Example 5.3: A field experiment to compare six melon cultivars with four replications each is arranged as follows:

Block 1	4	3	6	1	5	2
Block 2	2	1	5	6	4	3
Block 3	6	5	3	4	1	2
Block 4	1	4	6	3	2	5



Fig. 5.1: A randomized complete block design.

The design was generated by

```
proc plan seed=982261;
factors block=4 ordered cultivar=6;
run;
```

The design is called a **randomized complete block design (RCBD)**. Note that each treatment (melon cultivar) appears once in each block.

Not only does blocking avoid very uneven patterns of allocation to experimental units, it can also achieve a gain in accuracy. This gain will accrue if conditions within a block are more homogeneous than between blocks. In other words, blocks should be arranged so that **homogeneity within blocks is maximized**. The gain in accuracy is due to the fact that treatment comparisons can be done within a block.

Example 5.3: In field experiments, blocks are formed by adjacent plots, so that homogeneity among plots within a block is maximized. If there is a gradient in some important environmental factor, homogeneity within blocks is maximized by arranging blocks along the gradient. Plots within a block should be arranged to achieve the same goal, i.e., they should be placed orthogonal to the gradient (along "isolines") so each block is equally affected by the gradient as in Fig. 5.1. A gradient within a plot can be tolerated so long as the mean level of the environmental variable within a plot is the same for each plot within a block.

Blocking is useful even when there is no clear gradient, because adjacent plots tend to be more similar than distant plots. Also, blocking works as an insurance against unforeseen accidents. For example, if a disease enters the field from one side, damage may be restricted to a block or two, while the other blocks remain in tact. In this case, conditions within blocks will be more even than between blocks.

Example 5.4: In feeding experiments, blocks may be formed by groups of animals, where animals are grouped so that animals within a group are as homogeneous as possible. For

example, animals may be grouped according to their weight. Alternatively, animals from the same litter (born to the same mother) may be regarded as a block.

Example 5.5: A laboratory experiment is designed to compare the growth rate of bacteria on 20 different media. The media are brought out onto petri dishes and are then inoculated with bacteria. In the lab there is only one growth chamber, in which the petri dishes can be laid out. The capacity of the chamber is limited to 20 petri dishes. To test a minimum number of replications, the experiment must be replicated over time. It is natural to use time as a blocking factor. For each date, 20 petri dishes would be prepared, one for each growth medium. Each date corresponds to a block. By blocking experimental units in this way, differences between dates among experimental conditions can be controlled and separated from experimental error, thus improving the accuracy of the experiment.

Exercise 5.3: You want to test seven types of fertilizer for rice in four complete blocks. Generate a randomized field plan for this experiment using PROC PLAN.

The Latin square design

The randomized complete block design will provide control for one source of local variation. If there are several gradients or sources of local variation, one may combine different blocking systems.

Example 5.6 (non-biological, but useful nonetheless; taken from Mead et al., 1993): Suppose you want to compare four makes (brands) of tyre (A, B, C, and D, say) using just one car. To make the test fair, you want to test all four makes of tyre simultaneously, so on a trial, each make can be tested in one of the four possible positions. To obtain replication, you need to run several trials. Now the wear of a tyre will be dependent on the position, so it is not a good idea to test a given make of tyre in the same position on each trial. It is better to change positions in each trial. The task at the design stage, therefore, is to find an arrangement so that each make is tested once in each position and that each make is tested only once in any one trial. One possible arrangement is as follows:

	Trial 1	Trial 2	Trial 3	Trial 4
Front offside	A	B	C	D
Front near	C	A	D	B
Rear offside	D	C	B	A
Rear near	B	D	A	C

This is a **Latin square design**. Note that each make appears once in each position and once in each trial. There are two sources of local variation: (1) position on the car and (2) number of trial. The Latin square design above allows control of both sources of error by blocking.

Example 5.7: Four diets (A, B C, D) are to be tested on four cows. Replication will be available only if each cow tests more than one diet. One option is to have each cow test each diet. Cows here are an important source of variation, since differences in weight gain are expected among cows. Thus, we may use cows as one blocking factor by having each cow test each diet. Now one cow can only test one diet at a time, so we would need four trials replicated in time. If on the first trial, each cow tests the first diet, on the second, each cow

tests the second, etc., diet effects are confounded with time. For example, weather may differ between different trials, so differences in weight gain can be caused just by environmental variation associated with time. Clearly, time is a second important source of variation to be controlled. Blocking treatments in time to control this source of variation, we will want to test the whole set of treatments (diets) at each time. Again, a Latin square will achieve this goal.

	Cow			
	1	2	3	4
Period 1	A	B	C	D
Period 2	B	D	A	C
Period 3	C	A	D	B
Period 4	D	C	B	A

To avoid carry-over effects, one will need to allow for some time of adjustment between the four periods.

The two blocking variables in a Latin square may generally be denoted as **rows** and **columns** for obvious reasons. Periods are rows and cows are columns in Example 5.6. Randomization of a Latin square design proceeds in three independent steps:

- (1) Randomization of rows
- (2) Randomization of columns
- (3) Randomization of treatments

Each randomization is performed independently using random numbers. The following code, which is somewhat complex, does this for Example 5.5:

```
proc plan seed=614871;
factors trial=4 ordered position=4 ordered;
treatments make=4 cyclic;
output out=b;
trial    cvals=( 'Trial1' 'Trial2' 'Trial3' 'Trial4') random
position cvals=( 'Front offside' 'Front near'
                 'Rear offside' 'Rear near' )           random
make      cvals=( 'A' 'B' 'C' 'D' )           random;
run;

proc print data=b;
run;
```

Output:

Obs	trial	position	make
1	Trial3	Rear near	B
2	Trial3	Front offside	A
3	Trial3	Front near	C
4	Trial3	Rear offside	D
5	Trial4	Rear near	A
6	Trial4	Front offside	C

7	Trial4	Front near	D
8	Trial4	Rear offside	B
9	Trial2	Rear near	C
10	Trial2	Front offside	D
11	Trial2	Front near	B
12	Trial2	Rear offside	A
13	Trial1	Rear near	D
14	Trial1	Front offside	B
15	Trial1	Front near	A
16	Trial1	Rear offside	C

5.3 Replication and balance

Replication is a requirement in any experiment, for without replication it is impossible to disentangle systematic treatment effects and random environmental variation. Also, an analysis of variance can be performed only with replicated data, because without replication we cannot estimate the variance. **Generally, replicates correspond to randomization units**, e.g., plots within blocks (field experiment), pots (greenhouse), animals, litters, petri-dishes, etc. It is very important to note that, statistically speaking, an experimental unit becomes a replicate not by virtue of its physical properties (e.g., a "plot", "animal"), but by randomization, i.e., randomized allocation of treatments to experimental units.

In designed experiments it is often feasible to have the same number of replications per treatment. Such designs are said to be **balanced**. Balance is desirable mainly because treatment comparisons can be performed with the same accuracy for each treatment. That is, the standard error of a difference does not depend on the pair of treatments. Recall from section 4.1, that the **standard error of the difference** among two treatment (group) means in a one-way ANOVA set-up is

$$s.e.d. = s.e.(\hat{\mu}_1 - \hat{\mu}_2) = s.e.(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}) = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where s is the sample standard deviation (square root of ANOVA error mean square) and n_1 , n_2 are the sample sizes of treatments 1 and 2. If the sample size per treatment is constant, i.e.,

$$n_1 = n_2 = \dots = n$$

the design is balanced and the $s.e.d.$ is constant for each pair of treatments. It simplifies to

$$s.e.d. = \sqrt{\frac{2s^2}{n}}$$

In the pre-computer age, balance was needed for a second reason, i.e., computational ease. There are simple computational formulae for the sums of squares in the analysis of variance and for least squares estimates of treatment means only for balanced data. These simple equations are not applicable for unbalanced data.

5.4 Pseudo-replication and true replication

Example 5.8: An experiment is planned to compare three herbicides for cabbage. For each treatment (herbicide) there are ten cabbage plants. Yield is assessed on a plant basis, so there are ten measured yields per treatment.

Design (1) (Fig. 5.2):

The experimental area could be subdivided into three plots, one for each herbicide, and ten plants planted per plot. Randomization would involve randomly allocating the three herbicides to the three plots (Fig. 5.2). The experiment could be analysed by ANOVA regarding the ten plants as ten replications. This analysis would be highly misleading, however, because randomization did not extend to plants within a plot, i.e., the same treatment was applied to each plant within a plot. If there is a fertility gradient between plots, we might detect highly significant differences between "treatments", where, in fact, difference are due to soil differences, not due to treatments. The clue here is that plants within a plot are no real replications. True replications will be randomization units, i.e., plots in this case. The design in Fig. 5.2 has no true replication, since there is only one plot per treatment.

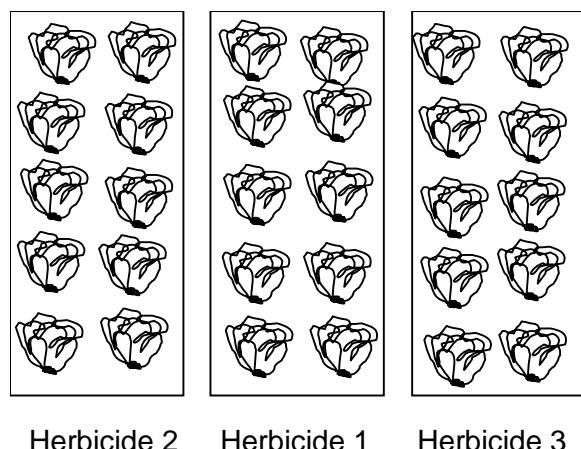


Fig. 5.2: A design without replication.

To fix the problem, we could regard the thirty plants as randomization units and allocate treatments to plants completely at random or according to a blocked design. This would yield a valid design from a statistical point of view, but there is a disadvantage in choosing plants instead of plots as randomization units. We need to expect border effects between adjacent plants receiving different treatments. For this reason, plants are seldom chosen as randomization units, except perhaps for large plant species (trees). To reduce border effects, a plot of several plants receives the same treatment and plots are randomized instead of plants. Thus, if we want a replicated design, several plots are needed per treatment. The minimum is two plots per treatment, as in Fig. 5.3.

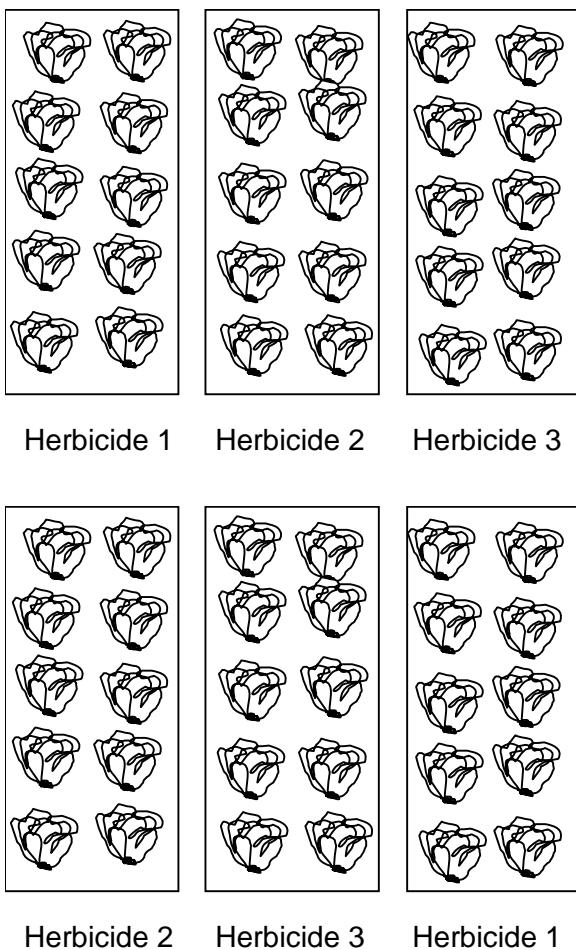


Fig. 5.3: A randomized complete block design with two "true" replications per treatment.

Statistical analysis would proceed in two steps:

- (1) Compute average yield of the ten plants on a plot
- (2) Perform a statistical analysis based on plot means

One should **not** do an ANOVA on the individual plant yields, for this would falsely regard plants as true replicates. Clearly, plants within a plot are **pseudo-replicates**.

5.5 Incomplete blocks

Often, the number of treatments is so large that the objective of obtaining homogeneous blocks cannot be achieved.

Example 5.9: If 100 cultivars are to be tested in complete blocks, a block would be made up of 100 plots. Heterogeneity increases with the block area, and experience shows that with 100 plots per block, blocking is virtually ineffective in controlling experimental error. For this reason, several alternative designs with incomplete blocks are often used. For example, plant breeders commonly use so-called lattice designs or α -designs to test new lines or cultivars. Both of these designs involve incomplete blocks.

Example 5.10: Suppose we plan an experiment, in which several tasters are to evaluate different types of yoghurt. To control experimental error, ideally, we would want to have each taster evaluate each yoghurt, so tasters could be used as a blocking variable (each taster would be a complete block). Suppose there are four tasters (1, 2, 3, 4) and four yoghurts (A, B, C, D). Each taster would have to test four yoghurts. It is known, however, that the quality of the tasters assessment will become inaccurate if too many products are tested by the same person. Suppose that each taster can test only three yoghurts. How should we design the experiment? Some thought reveals that a design as the following is inevitable:

	Taster			
	1	2	3	4
Yoghurts	A	A	A	B
	B	B	C	C
	C	D	D	D

Note that each yoghurt gets replicated three times, even though the blocks are incomplete. Also note, that each pair of treatments appears together in the same block the same number of times. The design belongs to the large class of **balanced incomplete block designs (BIBD)**. The balance makes sure that each treatment difference can be estimated with the same accuracy (s.e.d.).

Randomization of incomplete blocks involves three steps:

- (1) Randomized allocation of treatments to treatment labels
- (2) Randomization of block order
- (3) Randomization of treatments within blocks

If you plan to use incomplete blocks, I strongly suggest you consult a statistician.

The Latin square design, a design with two blocking factors (generally denoted as rows and columns), is quite restrictive in that the number of rows and columns and hence the number of replications must equal the number of treatments. There are many alternative designs with rows and columns, for which there may be incomplete rows and/or incomplete columns. Designs arranged in rows and columns are generally called **row-column-designs**. A Latin square is the simplest form of a row-column design. Space does not allow covering other row-column designs in detail. Suffice it to say that there is great flexibility in obtaining row-column designs.

5.6 Statistical analysis of experiments with one qualitative treatment factor

5.6.1 Completely randomized design

The completely randomized design may be analysed by one-way ANOVA, with which we are familiar from sections 4.1 and 4.2. The model is

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad (5.1)$$

where

y_{ij} = measurement of j -th replicate of i -th treatment

α_i = effect of i -th treatment

In designed experiments, the number of replicates is usually the same for each treatment, i.e., the data are balanced. In this case, the standard error of a difference (*s.e.d.*) can be estimated by

$$s.e.d. = \sqrt{\frac{2s^2}{n}}$$

where s^2 is the ANOVA error mean square and n is the number of replicates per treatment. The t-statistic for comparing two means is given by

$$t_{\text{exp}} = \frac{|\bar{d}|}{s.e.d.}$$

where

$$\bar{d} = \bar{y}_{i_1} - \bar{y}_{i_2}$$

The t-test will be judged significant when

$$t_{\text{exp}} = \frac{|\bar{d}|}{s.e.d.} > t_{\text{tab}}$$

where t_{tab} is the tabular t-value with error d.f. and significance level α . This can be rearranged to yield

$$|\bar{d}| > t_{\text{tab}} \times s.e.d.$$

Thus, a difference that exceeds the critical difference value of $t_{\text{tab}} \times \text{s.e.d.}$, is significant by the t-test. The critical difference is usually referred to as **least significant difference (LSD)**:

$$LSD = t_{\text{tab}} \times s.e.d.$$

It needs to be stressed that a common LSD is available for balanced data only.

Example 4.3 (Mead et al., p. 52; **melons.dat**): An experiment was performed to compare four melon varieties. Each variety was tested on six plots. The allocation of treatments (varieties) to experimental units (plots) was completely at random (**completely randomized design**; see Chapt. 5). The yields were as follows:

Variety	v1	v2	v3	v4
Yields	25.12	40.25	18.30	28.55
	17.25	35.25	22.60	28.05
	26.42	31.98	25.90	33.20
	16.08	36.52	15.05	31.68
	22.15	43.32	11.42	30.32
	15.92	37.10	23.68	27.58
Mean (\bar{y}):	20.49	37.40	19.49	29.90

The objective of the analysis is to compare the six variety means. The ANOVA table is as follows:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1291.477146	430.492382	23.42	<.0001
Error	20	367.653150	18.382658		
Corrected Total	23	1659.130296			

From the ANOVA we find $s^2 = 18.32$. For 20 error d.f. the tabular t-value is 2.086 (see Tab. 4.I). Thus

$$LSD = 2.086 \times \sqrt{\frac{2 \times 18.32}{6}} = 5.16$$

The means can be compiled into a table as follows:

Variety	Mean
v1	20.49
v2	37.40
v3	19.49
v4	29.90
LSD(5%)	5.16

It is easily checked that all pairwise comparisons are significantly different, except the one between varieties A and C.

Mean comparisons can be further facilitated by the so-called lines or letters display, in which means which are not significantly different, are followed by the same letter. The letters free the reader from the need to compute treatment differences.

To generate a lines display, the first step is to arrange treatments **ordered by means** (!) in a cross tabulation. The cells are filled with the mean differences. The differences are compared to the LSD and significant differences marked with an asterisk (*).

	v3	v1	v4	v2
	19.49	20.49	29.90	37.40
v3		1.00	10.41*	17.91*
V1			9.41*	16.91*
v4				7.50*
v2				

$$\text{LSD}(5\%) = 5.16$$

All comparisons are performed above the diagonal of the cross-tabulation (convince yourself that all comparisons do, in fact, appear above the diagonal). Next, treatment labels are arranged horizontally ordered by means:

v3 v1 v4 v2

Now we go through the cross-tabulation row-by-row. For each row, we draw a line underneath the ordered treatment labels, connecting treatments that are not significantly different. For each row, the line starts at the treatment corresponding to that row. For the first row (C), the line starts at "C". The line can be drawn up to "A", because A and C are not significantly different. It cannot be drawn further, however, because C is significantly different from D and B.

v3 v1 v4 v2

For the next row (A), we underline A. The line cannot be drawn further, because A differs significantly from both D and B.

v3 v1 v4 v2

The next row is for D. We underline D but cannot go further because D differs significantly from B.

v3 v1 v4 v2

Finally, we underline B. The line cannot go further because B is the last treatment.

v3	v1	v4	v2
—	—	—	—

The lines have the following interpretation: Treatments connected by a line are not significantly different. The second line is obviously redundant, because it is contained in or fully covered by the first line. Thus, the second line can be dropped.

v3	v1	v4	v2
—	—	—	—

Finally, we assign a small-case letter to each of the remaining lines and transfer the letters to the means table.

v3	v1	v4	v2
—	a	—	b
		—	c

Variety	Mean
v1	20.49 ^a
v2	37.40 ^c
v3	19.49 ^a
v4	29.90 ^b
LSD(5%)	5.16

Means followed by the same letter are not significantly different at $\alpha = 5\%$.

Lines displays are automatically produced by the GLM procedure, when the LSD option is used with the MEANS statement. The code for the melon data is as follows:

```

data;
input
yield      variety$;
datalines;
25.12      v1
17.25      v1
26.42      v1
16.08      v1
22.15      v1
15.92      v1
40.25      v2
35.25      v2
31.98      v2
36.52      v2
43.32      v2
37.10      v2
18.30      v3
22.60      v3
25.90      v3
15.05      v3

```

```

11.42      v3
23.68      v3
28.55      v4
28.05      v4
33.20      v4
31.68      v4
30.32      v4
27.58      v4
;
proc glm;
class variety;
model yield=variety;
lsmeans variety/pdiff;
run;

```

The \$-sign needs to be placed after the variable name for variety, because varieties are coded by A, B, C, and D (alpha-numeric coding) rather than by numbers (numeric coding).

Exercise 5.4 (Mead et al., p. 54): The percentage moisture content is determined from 10 samples for each of four different soils (also see **moisture.dat**). Compute an ANOVA and perform multiple comparisons using the LSD test.

Soil	S1	S2	S3	S4
Percentage moisture	12.8	8.1	9.8	16.4
	13.4	10.3	10.6	8.2
	11.2	4.2	9.1	15.1
	11.6	7.8	4.3	10.4
	9.4	5.6	11.2	7.8
	10.3	8.1	11.6	9.2
	14.1	12.7	8.3	12.6
	11.9	6.8	8.9	11.0
	10.5	6.9	9.2	8.0
	10.4	6.4	6.4	9.8
Mean ($\bar{y}_{i\bullet}$):	11.56	7.69	8.94	10.85

It should re-iterated that a common LSD as well as a lines/letters display is available only with balanced data. Thus, the MEANS statement should be used only with balanced data. For unbalanced data, one should generally use the LSMEANS statement (Section 4.2). With the procedure GLIMMIX, a lines display can be generated with the LSMEANS statement. But when the standard errors of a difference (s.e.d.) are not constant, the lines display may sacrifice some significance statements. The code for the Example 4.3 (which is balanced) is as follows:

```

proc glimmix;
class variety;
model yield=variety;
lsmeans variety/pdiff lines;
run;

```

variety Least Squares Means

variety	Estimate	Standard		DF	t Value	Pr > t
		Error				
v1	20.4900	1.7504		20	11.71	<.0001
v2	37.4033	1.7504		20	21.37	<.0001
v3	19.4917	1.7504		20	11.14	<.0001
v4	29.8967	1.7504		20	17.08	<.0001

Differences of variety Least Squares Means

variety	_variety	Estimate	Standard		DF	t Value	Pr > t
			Error				
v1	v2	-16.9133	2.4754		20	-6.83	<.0001
v1	v3	0.9983	2.4754		20	0.40	0.6910
v1	v4	-9.4067	2.4754		20	-3.80	0.0011
v2	v3	17.9117	2.4754		20	7.24	<.0001
v2	v4	7.5067	2.4754		20	3.03	0.0066
v3	v4	-10.4050	2.4754		20	-4.20	0.0004

T Grouping for variety Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

variety	Estimate
v2	37.4033
v4	29.8967
v1	20.4900
v3	19.4917

Exercise 5.4 (continued): Delete observations with moisture > 12 and use PROC GLIMMIX to derive a lines display. Verify that this is not possible with GLM.

5.6.2 Randomized complete block design

The linear model needs to have a block effect in addition to the treatment effect to account for the design, i.e.,

$$y_{ij} = \mu + b_j + \alpha_i + e_{ij} \quad (5.2)$$

where

y_{ij} = measurement of j -th replicate of i -th treatment

α_i = effect of i -th treatment

b_j = effect of j -th block

Since main interest is in treatment effects, the analysis of variance is based on the following sequence of models:

Model	Error-SS	Reduction in SS
$y_{ij} = \mu + e_{ij}$	$SS(\mu)$	
$y_{ij} = \mu + b_j + e_{ij}$	$SS(\mu, b_j)$	$RSS(b_j \mu) = SS(\mu) - SS(\mu, b_j)$
$y_{ij} = \mu + b_j + \alpha_i + e_{ij}$	$SS(\mu, b_j, \alpha_i)$	$RSS(\alpha_i \mu, b_j) = SS(\mu, b_j) - SS(\mu, b_j, \alpha_i)$

For unbalanced data, it is important that blocks are fitted first and treatments last. **Generally, the appropriate test of an effect is obtained by fitting that effect last in a sequence.** For balanced data, the order of fitting is immaterial, because the resulting SS are the same, i.e.,

$$RSS(\alpha_i | \mu) = RSS(\alpha_i | \mu, b_j)$$

$$RSS(b_j | \mu) = RSS(b_j | \alpha_i, \mu).$$

The sums of squares are compiled into an ANOVA table as usual:

Source	d.f.	SS	MS
Blocks, ignoring treatments	$r-1$	$RSS(b_j \mu)$	$RSS(b_j \mu)/(r-1)$
Treatments, adj. for blocks	$t-1$	$RSS(\alpha_i \mu, b_j)$	$RSS(\alpha_i \mu, b_j)/(t-1)$
Error	$N-t-r+1$	$SS(\mu, b_j, \alpha_i)$	$s^2 = SS(\mu, b_j, \alpha_i)/(N-t-r+1)$

r = number of blocks; t = number of treatments; N = total number of observations

Balanced data

For balanced data, least square treatment means are equal to the simple sample means $\bar{y}_{i\bullet}$, and the s.e.d. is constant for each pair of treatments. Thus, multiple comparisons by an LSD test are done as in case of the completely randomized design. We can use the MEANS statement to perform the comparisons.

Example 5.11: An RCB experiment was conducted to assess the yield (kg/ha) of rice cultivar IR8 at six different seeding densities (kg/ha) (Gomez and Gomez, 1984) (**ir8.dat**):

Density (kg/ha)	Block			
	1	2	3	4
25	5113	5398	5307	4678
50	5346	5952	4719	4264
75	5272	5713	5483	4749
100	5164	4831	4986	4410
125	4804	4848	4432	4748
150	5254	4542	4919	4098

Important: The treatment factor is a quantitative variable. Thus, a regression analysis would be useful. In fact, regression is more efficient than ANOVA followed by multiple comparison of means, whenever the treatment variable is quantitative. We will come back to the present example when introducing polynomial regression in section 5.8.

The analysis is performed in GLM as follows:

```

data;
input density block yield;
datalines;
25    1    5113
25    2    5398
25    3    5307
25    4    4678
50    1    5346
50    2    5952
50    3    4719
50    4    4264
75    1    5272
75    2    5713
75    3    5483
75    4    4749
100   1    5164
100   2    4831
100   3    4986
100   4    4410
125   1    4804
125   2    4848
125   3    4432
125   4    4748
150   1    5254
150   2    4542
150   3    4919
150   4    4098
;
proc glm;
class block density;
model yield=block density;
means density/lsd;
run;

```

Note that the model statement contains the block term in addition to the treatment term.

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	3142691.667	392836.458	3.55	0.0165
Error	15	1658376.167	110558.411		
Corrected Total	23	4801067.833			
		R-Square	Coeff Var	Root MSE	yield Mean
		0.654582	6.704258	332.5032	4959.583
Source	DF	Type I SS	Mean Square	F Value	Pr > F
block	3	1944360.833	648120.278	5.86	0.0074
density	5	1198330.833	239666.167	2.17	0.1128
Source	DF	Type III SS	Mean Square	F Value	Pr > F
block	3	1944360.833	648120.278	5.86	0.0074
density	5	1198330.833	239666.167	2.17	0.1128

The output contains two sets of SS, namely Type I SS and Type III SS. These are identical in the present case, because the data are balanced. Generally, the Type I SS corresponds to the sequential reduction in error SS. The order in which terms are fitted is determined by the order in which terms are listed in the MODEL statement. Type III SS are constructed in a different way. We will discuss Type III SS later in conjunction with experiments involving more than one treatment factor. Here, we merely focus in Type I SS.

The F-test for treatments is not significant. Nevertheless, we look at the multiple t-tests:

Alpha	0.05
Error Degrees of Freedom	15
Error Mean Square	110558.4
Critical Value of t	2.13145
Least Significant Difference	501.14

Means with the same letter are not significantly different.

t Grouping		Mean	N	density
	A	5304.3	4	3
	A			
B	A	5124.0	4	1
B	A			
B	A	5070.3	4	2
B	A			
B	A	4847.8	4	4
B				
B		4708.0	4	5
B				
B		4703.3	4	6

From the output, we obtain the following table of means:

Density (kg/ha)	Mean (kg/ha)
25	5124.0 ^{ab}
50	5070.3 ^{ab}
75	5304.3 ^a
100	4847.8 ^{ab}
125	4708.0 ^b
150	4703.3 ^b

LSD(5%) 501.1

Means followed by the same letter are not significantly different at the 5% level of significance.

Unbalanced data

The expected value of an observation in the RCB design is

$$E(y_{ij}) = \eta_{ij} = \mu + b_j + \alpha_i$$

It is natural to define the treatment mean as the average of η_{ij} across blocks, i.e.,

$$\bar{\eta}_{i\bullet} = \mu + \alpha_i + \bar{b}_\bullet = \mu + \alpha_i + \frac{b_1 + b_2 + \dots + b_r}{r}$$

When it comes to the estimation of $\bar{\eta}_{i\bullet}$, it matters whether or not the data are balanced. In the case of balanced data the least squares estimator for this mean is

$$\hat{\bar{\eta}}_{i\bullet} = \bar{y}_{i\bullet} ,$$

which is just the simple mean of observations for the i -th treatment. This simple mean is no longer appropriate, however, when there are missing data. This is best illustrated using a simple example.

Example 5.12: Consider the following hypothetical results of an RCB experiment with three cultivars and two blocks:

	Cultivar		
	1	2	3
Block	1	10	20
	2	20	30
	3	60	70
Cultivar mean $\bar{y}_{i\cdot}$:	30	40	50

Treatment 3 has the largest mean and would be judged the best. We have simplified the data by assuming that there is no experimental error and that the additive model holds exactly. Thus, differences among treatments are exactly the same in each block. No one will doubt that the simple mean $\bar{y}_{i\cdot}$ is a useful measure for the performance of a cultivar. As we have noted before, the simple means are least squares estimates of $\bar{\eta}_i$, in case of balanced data.

Now assume that the observation for cultivar 3 in block 3 is missing, i.e., the data are

	Cultivar		
	1	2	3
Block	1	10	20
	2	20	30
	3	60	70
Cultivar mean $\bar{y}_{i\cdot}$:	30	40	35

We have computed simple means as before. Due to the missing value, the mean for cultivar 3 has now changed. Its mean now lies between those for cultivars 1 and 2. Thus, were we to base our assessment on simple means, cultivar 3 would be judged differently compared to the balanced case. In fact, the comparison of cultivars would not be fair, because an observation from the most favourable block is missing for cultivar 3. Obviously, simple (unadjusted) means are misleading in the case of missing data.

A natural thing to do is to compare cultivars within blocks. The hypothetical data are such that, e.g., the difference of cultivar 3 minus cultivar 1 is 20 in blocks 1 and 2. Thus, we would expect the difference to be the same in block 3. We have no observation for cultivar 3 in block 3 and thus cannot verify this based on observed data. Instead, we can ask: "What would have been the most likely yield of cultivar 3 in block 3?"

Since the yield of cultivar 1 is 60, we would expect a yield of $60+20 = 80$ for cultivar 3. To obtain a corrected or adjusted mean for cultivar 3, we can plug in the imputed value into our table and compute simple means, as before:

	Cultivar			
	1	2	3	imputed value
Block	1	10	20	30
	2	20	30	40
	3	60	70	80
Cultivar mean:	30	40	50	adjusted mean!

Due to the simplicity of the example, the same result is found for the comparison of cultivars 2 and 3.

Example 5.13: The first example was rather artificial because absence of experimental error was assumed. Now we add some error by perturbing the original example as follows:

	Cultivar		
	1	2	3
Block	1	12	18
	2	18	32
	3	57	73
Cultivar mean $\bar{y}_{i \cdot}$:	29	41	35

Differences among cultivars now are not the same from block to block due to (simulated) experimental error. The best we can do regarding the comparison of cultivar 3 with the other cultivars is to analyse the first two blocks, discarding the third:

	Cultivar		
	1	2	3
Block	1	12	18
	2	18	32
Cultivar mean $\bar{y}_{i \cdot}$:	15	25	35

The difference of cultivar 3 minus cultivar 1 is 20, while the difference of cultivar 3 minus cultivar 2 is 10. The difference between cultivar 1 and 2 is 10. To impute the missing value, this result can be used. The most plausible value for the missing cell in the third block is the one that is in best agreement with the differences computed from the first two blocks. Things are more complicated now, because the difference of cultivars 1 and 2 is 8 in block 3, which is not the same as the difference found from the means based on the first two blocks. An intuitive approach is to plug in a value for the missing cell such that the average of the differences 3–1 and 3–2 is the same in block 3 as that obtained for means from the two first blocks. For the further development it is helpful to denote the imputed value as m :

	Cultivar		
	1	2	3
Block	1	12	18
	2	18	32
	3	57	73

We now compute differences using the variable m for the missing value:

	3 minus 1	3 minus 2	average
Means across blocks 1 and 2	20	10	15
Data in block 3	$m - 57$	$m - 73$	$m - 65$

Our requirement was that the average difference in block 3 is the same as the average difference based on means across blocks 1 and 2. Thus, our requirement yields

$$15 = m - 65 \Leftrightarrow m = 80$$

Using this imputed value to complete the table we find

	Cultivar			
	1	2	3	imputed value
Block	1	12	18	37
	2	18	32	33
	3	57	73	80

Cultivar mean:	29	41	50	adjusted mean!
----------------	----	----	----	----------------

Exercise 5.4b: An alternative procedure to impute the missing cell in example 5.13 is to consider the four "tetrads" that can be formed between the missing cell and three other cells. A tetrad corresponds to a two-by-two table of means for two cultivars and two blocks. There are four such tables involving block 3 and cultivar 3. For example, one possible tetrad is given by the following table:

	Cultivar	
	2	3
Block	2	32
	3	m

The imputed value based on this tetrad is $m = 74$. Impose the m from each of the four tetrads. Verify that the mean of these different imputed values for m equals 80.

We have looked at two simple examples, where we computed "adjusted means" by simple and intuitive algebra. A general method that will yield the same result in the present case, is the method of least squares. In the general case, no simple scalar expressions can be given for the least squares solutions for the parameters, so details will not be given here. Just assume that least squares solutions can be obtained using a computer and denote these by the hat-notation as $\hat{\mu}$, \hat{b}_j , and $\hat{\alpha}_i$. Then, the least squares estimator of the mean $\bar{\eta}_i$ is given by

$$\hat{\eta}_{i\bullet} = \hat{\mu} + \hat{\alpha}_i + \hat{b}_\bullet = \hat{\mu} + \hat{\alpha}_i + \frac{\hat{b}_1 + \hat{b}_2 + \dots + \hat{b}_r}{r}$$

This estimator is also known as the **least squares mean** or **adjusted mean**. In the case of balanced data, the least squares mean will coincide with the simple mean (unadjusted mean), but for unbalanced data this is no longer the case, i.e., we have

$$\hat{\eta}_{i\bullet} = \hat{\mu} + \hat{\alpha}_i + \hat{b}_\bullet = \hat{\mu} + \hat{\alpha}_i + \frac{\hat{b}_1 + \hat{b}_2 + \dots + \hat{b}_r}{r} \neq \bar{y}_{i\bullet}$$

The adjusted means computed for the two simple examples are, in fact, the least squares means. This will now be verified using SAS.

As we have seen before, simple (unadjusted) means can be obtained by the MEANS statement. To obtain least squares means (adjusted means), we need to use the LSMEANS statement. This is illustrated starting with the first example (complete data):

```
data;
input block cultivar yield;
datalines;
1 1 10
1 2 20
1 3 30
2 1 20
2 2 30
2 3 40
3 1 60
3 2 70
3 3 80
;
proc glm;
class cultivar block;
model yield=block cultivar;
means cultivar;
lsmeans cultivar;
run;
```

Output for MEANS statement:

General Linear Models Procedure

Level of		-----YIELD-----	
CULTIVAR	N	Mean	SD
1	3	30.0000000	26.4575131
2	3	40.0000000	26.4575131
3	3	50.0000000	26.4575131

Output for LSMEANS statement:

General Linear Models Procedure
Least Squares Means

CULTIVAR	YIELD LSMEAN
1	30.0000000
2	40.0000000
3	50.0000000

We see that in the balanced case MEANS and LSMEANS yield the same result. Next we do the same analysis with the last observation missing:

```
data;
input block cultivar yield;
datalines;
1 1 10
1 2 20
1 3 30
2 1 20
2 2 30
2 3 40
3 1 60
3 2 70
3 3 .
;
proc glm;
class cultivar block;
model yield=block cultivar;
means cultivar;
lsmeans cultivar;
run;
```

Output for MEANS statement:

General Linear Models Procedure

CULTIVAR	N	-----YIELD-----	
		Mean	SD
1	3	30.0000000	26.4575131
2	3	40.0000000	26.4575131
3	2	35.0000000	7.0710678

Output for LSMEANS statement:

General Linear Models Procedure
Least Squares Means

CULTIVAR	YIELD LSMEAN
1	30.0000000
2	40.0000000
3	50.0000000

Here, the least squares mean and the simple mean for cultivar 3 are not the same. Now on to the second example:

```

data;
input block cultivar yield;
datalines;
1 1 12
1 2 18
1 3 37
2 1 18
2 2 32
2 3 33
3 1 57
3 2 73
3 3 .
;
proc glm;
class cultivar block;
model yield=block cultivar;
means cultivar;
lsmeans cultivar;
run;

```

Output for MEANS statement:

General Linear Models Procedure

CULTIVAR	N	-----YIELD-----	
		Mean	SD
1	3	29.0000000	24.4335834
2	3	41.0000000	28.5832119
3	2	35.0000000	2.8284271

Output for LSMEANS statement:

General Linear Models Procedure Least Squares Means

CULTIVAR	YIELD LSMEAN
1	29.0000000
2	41.0000000
3	50.0000000

Again, the least squares (adjusted) mean and the simple (unadjusted) mean for cultivar 3 are not the same.

Exercise 5.5: Verify the above computations.

Example 5.11 (continued): We use the seeding density experiment with rice from above (**ir8.dat**), but assume that the last observation is missing, so the data are unbalanced. The missing observation is generated by replacing the last observation with a dot ("."), the symbol for missing data in SAS. For analysis, we can use exactly the same code as before. We want to fit treatments after blocks, so the treatment factor appears after the block factor in the MODEL statement.

```

data;
input density block yield;
datalines;
25    1    5113
25    2    5398
25    3    5307
25    4    4678
50    1    5346
50    2    5952
50    3    4719
50    4    4264
75    1    5272
75    2    5713
75    3    5483
75    4    4749
100   1    5164
100   2    4831
100   3    4986
100   4    4410
125   1    4804
125   2    4848
125   3    4432
125   4    4748
150   1    5254
150   2    4542
150   3    4919
150   4    .
;
proc glm;
class block density;
model yield=block density;

```

```
run;
```

Output:

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2398048.16763281	299756.02095410	2.58	0.0581
Error	14	1628418.78888893	116315.62777778		
Corrected Total	22	4026466.95652174			

R-Square	C.V.	Root MSE	YIELD Mean
0.595571	6.825051	341.05077009	4997.04347826

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BLOCK	3	1355255.98985508	451751.99661836	3.88	0.0327
DENSITY	5	1042792.17777778	208558.43555556	1.79	0.1789

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BLOCK	3	1485881.46111112	495293.82037037	4.26	0.0247
DENSITY	5	1042792.17777778	208558.43555556	1.79	0.1789

Since treatments are fitted after fitting blocks, the treatment sum of squares, $RSS(\alpha_i | b_j, \mu) = 1042792.18$ (look under Type I SS!), is corrected/adjusted for block effects. Thus, $RSS(\alpha_i | b_j, \mu)$ is sometimes referred to as "**corrected treatment sum of squares**".

To illustrate the importance of fitting order, we repeat the analysis, reversing the order of fitting for blocks and treatments. We now fit treatments prior to blocks. This is effected by interchanging the order of BLOCK and TRT in the MODEL statement.

```
proc glm;
class block density;
model yield=density block;
run;
```

Output:

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2398048.16763281	299756.02095410	2.58	0.0581
Error	14	1628418.78888893	116315.62777778		
Corrected Total	22	4026466.95652174			

R-Square	C.V.	Root MSE	YIELD Mean
0.595571	6.825051	341.05077009	4997.04347826

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DENSITY	5	912166.70652174	182433.34130435	1.57	0.2325
BLOCK	3	1485881.46111112	495293.82037037	4.26	0.0247
Source	DF	Type III SS	Mean Square	F Value	Pr > F
DENSITY	5	1042792.17777778	208558.43555556	1.79	0.1789
BLOCK	3	1485881.46111112	495293.82037037	4.26	0.0247

Since we have fitted treatments before blocks, the sum of squares for treatments, $RSS(\alpha_i|\mu) = 912166.7$ (look under Type I SS!) is not corrected for blocks. It is therefore denoted as "**uncorrected**" treatment sum of squares. Note that the uncorrected treatment SS is different from the corrected treatment SS, because treatment effects are confounded with block effects. Thus the uncorrected treatment SS is not appropriate for testing treatment effects. The SS for blocks, however, is corrected for treatments, so this analysis provides the correct test for blocks.

The least squares means may be compared by multiple t-tests. When data are unbalanced, there is no common s.e.d. and thus no common LSD for all comparisons. Instead, each comparison has its own s.e.d. and its own LSD. To obtain multiple t-tests from the LSMEANS statement in GLM, add the option /PDIFF as follows:

```
proc glm;
class block density;
model yield=block density;
lsmeans density/pdiff;
run;
```

Output:

The GLM Procedure
Least Squares Means

density	yield	LSMEAN Number
25	5124.00000	1
50	5070.25000	2
75	5304.25000	3
100	4847.75000	4
125	4708.00000	5
150	4757.98333	6

Least Squares Means for effect density
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: yield

i/j	1	2	3	4	5	6
1		0.8268	0.4672	0.2712	0.1065	0.1876
2	0.8268		0.3484	0.3718	0.1553	0.2569
3	0.4672	0.3484		0.0792	0.0269	0.0577
4	0.2712	0.3718	0.0792		0.5715	0.7391
5	0.1065	0.1553	0.0269	0.5715		0.8526
6	0.1876	0.2569	0.0577	0.7391	0.8526	

The output provides p-values for the pairwise t-tests. There is just one significant comparison. LSMEANS does not provide a lines display. The reason is that for unbalanced data, a lines display is not generally possible due to non-constancy of the s.e.d. From the p-values, one can try to generate a lines display by hand using the procedure stated in section 5.6.1, but this is not guaranteed to work in general (though it does most of the time).

GLM does not provide s.e.d.'s. To obtain these, use the MIXED procedure instead of the GLM procedure as follows:

```
proc mixed;
class block density;
model yield=block density;
lsmeans density/pdiff;
run;
```

Output:

Least Squares Means

Effect	density	Estimate	Standard Error	DF	t Value	Pr > t
density	25	5124.00	170.53	14	30.05	<.0001
density	50	5070.25	170.53	14	29.73	<.0001
density	75	5304.25	170.53	14	31.11	<.0001
density	100	4847.75	170.53	14	28.43	<.0001
density	125	4708.00	170.53	14	27.61	<.0001
density	150	4757.98	201.77	14	23.58	<.0001

Differences of Least Squares Means

Effect	density	_density	Estimate	Standard Error	DF	t Value	Pr > t
density	25	50	53.7500	241.16	14	0.22	0.8268
density	25	75	-180.25	241.16	14	-0.75	0.4672
density	25	100	276.25	241.16	14	1.15	0.2712
density	25	125	416.00	241.16	14	1.73	0.1065
density	25	150	366.02	264.18	14	1.39	0.1876
density	50	75	-234.00	241.16	14	-0.97	0.3484
density	50	100	222.50	241.16	14	0.92	0.3718
density	50	125	362.25	241.16	14	1.50	0.1553
density	50	150	312.27	264.18	14	1.18	0.2569
density	75	100	456.50	241.16	14	1.89	0.0792
density	75	125	596.25	241.16	14	2.47	0.0269
density	75	150	546.27	264.18	14	2.07	0.0577
density	100	125	139.75	241.16	14	0.58	0.5715
density	100	150	89.7667	264.18	14	0.34	0.7391
density	125	150	-49.9833	264.18	14	-0.19	0.8526

The s.e.d.'s are not constant for all comparisons. The s.e.d.'s are computed using a general method, which is not explained here (see Appendix A).

In summary: order of fitting is important for unbalanced data. For unbalanced data, use LSMEANS, for balanced data use MEANS.

Exercise 5.6: Verify the ANOVA results for the rice data (**ir8.dat**) using the SAS procedure GLM. Analyse both balanced and unbalanced data. Compare means by LSMEANS and MEANS. Show that for balanced data the results are identical, while for unbalanced data they are not. Note that for unbalanced data the correct analysis is provided by LSMEANS.

Exercise 5.7: Assume the tasting experiment from Example 5.10 yields the following rating scores between 1 and 10:

Yoghurt	Taster			
	1	2	3	4
A	8	5	9	
B	6	2		2
C	4		4	3
D		4	7	4

Perform an ANOVA. Verify that the order of fitting terms matters with respect to the partitioning of sums of squares. Use the LSMEANS statement to compare adjusted means. Compare this to simple means computed using the MEANS statement (these are inappropriate here!).

5.6.3 Latin square

The linear model needs to have effects for rows and columns in addition to the treatment effect to account for the design:

$$y_{ijk} = \mu + r_j + c_k + \alpha_i + e_{ijk} \quad (5.3)$$

where

y_{ijk} = measurement of i -th treatment in j -th row and k -th column

α_i = effect of i -th treatment

r_j = effect of j -th row

c_k = effect of k -th column

Since main interest is in treatment effects, the analysis of variance is based on the following sequence of models:

Model	Error-SS	Reduction in SS
$y_{ijk} = \mu + e_{ijk}$	$SS(\mu)$	
$y_{ijk} = \mu + r_j + e_{ijk}$	$SS(\mu, r_j)$	$RSS(r_j \mu) = SS(\mu) - SS(\mu, r_j)$
$y_{ijk} = \mu + r_j + c_k + e_{ijk}$	$SS(\mu, r_j, c_k)$	$RSS(c_k \mu, r_j) = SS(\mu, r_j) - SS(\mu, r_j, c_k)$
$y_{ijk} = \mu + r_j + c_k + \alpha_i + e_{ijk}$	$SS(\mu, r_j, c_k, \alpha_i)$	$RSS(\alpha_i \mu, r_j, c_k) = SS(\mu, r_j, c_k) - SS(\mu, r_j, c_k, \alpha_i)$

For unbalanced data (missing values in the Latin square), it is important that blocks (rows and columns) are fitted first and treatments last. Generally, the order of fitting rows and columns is unimportant, so, long as treatments are fitted last, since we are interested only in treatment effects.

The sums of squares are compiled into an ANOVA table as usual:

Source	d.f.	SS	MS
Rows	$t-1$	$RSS(r_j \mu)$	$RSS(r_j \mu)/(t-1)$
Columns	$t-1$	$RSS(c_k \mu, r_j)$	$RSS(c_k \mu, r_j)/(t-1)$
Treatments	$t-1$	$RSS(\alpha_i \mu, r_j, c_k)$	$RSS(\alpha_i \mu, r_j, c_k)/(t-1)$
Error	$N-3t+2$	$SS(\mu, r_j, c_k, \alpha_i)$	$s^2 = SS(\mu, r_j, c_k, \alpha_i)/(N-3t+2)$

t = number of treatments = number of rows = number of columns

N = total number of observations

Note that the d.f. for rows, columns and treatments are identical for the Latin square design, because the numbers of treatments (t) equals that of rows and columns for this design. For balanced data, least squares means are identical to simple means, and a common LSD may be computed as for the completely randomized design (CRD) and the randomized complete block design (RCBD). This is available via the MEANS statement. Also, the order of fitting terms is immaterial, since the partitioning of SS is the same regardless of fitting order.

Exercise 5.8 (Example 5.7): Four diets (A, B C, D) were tested on four cows. Each cow tested each diet. Diets were tested during four periods. Cow and period were used as blocking factors in a Latin square design. Milk yields in pounds were as follows:

	Cow			
	1	2	3	4
Period 1	A 192	B 195	C 292	D 249
Period 2	B 190	D 203	A 218	C 210
Period 3	C 214	A 139	D 245	B 163
Period 4	D 221	C 152	B 204	A 134

Perform an ANOVA with subsequent multiple comparisons by the LSD test. Show that the results are as follows:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
cow	3	9929.187500	3309.729167	24.47	0.0009
period	3	6539.187500	2179.729167	16.12	0.0028
diet	3	8607.687500	2869.229167	21.22	0.0014

Alpha	0.05
Error Degrees of Freedom	6
Error Mean Square	135.2292
Critical Value of t	2.44691
Least Significant Difference	20.12

Means with the same letter are not significantly different.

t Grouping	Mean	N	diet
A	229.500	4	D
A	217.000	4	C
B	188.000	4	B
B	170.750	4	A

Hint: To read the data into a SAS dataset, you need to define four columns labelled "cow", "period", "diet", and "yield". Each observation is represented by a row, so the dataset must have 16 rows.

5.7 Treatment structure - contrasts

Example 5.12: Sokal & Rohlf present results on a tissue culture experiment. The purpose was to study the effect of the addition of different sugars on length (in ocular units $\times 0.114 = \text{mm}$) of pea sections grown in tissue culture with auxin present (**sugar.dat**). Treatments (sugars)

were randomly allocated to experimental units (petri dishes). The results were as follows:

Observation	Treatment				
	1 control	2 2% glucose	3 2% fructose	4 1% glucose + 1% fructose (mixture)	5 2% sucrose
1	75	57	58	58	62
2	67	58	61	59	66
3	70	60	56	58	65
4	75	59	58	61	63
5	65	62	57	57	64
6	71	60	56	56	62
7	67	60	61	58	65
8	67	57	60	57	65
9	76	59	57	57	62
10	68	61	58	59	67

In addition to comparing all treatment means, the researchers were interested in the following two comparisons:

Control vs. average of all sugar treatments
Mixture vs. pure sugars (fructose and glucose)

These types of mean comparisons are termed **linear contrasts**.

Treatment means are estimated by simple sample means $\bar{y}_{i\bullet}$. The two contrasts may be expressed as follows:

Description of contrast	Contrast
Control vs. average of all sugar treatments	$\bar{y}_{1\bullet} - \frac{\bar{y}_{2\bullet} + \bar{y}_{3\bullet} + \bar{y}_{4\bullet} + \bar{y}_{5\bullet}}{4}$
Mixture vs. pure sugars (fructose and glucose)	$\bar{y}_{4\bullet} - \frac{\bar{y}_{2\bullet} + \bar{y}_{3\bullet}}{2}$

The observed sample means are as follows:

Treatment	1	2	3	4	5
Mean ($\bar{y}_{i\bullet}$)	70.1	59.3	58.2	58.0	64.1

The estimated contrasts "Control vs. average of all sugar treatments" is:

$$L_1 = 70.1 - \frac{59.3 + 58.2 + 58.0 + 64.1}{4} = 10.20$$

Thus, it appears that sugars, on average, depress growth of pea sections. The contrast "Mixture vs. pure sugars" is computed as

$$L_2 = 58.0 - \frac{59.3 + 58.2}{2} = -0.75$$

The difference is rather small, indicating a slight advantage for the pure sugars.

The two examples can be related to the general expression of a linear contrast:

$$L = \sum_{i=1}^t c_i \bar{y}_i$$

In this expression, c_i ($i = 1, \dots, t$) are the contrast coefficients. For the example, the contrast coefficients are as follows:

Description of contrast	Coefficients				
	c_1	c_2	c_3	c_4	c_5
Control vs. average of all sugar treatments	1	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
Mixture vs. pure sugars (fructose and glucose)	0	$-\frac{1}{2}$	$-\frac{1}{2}$	1	0

A contrast may be tested by a t-test, where the t-statistic is of the form

$$t_{obs} = \frac{|L|}{s.e.(L)}$$

where $s.e.(L)$ is the estimate standard error of the contrast L . For balanced data, the standard error is

$$s.e.(L) = \sqrt{\frac{s^2}{n} \sum_{i=1}^t c_i^2}$$

Remark: Note that a pairwise comparison is a contrast with coefficients 1 and -1 for the two treatments to be compared and coefficients equal to zero for the other treatments.

The ANOVA for the sugar data yields a variance estimate of $s^2 = 5.46$. With this, the standard errors are computed as follows:

$$s.e.(L_1) = \sqrt{\frac{5.46}{10} \left[1^2 + \left(-\frac{1}{4}\right)^2 + \left(-\frac{1}{4}\right)^2 + \left(-\frac{1}{4}\right)^2 + \left(-\frac{1}{4}\right)^2 \right]} = 0.826$$

$$s.e.(L_2) = \sqrt{\frac{5.46}{10} \left[0^2 + \left(-\frac{1}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 + 1^2 + 0^2 \right]} = 0.905$$

From this, the t-statistics are computed as follows:

Description of contrast	L	s.e.(L)	t_{obs}
Control vs. average of all sugar treatments	10.20	0.826	12.35
Mixture vs. pure sugars (fructose and glucose)	-0.75	0.905	0.83

The experiment has 45 error d.f., so the tabular t-value is $t_{tab} = 2.014$. Thus, the first contrast is significant, while the second is not. Thus, sugars depress growth, on average, and there is no mixing effect for fructose and glucose. To implement contrast estimation in PROC GLM, one may use the ESTIMATE statement as follows:

```

data;
input trt length;
datalines;
1 75
1 67
<more data>
5 62
5 67
;
proc glm;
class trt;
model length=trt;
means trt;
estimate 'control vs. all sugars' trt 4 -1 -1 -1 -1 /divisor=4;
estimate 'mixture vs. pure'      trt 0 -1 -1  2  0 /divisor=2;
run;

```

Note that the ESTIMATE statement accepts only integer-valued coefficients. Thus, coefficients are expanded by their common denominator. The common denominator is specified in the DIVISOR= option, while the numerator is listed behind the label of the treatment variable.

Exercise 5.9: Compute the contrasts L_1 (Control vs. average of all sugar treatments) and L_2 [Mixture vs. pure sugars (fructose and glucose)] for the sugar data. In addition, perform pairwise comparisons among the three treatments with pure sugars at 2% concentration using the ESTIMATE statement.

Exercise 5.10: The following table gives the nitrogen content in milligrams of red clover plants inoculated with cultures of *Rhizobium trifolii* (see **rhizobium.dat**). There were five different *Rhizobium* strains, which were tested individually (treatments 1-5). In addition, the composite mixture of all five strains was also tested (treatment 6). The trial was laid out as a completely randomized design.

3Dok1	3Dok5	3Dok4	3Dok7	3Dok13	Composite
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
32.0	24.3	15.8	18.6	14.2	20.8

Perform an ANOVA and test, whether the mean of the five strains tested individually differs significantly from the composite.

Final remark: The above equations for contrasts, standard errors, etc. are valid for balanced data. They apply equally to the CRD, the RCBD and Latin squares. When the data are unbalanced, the underlying principles remain the same, but more general methods of computation are needed, which have not been discussed here for brevity (see Appendix A). The nice thing for the user of SAS is that the same statements can be used as for balanced data to estimate and test contrasts.

5.8 Statistical analysis of experiments with one quantitative treatment factor

Many experiments are concerned with quantitative treatment variables.

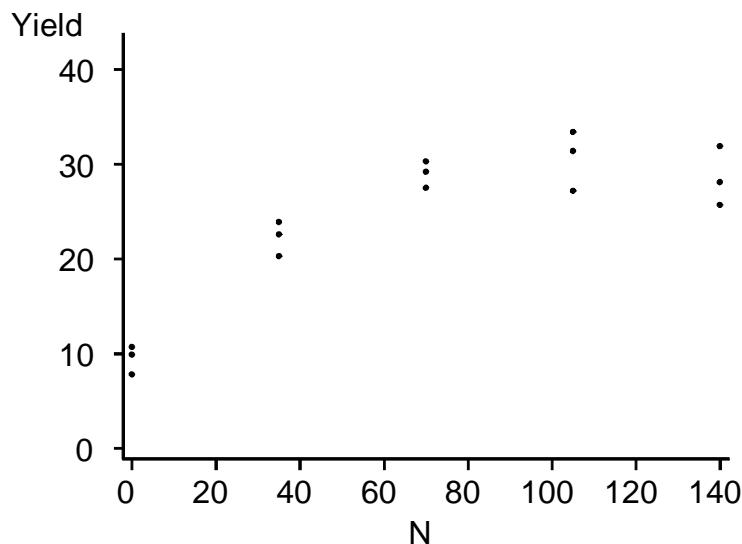
Example 5.13: An experiment was conducted to assess the influence of different quantities of N-fertiliser (N) on root dry matter yield of sugar beets. The experiment was laid out in complete blocks (Petersen, 1994). Here, we will assume for simplicity that the experiment has been completely randomised. In an exercise, the fully appropriate analysis taking into account the blocking structure will be considered.

Table: Root dry matter (t/ha) of sugar beet depending on the amount of N-fertiliser.

Replication	Fertiliser level (kg N /ha)				
	0	35	70	105	140
1	9.9	20.3	27.5	31.4	28.1
2	7.8	22.6	30.3	27.2	25.7
3	10.7	23.9	29.2	33.4	31.9

The objective of the analysis is to study the effect of fertiliser on yield.

The following display plots the raw data against fertiliser level.



The yield increases with N up to a level of 105 kg N/ha, with a subsequent slight yield depression.

It is possible to analyse such experiments using the same methods as those used for qualitative treatment variables (analysis of variance followed by multiple treatment comparisons), because several observations are available per level of the treatment factor. While such analyses are not downright wrong (though some statisticians consider them as "a misuse of statistics"), they are not usually fully efficient, because the quantitative nature of the treatment variable is not exploited. A quantitative variable can always be down-scaled to a qualitative level, but this entails a loss of information. To fully exploit the quantitative information, one may perform regression analysis. Here, we will only consider polynomial regression as the simplest form of regression. For other nonlinear regression techniques, including "intrinsically

"nonlinear regression" (logistic regression, etc.), you may refer to standard texts such as Ratkowsky (1983, 1989) and Seber and Wild (1989).

To exploit the quantitative nature of the treatment variable, one may regress the response variable (yield) on the treatment variable (fertiliser level). The association of response and treatment variable can be described by a suitable function, for example by a regression line of the form $Y = \mu + \beta_1 X$. The corresponding model for an observation is

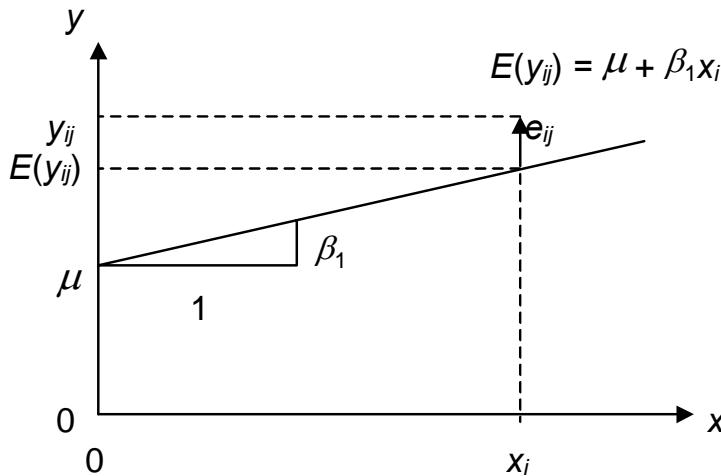
$$y_{ij} = \mu + \beta_1 x_i + e_{ij}$$

where y_{ij} = j -th observation ($j = 1, \dots, n_i$) for i -th ($i = 1, \dots, t$) level of the treatment variable, μ = intercept of regression line, β_1 = slope of regression line and x_i = quantitative value of i -th level of treatment variable. Note that in this model there are several (n_i) observations y_{ij} for the same level of the treatment variable. For a given treatment level, the observations y_{ij} are randomly distributed around the expected value of

$$E(y_{ij}) = \mu + \beta_1 x_i \quad (5.4)$$

It is assumed that deviations from the expected value follow a normal distribution with zero mean and variance σ^2 , i.e.

$$e_{ij} \sim N(0, \sigma^2)$$



The slope of the regression equation (5.4) has the following useful interpretation: **If the level of the treatment variable is increased by one unit, the expected value of the response variable increases by β_1 units.**

Quite often, the relationship between response and treatment variable is nonlinear, as in the fertiliser example. The simplest way to model a nonlinear response is to fit a second degree polynomial of the form:

$$y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$$

If this does not fit, higher powers of x_i can be added. Note that a line is a 1st degree polynomial. A polynomial is a linear model in that it is linear in the regressor variables x_i, x_i^2, x_i^3, \dots . It is this property of polynomials that make them so convenient in practice: the

statistical analysis can proceed within a linear regression framework (although a nonlinear response may be fitted). In most applications, it will be sufficient to use polynomials up to 3rd degree. A second degree (quadratic) polynomial is a parabola with one maximum or one minimum:

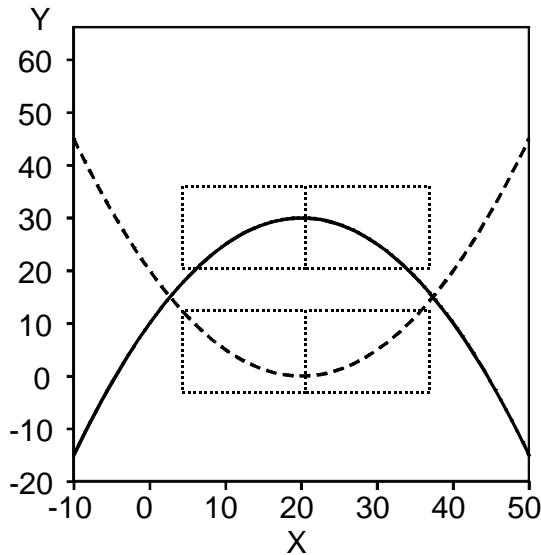


Fig. 5.1: Two quadratic polynomials (parabola). Dotted rectangles: suitable parts of parabola to describe response patterns common in plant research.

Nonlinear responses with continuously increasing or decreasing slope, possibly showing a maximum or minimum, can usually be modelled using a quadratic polynomial, provided the response shows no point of inflection, i.e. if the response is not sigmoidal. This is indicated in the above graph by the dotted rectangles, which pick the suitable "arms" of the parabola. Four numerical examples are shown in Fig. 5.2.

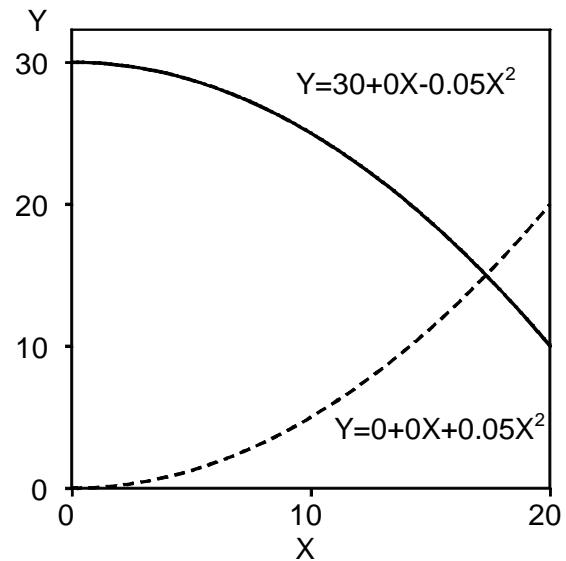
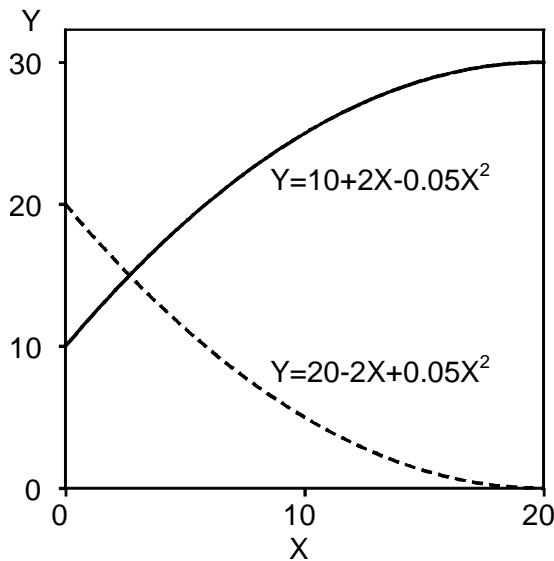


Fig. 5.2: Four quadratic polynomials.

If the response is sigmoidal in shape, a third degree (cubic) polynomial may be a suitable model. Two examples are given in Fig. 5.3.

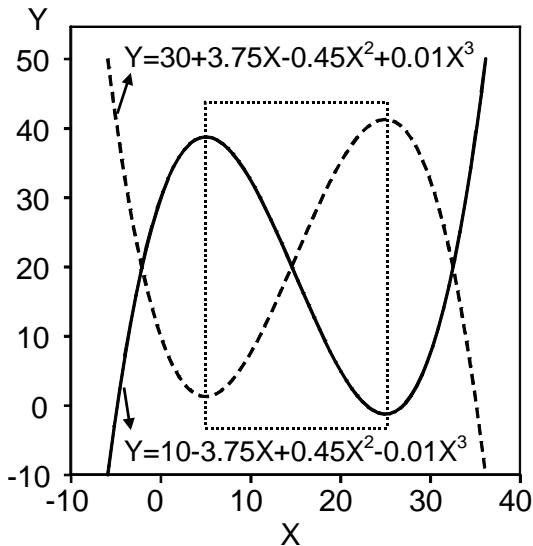


Fig. 5.3: Two cubic polynomials. Dotted rectangles: suitable parts to describe response patterns common in plant research.

I do not recommend to generally consider higher degree polynomials. If no polynomial up to third degree fits the data, it is usually preferable to fit an intrinsically non-linear model (Seber and Wild, 1989), e.g. a logistic regression model, or to use other methods, e.g. approaches based on multiple linear contrasts.

An ANOVA table can be constructed based on the principle of sequential model building, according to which one proceeds from the simplest to the most complex model. For polynomial regression, one starts with the simplest model, i.e., the model with just an intercept term. Then, more terms are successively added, adhering to the sequence "linear", "quadratic", "cubic", etc. In other words, for a term of order k to be added, all terms up to order $k-1$ should be in the model. This principle is related to a concept, which John Nelder (1994, 2000) dubbed **marginality**. To illustrate the underlying philosophy, consider the following counter-examples.

Example 5.14: Assume that we fit a linear term without an intercept:

$$E(y_{ij}) = \beta_1 x_i$$

This regression line passes through the origin. The model implies that the expected response is zero at $x_i = 0$. Such a model will not usually make sense biologically, except in rare cases. For example, the response of yield to increased fertiliser doses will be such that there is a non-zero yield for a fertiliser dose of $x_i = 0$. The relevant question to ask is whether there is an increase in yield for $x_i > 0$, starting from the non-zero yield at $x_i = 0$. To answer this question, we need to ask whether the full model

$$E(y_{ij}) = \mu + \beta_1 x_i$$

significantly improves the fit (reduces the error SS) relative to the reduced model

$$E(y_{ij}) = \mu$$

Thus, we fit the linear term β_1 only after having fitted the intercept term μ .

Example 5.15: Assume we were to fit a quadratic term without having fitted a linear term. Thus the full model would read

$$E(y_{ij}) = \mu + \beta_2 x_i^2$$

This model has a minimum or maximum at $x_i = 0$, depending on the sign of β_2 . While it is not uncommon for a response to show a maximum or minimum, it is very unlikely that this optimum occurs exactly at $x_i = 0$. For example, the response of yield to increased amounts of fertiliser will usually show a positive slope at $x_i = 0$, and the yield will proceed to a maximum at some fertiliser dose $x_i > 0$. For a quadratic polynomial to have an optimum, which does **not** lie at $x_i = 0$, the model must have a linear term! Thus, we need to fit the full model

$$E(y_{ij}) = \mu + \beta_1 x_i + \beta_2 x_i^2$$

Now what is an appropriate reduced model corresponding to this full model? The above discussion suggests that the appropriate reduced model is

$$E(y_{ij}) = \mu + \beta_1 x_i$$

This model, if compared to the further reduced model $E(y_{ij}) = \mu$, will detect if there is any response to increased fertiliser level. Adding the quadratic term will then show whether the slope of the response is constant over the investigated x-domain (linear model) or changing (quadratic), possibly with a maximum occurring somewhere in the investigated x-domain. Conversely, if no response is detected in the first place (linear term not significant), it makes little sense to ask for significance of a quadratic term, for this would imply the unrealistic model $E(y_{ij}) = \mu + \beta_2 x_i^2$. [If one is willing to consider a quadratic model despite a non-significant linear term, which requires a non-statistical justification, then the linear term should be retained regardless of its non-significance.]

Now back to the construction of an ANOVA table. We start with the simplest model, and successively add terms, observing the marginality principle. The number of polynomial terms is restricted by the number of treatment levels (t). For example, if there are two x-levels, we can only fit a linear model, because for two points in an (x,y) -plane, there is one line that exactly passes through these two points (while there is an infinite number of quadratic polynomials passing through these same two points). Similarly, there is always exactly one quadratic polynomial passing through three points, one cubic polynomial passing through four points, etc. Thus, the maximal number of polynomial terms is $(t - 1)$, where t is the number of treatments.

For each model in the sequence (intercept, linear, quadratic, etc.), we estimate parameters by least squares and compute the residual error sum of squares (SS). **The reduction in error SS caused by the addition of a term is equivalent to the SS for that term appearing in the ANOVA table.** The reduction can be represented conveniently using the familiar $RSS(\cdot)$ -notation (Chapter 5). The sequence of models is as follows:

Model	SS_{error}	Reduction in SS
$y_{ij} = \mu + e_{ij}$	$SS(\mu)$	
$y_{ij} = \mu + \beta_1 x_i + e_{ij}$	$SS(\mu, \beta_1)$	$RSS(\beta_1 \mu) = SS(\mu) - SS(\mu, \beta_1)$
$y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$	$SS(\mu, \beta_1, \beta_2)$	$RSS(\beta_2 \mu, \beta_1) = SS(\mu, \beta_1) - SS(\mu, \beta_1, \beta_2)$
$y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_{ij}$	$SS(\mu, \beta_1, \beta_2, \beta_3)$	$RSS(\beta_3 \mu, \beta_1, \beta_2) = SS(\mu, \beta_1, \beta_2) - SS(\mu, \beta_1, \beta_2, \beta_3)$
etc.	.	.

Note that the simplest ("zero-degree") polynomial, $y_{ij} = \mu + e_{ij}$, is equivalent to the reduced model in a one-way ANOVA for a qualitative factor. Similarly, it turns out that the full model $y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_{t-1} x_i^{t-1} + e_{ij}$ is equivalent to the full model $y_{ij} = \mu + \alpha_i + e_{ij}$ for a one-way ANOVA of a qualitative factor. Note that there is exactly one polynomial of degree $(t-1)$, which passes through t points. As pointed out above, it will not usually be useful to consider terms higher than cubic, so for a larger number of treatment levels it may not be necessary to fit the whole sequence up to order $(t-1)$. We will come back to this point later. The polynomial with $(t-1)$, being the last one on the model fitting sequence, is called the **saturated model**, because no more polynomial terms can be added. The SS corresponding to the different model terms are assembled in an ANOVA table:

Source	degrees of freedom	SS
β_1	1	$RSS(\beta_1 \mu)$
β_2	1	$RSS(\beta_2 \beta_1, \mu)$
.	.	.
.	.	.
β_{t-1}	1	$RSS(\beta_{t-1} \beta_{t-2}, \dots, \beta_1, \mu)$
Error	$N-t$	$SS(\mu, \beta_1, \beta_2, \dots, \beta_{t-1})$

N = total no. of observations; t = number of treatments

The degrees of freedom (d.f.) for a source of variation (model term) is given by the difference in the number of (free) parameters for the models with and without the corresponding model term. Now the sequential improvement of model fit by the addition of terms can be tested. We start with the model $y_{ij} = \mu + e_{ij}$. If there is an association with x_i , the addition of the linear term $\beta_1 x_i$ should result in a significant improvement of fit and thus in a significant reduction in error SS. If the response of y to increasing x_i is non-linear, addition of a quadratic term should further improve the fit, etc.

Example 5.13: For the fertiliser trial the following graphs show the fitted curves for a sequence of models ranging from $y_{ij} = \mu + e_{ij}$ to $y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_{ij}$. The addition of a linear term results in a clear improvement of fit, which is seen from the reduction in error SS from 968.25 to 316.78. Obviously, however, the response is non-linear, as revealed by a glance at the scatter plot. Not surprisingly, therefore, addition of a quadratic term leads to a further clear reduction in error SS to 56.28. By contrast, adding a cubic term (x^3) does not markedly change the fit: The error SS only goes down to 54.69, and the fitted cubic curve is virtually the same for the quadratic fit.

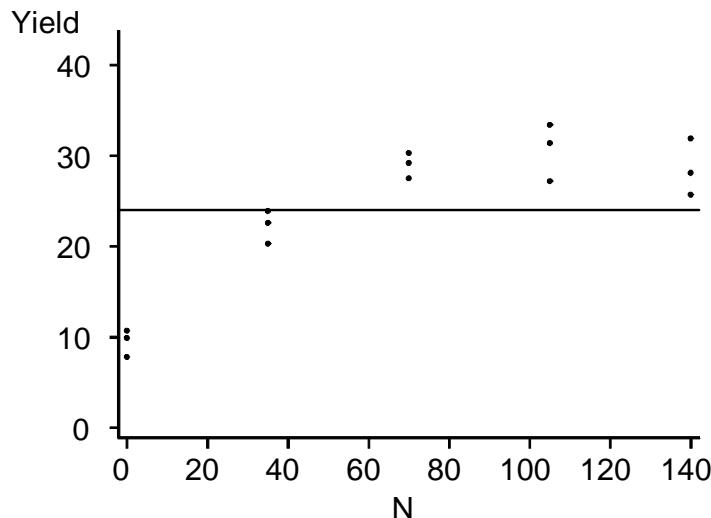


Fig. 5.4: Fit of intercept model $y_{ij} = \mu + e_{ij}$. $SS(\mu) = 968.25$.

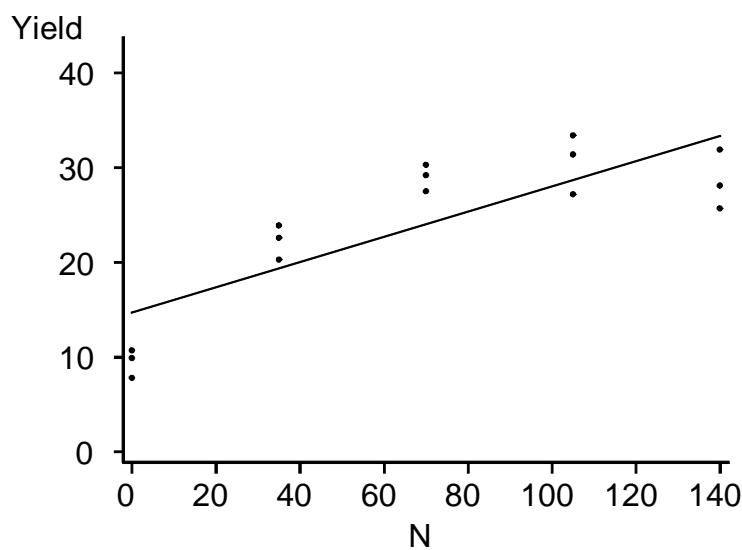


Fig. 5.5: Fit of linear model $y_{ij} = \mu + \beta_1 x_i + e_{ij}$. $SS(\mu, \beta_1) = 316.78$.

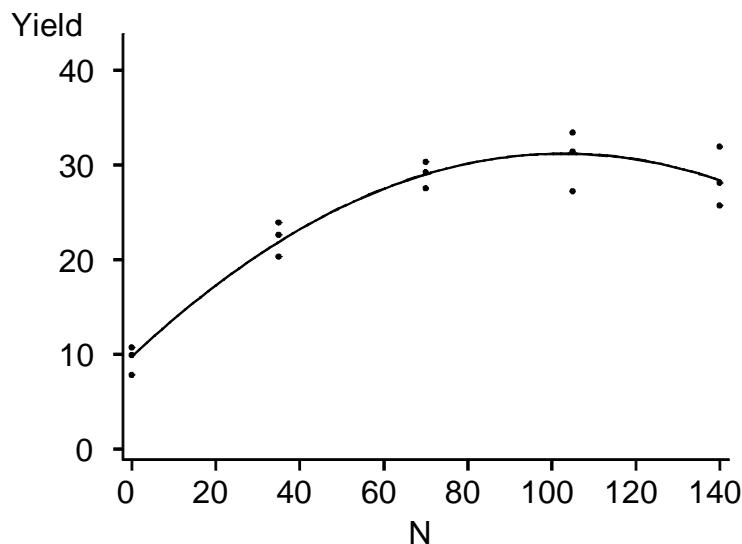


Fig. 5.6: Fit of quadratic model $y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$. $SS(\mu, \beta_1, \beta_2) = 56.25$.

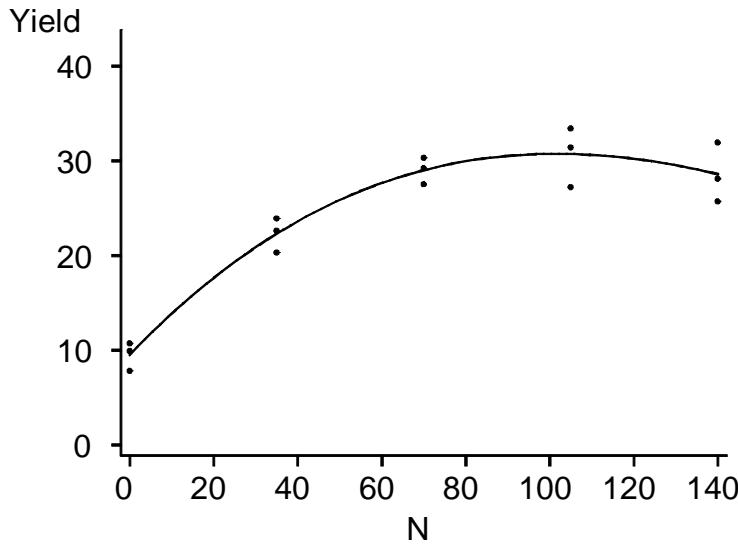


Fig. 5.7: Fit of cubic model $y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_{ij}$. $SS(\mu, \beta_1, \beta_2, \beta_3) = 54.69$.

Since there are $t = 5$ treatment levels, we can fit terms up to 4-th order (quartic). The SS for the model sequence are listed below:

Model	Error SS	Reduction
$y_{ij} = \mu + e_{ij}$	968.25	
$y_{ij} = \mu + \beta_1 x_i + e_{ij}$	316.78	651.47
$y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$	56.28	260.50
$y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_{ij}$	54.691	1.59
$y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + e_{ij}$	54.687	0.004
		Sum: 913.56

The sequential SS are compiled in an ANOVA table:

Source	d.f.	SS	MS	F	p-value
β_1 (linear)	1	651.47	651.47	119.13	<0.0001
β_2 (quadratic)	1	260.50	260.50	47.64	<0.0001
β_3 (cubic)	1	1.59	1.59	0.29	0.6019
β_4 (quartic)	1	0.004	0.004	<0.01	0.9793
Error	10	54.687	5.469		

The critical F -value for all four F-tests is $F_{tab}(1, N-t=10, \alpha = 5\%) = 4.96$. The linear and quadratic terms are significant, while the cubic and quartic terms are not significant. This result of the F-tests can also be inferred from the p-values, which are produced by linear model packages.

Rather than exploiting the quantitative information on the treatment factor, we could have performed an ANOVA treating the factor as qualitative, followed by multiple comparison of means. This, however, does not yield as much information as the regression analysis and is wasteful of parameters ($t = 5$ treatment means need to be fitted rather than three model terms for a quadratic polynomial). Also, we cannot interpolate, i.e., we cannot determine what would have been the yield at fertiliser levels intermediate between two of the tested levels. The "qualitative" one-way ANOVA is as follows:

Source	d.f.	SS	MS	F	p-value
Treatments	4	913.563	228.391	41.76	<0.0001
Error	10	54.687	5.469		
Treatment (kg N/ha)		Mean			
0		9.5 ^a			
35		22.3 ^b			
70		29.0 ^c			
105		30.7 ^c			
140		28.6 ^c			
LSD(5%)		4.25			

The ANOVA rejects the global null of no treatment differences. The subsequent comparison of means reveals an increase of yield up to a fertiliser dose of 105 kg/ha, but it is not possible to derive a quantitative functional expression for the dose-response relationship. Specifically, it is not clear from the mean comparisons that the slope of the response decreases with increasing fertiliser dose (equivalent "diminishing returns on investment" in economics) and that there is even a yield reduction at high doses.

Note that in the example the SS for the four polynomial coefficients exactly add up to the treatment sum of squares (913.563). Similarly, the degrees of freedom of coefficients β_1 through β_4 add up to the treatment d.f. $t - 1 = 4$. We may say that the treatment d.f. and the treatment SS are split up into linear, quadratic, cubic, and quartic components. Consequently, the full model for a quartic (4th degree) polynomial yields the same error SS as a "qualitative" ANOVA for the fertiliser factor. The splitting of treatment d.f. and SS can be visualised by the following structure in an ANOVA table:

Source	d.f.	SS	MS	F	p-value
Treatments	4	913.563	228.391	41.76	<0.0001
β_1 (linear)	1	651.47	651.47	119.13	<0.0001
β_2 (quadratic)	1	260.50	260.50	47.64	<0.0001
β_3 (cubic)	1	1.59	1.59	0.29	0.6019
β_4 (quartic)	1	0.004	0.004	<0.01	0.9793
Error	10	54.687	5.469		

Lack-of-fit test

In the preceding section we have fitted the complete model sequence up to the term of order $(t - 1)$. The polynomial of degree $(t - 1)$ is the **saturated model** of the sequence and is equivalent to the full model of a "qualitative" ANOVA based on the model

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

The sequence of models can be modified by skipping all terms, starting from some model order k . For example, after fitting the linear term ($k = 1$), we can proceed directly to the saturated model. The resulting model sequence is as follows:

Model

$$y_{ij} = \mu + e_{ij}$$

$$y_{ij} = \mu + \beta_1 x_i + e_{ij}$$

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad (\text{saturated model})$$

Note that the first two models are special (reduced) cases of the saturated model: For model the second model we have

$$\alpha_i = \beta_1 x_i$$

while for the first model $\alpha_i = 0$. For what follows, it will be useful to re-express the saturated model in a slightly different, but equivalent form. The important feature of the saturated model is that the mean (expected value) for each treatment is free to vary, independently of means for the other treatments. Note that under the reduced model(s), the means/expected values of all treatments are interdependent due to the polynomial response. For example, if means are 20 at $x_1 = 0$ and 30 at $x_2 = 35$, the mean must be 40 at $x_3 = 70$ under the linear model ($y_{ij} = \mu + \beta_1 x_i + e_{ij}$). The fact that means can be freely chosen under the saturated model, can be modelled by a systematic deviation from the linear regression, if we make the assumption that the deviation is free to vary among treatments. We may write

$$\alpha_i = \beta_1 x_i + \delta_i$$

and thus

$$y_{ij} = \mu + \beta_1 x_i + \delta_i + e_{ij}$$

where δ_i is a systematic deviation from the regression for the i -th treatment. Both formulations of the saturated model are equivalent. The second form better reflects the hierarchical structure of the model sequence. The re-parameterized sequence is:

Model	SS_{error}	Reduction in SS
$y_{ij} = \mu + e_{ij}$	$SS(\mu)$	
$y_{ij} = \mu + \beta_1 x_i + e_{ij}$	$SS(\mu, \beta_1)$	$RSS(\beta_1 \mu) = SS(\mu) - SS(\mu, \beta_1)$
$y_{ij} = \mu + \beta_1 x_i + \delta_i + e_{ij}$	$SS(\mu, \beta_1, \delta_i)$	$RSS(\delta_i \mu, \beta_1) = SS(\mu, \beta_1) - SS(\mu, \beta_1, \delta_i)$

If there is a significant reduction in SS between the model $y_{ij} = \mu + \beta_1 x_i + e_{ij}$ and the saturated model $y_{ij} = \mu + \beta_1 x_i + \delta_i + e_{ij}$, it can be concluded that a linear model does not adequately describe the data, i.e. there is a significant departure from linearity or, more generally, a significant **lack-of-fit**. To fully describe the data, a systematic deviation ("lack-of-fit effect") δ_i is needed. If, by contrast, the reduction in SS is non-significant, the linear model fits well, and there is justification to conclude that the response is (at least approximately) linear. The least we can say in this latter situation is that no departure from linearity can be detected. It follows from all this that the reduction in SS between the two models can be used to test for departure from linearity (more generally for lack-of-fit of a given reduced model). The ANOVA table for the model sequence is:

Source	d.f.	SS
Treatments	$t - 1$	
β_1 (linear)	1	$RSS(\beta_1 \mu)$
δ_i (lack-of-fit)	$t - 2$	$RSS(\delta_i \mu, \beta_1)$
Error	$N - t$	$SS(\mu, \beta_1, \delta_i)$

N is the total number of observations, while t is the number of treatment levels. Note that, again, the d.f. for a model term are equal to the difference in the number of free parameters for the model with the term and for the model without the term. Thus, at each step of the sequence, the d.f. are computed as the increase in the number of free parameters. Of course, we can expand the sequence by adding a quadratic, cubic, ... term, following the linear term, again adhering to the marginality principle. We can keep adding polynomial terms until the lack-of-fit test becomes non-significant, i.e., until no further lack-of-fit is detected.

Example 5.13: For the fertiliser trial we find the following error SS based on a sequence with a linear term:

Model	SS_{error}	Reduction in SS
$y_{ij} = \mu + e_{ij}$	968.25	
$y_{ij} = \mu + \beta_1 x_i + e_{ij}$	316.78	651.47
$y_{ij} = \mu + \beta_1 x_i + \delta_i + e_{ij}$	54.687	262.09

This yields the following ANOVA table:

Source	d.f.	SS	MS	F	p-value
Treatments	4				
β_1 (linear)	1	651.47	651.47	119.13	<0.0001
δ_i (lack-of-fit)	3	262.09	87.36	15.98	0.0004
Error	10	54.687	5.47		

The lack-of-fit test is significant [$F_{Exp} = 15.18 > F_{Tab}(t-k-1 = 3, N-t = 10, \alpha = 5\%) = 3.71$], so there is a significant departure from linearity. This finding agrees with the impression from the scatter plot and the ANOVA for the polynomial regression, in which the quadratic term was significant. We now add the quadratic term and find:

Model	SS_{error}	Reduction in SS
$y_{ij} = \mu + e_{ij}$	968.25	
$y_{ij} = \mu + \beta_1 x_i + e_{ij}$	316.78	651.47
$y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$	56.28	260.50
$y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + \delta_i + e_{ij}$	54.687	1.59

ANOVA table:

Source	d.f.	SS	MS	F	p-value
Treatment	4				
β_1 (linear)	1	651.47	651.47	119.13	<0.0001
β_2 (quadratic)	1	260.50	260.50	47.64	<0.0001
δ_i (lack-of-fit)	2	1.59	0.80	0.15	0.8864
Error	10	54.687	5.47		

After adding the quadratic term, the lack-of-fit test is no longer significant [$F_{exp} = 0.15 < F_{Tab}(t-k-1 = 2, N-t = 10, \alpha = 5\%) = 4.10$]. We can conclude that the quadratic model adequately describes the dose-response relationship.

Of course there is a close relation between polynomial regression and lack-of-fit testing: The sum of squares for the polynomial terms of order ($k+1$) up to order ($t-1$) is identical to the SS for the lack-of-fit of a k -th degree polynomial. This can be demonstrated for the case of a 1st degree polynomial.

Polynomial regression:

Source	d.f.	SS
Treatments	4	
β_1 (linear)	1	651.47
β_2 (quadratic)	1	260.50
β_3 (cubic)	1	1.59
β_4 (quartic)	1	0.004
Error	10	54.687

Lack-of-fit-test:

Source	d.f.	SS
Treatments	4	
β_1 (linear)	1	651.47
β_2 (quadratic)	1	260.50
β_3 (cubic)	1	1.59
β_4 (quartic)	1	0.004
δ_i (lack-of-fit)	3	262.09
Error	10	54.687

The main advantage of the lack-of-fit test is that it is unnecessary to fit all polynomial terms up to order ($t-1$) by default. This advantage becomes particularly relevant when t is large.

SAS hints

To explain the use of SAS for polynomial regression as described above, it is first necessary to point out the purpose of the CLASS statement. To analyse the fertiliser example by a "qualitative" one-way ANOVA based on the model

$$y_{ij} = \mu + \alpha_i + e_{ij} ,$$

we would use the following code:

```
data;
input fert yield;
datalines;
0    9.9
0    7.8
0    10.7
35   20.3
35   22.6
35   23.9
70   27.5
70   30.3
70   29.2
105  31.4
105  27.2
105  33.4
140  28.1
140  25.7
140  31.9
;
proc glm;
class fert;
model yield=fert;
run;
```

The variable FERT is the treatment variable. Each level of the variable FERT corresponds to a different fertilizer treatment. There are five treatments, and so the variable FERT has five levels. For each treatment level, the model has a separate effect. By listing the FERT variable in the CLASS statement, GLM is prompted to actually generate a separate effect for each level of FERT ($\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$).

Output:

```
General Linear Models Procedure
      Class Level Information

      Class    Levels    Values
      FERT      5     0 35 70 105 140

      Number of observations in data set = 15
```

General Linear Models Procedure

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	913.56266667	228.39066667	41.76	0.0001
Error	10	54.68666667	5.46866667		
Corrected Total	14	968.24933333			

	R-Square	C.V.	Root MSE	Y Mean
	0.943520	9.746533	2.33851805	23.99333333
Source	DF	Type I SS	Mean Square	F Value
FERT	4	913.56266667	228.39066667	41.76
Source	DF	Type III SS	Mean Square	F Value
FERT	4	913.56266667	228.39066667	41.76

To verify the functionality of the CLASS statement, we use the same GLM code as before, but drop the CLASS statement:

```
proc glm;
model yield=fert;
run;
```

Output:

General Linear Models Procedure					
Number of observations in data set = 15					
General Linear Models Procedure					
Dependent Variable: YIELD					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	651.46800000	651.46800000	26.73	0.0002
Error	13	316.78133333	24.36779487		
Corrected Total	14	968.24933333			
	R-Square	C.V.	Root MSE	Y Mean	
	0.672831	20.57394	4.93637467	23.99333333	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
FERT	1	651.46800000	651.46800000	26.73	0.0002
Source	DF	Type III SS	Mean Square	F Value	Pr > F
FERT	1	651.46800000	651.46800000	26.73	0.0002
Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate	
INTERCEPT	14.67333333	6.65	0.0001	2.20761386	
FERT	0.13314286	5.17	0.0002	0.02575013	

There are some notable differences in the two outputs. The second output does not report levels of the FERT variable under a heading "CLASS LEVEL INFORMATION". More

importantly, the ANOVAs are different. The FERT variable has only one d.f. in the second output (without the CLASS statement), while it had 4 d.f. when the CLASS statement is used (first output). Finally, parameter estimates are printed at the end of the second output, and there is only one parameter estimate corresponding to the FERT variable. Obviously, we have fitted a different model, clearly not the model $y_{ij} = \mu + \alpha_i + e_{ij}$. But what model has been fitted? The answer is that we have fitted the simple linear regression model

$$y_{ij} = \mu + \beta_1 x_i + e_{ij}$$

which only has one parameter (β_1) corresponding to the treatment variable FERT. In fact, levels of the variable FERT correspond to the regressor variable x_i in the linear regression model. What we find under "FERT" in the list of parameter estimates, is the least squares estimate for β_1 . Thus, by listing FERT in the MODEL statement, but **not** in the CLASS statement, FERT is regarded as a quantitative variable, and a linear regression is fitted. If, by contrast, we also list FERT in the CLASS statement, FERT is regarded as a qualitative variable, and a separate effect is generated for each level of FERT.

Now consider fitting the quadratic model

$$y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$$

This is a multiple linear regression model with regressor variables x_i and x_i^2 . It is fitted by

```
proc glm;
model yield=fert fert*fert;
run;
```

Output:

General Linear Models Procedure					
Dependent Variable: YIELD					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	911.97180952	455.98590476	97.23	0.0001
Error	12	56.27752381	4.68979365		
Corrected Total	14	968.24933333			
R-Square		C.V.	Root MSE	Y Mean	
0.941877		9.025812	2.16559314	23.99333333	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
FERT	1	651.46800000	651.46800000	138.91	0.0001
FERT*FERT	1	260.50380952	260.50380952	55.55	0.0001
Source	DF	Type III SS	Mean Square	F Value	Pr > F
FERT	1	516.06920416	516.06920416	110.04	0.0001
FERT*FERT	1	260.50380952	260.50380952	55.55	0.0001

Note the difference between Type I and Type III SS. The sequential SS discussed in the previous section correspond to Type I SS. Since the FERT term is listed prior to the FERT*FERT term in the MODEL statement, we obtain the correct fitting order (linear before quadratic). Type III is not generally appropriate for polynomial regression. In the case at hand, it gives the reduction in SS for a model term after fitting all other terms, except the intercept. Thus, the Type III SS for x_i is the reduction in SS due to fitting x_i , after having fitted x_i^2 . This fitting order violates the marginality principle and is not appropriate.

Now on to a forth-degree (quartic) polynomial:

```
proc glm;
model yield=fert
    fert*fert
    fert*fert*fert
    fert*fert*fert*fert;
run;
```

Output:

General Linear Models Procedure

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	913.56266667	228.39066667	41.76	0.0001
Error	10	54.68666667	5.46866667		
Corrected Total	14	968.24933333			
	R-Square	C.V.	Root MSE		Y Mean
	0.943520	9.746533	2.33851805	23.99333333	

Source	DF	Type I SS	Mean Square	F Value	Pr > F
FERT	1	651.46800000	651.46800000	119.13	0.0001
FERT*FERT	1	260.50380952	260.50380952	47.64	0.0001
FERT*FERT*FERT	1	1.58700000	1.58700000	0.29	0.6019
FERT*FERT*FERT*FERT	1	0.00385714	0.00385714	0.00	0.9793

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FERT	1	24.91377060	24.91377060	4.56	0.0586
FERT*FERT	1	0.70427476	0.70427476	0.13	0.7272
FERT*FERT*FERT	1	0.00321239	0.00321239	0.00	0.9811
FERT*FERT*FERT*FERT	1	0.00385714	0.00385714	0.00	0.9793

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
INTERCEPT	9.466666667	7.01	0.0001	1.35014403
FERT	0.459761905	2.13	0.0586	0.21540411
FERT*FERT	-0.002772109	-0.36	0.7272	0.00772467
FERT*FERT*FERT	0.000002138	0.02	0.9811	0.00008821
FERT*FERT*FERT*FERT	0.000000008	0.03	0.9793	0.00000031

Again, the Type III SS yield nonsensical results. For example, the Type III SS for x_i now is the reduction due to fitting x_i , after having fitted all terms of higher order. This, again, violates the marginality principle. Also note, that according to the Type III SS, all model terms are non-significant, while the sequential Type I SS show a clear significance of the linear and quadratic terms.

Finally, consider the lack-of-fit model

$$y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + \delta_i + e_{ij}$$

To fit the lack-of-fit term, we need to use the CLASS statement. Specifically, we need to use a variable which has the same levels as the treatment variable so that a separate lack-of-fit effect δ_i is generated for each level of the treatment variable. This is basically the same situation as in the "qualitative" one-way ANOVA, where a separate effect needs to be assigned to each treatment. For polynomial regression, we have stored the treatment variable under the label FERT. We could be tempted to use this variable to generate the δ_i effects by simply listing it in the CLASS statement, but this would preclude the use of FERT to fit the linear term β_1 ! For this reason, we need to duplicate the FERT-variable and list only the label of the duplicated variable in the CLASS statement. We will store the duplicate under the label LACKFIT. The LACKFIT label is listed in both the CLASS and the MODEL statements, while the FERT variable only appears in the MODEL statement. To obtain the correct order of fitting, LACKFIT, which models the lack-of-fit effect, δ_i , appears last in the MODEL statement.

```

data;
input fert yield;
lackfit=fert;
datalines;
0    9.9
0    7.8
0    10.7
35   20.3
35   22.6
35   23.9
70   27.5
70   30.3
70   29.2
105  31.4
105  27.2
105  33.4
140   28.1
140   25.7
140   31.9
;
proc glm;
class lackfit;
model yield=fert fert*fert lackfit;
run;

```

Output:

General Linear Models Procedure
Class Level Information

Class	Levels	Values
LACKFIT	5	0 35 70 105 140

Number of observations in data set = 15

General Linear Models Procedure

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	913.56266667	228.39066667	41.76	0.0001
Error	10	54.68666667	5.46866667		
Corrected Total	14	968.24933333			

R-Square	C.V.	Root MSE	Y Mean
0.943520	9.746533	2.33851805	23.99333333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
FERT	1	651.46800000	651.46800000	119.13	0.0001
FERT*FERT	1	260.50380952	260.50380952	47.64	0.0001
LACKFIT	2	1.59085714	0.79542857	0.15	0.8664

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FERT	0	0.00000000	.	.	.
FERT*FERT	0	0.00000000	.	.	.
LACKFIT	2	1.59085714	0.79542857	0.15	0.8664

The lack-of-fit test appears under the label LACKFIT for the Type I SS. Again, Type III SS yield nonsensical results. The SS for the linear and quadratic terms are zero. This happens, because the SS for either of these terms is the reduction in error SS after fitting all effects except the one in question. Thus, the lack-of-fit effect is fitted before the linear and quadratic term. Fitting the lack-of-fit term implies allowing a separate effect for each treatment. The model cannot be more general. In fact, fitting the lack-of-fit term always corresponds to the saturated model. Already having fitted the saturated model, addition of a regression term cannot further reduce the error SS. An alternative way of putting this is to say that the lack-of-fit effect uses up all the treatment d.f.

Note that it is important to list LACKFIT last in the model statement to obtain the correct fitting order. To illustrate, consider the following call of GLM:

```
proc glm;
class lackfit;
model yield=lackfit fert fert*fert;
run;
```

Output:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LACKFIT	4	913.56266667	228.39066667	41.76	0.0001
FERT	0	0.00000000	.	.	.
FERT*FERT	0	0.00000000	.	.	.
Source	DF	Type III SS	Mean Square	F Value	Pr > F
LACKFIT	2	1.59085714	0.79542857	0.15	0.8664
FERT	0	0.00000000	.	.	.
FERT*FERT	0	0.00000000	.	.	.

Now, we obtain the treatment SS under the label LACKFIT, because LACKFIT is fitted first. This is equivalent to doing a qualitative one-way ANOVA. There are no d.f. left for the regression terms, so their SS are zero. These SS are, of course, not the appropriate ones.

To fit the selected model (quadratic polynomial), one needs to drop the (non-significant) lack-of-fit term and use the code

```
proc glm;
class lackfit;
model yield=fert fert*fert/solution;
run;
```

At the end of the output, we find the least squares estimates of the parameters (μ, β_1, β_2) of the model $y_{ij} = \mu + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$:

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
INTERCEPT	9.692380952	8.24	0.0001	1.17669271
FERT	0.417768707	10.49	0.0001	0.03982528
FERT*FERT	-0.002033042	-7.45	0.0001	0.00027278

From this, the fitted polynomial is

$$Y = 9.69 + 0.418 * \text{FERT} - 0.00203 * \text{FERT}^2$$

Exercise 5.11: Verify all regression results for the fertiliser data, which are stored under **polynomial_fertiliser.dat**. Generate a plot of the quadratic model by the following code:

```
symbol interpol=rq value=dot;
proc gplot;
plot yield*fert;
run;
```

To obtain confidence limits around the regression line, replace the option INTERPOL=RQ with INTERPOL=RQCLM in the SYMBOL statement (RQ stands for quadratic regression, CLM for confidence limit of a mean, i.e. the expected value at a given value for x_i). This option produces valid confidence limits results only for data from a completely randomised design, but is inappropriate for other designs, e.g., the RCB design. Theoretical details are omitted here.

Exercise 5.12: As mentioned at the beginning of this chapter, the fertiliser data were, in fact, obtained from an experiment laid out in randomised complete blocks. The dataset available in **polynomial_fertiliser_rcb.dat** and contains coding for blocks. Do a polynomial regression and the lack-of-fit testing taking this design into account. Hint: All models remain unaltered with respect to the treatment structure. To account for blocking, just add a block effect. Thus, base the lack-of-fit test for a quadratic regression on the model

$$y_{ij} = \mu + b_j + \beta_1 x_i + \beta_2 x_i^2 + \delta_i + e_{ij}$$

where b_j is the effect of the j -th block. Further hint: Be sure to fit the block effect before all other effects. This will be important for unbalanced data in the same way as for an analysis with a qualitative treatment factor. Generate unbalanced data from the complete fertiliser data by deleting the last observation (last treatment, last replicate) and run GLM with different fitting orders for the block effect. Convince yourself that F-values depend on the order of fitting terms.

For presentation purposes, one will want to report a regression equation. The model implies a separate regression for each block, where the regressions just differ in their intercept. The intercept for the j -th block is

$$\mu + b_j$$

Thus, it seems reasonable to estimate the mean of the intercept term, given by

$$\mu + \frac{b_1 + b_2 + b_3}{3}$$

The mean intercept can be estimated using the ESTIMATE statement. If blocks are coded by BLOCK, the following SAS code does the job.

```
proc glm;
  class block lackfit;
  model yield=block fert fert*fert/solution;
  estimate 'mean intercept' intercept 3 block 1 1 1/divisor=3;
run;
```

Exercise 5.13: When discussing the randomised complete block design, the following example was used. An experiment was conducted to assess the yield (kg/ha) of rice cultivar IR8 at six different seeding densities (kg/ha) (Gomez and Gomez, 1984):

Density (kg/ha)	Block			
	1	2	3	4
25	5113	5398	5307	4678
50	5346	5952	4719	4264
75	5272	5713	5483	4749
100	5164	4831	4986	4410
125	4804	4848	4432	4748
150	5254	4542	4919	4098

Find the data under **ir8.dat**. We had analysed this experiment treating density as a qualitative factor. Now analyse the experiment using polynomial regression. Compare the results.

5.8.1 How many x-levels?

A typical design question is how many levels of the explanatory variable, x , should be tested. There is no simple answer, but generally there will be an optimum depending on the type of response. In most applications, three to six levels will be sufficient. There is no simple answer because the optimal design depends on the true dose-response relationship, which is usually unknown.

If the response is linear over a range from x_{min} to x_{max} , the best design is obtained by taking half the observations at x_{min} and the other half at x_{max} . This is so because the standard error of the slope estimate is inversely proportional to SS_x , the sum of squares of x . With simple linear regression, we have

$$s.e.(\hat{\beta}) = \sqrt{\frac{\sigma^2}{SS_x}}$$

For a given sample size, this is minimized by maximizing SS_x , and, if we require that x falls between x_{min} and x_{max} , SS_x is maximized by taking half the observations at x_{min} and the other half at x_{max} . The described design is also "G-optimal", i.e., the maximum of the variance of

$$\hat{\alpha} + \hat{\beta}x$$

over the interval $[x_{min}, x_{max}]$ is minimized.

The (G-) optimal design for a quadratic response, i.e., the design for which the variance of

$$\hat{\alpha} + \hat{\beta}_1x + \hat{\beta}_2x^2$$

over the interval $[x_{min}, x_{max}]$ is minimized, is obtained by taking the same number of observations at x_{min} , x_{max} and $(x_{min} + x_{max})/2$ (Rasch et al., 1998, Verfahrensbibliothek II).

In both cases, optimality of the design depends crucially on our knowledge of the form of true underlying response, and this is where things break down: Usually, we do not know *a priori*, whether the response is going to be linear, quadratic, cubic, or other. If, e.g., the true response is quadratic, but we have erroneously assumed a linear response and chosen the design optimal for that case, our design will not be optimal for the true underlying response.

These considerations indicate that the design of an experiment with a quantitative factor is not straightforward. A look at G-optimality and similar criteria is useful nonetheless, because it shows what should not be done: it is not a good idea to distribute design points (x -values) evenly so that there is just one observation per design point. It is usually better to concentrate on a few design points. If the response is expected to be convex or concave over the observed range, a response that can be well approximated by a quadratic, three or four x -points should suffice. If a sigmoidal shape is more likely (e.g., logistic), one may want to increase the number of points by one or two. Generally, the number of design points should be a little

higher (two or three levels more) than the number of parameters in the contemplated nonlinear regression models. The extra levels then allow testing the lack-of-fit.

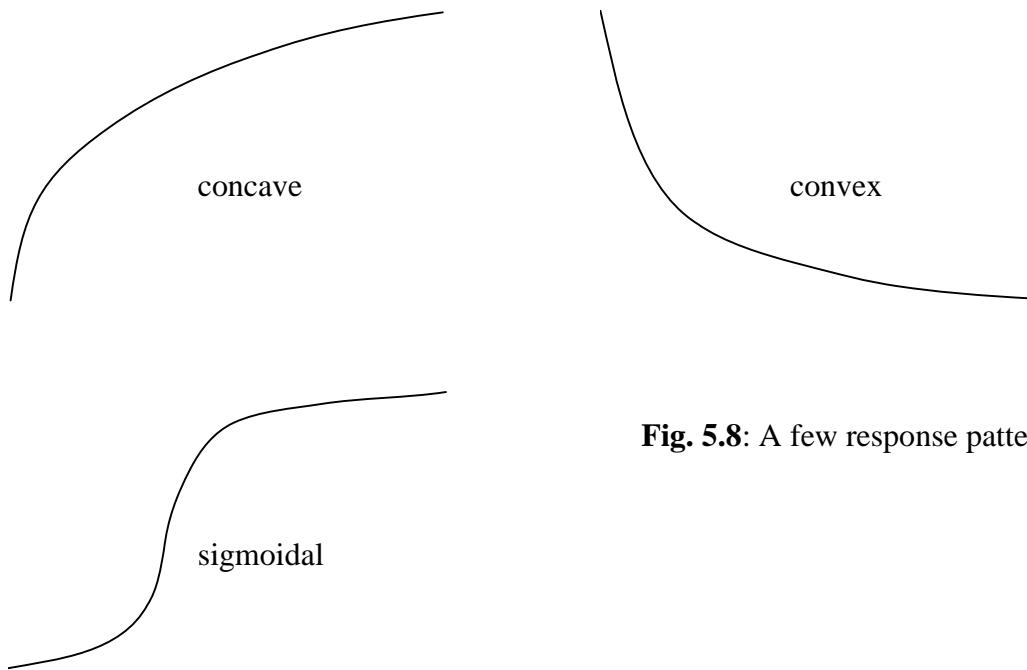


Fig. 5.8: A few response patterns

5.9 Analysis of covariance

An important aspect of experimental design is error control, i.e., minimisation of the experimental error variance σ^2 . This can be achieved in a number of ways. First of all, the experimenter will do everything possible to keep the experimental conditions as constant as possible for all experimental units. Secondly, blocking often can be used as a means to further reduce the error variance. The objective is best achieved if conditions **within blocks** are as homogeneous as possible, with any differences among experimental units occurring mainly **between blocks**. For example, in a field experiment blocks will be laid out along a gradient of an important environmental variable (soil fertility, water supply, shadow, wind) associated with experimental units. To make efficient use of blocking it is necessary that important factors varying among experimental units can be identified in advance and that a grouping of units into homogeneous blocks is possible. These circumstances may not always hold in practice, so blocking is not guaranteed to actually reduce the error variance.

Analysis of covariance is a second statistical approach for error control. Similar to blocking, this method exploits covariate information on experimental units. However, it does not use grouping into homogeneous groups. Rather, it accepts that there is heterogeneity among experimental units with respect to some covariate (fertility or other). The idea is to quantitatively measure the covariate for each unit and then use this information in a **regression analysis**, which allows one to answer the question: "what would have been the value of the response, had all experimental units had the same value for the covariate. For example, we may ask: What would have happened, had all plots on a field had the same clay content. In order to apply analysis of covariance (ANCOVA) techniques, it is necessary that the covariate can be quantitatively measured and that the covariate is **not** related to/influenced by the treatment factors. The importance of the latter point cannot be over-emphasised. A fool-proof check of whether covariate and treatment are unrelated is to make sure that the covariate is measured **before** different treatments are applied. However, covariates can also

be useful if measured after application of treatments, as will be seen in the second example below (Example 5.17).

In many applications, blocking and analysis of covariance are combined in the same experiment. In case of blocking, ANCOVA may account for residual heterogeneity among experimental units within blocks, thus further reducing the error variance.

Example 5.16: A feeding experiment with pigs was conducted to assess the effect of four different types of feed on the daily weight gains. It is known that daily gains depend on the initial weight of the animals at the onset of the feeding period. Differences in initial weight cannot be avoided. Some statistical method can be used to control this source of experimental error. Blocking is one obvious option here, were initial weights within blocks (groups) are as homogeneous as possible. A second, more commonly employed option is to use initial weight as a covariate in ANCOVA.

Example 5.17: 11 varieties of lima beans were compared form ascorbic acid content (Steel and Torrie, 1980: 411). From previous experience it was known that increase in maturity resulted in decrease in vitamin C (ascorbic acid) content. Since all varieties were not of the same maturity at harvest and since all plots of the same variety did not reach the same level of maturity on the same day, it was not possible to harvest all plots at the same stage of maturity. Hence, the percentage of dry matter based on 100 g of freshly harvested beans was observed as an index of maturity and used as a covariate (preceeding text verbatim from Steel and Torrie).

The basis of ANCOVA is a regression of the response variable (e.g., daily gain or ascorbic acid content) on the covariate (e.g., initial weight or maturity index). A regression is performed for each treatment. Provided that regression lines are parallel for different treatments (this assumption is important and must be tested!!!), the vertical distances among lines are suitable measures of treatment differences.

Example 5.18: The principles of ANCOVA will be explained using the following hypothetical example, which is rather extreme to demonstrate the salient features. The data pertain to two different feeds administered to six pigs each.

Feed 1		Feed 2	
Initial weight X (pound)	Gain Y (pound)	Initial weight X (pound)	Gain Y (pound)
41	0.89	67	0.94
23	0.76	73	1.01
18	0.75	55	0.88
33	0.89	67	0.92
25	0.71	72	0.99
30	0.88	53	0.81
Mean:	0.81		0.93

The two groups of animals for the two feeds differ markedly in initial weight. In real applications one would ensure by proper randomisation that initial weights are more evenly distributed! Despite randomisation, however, group differences in mean initial weight cannot

be entirely avoided. The effect of such differences will be demonstrated using the present extreme example. The following graph plots daily gains versus initial weights for both feeds. There is a clear positive association between response (gain) and covariate (initial weight), as highlighted by the regression lines.

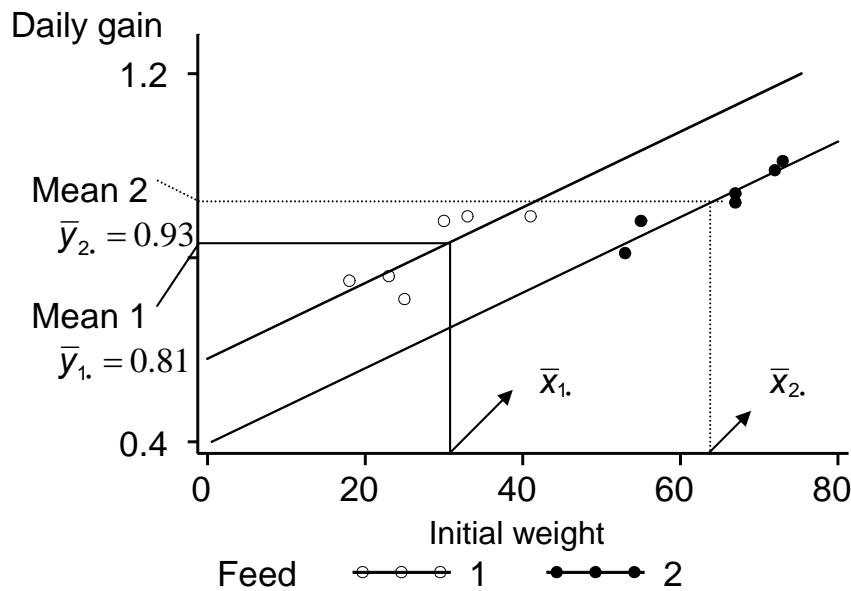


Fig. 5.9: Parallel regression lines and **simple means** of daily weight gain for pig feeding data.

Feed 1 has the smaller daily gains (mean: 0.81) compared to feed 2 (mean: 0.93), but also the smaller initial weights. The regression line for feed 1 lies above that for feed 2. **Thus, it can be concluded that feed 1 would have the better gains, if the initial weights for all animals and both feed were the same.** This is true despite the fact that feed 1 had the smaller gains in the experiment. The clue is the difference in initial weights, which needs to be accounted for.

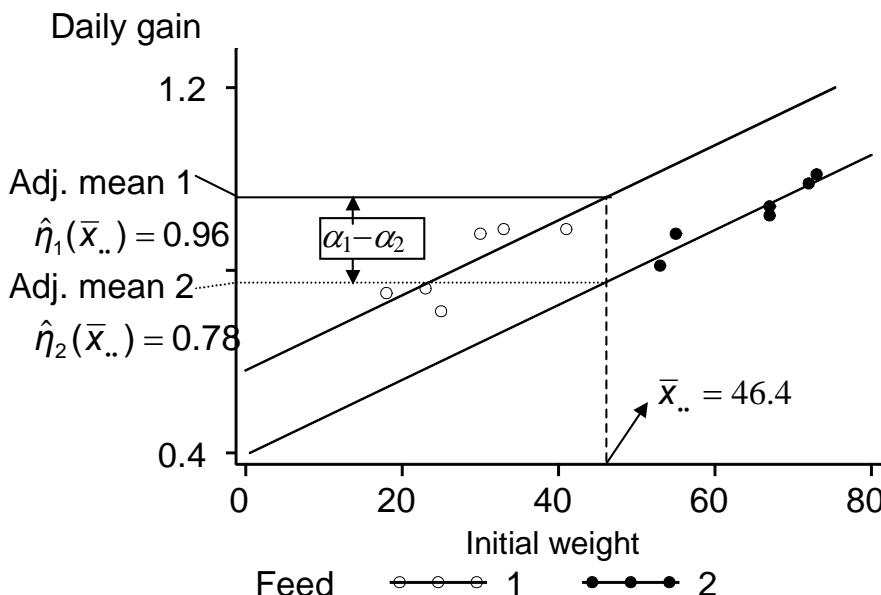


Fig. 5.10: Parallel regression lines and **adjusted means** of daily weight gain for pig feeding data (ANCOVA).

ANCOVA adjusts simple means for the effect of the covariate. The resulting **adjusted means** estimate the expected value of the response (gains) at a constant value for the covariate (initial weight). Of course, the same initial weight is assumed for both treatments. Usually, one takes the average of all initial weights as a point of reference. At this point, adjusted means have the smallest standard error. The average of all initial weights in our example is 46.4 pounds. Thus, for both feeds we estimate the daily gain to be expected at this initial weight based on the regression lines. This is illustrated in the figure above. The adjusted mean for feed 1 lies above the adjusted mean for feed 2 (while the unadjusted, simple mean for feed 2 was above that for feed 1!). The difference in adjusted means is identical to the vertical distance of the two parallel regression lines. The example clearly demonstrates that ignoring the covariate we would have obtained a biased result. The bias is so extreme that feed 2 would have been judged better than feed 1, while ANCOVA reveals the opposite result.

ANCOVA is based on the crucial assumption that regression lines are, in fact, parallel, and that the response is linear. The idea of ANCOVA can be extended to curvilinear responses, for example by adding a quadratic term, so long as the vertical distances among curves remain constant across the domain of the covariate. In any case, the assumption of parallelism must be tested as will be outlined below.

In addition to a correction for differences in values for the covariate, ANCOVA usually results in a reduction in error variance σ^2 . This is so because the portion of the experimental error, which can be modelled by the covariate, can be separated from the error variance, thus reducing the error term for (adjusted) mean comparisons. This gain in accuracy will be exemplified later using another example.

5.9.1 Models

The ANCOVA model is an extension of the simple ANOVA model for one qualitative treatment factor. Assuming a completely randomised design (the extension to other designs is straightforward and will be dealt with in exercises), the ANOVA model is

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

(μ = general mean; α_i = i -th treatment effect, y_{ij} = j -th replicate observation of i -th treatment). This model is extended by a regression term as follows:

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + e_{ij} \quad (5.5)$$

where x_{ij} is the value of the covariate for observation y_{ij} and β is a common slope for the regression lines. Errors e_{ij} are assumed to follow a normal distribution with zero mean and variance σ^2 as usual [$e_{ij} \sim N(0, \sigma^2)$]. This model describes a separate regression line for each treatment, with a common slope, but intercept depending on treatment. The intercept of the regression for the i -th treatment equals $\mu + \alpha_i$.

Based on the ANCOVA model (5.5), treatment differences are assessed by looking at the expected value for a given value of the covariate. The expected value of the i -th treatment at covariate value x_{ij} is

$$\eta_i(x_{ij}) = E(y_{ij}|x_{ij}) = \mu + \alpha_i + \beta x_{ij}$$

The **adjusted mean for the i -th treatment** is defined as the predicted value of the regression at the overall mean of the covariate:

$$\eta_i(\bar{x}_{\bullet\bullet}) = E(y_{ij} | \bar{x}_{\bullet\bullet}) = \mu + \alpha_i + \beta \bar{x}_{\bullet\bullet}$$

The adjusted means can be estimated by plugging in the least squares solutions for the model effects. This is achieved in GLM using LSMEANS (refer to SAS hints below).

Example 5.18: For the hypothetical feeding example, the least squares estimate of the adjusted means are as follows:

	Covariate $\bar{x}_{i\bullet}$	Simple mean Response $\bar{y}_{i\bullet}$	Adjusted mean Response $\hat{\eta}_i(\bar{x}_{\bullet\bullet})$
Feed 1	28.333	0.81	0.96
Feed 2	64.500	0.93	0.78
(see Fig. 5.8)			(see Fig. 5.9)
$(\bar{x}_{\bullet\bullet} = 46.417)$			

Note that Feed 2 has the better adjusted mean, while it has an inferior unadjusted (simple) mean. The mean for Feed 1 is adjusted upwards because the mean for the covariate is below average. Similarly, the mean for Feed 2 is adjusted downwards because the mean for the covariate is above average (see Figs. 5.8 and 5.9).

To test for significant treatment differences, consider the **reduced** model, relative to the above **full** model:

$$y_{ij} = \mu + \beta x_{ij} + e_{ij} \quad (5.6)$$

Under the reduced model, the regression lines of all treatments are superimposed, i.e., the vertical distance and hence treatment differences are zero. By contrast, under the full model, the regression lines, while parallel, differ in intercept, i.e., the vertical distances differ from zero, and hence there are treatment differences.

ANCOVA is based on the assumption of parallelism. To test this assumption, we need to consider a more general model, i.e., a model which allows slopes to differ among treatments. The extended model reads

$$y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + e_{ij} \quad (5.7)$$

where β_i is the slope for the i -th treatment. As before, the intercept of the regression for the i -th treatment equals $\mu + \alpha_i$.

Under model (5.7), there is an **interaction** between covariate and treatment variable. The set-up is similar to that encountered in Chapter 4 for the Child Health data. The null hypothesis that regression lines are parallel is

$$H_0: \beta_1 = \beta_2 = \dots = \beta_r. \quad (5.8)$$

Observe that (5.7) is a full model relative to the reduced model (5.5), just as (5.5) is a full model relative to the reduced model (5.6). Thus, models (5.6), (5.5), and (5.7) form a natural model building sequence.

For testing the null hypothesis of parallelism (5.8) and to emphasise that model (5.6) is nested within (5.7), it is useful to reparameterise the regression parameter β_i in much the same way as we have reparameterised the expected value μ_i in the one-way ANOVA model ($y_{ij} = \mu + \alpha_i + e_{ij}$) as a sum of general effect and treatment effect ($\mu_i = \mu + \alpha_i$). By analogy, we may write

$$\beta_i = \beta + \delta_i$$

where β is a common slope and δ_i is the treatment-specific deviation from the common slope. With this reparameterisation, model (5.7) can be re-expressed as

$$y_{ij} = \mu + \beta x_{ij} + \alpha_i + \delta_i x_{ij} + e_{ij}$$

The null hypothesis of parallelism then is

$$H_0: \delta_1 = \delta_2 = \dots = \delta_t$$

Note that the δ -effects play the role of a lack-of-fit effect. Also, $\delta_i x_{ij}$ may be interpreted as an interaction term. The sequence of models can be stated as follows:

Model	Description	Error SS	Reduction (RSS)	Degrees of freedom of reduction
$y_{ij} = \mu + e_{ij}$	Common mean	$SS(\mu)$		
$y_{ij} = \mu + \beta x_{ij} + e_{ij}$	Common line	$SS(\mu, \beta)$	$RSS(\beta/\mu) = SS(\mu) - SS(\mu, \beta)$	1
$y_{ij} = \mu + \beta x_{ij} + \alpha_i + e_{ij}$	Parallel lines	$SS(\mu, \beta, \alpha_i)$	$RSS(\alpha_i/\mu, \beta) = SS(\mu, \beta) - SS(\mu, \beta, \alpha_i)$	$t - 1$
$y_{ij} = \mu + \beta x_{ij} + \alpha_i + \delta_i x_{ij} + e_{ij}$	Non-parallel lines	$SS(\mu, \beta, \alpha_i, \delta_i)$	$RSS(\delta_i/\mu, \beta, \alpha_i) = SS(\mu, \beta, \alpha_i) - SS(\mu, \beta, \alpha_i, \delta_i)$	$t - 1$

t = number of treatments

ANOVA table:

Source	d.f.	SS
β (covariate, unadjusted)	1	$RSS(\beta/\mu)$
$\alpha_i \beta$ (treatments, adjusted)	$t - 1$	$RSS(\alpha_i/\mu, \beta)$
$\delta_i \beta, \alpha_i$ (lack-of-fit)	$t - 1$	$RSS(\delta_i/\mu, \beta, \alpha_i)$
Error	$N - 2t$	$SS(\mu, \beta, \alpha_i, \delta_i)$

t = number of treatments ; N = total number of observations

In the model building sequence, it is important that the covariate (β) be fitted before the treatment effect (α_i). This ensures that the F-test for treatments is adjusted for the covariate. In other words, the F-tests is for differences among **adjusted treatment means**. By contrast, if the covariate is fitted after the treatment effect, the F-test for treatment would compare **unadjusted treatment means**.

Example 5.16: In a feeding experiment with pigs, four different feeds were tested on ten animals each to assess the effect on daily weight gains (Y) (Snedecor and Cochran, 1967, p.440). Moreover, the initial weight was recorded as a covariate (X). We want to perform an F-test for treatment differences using initial weight as a covariate. The raw data are as follows:

Feed 1		Feed 2		Feed 3		Feed 4	
x	y	x	y	x	y	x	y
61	1.40	74	1.61	80	1.67	62	1.40
59	1.79	75	1.31	61	1.41	55	1.47
76	1.72	64	1.12	62	1.73	62	1.37
50	1.47	48	1.35	47	1.23	43	1.15
61	1.26	62	1.29	59	1.49	57	1.22
54	1.28	42	1.24	42	1.22	51	1.48
57	1.34	52	1.29	47	1.39	41	1.31
45	1.55	43	1.43	42	1.39	40	1.27
41	1.57	50	1.29	40	1.56	45	1.22
40	1.26	40	1.26	40	1.36	39	1.36

We find the following sequence:

Model	Error SS	Reduction
$y_{ij} = \mu + e_{ij}$	1.0228	
$y_{ij} = \mu + \beta x_{ij} + e_{ij}$	0.8731	$RSS(\beta \mu) = 0.1497$
$y_{ij} = \mu + \beta x_{ij} + \alpha_i + e_{ij}$	0.7043	$RSS(\alpha_i \beta, \mu) = 0.1688$
$y_{ij} = \mu + \beta x_{ij} + \alpha_i + \delta_i x_{ij} + e_{ij}$	0.6790	$RSS(\delta_i \alpha_i, \beta, \mu) = 0.0253$

Thus, we find the following ANOVA table:

Source	d.f.	SS	MS	F	p-value
β (covariate)	1	0.1497	0.1497	7.06	0.0122
α_i (treatments, adjusted)	3	0.1688	0.0563	2.65	0.0654
δ_i (lack-of-fit)	3	0.0253	0.0084	0.40	0.7556
Error	32	0.6790	0.0212		

The lack-of-fit term δ_i is not significant [$F_{exp} = 0.40 < F_{tab}(t-1=3, N-2t=32, \alpha = 5\%) = 2.90$], so the null hypothesis of parallel lines is not rejected. Hence, we may look at the test for (adjusted) treatment effects. This is not significant at the 5% level, though quite close, so there is some evidence of treatment differences. Note that the covariate is fitted before the treatment effect, so we are testing adjusted treatment means. The following adjusted means are found:

Feed	Adjusted mean
1	1.455
2	1.306
3	1.449
4	1.342

The pairwise comparisons of adjusted means cannot be performed with a common LSD. The reason is that the standard error of a difference differs among pairs of treatments. This is due to a dependence on the mean values of the covariate for the treatments (no formula is given here for brevity). Despite this problem, we can try to obtain a letters display. For the example, the LSMEANS statement produces the following output:

TRT	Y LSMEAN	Pr > T	H0: LSMEAN(i)=LSMEAN(j)			
			i/j	1	2	3
1	1.45499493	1 .	0.0253	0.9242	0.0881	
2	1.30676923	2 0.0253	.	0.0322	0.5843	
3	1.44889772	3 0.9242	0.0322	.	0.1030	
4	1.34233812	4 0.0881	0.5843	0.1030	.	

We try to derive a lines display manually, applying the standard procedure for balanced designs (see section 5.6). Note that in the cross-classification below, treatments are ordered by means. We test at $\alpha = 5\%$, so p-values < 0.05 indicate significance.

	[1] (1.455)	[3] (1.449)	[4] (1.342)	[2] (1.307)
[1] (1.455)		ns	ns	*
[3] (1.449)			ns	*
[4] (1.342)				ns
[2] (1.307)				

ns: not significant; *: significant at $\alpha = 5\%$.

Lines display:

[1]	[3]	[4]	[2]
—————	—————	—————	—————

Absorb redundant line and assign letters:

[1]	[3]	[4]	[2]
—————	—————(a)	—————	—————(b)

Means display:

Feed	Adjusted mean [§]
1	1.455 ^a
2	1.307 ^b
3	1.449 ^a
4	1.342 ^{ab}
Average s.e.d.	0.064

§ Means followed by the same letter
are not significantly different by a t-test at $\alpha = 5\%$.

SAS hints

To fit the model

$$y_{ij} = \mu + \beta x_{ij} + \alpha_i + \delta_i x_{ij} + e_{ij}$$

and obtain sequential reduction in error SS for the sequence of models (5.5), (5.6) and (5.7), we use the following code:

```
data;
input feed x y;
datalines;
1 61 1.40
1 59 1.79
1 76 1.72
1 50 1.47
1 61 1.26
1 54 1.28
1 57 1.34
1 45 1.55
1 41 1.57
1 40 1.26
2 74 1.61
2 75 1.31
2 64 1.12
2 48 1.35
2 62 1.29
2 42 1.24
2 52 1.29
2 43 1.43
2 50 1.29
2 40 1.26
3 80 1.67
3 61 1.41
3 62 1.73
3 47 1.23
3 59 1.49
3 42 1.22
3 47 1.39

```

```

3 42 1.39
3 40 1.56
3 40 1.36
4 62 1.40
4 55 1.47
4 62 1.37
4 43 1.15
4 57 1.22
4 51 1.48
4 41 1.31
4 40 1.27
4 45 1.22
4 39 1.36
;
proc glm;
class feed;
model y= x feed x*feed;
run;


```

Note that the term X*FEED in the model statement generates a separate lack-of-fit effect δ_i for each treatment (equivalent to an interaction term), because the FEED variable is listed in the CLASS statement and the term X*FEED appears as the last effect in the MODEL statement and hence is fitted last in the sequence. Also, FEED is fitted after X, so FEED yields test for adjusted means.

Output:

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	0.34381653	0.04911665	2.31	0.0498
Error	32	0.67896097	0.02121753		
Corrected Total	39	1.02277750			

R-Square	C.V.	Root MSE	Y Mean
0.336160	10.49252	0.14566238	1.38825000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X	1	0.14972299	0.14972299	7.06	0.0122
FEED	3	0.16878208	0.05626069	2.65	0.0654
X*FEED	3	0.02531146	0.00843715	0.40	0.7556

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X	1	0.12481514	0.12481514	5.88	0.0211
FEED	3	0.00905782	0.00301927	0.14	0.9338
X*FEED	3	0.02531146	0.00843715	0.40	0.7556

The sequential SS are found under Type I SS. Note that Type III are not appropriate for our analysis. The lack-of-fit test is not significant, so we drop the lack-of-fit term and compute adjusted means based on the reduced ANCOVA model (5.6)

$$y_{ij} = \mu + \beta x_{ij} + \alpha_i + e_{ij}$$

To do all pairwise comparisons by a t-test, we need to switch to the GLIMMIX procedure and use the following code:

```
proc glimmix;
class feed;
model y= x feed;
lsmeans feed/pdiff lines;
run;
```

The LINES option provides the lines (letter) display:

```
T Comparison Lines for Least Squares Means of feed

LS-means with the same letter are not significantly different.

      LSMEAN
      y LSMEAN    feed   Number

      A 1.4549949    1       1
      A
      A 1.4488977    3       3
      A
      B  A 1.3423381    4       4
      B
      B     1.3067692    2       2
```

The output of LSMEANS in GLM does not provide standard errors of differences. This can be obtained by using the MIXED or GLIMMIX procedures instead of GLM:

```
proc mixed;
class feed;
model y= x feed;
lsmeans feed/pdiff;
run;
```

Differences of Least Squares Means

Effect	feed	_feed	Estimate	Standard			
				StdErr	DF	tValue	Probt
feed	1	2	0.1482	0.06345	35	2.34	0.0253
feed	1	3	0.006097	0.06363	35	0.10	0.9242
feed	1	4	0.1127	0.06421	35	1.75	0.0881
feed	2	3	-0.1421	0.06373	35	-2.23	0.0322
feed	2	4	-0.03557	0.06441	35	-0.55	0.5843
feed	3	4	0.1066	0.06364	35	1.67	0.1030

The average of the standard errors can be added to a table of adjusted means for descriptive purposes. The average is quite meaningful, because the individual standard errors differ only marginally.

Exercise 5.14: Reproduce all results for the pig-feeding experiment (data in **ancova_pig.dat**).

Exercise 5.15: So far, ANCOVA has been considered for data from a completely randomised design. The technique is also available for other designs. A randomized complete block design was used to test 11 varieties of lima beans and compare their ascorbic acid content (Y) (Steel and Torrie, 1980: 411; **Example 5.17**). From previous experience it was known that increase in maturity resulted in a decrease of ascorbic acid content. Since all varieties were not of the same maturity at harvest and since all plots of the same variety did not reach the same level of maturity on the same day, it was not possible to harvest all plots at the same stage of maturity. Hence, the percentage of dry matter based on 100 g of freshly harvested beans (X) was observed as an index of maturity and used as a covariate (preceding text verbatim from Steel and Torrie). The data are as follows (Steel and Torrie, 1980: 412; see **ancova_lima.Bean.dat**):

Variety	Replicate (block)									
	1		2		3		4		5	
	X	Y	X	Y	X	Y	X	Y	X	Y
1	34.0	93.0	33.4	94.8	34.7	91.7	38.9	80.8	36.1	80.2
2	39.6	47.3	39.8	51.5	51.2	33.3	52.0	27.2	56.2	20.6
3	31.7	81.4	30.1	109.0	33.8	71.6	39.6	57.5	47.8	30.1
4	37.7	66.9	38.2	74.1	40.3	64.7	39.4	69.3	41.3	63.2
5	24.9	119.5	24.0	128.5	24.9	125.6	23.5	129.0	25.1	126.2
6	30.3	106.6	29.1	111.4	31.7	99.0	28.3	126.1	34.2	95.6
7	32.7	106.1	33.8	107.2	34.8	97.5	35.4	86.0	37.8	88.8
8	34.5	61.5	31.5	83.4	31.1	93.9	36.1	69.0	38.5	46.9
9	31.4	80.5	30.5	106.5	34.6	76.7	30.9	91.8	36.8	68.2
10	21.2	149.2	25.3	151.6	23.5	170.1	24.8	155.2	24.6	146.1
11	30.8	78.7	26.4	116.9	33.2	71.8	33.5	70.3	43.8	40.9

X = percentage of dry matter based on 100 g of freshly harvested beans

Y = ascorbic acid content in mg per 100 g dry weight

Perform an ANCOVA using Y as a response variable and X as a covariate. Can an ANCOVA be performed? Are there significant differences among varieties? **Hint:** You can use the same procedures as those used so far. The only modification is to add a block effect to the model. For example, the modified model (5.7) reads

$$y_{ij} = \mu + b_j + \beta x_{ij} + \alpha_i + \delta_i x_{ij} + e_{ij}$$

where b_j is the effect of the j -th block. Be sure to fit blocks before the other effects. Derive a lines display for mean comparisons (warning: this is tedious if done manually, though straightforward, due to the large number of treatments. You can use the LINES option in GLM do let the computer do it).

Exercise 5.16: A field experiment was performed to compare four cultivars of sugar beets (Munzert, p. 132). The layout was a randomized complete block design. The number of plants per plot was preplanned to a fixed size, but seed emergence was not complete, so some of the seeded plants were missing. Total plot yield is known to depend on the plant density, so the number of missing plants (MISS) on a plot was recorded as a covariate, along with plot yield (YIELD; kg/20 m²) (**beet.dat**).

Block	Cultivar	Miss	Yield
1	1	9	118
1	2	11	128
1	3	7	116
1	4	10	117
2	1	10	124
2	2	7	131
2	3	8	112
2	4	9	122
3	1	12	115
3	2	13	120
3	3	11	108
3	4	13	112
4	1	9	119
4	2	11	121
4	3	10	110
4	4	12	111

Do an analysis of covariance to compare the four sugar beet cultivars. In addition to the usual ANCOVA adjusted means, which are adjusted at $x = \bar{x}_{..}$, compute adjusted means at zero missing plants ($x = 0$). SAS hint: LSMEANS CULTIVAR/AT MISS=0;

6. Factorial experiments

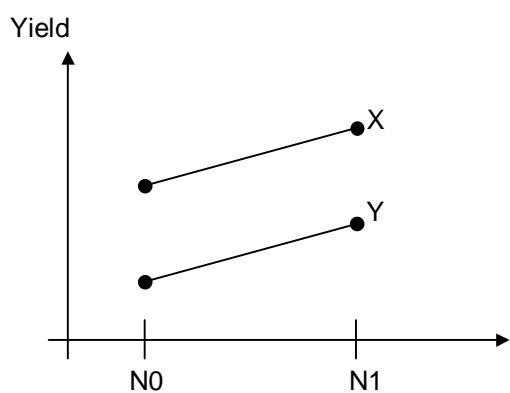
6.1 Interaction

In Chapter 5 we have looked at experiments, in which levels of one factor were systematically varied. To broaden the scope of an experiment, one may investigate more than one factor. Such experiments are called factorial experiments.

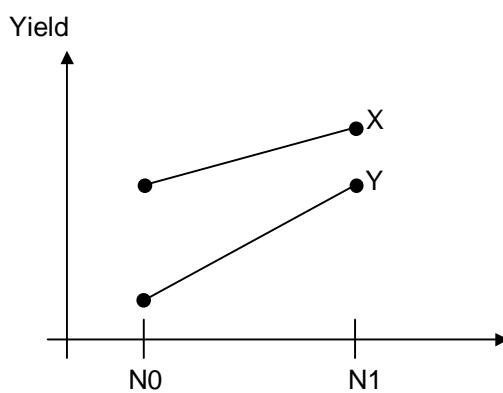
A main purpose of factorial experiments is the analysis of **interaction**. We speak of interaction in a statistical sense when differences among the levels of one factor depend on the levels of another factor.

Example 6.1: The following figure shows the response of two cultivars X and Y to two different fertilizer treatments N0 and N1.

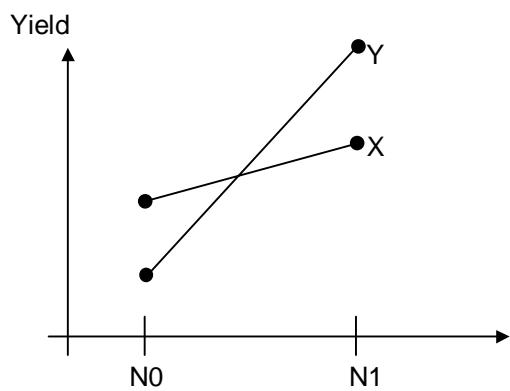
(a)



(b)



(c)



In Fig. (a) yield increases from N0 to N1 are the same for both cultivars and hence the lines connecting the dots for both cultivars run parallel, i.e., their vertical distance is constant. This implies that the yield differences among cultivars are independent of the level of N-fertilisation. In this case there are no interactions. By contrast, in figures (b) and (c) there is interaction. In Fig. (c), the interaction is so strong, that the rank order of both cultivars is different between both N-levels. This type of interaction is called **rank interaction**.

6.2 Mean comparisons

When both treatment factors are qualitative, one will want to compare treatment means. In an experiment with two factors, there are different types of mean comparison. Which of these is appropriate mainly depends on whether or not interactions are significant.

Example 6.2 (hypothetical): Assume that an experiment is performed with two cultivars (X and Y) and two types of fertilizer (N0 and N1), giving rise to $2 \times 2 = 4$ treatments. Each of the four treatments is tested on three plots ($n = 3$). Further assume that there are **significant interactions** and that the four treatment means for yield (t/ha) are as follows:

		Fertilizer		Marginal means for cultivars
		N0	N1	
Cultivar	X	1	2	1.5
	Y	4	3	3.5
Marginal means for fertilizers		2.5	2.5	

The four treatment means will be referred to as **cell means**. The cell means show that for each cultivar there are differences among the two fertilizers. Also, the response differs among cultivars. Specifically, cultivar X performs better with fertilizer N1, while cultivar Y fares better with fertilizer N0.

In addition to simple treatment means, one can compute **marginal means**. For example, we may compute **marginal means** for each fertilizer across the two cultivars. The marginal means will be based on $2n = 6$ observations. The marginal means are the same for both fertilizers. From this, we might be tempted to conclude that there is no difference among the two fertilizers. The conclusion is obviously wrong, however, because for each cultivar there is a difference among the cell means. **Thus, marginal means are misleading in the presence of interaction.**

Similarly, for each cultivar we can compute marginal means across fertilizers. The difference in marginal means is 2 t/ha in favor of cultivar Y. One might conclude that the advantage of cultivar Y is 2 t/ha for each fertilizer, but this conclusion is false because there is significant interaction. Specifically, the advantage is 3 t/ha when fertilizer N0 is applied, while the advantage is only 1 t/ha when fertilizer N1 is used. Thus, while it is true that the "average advantage" is 2 t/ha, this average advantage is of little practical relevance.

Example 6.2 shows that **a comparison of marginal means is meaningless in case there is interaction**. Treatment comparisons should be solely based on cell means (fertilizers separately for each cultivar; cultivars separately for each fertilizer), when interaction is significant.

Example 6.3 (hypothetical): Assume, as in Example 6.2, that an experiment is performed with two cultivars (X and Y) and two types of fertilizer (N0 and N1). The trait of interest is yield (t/ha). Each of the four treatments is tested on three plots ($n = 3$). Further assume that there is **no significant interaction** and that the four treatment means are as follows:

		Fertilizer		Marginal means for cultivars
		N0	N1	
Cultivar	X	1	2	1.5
	Y	3	4	3.5
Marginal means for fertilizers		2	3	

The marginal means for fertilizer show a difference of 1 t/ha in favour of N1. This difference applies not only to the marginal means, but also to the cell means: The fertilizer difference is 1 t/ha for both cultivars. This is the case because there is no interaction. Similarly, marginal means show an advantage of 2 t/ha for cultivar Y. The advantage is also 2 t/ha for cell means, no matter whether fertilizer N0 or N1 is applied.

Obviously, in this case, it makes sense to compare marginal means, because there is no interaction. In fact it is better to compare marginal means here for two reasons:

- (1) The analysis is simpler
- (2) Marginal means are based on $2n = 6$ observations, while cell means are computed from $n = 3$ observations. Thus, marginal means are more accurate, and so are comparisons among marginal means.

Example 6.3 shows that **a comparison of marginal means is useful and in fact preferable to a comparison of cell means when there are no interactions.**

The exposition so far has shown that a crucial question in the analysis of an experiment with two factors is whether or not the interaction is significant. If there is interaction we compare cell means, holding one of the two factors constant at a time. For example, we compare cultivars separately for each fertilizer level. Similarly, we may compare fertilizers separately for each cultivar. Conversely, if there is no interaction, we may just compare marginal means for fertilizer and marginal means for cultivar.

6.3 Linear model

In what follows we will assume for simplicity that the experiment was completely randomized. The linear model for an experiment with factors fertilizer and cultivar can be written as:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

where

- y_{ijk} = yield of k -th replicate of i -th fertilizer with j -th cultivar
- μ = intercept
- α_i = main effect for i -th fertilizer
- β_j = main effect for j -th cultivar

$(\alpha\beta)_{ij}$ = interaction of i -th fertilizer and j -th cultivar

e_{ijk} = error of y_{ijk}

Depending on the pattern of response to the two factors, some of the model effects may be dropped, as shown in Fig. 6.1.

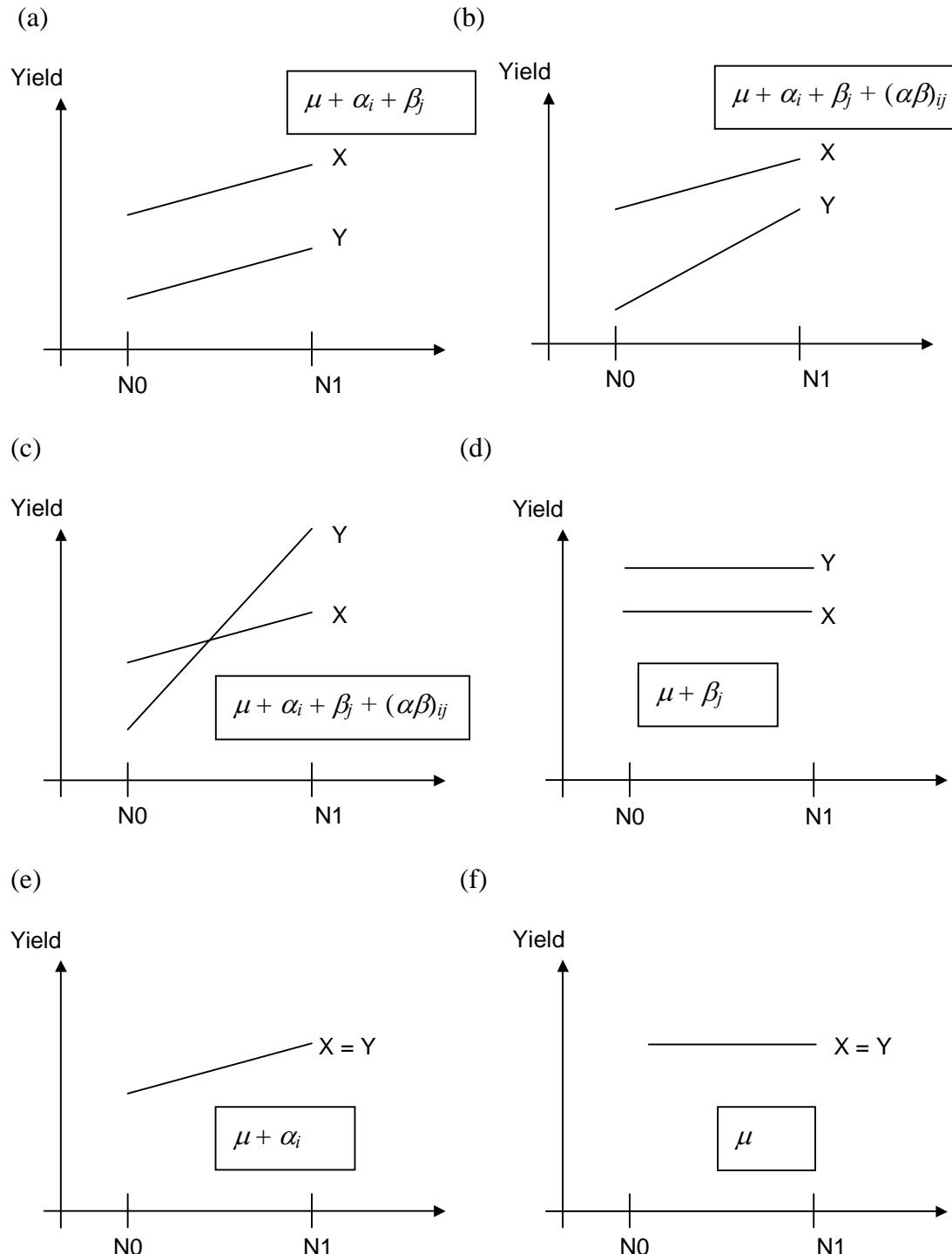


Fig. 6.1: Some examples for response patterns in a factorial experiment with two cultivars (X, Y) and two fertilizers (N0, N1). In each case the appropriate linear model (systematic part) is stated. α : fertilizer; β : cultivar.

6.3.1 Analysis of variance

An ANOVA based on model (6.1) will reveal which of the effects is significant and thus needed to adequately model the data. When performing an ANOVA, it is important to note that the primary test of interest is that for interaction. Only if interaction is not significant does it make sense to test main effects (see Chapter 4). A significant main effect implies significant differences of the corresponding marginal means (Tab. 6.1).

Tab. 6.1: Useful mean comparisons in a factorial experiment with fertilizer (α_i) and cultivar (β_j), depending on which terms are significant.

Model (significant terms only)	Example in Fig. 6.1	Type of mean comparisons
$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	(b), (c)	Simple cultivar \times fertilizer means (cell means)
$\mu + \alpha_i + \beta_j$	(a)	Marginal fertilizer means Marginal cultivar means
$\mu + \beta_j$	(d)	Marginal cultivar means
$\mu + \alpha_i$	(e)	Marginal fertilizer means
μ	(f)	None

To decide on the type of mean comparison, we need to first select an appropriate model, and this may be done by analysis of variance. Fig. 6.2 gives a simple decision tree for selecting the type of comparison based on the analysis of variance. The most important feature of the tree is that the test for interaction is the first step of the analysis.

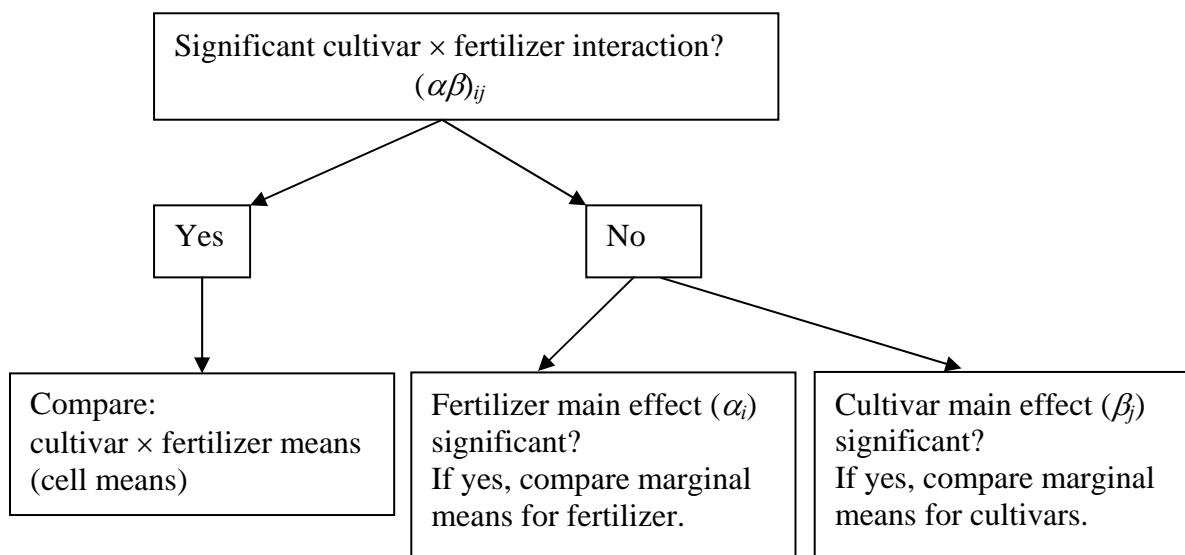


Fig. 6.2: Decision tree for analysis of two-factorial experiment with factors fertilizer (α_i) and cultivar (β_j).

The analysis of variance is based on the sequence of models given in Table 6.2. In fact, two sequences need to be considered, except when the data are balanced. In either case, main effects need to be fitted before the interaction term. In the sense of Nelder (1994) the interaction is marginal to main effects: It makes no sense to fit the term $(\alpha\beta)_{ij}$ first, because the resulting model would be saturated, i.e., a separate mean would be fitted for each factorial combination. Having fitted the term $(\alpha\beta)_{ij}$, adding a main effect cannot further reduce the error SS. Thus, main effects need to be fitted before the interaction. For unbalanced data, it matters which of the two main effects is fitted first. Therefore, two sequences need to be considered for unbalanced data.

Table 6.2: Sequence of models for analysis of two-factorial experiment with fertilizer (α) and cultivar (β).

(a) Factor α (fertilizer) fitted first:

Model	Error SS	Reduction
μ	$SS(\mu)$	
$\mu + \alpha_i$	$SS(\mu, \alpha_i)$	$RSS(\alpha_i \mu)$
$\mu + \alpha_i + \beta_j$	$SS(\mu, \alpha_i, \beta_j)$	$RSS(\beta_j \mu, \alpha_i)$
$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	$SS[\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij}]$	$RSS[(\alpha\beta)_{ij} \mu, \alpha_i, \beta_j]$

(b) Factor β (cultivar) fitted first:

Model	Error SS	Reduction
μ	$SS(\mu)$	
$\mu + \beta_j$	$SS(\mu, \beta_j)$	$RSS(\beta_j \mu)$
$\mu + \alpha_i + \beta_j$	$SS(\mu, \alpha_i, \beta_j)$	$RSS(\alpha_i \mu, \beta_j)$
$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	$SS[\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij}]$	$RSS[(\alpha\beta)_{ij} \mu, \alpha_i, \beta_j]$

The two sequences give rise to two ANOVA tables (Table 6.3).

Table 6.3: ANOVA tables for analysis of two-factorial experiment with fertilizers (α) and cultivars (β). It is assumed that all factorial combinations are observed.

(a) Factor α (fertilizer) fitted first:

Source (prose)	Source (model)	d.f.	SS
Fertilizer, ignoring cultivar (main effect)	α_i , ignoring β_j	$a - 1$	$RSS(\alpha_i \mu)$
Cultivar, adjusted for fertilizer (main effect)	β_j , adjusted for α_i	$b - 1$	$RSS(\beta_j \mu, \alpha_i)$
Interaction cultivar by fertilizer	$(\alpha\beta)_{ij}$	$(a - 1)(b - 1)$	$RSS[(\alpha\beta)_{ij} \mu, \alpha_i, \beta_j]$
Error		$N - ab$	$SS[\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij}]$

(b) Factor β (cultivar) fitted first:

Source (prose)	Source (model)	d.f.	SS
Cultivar, ignoring fertilizer (main effect)	β_j , ignoring α_i	$b - 1$	$RSS(\beta_j \mu)$
Fertilizer, adjusted for cultivar (main effect)	α_i , adjusted for β_j	$a - 1$	$RSS(\alpha_i \mu, \beta_j)$
Interaction cultivar by fertilizer	$(\alpha\beta)_{ij}$	$(a - 1)(b - 1)$	$RSS[(\alpha\beta)_{ij} \mu, \alpha_i, \beta_j]$
Error		$N - ab$	$SS[\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij}]$

a = number of fertilizer levels

b = number of cultivars

N = total number of observations

The first table is needed to test the main effect for fertilizer (adjusted for cultivar), while the second is needed to test the main effect for cultivar (adjusted for fertilizer). These tests are relevant only when the interaction is not significant. Note that both tables produce the same test for interactions.

For **balanced data**, the two tables are the same because in that case

$$RSS(\alpha_i|\mu, \beta_j) = RSS(\alpha_i|\mu) \text{ and}$$

$$RSS(\beta_j|\mu, \alpha_i) = RSS(\beta_j|\mu)$$

Thus, the order of fitting of main effects is immaterial (though the interaction always needs to be fitted after main effects in either case).

Example 6.4 (Steel and Torrie, 1980): Wilkinson (1954, Ph.D. thesis, University of Wisconsin, Madison) reports the results of an experiment to study the influence of time of bleeding, factor A, and diethylstilbestrol (an estrogenic compound), factor B, on plasma phospholipid in lambs. Five lambs were assigned at random to each of four treatment groups;

treatment combinations are for morning and afternoon times of bleeding with and without diethylstilbestrol treatment (completely randomized design). The data are shown in Table 6.4.

Table 6.4: The influence of time of bleeding and diethylstilestrol on phospholipids in lambs (**lambs.dat**).

Time of bleeding			
A.M.		P.M.	
Control	Treated	Control	Treated
8.53	17.53	39.14	32.00
20.53	21.07	26.20	23.80
12.53	20.80	31.33	28.87
14.00	17.33	45.80	25.06
10.80	20.07	40.20	29.33
Mean	13.28	19.36	36.53
			27.81

Inspection of the cell means suggests that there is strong interaction: The lambs treated with diethylstilbestrol show an increase of phospholipids compared to the control, when time of bleeding was in the morning, while the treated lambs show a decline compared to the control in the afternoon. The significance of this interaction may be tested by ANOVA.

The sequential SS for the lamb data are shown in Table 6.5.

Table 6.5: Sequential SS for lamb data.

(a) Factor time (α) fitted first:

Type of SS	Error SS	Value	Reduction
$SS(\mu)$		1919.33	
$SS(\mu, \alpha_i)$		662.58	$RSS(\alpha_i \mu) = 1256.75$
$SS(\mu, \alpha_i, \beta_j)$		653.87	$RSS(\beta_j \mu, \alpha_i) = 8.71$
$SS[\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij}]$		379.92	$RSS[(\alpha\beta)_{ij} \mu, \alpha_i, \beta_j] = 273.95$

(b) Factor treatment (β) fitted first:

Error SS		
Type of SS	Value	Reduction
$SS(\mu)$	1919.33	
$SS(\mu, \beta_j)$	1910.62	$RSS(\beta_j \mu) = 8.71$
$SS(\mu, \alpha_i, \beta_j)$	653.87	$RSS(\alpha_i \mu, \beta_j) = 1256.75$
$SS[\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij}]$	379.92	$RSS[(\alpha\beta)_{ij} \mu, \alpha_i, \beta_j] = 273.95$

The sequential SS in Table 6.5 show that the order of fitting of main effects is immaterial; the resulting reductions in error SS are the same. This independence of the order of fitting is a result of the balancedness of the data. The ANOVA is as follows (modified SAS output):

Source	DF	Type I SS	Mean Square	F Value	Pr > F
time	1	1256.75	1256.75	52.93	<.0001
trt	1	8.71	8.71	0.37	0.5532
interaction	1	273.95	273.95	11.54	0.0037
Error	16	379.92	23.75 $\Rightarrow s^2$		

6.3.2 Mean comparisons

The interaction is highly significant ($p = 0.0037$). Thus, it is not useful to compare marginal means, and we therefore ignore the test of main effects. Instead, we may compare the two treatments separately for each time (morning, afternoon). For **balanced data**, an LSD can be computed from the ANOVA table. The LSD for comparing two treatment means is

$$LSD = t_{tab} \times s.e.d.$$

where

$$s.e.d. = \sqrt{\frac{2s^2}{n}} \quad (\text{standard error of a difference})$$

$$s^2 = \frac{SS[\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij}]}{N - ab} \quad (\text{error MS; the usual ANOVA estimate of the error variance } \sigma^2)$$

n = number of replications per treatment

t_{tab} = tabular t-value with error d.f. = $N - ab + 1$)

For the lamb data, we find

$$N = 20, a = b = 2, n = 5$$

Error d.f. = 16

$t_{tab} = 2.120$

$s^2 = 23.75$ (taken from ANOVA table)

$$s.e.d. = \sqrt{\frac{2 \times 23.75}{5}} = 3.08$$

$$LSD = 2.120 \times 3.08 = 6.53$$

	Time of bleeding	
	Morning	Afternoon
Control	13.28	36.53
Diethylstilestrol	19.36	27.81

$$LSD(5\%) = 6.53$$

The difference among control and Diethylstilestrol means is not significant in the morning, but there is a significant difference in the afternoon.

SAS hints

Two-way ANOVA

```
data;
input
time$ trt$ phospholipid;
datalines;
am control 8.53
am control 20.53
am control 12.53
am control 14.00
am control 10.80
am treated 17.53
am treated 21.07
am treated 20.80
am treated 17.33
am treated 20.07
pm control 39.14
pm control 26.20
pm control 31.33
pm control 45.80
pm control 40.20
pm treated 32.00
pm treated 23.80
pm treated 28.87
pm treated 25.06
pm treated 29.33
;
proc glm;
```

```

class time trt;
model phospholipid=time trt time*trt;
run;

```

To compute cell means, use the LSMEANS statement. Note that cell means are least squares estimates of

$$\eta_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

```

proc glm;
class time trt;
model phospholipid=time trt time*trt;
lsmeans time*trt/pdiff;
run;

```

Output:

```

The GLM Procedure
Least Squares Means

          phospholipid      LSMEAN
time     trt           LSMEAN    Number
          am       control   13.2780000   1
          am       treated   19.3600000   2
          pm       control   36.5340000   3
          pm       treated   27.8120000   4

```

```

Least Squares Means for effect time*trt
Pr > |t| for H0: LSMean(i)=LSMean(j)

```

```
Dependent Variable: phospholipid
```

i/j	1	2	3	4
1		0.0660	<.0001	0.0002
2	0.0660		<.0001	0.0145
3	<.0001	<.0001		0.0121
4	0.0002	0.0145	0.0121	

Instead of GLM, we can use the MIXED procedure, which in the case at hand does essentially the same computations as GLM, but produces a more useful output for computing the LSD.

```

proc mixed;
class time trt;
model phospholipid=time trt time*trt;
lsmeans time*trt/pdiff;
run;

```

Output:

Effect	time	trt	Estimate	Error	DF	t Value	Pr > t
time*trt	am	control	13.2780	2.1792	16	6.09	<.0001
time*trt	am	treated	19.3600	2.1792	16	8.88	<.0001
time*trt	pm	control	36.5340	2.1792	16	16.76	<.0001
time*trt	pm	treated	27.8120	2.1792	16	12.76	<.0001

Differences of Least Squares Means

Effect	time	trt	_time	_trt	Estimate	Standard Error	DF	t Value		Pr > t
								t Value	Pr > t	
time*trt	am	control	am	treated	-6.0820	3.0819	16	-1.97	0.0660	
time*trt	am	control	pm	control	-23.2560	3.0819	16	-7.55	<.0001	
time*trt	am	control	pm	treated	-14.5340	3.0819	16	-4.72	0.0002	
time*trt	am	treated	pm	control	-17.1740	3.0819	16	-5.57	<.0001	
time*trt	am	treated	pm	treated	-8.4520	3.0819	16	-2.74	0.0145	
time*trt	pm	control	pm	treated	8.7220	3.0819	16	2.83	0.0121	

The bottom of the output lists all pairwise differences with standard errors of a difference (s.e.d.) and results of t-tests. Not all comparisons are relevant here! We are interested only in comparing two treatments at the same time, or different times for the same treatment. Clearly, due to its length, the list of pairwise comparisons is not convenient for reporting. The main advantage of the output is that it provided the s.e.d. (3.08199) and the error d.f. (DF=16). This is all we need to compute the LSD. You can read the tabular t-value from a table and compute the LSD by hand. Alternatively, you can let SAS do the job. The table of differences can be stored in to a SAS dataset using the **output delivery system** (ODS). The commands are as follows:

```

quit;
ods output diffs=diffs;
proc mixed;
class time trt;
model phospholipid=time trt time*trt;
lsmeans time*trt/pdiff;
run;

```

Important hint: Often, it is necessary to type the QUIT; command before running ODS. This closes any preceding call of another procedure. Sometimes, without the quite statement, no ODS output is generated. A call of PROC PRINT shows the data set DIFFS resulting from the ODS statements.

```

proc print data=diffs;
run;

```

Output:

Obs	Effect	time	trt	_time	_trt	Estimate	StdErr	DF	tValue	ProbT
1	time*trt	am	control	am	treated	-6.0820	3.0819	16	-1.97	0.0660
2	time*trt	am	control	pm	control	-23.2560	3.0819	16	-7.55	<.0001
3	time*trt	am	control	pm	treated	-14.5340	3.0819	16	-4.72	0.0002
4	time*trt	am	treated	pm	control	-17.1740	3.0819	16	-5.57	<.0001
5	time*trt	am	treated	pm	treated	-8.4520	3.0819	16	-2.74	0.0145
6	time*trt	pm	control	pm	treated	8.7220	3.0819	16	2.83	0.0121

The output is about the same as in the printed output of MIXED. However, some of the labels have changed, and this is important for further processing. For example, the *s.e.d.* appeared

under "Standard error" in the original output. It is now printed under "StdErr". To compute the LSD, we need to obtain the tabular t-value for DF=16 d.f. The following code does this:

```
data ttab;
DF=16;
alpha=0.05;
ttab=tinv(1-alpha/2, DF);

proc print; run;
```

Output:

Obs	DF	alpha	ttab
1	16	0.05	2.11991

This computation of t_{tab} can be integrated into the data set DIFFS, and the LSD can be computed as follows (note that DIFFS contains the d.f. under the variable DF):

```
data diffss;
set diffss;
alpha=0.05;
ttab=tinv(1-alpha/2, DF);
lsd=ttab*stderr;

proc print data=diffss;
run;
```

Output:

	E	s	t	s	t	v	p	a	t	l	s	d	
	f	f	t	i	t	m	d	a	r	l	t	l	
0	e	c	i	m	r	i	a	E	l	o	p	t	
b	t	m	t	r	m	t	r	D	u	b	h	a	
s	t	e	t	e	t	e	r	F	e	t	a	b	
1	time*trt	am	control	am	treated	-6.0820	3.0819	16	-1.97	0.0660	0.05	2.11991	6.53333
2	time*trt	am	control	pm	control	-23.2560	3.0819	16	-7.55	<.0001	0.05	2.11991	6.53333
3	time*trt	am	control	pm	treated	-14.5340	3.0819	16	-4.72	0.0002	0.05	2.11991	6.53333
4	time*trt	am	treated	pm	control	-17.1740	3.0819	16	-5.57	<.0001	0.05	2.11991	6.53333
5	time*trt	am	treated	pm	treated	-8.4520	3.0819	16	-2.74	0.0145	0.05	2.11991	6.53333
6	time*trt	pm	control	pm	treated	8.7220	3.0819	16	2.83	0.0121	0.05	2.11991	6.53333

From the output we find LSD = 6.5333.

A very convenient to compare simple means us to use the SLICE statement of the GLIMMIX procedure. For example, if you want to compare treatments separately for each time point, i.e., you want to look a slices by time, you can use this code:

```
proc glimmix;
class trt time;
```

```

model phospholipid=trt|time;
slice trt*time/sliceby=time lines;
run;

```

F Test for trt*time Least Squares Means Slice

Slice	Num	Den	F Value	Pr > F
	DF	DF		
time am	1	16	3.89	0.0660

T Grouping for trt*time Least Squares Means Slice (Alpha=0.05)

LS-means with the same letter are not significantly different.

Slice	trt	Estimate	
time am	treated	19.3600	A
time am			A
time am	control	13.2780	A

F Test for trt*time Least Squares Means Slice

Slice	Num	Den	F Value	Pr > F
	DF	DF		
time pm	1	16	8.01	0.0121

The GLIMMIX Procedure

T Grouping for trt*time Least Squares Means Slice (Alpha=0.05)

LS-means with the same letter are not significantly different.

Slice	trt	Estimate	
time pm	control	36.5340	A
time pm			A
time pm	treated	27.8120	B

For each slice, the output first shows so-called simple F-tests. These test if there is a difference among means in that slice. Next comes a table of means together with a lines display. But you will not obtain an LSD here.

6.3.3 Unbalanced data

Example 6.5: (Dr. Tafaj, Fachgebiet Tierernährung, Universität Hohenheim). A feeding experiment was conducted with 72 pigs to study the effect of two types of additives: copper (mg/kg diet) and phytase (U/kg diet; U = μ mol substrate per minute = unit for activity of enzyme) (**pigs.dat**). Each additive was tested in three levels, giving rise to a total of nine treatments. The pigs were randomly allocated to treatments (completely randomized design). The treatments were as follows:

	Phytase (U/kg)			
	0	250	500	
Copper mg/kg	20			
	80			
	175			

Additives of microbial phytase to feed rations are thought to increase the availability of zink (Pallauf, 1992). Zink, in turn, can interact with copper and reduce the availability of copper (O'Dell, 1997). The objective of the experiment was to test whether an improved availability of zink by addition of microbial phytase reduces copper availability and whether this reduction can be offset by addition of copper. Among others, the following data were recorded on each pig:

Phyt = Amount of phytase added (U/kg diet)
 Cu = Amount of copper added (mg/KG diet)
 CuS = Copper concentratum in blood serum (mg/L)
 CuL = total copper content in liver (mg)
 CuG = copper concentration in bile (mg/L)

The two treatment factors are quantitative, so polynomial regression could be considered. Nevertheless, we will analyse the two factors as if they were qualitative (but see Exercise 6.5). Of the three response variables (CuS, CuL, CuG), we will analyse CuG, the copper concentration in the bile. Two observations are missing for this variable, so the data are unbalanced. The sequential SS are listed in Table 6.6.

Table 6.6: Sequential SS for pig data (CuG).

(a) Factor copper (α) fitted first:

Type of SS	Error SS		
$SS(\mu)$	747.22		
$SS(\mu, \alpha_i)$	430.42	$RSS(\alpha_i \mu) = 316.80$	
$SS(\mu, \alpha_i, \beta_j)$	421.15	$RSS(\beta_j \mu, \alpha_i) = 9.26$	
$SS[\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij}]$	391.87	$RSS[(\alpha\beta)_{ij} \mu, \alpha_i, \beta_j] = 29.28$	

(b) Factor phytase (β) fitted first:

Error SS			
Type of SS	Value	Reduction	
$SS(\mu)$	747.22		
$SS(\mu, \beta_j)$	737.42	$RSS(\beta_j \mu) = 9.80$	
$SS(\mu, \alpha_i, \beta_j)$	421.15	$RSS(\alpha_i \mu, \beta_j) = 316.27$	
$SS[\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij}]$	391.87	$RSS[(\alpha\beta)_{ij} \mu, \alpha_i, \beta_j] = 29.28$	

It should be noted that the reductions in error SS are not the same for both orders of fitting in Table 6.6. This is because the data are unbalanced. The ANOVA Tables for both sequences are as follows:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Cu	2	316.8005303	158.4002652	25.06	<.0001
Phyt	2	9.2636248	4.6318124	0.73	0.4847
Phyt*Cu	4	29.2771827	7.3192957	1.16	0.3380
Error	62	391.8742339	6.3205522		

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Phyt	2	9.7963876	4.8981938	0.77	0.4651
Cu	2	316.2677676	158.1338838	25.02	<.0001
Phyt*Cu	4	29.2771827	7.3192957	1.16	0.3380
Error	62	391.8742339	6.3205522		

The interaction is not significant, so we may look at the tests for main effects. The correct test for phytase (adjusted for copper) is obtained from the first table and is not significant. The second table shows that the copper main effect is significant. Thus, phytase does not seem to reduce the copper concentration in the bile. However, addition of copper results in a significant change of copper availability in the bile. In other words: the copper availability is affected more by the copper supplementation than by phytase. The ANOVA suggests that the model can be reduced to

$$\mu + \alpha_i \quad (6.1)$$

where α_i is the copper main effect. This is just the one-way ANOVA model for copper. The copper means are estimated by fitting the reduced model. The least squares estimates of the means (6.1) are:

Cu-Level	CuG (mg/kg)
20	3.19
80	3.94
175	7.99

Not unexpectedly, copper availability in the bile increased with copper added.

A common LSD cannot be computed, because the *s.e.d.* differs among pairs. In the case at hand it is given by

$$s.e.d.(\hat{\alpha}_i - \hat{\alpha}_{i'}) = s \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$$

where n_i is the number of observations for the i -th copper level and s is the square root of the ANOVA error mean square for the reduced model (6.1). We find the following

Comparison	s.e.d.
20 vs. 80	0.7341
20 vs. 175	0.7341
80 vs. 175	0.7263

Pairwise t-tests are performed by comparing

$$t_{obs} = \frac{|\hat{\alpha}_i - \hat{\alpha}_{i'}|}{s.e.d.(\hat{\alpha}_i - \hat{\alpha}_{i'})}$$

to a t-distribution with d.f. equal to the d.f. of the error MS.

SAS hints

The two model sequences are fitted by

```
proc glm;
  class phyt cu;
  model CuG=cu phyt cu*phyt;
  run;

proc glm;
  class phyt cu;
  model CuG=phyt cu cu*phyt;
  run;
```

Pairwise differences based on the reduced model may be tested by

```
proc glm;
  class phyt Cu;
  model CuG=cu;
  lsmeans cu/pdiff;
  run;
```

or

```

proc mixed;
class phyt Cu;
model CuG=cu;
lsmeans cu/pdiff;
run;

```

Output:

Differences of Least Squares Means

Effect	Cu	_Cu	Estimate	Standard		t Value	Pr > t
				Error	DF		
Cu	20	80	-0.7479	0.7341	68	-1.02	0.3119
Cu	20	175	-4.8009	0.7341	68	-6.54	<.0001
Cu	80	175	-4.0529	0.7263	68	-5.58	<.0001

There is no significant difference between the two low doses, while the two low doses are significantly different from the highest dose. Thus a high level of copper supplementation in the diet (175 mg/kg diet) can reduce the negative effect of phytase addition on the copper availability, though no such negative effect was detected for the bile (effect of phytase is not significant). Phytase effects were detected, however, for copper in the liver (see Exercise 6.2).

6.3.4 More on comparison of means

As stated above (e.g., Tab. 6.1), the appropriate types of mean comparisons depend on the model selected by the two-way ANOVA. This is elaborated in Table 6.7. As an example, consider the case where the interaction is not significant, but both main effects are. The selected model for a completely randomized design is

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

Thus, the expected value of an observation for the i -th level of factor A and the j -th level of factor B is

$$\mu + \alpha_i + \beta_j$$

The marginal mean for A is defined as the average of this expected value across levels of B:

$$\eta_i = \frac{\sum_{j=1}^b (\mu + \alpha_i + \beta_j)}{b} = \mu + \alpha_i + \bar{\beta}_i$$

where b is the number of levels for factor B and

$$\bar{\beta}_i = \frac{\sum_{j=1}^b \beta_j}{b}$$

Generally, the means are estimated by least squares, i.e., parameters in the expression for the means are replaced by their least squares estimates/solutions:

Table 6.7: Useful mean comparisons in factorial experiment with factors A (α_i) and B (β_j), depending on which terms are significant. Table assumes a completely randomized design. For block designs and row-column designs, results are basically the same (just need to add block effects and rows/column effect, respectively).

Model (significant terms only)	Type of mean comparisons	Model expression for means	Estimate for balanced data	s.e.d. for balanced data [§]
$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	Simple A \times B (cell) means	$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	$\bar{y}_{ij\bullet}$	$s\sqrt{\frac{2}{n}}$
$\mu + \alpha_i + \beta_j$	Marginal A means	$\mu + \alpha_i + \bar{\beta}_\bullet$	$\bar{y}_{i\bullet\bullet}$	$s\sqrt{\frac{2}{nb}}$
	Marginal B means	$\mu + \bar{\alpha}_\bullet + \beta_j$	$\bar{y}_{\bullet j\bullet}$	$s\sqrt{\frac{2}{na}}$
$\mu + \alpha_i$	Marginal A means	$\mu + \alpha_i$	$\bar{y}_{i\bullet\bullet}$	$s\sqrt{\frac{2}{nb}}$
	Marginal B means	$\mu + \beta_j$	$\bar{y}_{\bullet j\bullet}$	$s\sqrt{\frac{2}{na}}$
μ	None	-	-	-

§ s = square root of ANOVA mean square for error

n = number of replicated per treatment (A \times B combination); in block designs, we denote the number of replicates by r .

a = number of levels for factor A

b = number of levels for factor B

$$\hat{\eta}_i = \frac{\sum_{j=1}^b (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)}{b}$$

(hat notation indicates least squares solution of parameter). It turns out that for balanced data, the least squares estimate of the marginal mean takes a particularly simple form:

$$\hat{\eta}_i = \bar{y}_{i\bullet\bullet}$$

where

$$\bar{y}_{i\bullet\bullet} = \frac{\sum_{j=1}^b \sum_{k=1}^n y_{ijk}}{nb},$$

i.e., the least squares marginal mean of the i -th level of factor A is identical to the arithmetic mean of all observations of the i -th level of factor A. **It is generally true for balanced data that least squares means are the same as arithmetic means. This is no longer true for unbalanced data, where generally the least squares means should be used.**

For unbalanced data, the *s.e.d.* depends on the pair of treatments, so it is not possible to compute a common LSD. For balanced data, the *s.e.d.* is constant, so a common LSD may be computed from

$$\text{LSD} = t_{tab} \times s.e.d.$$

Where the *s.e.d.* is computed as given in Table 6.7 and t_{tab} is the tabular t-value with error d.f. from the two-way ANOVA.

All of the above remains valid for block designs and row-column designs, except that block effects/row and column effects need to be added. For example, if the experiment is laid out in blocks, the model with interaction is

$$y_{ijk} = \mu + b_k + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

where b_k is the effect of the k -th block ($k = 1, \dots, r$). When interactions are not significant, while both main effects are significant, the expected value of an observation is

$$\mu + b_k + \alpha_i + \beta_j$$

The marginal mean for factor A is obtained by averaging over blocks and levels of factor B:

$$\eta_i = \frac{\sum_{j=1}^b \sum_{k=1}^r (\mu + b_k + \alpha_i + \beta_j)}{rb} = \mu + \bar{b}_\bullet + \alpha_i + \bar{\beta}_\bullet$$

where r denotes the number of blocks. Again, for balanced data, least squares means are the same as arithmetic means $\hat{\eta}_i = \bar{y}_{i\bullet\bullet}$, and LSDs can be computed using the *s.e.d.*'s given in

Table 6.7.

SAS hints

It is generally recommended to compute means by the LSMEANS statement. A letters display needs to be computed manually (there is a SAS macro to do this semi-automatically available at www.uni-hohenheim.de/bioinformatik/). In case of balanced data and non-significant interactions, letters displays for marginal means may be obtained using the MEANS statement as follows (assuming a completely randomized design), were factors A and B are coded by variables a and b:

```
/*both main effects significant, data balanced*/
proc glm;
class a b;
model y=a b;
means a b/lsd;
run;

/*only main effect A significant, data balanced*/
proc glm;
class a b;
model y=a;
means a/lsd;
run;

/*only main effect B significant, data balanced*/
proc glm;
class a b;
model y=b;
means b/lsd;
run;
```

Exercise 6.1: Reproduce all computations for the lamb data (**lamb.dat**; Example 6.1).

Exercise 6.2: Perform a two-way ANOVA for the traits CuS and CuL in the pigs data (**pigs.dat**). Compute LSDs were appropriate (balanced data). Inspect diagnostic plots of studentized residual using PROC GLIMMIX and ODS GRAPHICS as follows (find the plots in the RESULTS panel of the Explorer window on the left-hand side of the SAS Editor).

```
ods graphics on;
proc glimmix plots=studentpanel;
<statements>
run;
ods graphics off;
```

Exercise 6.3 (Köhler et al., 1984): A greenhouse experiment was conducted to assess the effect of two factors on the yield of a wine variety (**fertilizer_plant_protection.dat**). Factor A: Fertilizer (D1, D2); Factor B: Plant protection (P1, P2, P3). The experiment was laid out according to a completely randomized design.

		Plant protection		
		P1	P2	P3
Fertiliser	D1	21.3	22.3	23.8
		20.9	21.6	23.7
		20.4	21.0	22.6
Mean:		20.9	21.6	23.4
D2		12.7	12.0	14.5
		14.9	14.2	16.7
		12.9	12.1	14.5
Mean:		13.5	12.8	15.7

Yield

•: P1; o: P2; □: P3

The above figure shows that the fertiliser effect is similar for the three plant protection measures, suggesting lack of interaction. Note that the abscissa of the above figure does not have a quantitative scale, and that we have just "connected the dots" to denote which data points belong to the same level of plant protection. The lines do not correspond to a regression. Show that the following ANOVA table holds, regardless of the order in which main effects are fitted (before interaction; note: data are balanced).

Source	d.f.	SS	F	p-value
Fertiliser	1	296.87	312.49	0.0001
Plant protection	2	17.78	9.36	0.0036
Interaction	2	1.69	0.89	0.4371
Error	12	11.41		

Perform multiple comparisons of appropriate means by t-tests. Compute LSDs were appropriate.

6.4 Split-plot designs

Factorial experiments may be laid out using the same designs as used for experiments with one factor. Each factorial combination is a treatment, and treatments are randomized in the same way as in experiments with a single factor. In field experiments, however, complete randomization or randomization according to an RCBD may be impractical for technical or other reasons.

Example 6.6: An experiment was performed with six N-fertilizer levels (N1, N2, N3, N4, N5, N6) and four rice varieties (V1, V2, V3, V4) (Gomez & Gomez, 1984). The experiment was laid out as a split-plot design as follows:

Step 1: Each block was partitioned into six **main-plots**. Fertilizer treatments were randomly allocated to main plots within blocks. Randomization was done separately for each of three blocks.

N2	N3	N4	N1	N6	N5
----	----	----	----	----	----

Block I

N5	N6	N2	N4	N1	N3
----	----	----	----	----	----

Block II

N3	N1	N6	N5	N4	N2
----	----	----	----	----	----

Block III

Step 2: Each main plot was **split** into four **sub-plots** or **split-plots** (**Thus the term split-plot design**). The four varieties were randomly allocated to sub-plots or split-plots within a main plot. Randomization was done separately for each main-plot.

N2V2	N3V2	N4V3	N1V4	N6V2	N5V2		N5V4	N6V1	N2V3	N4V4	N1V1	N3V1	
N2V4	N3V1	N4V2	N1V3	N6V1	N5V4		N5V3	N6V3	N2V2	N4V2	N1V3	N3V2	
N2V1	N3V4	N4V1	N1V2	N6V3	N5V1		N5V2	N6V2	N2V4	N4V1	N1V2	N3V4	
N2V3	N3V3	N4V4	N1V1	N6V4	N5V3		N5V1	N6V4	N2V1	N4V3	N1V4	N3V3	
Block I						Block II							
N3V3	N1V3	N6V4	N5V2	N4V4	N2V1								
N3V2	N1V2	N6V2	N5V1	N4V2	N2V3								
N3V4	N1V1	N6V1	N5V4	N4V1	N2V4								
N3V1	N1V4	N6V3	N5V3	N4V3	N2V2								
Block III													

It is useful here to randomize N-levels in main plots for several reasons. Rice is an irrigated crop, so plots need to be irrigated/flooded. Plots may be bordered by bunds to minimize water flow between plots both above and below ground. Nevertheless, bunds cannot prevent leaching all together. Thus, if two adjacent plots receive different levels of fertilizer, fertilizer from the high fertilizer plot may leach to the low fertilizer plot. To minimize such border effects, gaps or border strips need to be inserted between plots. If treatments are completely randomized, a lot of experimental area will be devoted to border strips. To reduce the area lost by border strips, one may group (sub-)plots receiving the same fertilizer treatment (a group of sub-plots defines a main-plot). Obviously, no border strips are needed between such plots, because the amount of fertilizer is the same for each plot. The experimental design to do this is the split-plot design.

The randomization for the above split-plot design is obtained as follows:

```
proc plan seed=74325218;
factors block=3 ordered N=6 V=4;
run;
```

Output:

The PLAN Procedure

Factor	Select	Levels	Order
block	3	3	Ordered
N	6	6	Random
V	4	4	Random

block N ---V---

1	2	2 4 1 3
	3	2 1 4 3
	4	3 2 1 4
	1	3 2 1 4
	6	2 1 3 4
	5	2 4 1 3
2	5	4 3 2 1
	6	1 3 2 4
	2	3 2 4 1
	4	4 2 3 1
	1	1 3 2 4
	3	1 2 4 3
3	3	3 2 4 1
	1	3 2 1 4
	6	4 2 1 3
	5	2 1 4 3
	4	4 2 1 3
	2	1 3 4 2

Exercise 6.4: A researcher wants to perform an experiment to test the effect of three soil preparation methods and of five varieties of sorghum on yield. Due to technical reasons, the experiment is to be laid out as a split-plot design with soil preparation methods as main-plots. The researcher plans to have seven replications per treatment. Main plots are to be laid out in complete blocks. Generate a randomized split-plot design for this experiment using PROC PLAN.

6.4.1 Linear model

If the design were a randomized block design, we would analyse according to the following model:

$$y_{ijk} = \mu + b_k + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

where e_{ijk} is the error term corresponding to a plot. A justification for taking this error term random is the randomization of the design. The error term corresponds to the **randomization unit**, i.e., to the plot. In a split-plot design, there are two randomization units, i.e., the main-plot and the sub-plot. Accordingly, the linear model should have two error terms, one for main-plots and one for sub-plots.

Main-plots: Each main-plot corresponds to a combination of level of the main plot factor and block. An example is given below:

N3	N1	N6	N5	N4	N2
----	----	----	----	----	----

Block I

main-plot for $i = 6$ (6-th N-level) and
 $k = 1$ (Block 1)

Thus, for each block \times N combination the model needs to have a separate effect. We denote this effect by f . Specifically, we let

f_{ik} = main-plot error
= error of main-plot in k -th block for i -th N-level

It is assumed that main-plot errors f_{ik} follow a normal distribution with mean 0 and variance σ_f^2 . A short-hand for this assumption is

$$f_{ik} \sim N(0, \sigma_f^2)$$

Similarly, **sub-plots** corresponds to combinations of block \times N-level \times variety. An example is as follows:

N3V3	N1V3	N6V4	N5V2	N4V4	N2V1
N3V2	N1V2	N6V2	N5V1	N4V2	N2V3
N3V4	N1V1	N6V1	N5V4	N4V1	N2V4
N3V1	N1V4	N6V3	N5V3	N4V3	N2V2

Block I

sub-plot for $i = 6$ (6-th N-level),
 $j = 3$ (variety 3), and
 $k = 1$ (Block 1)

Thus, the sub-plot effect needs to be indexed by block, N-level and variety:

e_{ijk} = sub-plot error
= effect of sub-plot in k -th block with i -th N-level and j -th cultivar

The sub-plot error is assumed to follow a normal distribution with zero mean and variance σ^2 :

$$e_{ijk} \sim N(0, \sigma^2)$$

The complete model is written as follows:

$$y_{ijk} = \mu + b_k + \alpha_i + \beta_j + (\alpha\beta)_{ij} + f_{ik} + e_{ijk}$$

6.4.2 Analysis of variance

The ANOVA of linear models up to now had one aspect in common: there was just one error term in the linear model, and thus, all F-tests were based on the ratio of a mean square (MS) for the effect of interest and the residual error mean square. The split-plot design differs in that there are two error terms: main-plot error and sub-plot error. As a result, the ANOVA has two error MS:

Source	d.f. [§]	SS	MS ^{\$}
Blocks (b_k)	($r-1$)	$RSS(b_k \mu)$	
Main effect A (main plot factor)	($a-1$)	$RSS(\alpha_i b_k, \mu)$	
Error (A) (main plot error)	($a-1)(r-1$)	$RSS(f_{ik} \alpha_i, b_k, \mu)$	s_a^2
Main effect B (sub-plot factor)	($b-1$)	$RSS(\beta_j f_{ik}, \alpha_i, b_k, \mu)$	
Interaction A \times B	($a-1)(b-1$)	$RSS[(\alpha\beta)_{ij} \beta_j, f_{ik}, \alpha_i, b_k, \mu]$	
Error (B) (sub-plot error)	$a(b-1)(r-1)$	$SS[(\alpha\beta)_{ij}, \beta_j, f_{ik}, \alpha_i, b_k, \mu]$	s_b^2

§ assuming balanced data. For unbalanced data, the d.f. are computed differently. r = no. of blocks; a = no. of levels for factor A; b = no. of levels for factor B

\$ s_a^2 and s_b^2 symbolize the mean squares for main-plot error and sub-plot error.

We could compute main-plot means for factor A and analyse these according to a randomized complete block design. The error term for this test must be equivalent to that of a block design. It can be shown that $RSS(f_{ik}|\alpha_i, b_k, \mu)$ ("Error A") is, in fact, the appropriate error term. By contrast, the main effect for the sub-plot factor and the interaction need to be tested against a different error term, i.e., "Error B" in the above ANOVA Table. To see why this is so, we need to look at the expected values for the different mean squares in the ANOVA table. This was also done in Section 4.1, where the one-way ANOVA was discussed. It was shown there that the groups MS and the error MS have expected values that differ in a term involving only the treatment effects. Under the null hypothesis of no treatment effects, this term vanished, so the F-statistic was expected to be about unity under the null hypothesis. The same principles may now be applied to the split-plot ANOVA. For simplicity, we restrict attention to the balanced case. The unbalanced case will be discussed later. The expected MS are given in Table 6.8.

Table 6.8: Expected mean squares in the split-plot ANOVA, assuming balanced data.

Source	d.f.	Expected $MS - E(MS)$	H_0 tested by MS	Expected MS under H_0
Blocks	$(r-1)$	$\sigma^2 + b\sigma_f^2 + \frac{ab}{r-1} \sum_{k=1}^r (b_k - \bar{b}_\bullet)^2$	$b_1 = b_2 = \dots$	$\sigma^2 + b\sigma_f^2$
Main effect A	$(a-1)$	$\sigma^2 + b\sigma_f^2 + \frac{br}{a-1} \sum_{i=1}^a (\alpha_i - \bar{\alpha}_\bullet)^2 \quad (*)$	$\alpha_1 = \alpha_2 = \dots$	$\sigma^2 + b\sigma_f^2$
Error (A)	$(a-1)(r-1)$	$\sigma^2 + b\sigma_f^2$	None	$\sigma^2 + b\sigma_f^2$
Main effect B	$(b-1)$	$\sigma^2 + \frac{ar}{(b-1)} \sum_{j=1}^b (\beta_j - \bar{\beta}_\bullet)^2 \quad (*)$	$\beta_1 = \beta_2 = \dots$	σ^2
Interaction	$(a-1)(b-1)$	$\sigma^2 + \frac{r}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b [(\alpha\beta)_{ij} - (\bar{\alpha\beta})_{i\bullet} - (\bar{\alpha\beta})_{\bullet j} + (\bar{\alpha\beta})_{\bullet\bullet}]^2$	$(\alpha\beta)_{11} = (\alpha\beta)_{12} = \dots$	σ^2
Error (B)	$a(b-1)(r-1)$	σ^2	None	σ^2

(*) Assuming there is no interaction!

r = no. of blocks

a = no. of levels for factor A

b = no. of levels for factor B

To construct an F-test for a term (main effects, interaction), we need to find an appropriate error term. The clue here is that the expected value of the error MS should be the same as the expected value of the treatment MS under the null hypothesis (H_0). For example, the expected value of the interaction MS equals σ^2 under the H_0 of no interaction. This is the same as the expected value of the split-plot error MS . Thus, the F-test for interaction is based on the ratio of interaction MS and split-plot error MS . Similarly, the sub-plot main effect is tested against the split-plot error, while the main plot effect is tested against the main-plot error (Table 6.8).

Example 6.6: The rice data (kg/ha) were as follows (**rice.dat**) (data slightly modified from Gomez & Gomez, 1984):

		Block		
		1	2	3
N1	V1	4520	4208	4030
	V2	4034	5044	3840
	V3	3554	2674	3304
	V4	4216	4212	5016
N2	V1	5598	5256	6162
	V2	6682	5948	5316
	V3	4948	6094	5286
	V4	5372	4694	4382
N3	V1	5806	6600	6794
	V2	5738	6307	6732
	V3	5974	5904	6104
	V4	4276	5924	4236
N4	V1	6192	7146	6860
	V2	6869	7072	6744
	V3	5522	5970	6550
	V4	2504	5126	3818
N5	V1	7470	7578	7642
	V2	7862	6324	6666
	V3	7260	6392	6410
	V4	1594	1690	2856
N6	V1	8542	9012	8548
	V2	6318	7567	5736
	V3	5684	7302	5210
	V4	2338	1560	1744

The ANOVA is as follows (note that data are balance, so the order of fitting is immaterial):

Source	d.f.	SS	MS	F	p-value
Main plot stratum:					
Blocks	2	1084820	542410		
N	5	30480453	6096091	10.98	<0.0001
Error (A)	10	5549527	554953	$= E_a$	

Sub plot stratum:

Source	d.f.	SS	MS	F	p-value
Variety	3	89885035	29961678	85.74	<0.0001
Interaction	15	69378044	4625203	13.24	<0.0001
Error (B)	36	12579905	349442	= E_b	

Both main effects and interaction are significant.

SAS hints

The random effect for the main plot error needs to be listed in the RANDOM statement of GLM. Since the main-plot error is indexed by blocks and N-level, the effect is coded formally as an "interaction" between block and N. GLM will compute the expected mean squares when the /TEST option is used on the RANDOM statement.

```

data;
input
n v block yield;
datalines;
1 1 1 4520
1 1 2 4208
<more data>
6 4 2 1560
6 4 3 2744
;
proc glm;
class n v block;
model yield=block n block*n v n*v;
random block*n/test;
run;

```

Output:

Source	Type III Expected Mean Square
block	Var(Error) + 4 Var(n*block) + Q(block)
n	Var(Error) + 4 Var(n*block) + Q(n,n*v)
n*block	Var(Error) + 4 Var(n*block)
v	Var(Error) + Q(v,n*v)
n*v	Var(Error) + Q(n*v)

It is easily verified that the expected MS computed by GLM are the same as those given in Table 6.8. GLM computes expected mean squares based on Type III SS. For balanced data,

these coincide with Type I SS (sequential SS). To generate expected MS based on Type I SS, we need to add the options E1 and SS1 in the MODEL statement:

```
proc glm;
class n v block;
model yield=block n block*n v n*v/e1 ss1;
random block*n/test;
run;
```

The expected *MS* are the same here because the data are balanced. For unbalanced data, the results differ, and I would recommend Type I SS. The split-plot ANOVA is computed as follows:

```
The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: yield

Source          DF      Type I SS      Mean Square      F Value      Pr > F
block           2       1084820        542410        0.98        0.4095
*   n            5       30480453        6096091       10.98       0.0008

Error: MS(n*block)      10       5549527        554953
* This test assumes one or more other fixed effects are zero.

Source          DF      Type I SS      Mean Square      F Value      Pr > F
n*block         10       5549527        554953        1.59        0.1499
*   v            3       89885035        29961678       85.74       <.0001
n*v            15       69378044        4625203       13.24       <.0001

Error: MS(Error)      36       12579905        349442
* This test assumes one or more other fixed effects are zero.
```

The fixed effects assumed to be zero for the test of main effects in the output (see comment marked with an asterisk - *) are the interaction effects (compare this to Table 6.8). This reflects the fact that a test of main effects makes sense only when there are no interactions.

6.4.3 Mean comparisons

Balanced sata

Table 6.9: Standard errors of a difference (*s.e.d.*) and associated error d.f. for balanced split-plot design.

Comparisons	Standard error of a difference (<i>s.e.d.</i>)	Error d.f.
Marginal A means	$s_a \sqrt{\frac{2}{rb}}$	$(a-1)(r-1)$
Marginal B means	$s_b \sqrt{\frac{2}{ra}}$	$a(b-1)(r-1)$
AB means at same level of A	$s_b \sqrt{\frac{2}{r}}$	$a(b-1)(r-1)$
AB means at same level of B	$\sqrt{\frac{2[(b-1)s_b^2 + s_a^2]}{rb}}$	$df_{satterth}$

a (b) = no. of levels of factor A (B); r = no. of blocks;

s_a^2 (s_b^2) = main-plot (sub-plot) error MS

The *s.e.d.* appropriate for mean comparisons depends in the type of comparison. Table 6.8 shows the *s.e.d.*'s and associated error d.f. for the balanced case. Note that in the balanced case cell means are identical to least squares means as for other balanced designs (Section 6.3). For either comparison, the LSD is computed as

$$LSD = t_{tab} \times s.e.d.$$

in the usual way. The *s.e.d.* for comparing AB means at the same level of B are computed from a linear combination of ANOVA MS : s_a^2 and s_b^2 . Consequently, the error d.f. are a weighted mean of the d.f. associated with these two MS . According to the **Satterthwaite method** the d.f. associated with a linear combination of mean squares

$$c_1 MS_1 + c_2 MS_2 + \dots$$

is given by

$$df_{satterth} = \frac{(c_1 MS_1 + c_2 MS_2 + \dots)^2}{\frac{(c_1 MS_1)^2}{df_1} + \frac{(c_2 MS_2)^2}{df_2} + \dots}$$

where c_i are the coefficients of MS_i and df_i are the error d.f. associated with mean squares MS_i . For the comparison of AB means at the same level of B we find the following d.f.:

$$df_{satterth} = \frac{\left(\frac{2s_a^2}{rb} + \frac{2(b-1)s_b^2}{rb} \right)^2}{\frac{\left(\frac{2s_a^2}{rb} \right)^2}{(a-1)(r-1)} + \frac{\left(\frac{2(b-1)s_b^2}{rb} \right)^2}{a(b-1)(r-1)}}$$

SAS hints

Warning: PROC GLM does not compute appropriate *s.e.d.*! Thus, PROC MIXED should be used to do all computations. The code is essentially the same, except for one major difference. In GLM, a random effect is listed both in the MODEL statement and the RANDOM statement. By contrast, in MIXED a random effect is listed only in the RANDOM statement. For balanced data, the NOBOUND option needs to be used to produce the same results as with GLM (The nobound option allows negative variance component estimates. Details cannot be explained here).

```
proc mixed;
class n v block;
model yield=block n v n*v;
random block*n;
run;
```

MIXED uses a different strategy to compute F-tests (Appendix B). In the case at hand, the same result is obtained as with GLM:

Type 3 Tests of Fixed Effects

Effect	Num	Den		Pr > F
	DF	DF	F Value	
block	2	10	0.98	0.4095
n	5	10	10.98	0.0008
v	3	36	85.74	<.0001
n*v	15	36	13.24	<.0001

Type III F-tests are computed by default. To obtain Type I F-tests, use the following code:

```
proc mixed;
class n v block;
model yield=block n v n*v/htype=1;
random block*n;
run;
```

Output:

Type 1 Tests of Fixed Effects

Effect	Num	Den	F Value	Pr > F
	DF	DF		
block	2	10	0.98	0.4095
n	5	10	10.98	0.0008
v	3	36	85.74	<.0001
n*v	15	36	13.24	<.0001

The two ANOVA tables are the same because the data are balanced. Means are computed by the LSMEANS statement (there is no MEANS statement in MIXED). Since interactions are significant, we compute and compare N × variety means. Satterthwaite d.f. are obtained using the DDFM=SATTERTH option in the model statement:

```
proc mixed;
class n v block;
model yield=block n v n*v/ddfm=satterth;
random block*n;
lsmeans n*v/pdiff;
run;
```

Output (means):

Least Squares Means

Effect	n	v	Estimate	Standard		t Value	Pr > t
				Error	DF		
n*v	1	1	4252.67	365.52	36	11.63	<.0001
n*v	1	2	4306.00	365.52	36	11.78	<.0001
n*v	1	3	3177.33	365.52	36	8.69	<.0001
n*v	1	4	4481.33	365.52	36	12.26	<.0001
n*v	2	1	5672.00	365.52	36	15.52	<.0001
n*v	2	2	5982.00	365.52	36	16.37	<.0001
n*v	2	3	5442.67	365.52	36	14.89	<.0001
n*v	2	4	4816.00	365.52	36	13.18	<.0001
n*v	3	1	6400.00	365.52	36	17.51	<.0001
n*v	3	2	6259.00	365.52	36	17.12	<.0001
n*v	3	3	5994.00	365.52	36	16.40	<.0001
n*v	3	4	4812.00	365.52	36	13.16	<.0001
n*v	4	1	6732.67	365.52	36	18.42	<.0001
n*v	4	2	6895.00	365.52	36	18.86	<.0001
n*v	4	3	6014.00	365.52	36	16.45	<.0001
n*v	4	4	3816.00	365.52	36	10.44	<.0001
n*v	5	1	7563.33	365.52	36	20.69	<.0001
n*v	5	2	6950.67	365.52	36	19.02	<.0001
n*v	5	3	6687.33	365.52	36	18.30	<.0001
n*v	5	4	2046.67	365.52	36	5.60	<.0001

n*v	6	1	8700.67	365.52	36	23.80	<.0001
n*v	6	2	6540.33	365.52	36	17.89	<.0001
n*v	6	3	6065.33	365.52	36	16.59	<.0001
n*v	6	4	1880.67	365.52	36	5.15	<.0001

There are 24 treatments and a total of $24*23/2 = 256$ pairwise comparisons, which are listed in a line-by-line fashion:

Differences of Least Squares Means									
Effect	n	v	<u>n</u>	<u>v</u>	Standard		DF	t Value	Pr > t
					Estimate	Error			
n*v	1	1	1	2	-53.3333	482.66	36	-0.11	0.9126
n*v	1	1	1	3	1075.33	482.66	36	2.23	0.0322
n*v	1	1	1	4	-228.67	482.66	36	-0.47	0.6385
n*v	1	1	2	1	-1419.33	516.93	36	-2.75	0.0094
<252 more comparisons!>									

Clearly, this output is a mess! Only a fraction of the 252 pairwise comparisons is of interest. Specifically, we are interested only in comparing N-means for the same variety and variety means for the same level of N. Note that the d.f. for the fourth comparison are a broken number; these are the Satterthwaite d.f., which are needed here, since we are comparing N*V means at the same level of V. To pick the relevant comparisons, output of the pairwise difference may be written into a SAS file using ODS (do not forget to type a quit command before using ODS!):

```
quit;
ods output diffs=diffs;
proc mixed data=t;
class n v block;
model yield=block n v n*v/ddfm=satterth;
random block*n;
lsmeans n*v/pdiff;
run;

proc print data=diffs;
run;
```

Output in dataset "DIFFS":

Effect	n	v	<u>n</u>	<u>v</u>	Estimate	StdErr	DF	tValue	Probt
n*v	1	1	1	2	-53.3333	482.66	36	-0.11	0.9126
n*v	1	1	1	3	1075.33	482.66	36	2.23	0.0322
n*v	1	1	1	4	-228.67	482.66	36	-0.47	0.6385
n*v	1	1	2	1	-1419.33	516.93	36	-2.75	0.0094
<252 more comparisons!>									

Note that there is a slight change in the coding of some variables (standard error, t-value, p-value). Also, note that for the first two comparisons, both treatments have the same N-level (N=1), while for the fourth comparison, the N-levels differ (N=1 and N=2). To pick comparisons for which both means have the same level of N, we use the following IF-statement inside a datastep:

```
if n = _n;
```

The full set of statements, including computation of the LSD (see Section 6.3) is as follows:

```
data N_constant;
set diffs;
if n = _n;
ttab=tinv(0.975,df);
lsd=ttab*stderr;

proc print data=N_constant;
run;
```

Output:

Effect	n	v	_n	_v	Estimate	StdErr	DF	tValue	Probt	ttab	lsd
n*v	1	1	1	2	-53.3333	482.66	36	-0.11	0.9126	2.02809	978.881
n*v	1	1	1	3	1075.33	482.66	36	2.23	0.0322	2.02809	978.881
<34 further comparisons, where N-level is the same for both treatments>											

The LSD is the same for all comparisons (LSD = 978.9) involving the same N-level because the data are balanced. The LSD for comparisons with the same variety are computed similarly:

```
data variety_constant;
set diffs;
if v = _v;
ttab=tinv(0.975,df);
lsd=ttab*stderr;

proc print data=variety_constant;
run;
```

Output:

Obs	Effect	n	v	_n	_v	Estimate	StdErr	DF	tValue	Probt	ttab	lsd
1	n*v	1	1	2	1	-1419.33	516.93	36	-2.75	0.0094	2.02809	1048.38
2	n*v	1	1	3	1	-2147.33	516.93	36	-4.15	0.0002	2.02809	1048.38
<58 further comparisons, where variety is the same for both treatments>												

The LSD is LSD = 898.1 for comparisons involving the same variety.

The results may be summarized as follows:

	N1 (0)	N2 (60)	N3 (90)	N4 (120)	N5 (150)	N6 (180)
V1	4253	5672	6400	6733	7563	8701
V2	4306	5982	6259	6895	6951	6540
V3	3177	5443	5994	6014	6687	6065
V4	4481	4816	4812	3816	2047	1881

LSD($\alpha = 5\%$) = 1048.4 (comparisons in the same row)

LSD($\alpha = 5\%$) = 978.8 (comparisons in the same column)

If you want to produce a letters display, you can use the GLIMMIX procedure (but this does not produce an LSD automatically – you'd need to use the same approach as just described for MIXED):

```
proc glimmix data=rice;
class block n v;
model yield=block n v n*v/htype=1;
random block*n;
slice n*v/pdiff sliceby=n lines;
slice n*v/pdiff sliceby=v lines;
run;
```

Output (just showing the first slice for n):

T Grouping for n*v Least Squares Means Slice (Alpha=0.05)

LS-means with the same letter are not significantly different.

Slice	v	Estimate	
n 1	4	4481.33	A
n 1			A
n 1	2	4306.00	A
n 1			A
n 1	1	4252.67	A
n 1			A
n 1	3	3177.33	B

Exercise 6.4 (Mead et al., 1993, p. 133): The response of six varieties of lettuce, grown in frames, to various uncovering dates was investigated in a split-plot experiment with four blocks (**lettuce.dat**). The main-plot treatments were three uncovering dates and each main-plot was split into six split-plots for the six varieties. Perform an ANOVA followed by mean comparisons using an LSD test.

Uncovering date	Variety	Block			
		I	II	III	IV
x	A	11.8	7.5	9.7	6.4
	B	8.3	8.4	11.8	8.5
	C	9.2	10.6	11.4	7.2
	D	15.6	10.8	10.3	14.7
	E	16.2	11.2	14.0	11.5
	F	9.9	10.8	4.8	9.8
y	A	9.7	8.8	12.5	9.4
	B	5.4	12.9	11.2	7.8
	C	12.1	15.7	7.6	9.4
	D	13.2	11.3	11.0	10.7
	E	16.5	11.1	10.8	8.5
	F	12.5	14.3	15.9	7.5
z	A	7.0	9.1	7.1	6.3
	B	5.7	8.4	6.1	8.8
	C	3.3	6.9	1.0	2.6
	D	12.6	15.4	14.2	11.3
	E	12.6	12.3	14.4	14.1
	F	10.2	11.6	10.4	12.2

Unbalanced data

When data are unbalanced, no common LSD can be computed. If MIXED is used, LSMEANS computes socalled **weighted** least squares means rather than ordinary least squares means as computed in GLM. Also, the F-tests obtained by MIXED differ slightly from those obtained in GLM because a weighted least squares analysis is performed. Details are omitted here. The good news for the user of MIXED is that the same code can be used as for balanced data. Only change: drop the NOBOUND option.

6.5 Factorial experiments with quantitative factors

When at least one of the factors in a factorial experiment is quantitative, one should consider polynomial regression, as discussed for single factor experiments (Section 5.8). For two quantitative factors, it is useful to look at so-called **response surface methodology** (see book by Dean and Voss; also see Mead et al., Section 16.7).

Example 6.6: The N-levels for the six N-treatment in the rice data (**rice.dat**) were as follows:

N	N-dose	Label in model
1	0	x_1
2	60	x_2
3	90	x_3
4	120	x_4
5	150	x_5
6	180	x_6

We can consider fitting a polynomial.

The ANOVA has revealed significant interaction between variety and N-level, so the regression curves are not parallel. Thus, we consider fitting the following model:

$$y_{ijh} = \mu + b_h + \alpha_i + \gamma_{i1}x_j + f_{ih} + e_{ijh}$$

where

γ_{i1} = slope of i -th variety for the linear regression on N-dose
 x_j = j -th N-dose

This model implies that each variety has its own regression line. To test the lack of fit, we need a deviation from the regression for each variety at each N-level. Thus, the lack-of-fit term needs to be indexed by both i and j . The model is

$$y_{ijk} = \mu + b_k + \alpha_i + \boxed{\gamma_{i1}x_j + \delta_{ij}} + f_{ik} + e_{ijk}$$

where

δ_{ij} = lack-of-fit effect

We may add polynomial terms until the lack-of-fit test is not significant. For example, the lack-of-fit test for the quadratic model is obtained by the following model:

$$y_{ijk} = \mu + b_k + \alpha_i + \boxed{\gamma_{i1}x_j + \gamma_{i2}x_j^2} + \delta_{ij} + f_{ik} + e_{ijk}$$

SAS hints

The N-dose is stored under the variable name NDOSE. To obtain the correct test for regression terms, we may use the HTYPE=1 option. Also, the lack-of-fit effect needs to be fitted after the regression terms. The SAS code using MIXED is as follows:

```

data;
input
block n v n_amount yield;
lackfit=n;
datalines;
    1 1 1 0 4520
    1 1 2 0 4034
<more data>
    3 6 3 180 5210
    3 6 4 180 1744
;
proc mixed data=t;
class lackfit v block;
model yield=block v v*n_amount lackfit*v/htype=1;
random block*n;
run;


$$\mu + b_k + \alpha_i + \gamma_{i1}x_j + \delta_{ij} + f_{ik} + e_{ijk}$$


```

Output:

Type 1 Tests of Fixed Effects

Effect	Num	Den		Pr > F
	DF	DF	F Value	
block	2	10	0.98	0.4095
v	3	36	85.74	<.0001
n_amount*v	4	36	52.59	<.0001
lackfit*v	16	36	2.69	0.0068

The lack-of-fit is significant, so we add a quadratic term:

```

proc mixed data=t;
class n v block;
model yield=block v v*n_amount v*n_amount*n_amount lackfit*v/htype=1;
random block*n;
run;

```

Output:

Type 1 Tests of Fixed Effects

Effect	Num	Den		Pr > F
	DF	DF	F Value	
block	2	10	0.98	0.4095
v	3	36	85.74	<.0001
n_amount*v	4	36	52.59	<.0001
n_amount*n_amount*v	4	36	8.11	<.0001
lackfit*v	12	36	0.89	0.5665

Now the lack of fit is not significant, i.e., the quadratic response fits well. The remaining task is to fit the quadratic regression curves for each variety. To do this, we fit the quadratic model without the lack-of-fit effect:

$$y_{ijk} = \mu + b_h + \alpha_i + \gamma_{i1}x_j + \gamma_{i2}x_j^2 + f_{ik} + e_{ijk}$$

The expected value of an observation is obtained from the model by "stripping off" the random effects:

$$\eta_{ijk} = \mu + b_h + \alpha_i + \gamma_{i1}x_j + \gamma_{i2}x_j^2$$

Obviously, the intercept depends not only on varieties, but also on blocks. Thus, we may compute the average across blocks for convenience:

$$\bar{\eta}_{ij\bullet} = \mu + \bar{b}_\bullet + \alpha_i + \gamma_{i1}x_j + \gamma_{i2}x_j^2$$

The intercept of the i -th variety is

$$\mu + \bar{b}_\bullet + \alpha_i$$

The linear and quadratic terms are obtained directly by using the SOLUTION option. The intercept terms may be computed using the ESTIMATE statement as follows:

```
proc mixed data=t;
class n v block;
model yield=block v v*n_amount v*n_amount*n_amount/solution;
estimate 'intercept variety 1' intercept 3 v 3 0 0 0 block 1 1 1 /divisor=3;
estimate 'intercept variety 2' intercept 3 v 0 3 0 0 block 1 1 1 /divisor=3;
estimate 'intercept variety 3' intercept 3 v 0 0 3 0 block 1 1 1 /divisor=3;
estimate 'intercept variety 4' intercept 3 v 0 0 0 3 block 1 1 1 /divisor=3;
random block*n;
run;
```

Output:

Effect	block	v	Standard				
			Estimate	Error	DF	t Value	Pr > t
Intercept			4499.12	364.62	13	12.34	<.0001
block	1		-46.3750	196.98	13	-0.24	0.8175
block	2		234.08	196.98	13	1.19	0.2559
block	3		0
v	1		-238.22	471.95	45	-0.50	0.6162
v	2		-283.01	471.95	45	-0.60	0.5517
v	3		-1360.87	471.95	45	-2.88	0.0060
v	4		0
n_amount*v	1		18.8140	8.1759	45	2.30	0.0261
n_amount*v	2		35.6170	8.1759	45	4.36	<.0001
n_amount*v	3		45.2917	8.1759	45	5.54	<.0001
n_amount*v	4		14.6149	8.1759	45	1.79	0.0806
n_amount*n_amount*v	1		0.02583	0.04331	45	0.60	0.5539
n_amount*n_amount*v	2		-0.1248	0.04331	45	-2.88	0.0060
n_amount*n_amount*v	3		-0.1605	0.04331	45	-3.71	0.0006
n_amount*n_amount*v	4		-0.1764	0.04331	45	-4.07	0.0002

Label	Estimate	Standard		t Value	Pr > t
		Error	DF		
intercept variety 1	4323.47	346.43	13	12.48	<.0001
intercept variety 2	4278.68	346.43	13	12.35	<.0001
intercept variety 3	3200.82	346.43	13	9.24	<.0001
intercept variety 4	4561.69	346.43	13	13.17	<.0001

From this result, the estimated quadratic regression equations are:

$$\text{Variety V1: } \text{YIELD} = 4323 + 18.81 \times N + 0.02583 \times N^2$$

$$\text{Variety V2: } \text{YIELD} = 4279 + 35.62 \times N - 0.12481 \times N^2$$

$$\text{Variety V3: } \text{YIELD} = 3201 + 45.29 \times N - 0.16049 \times N^2$$

$$\text{Variety V4: } \text{YIELD} = 4462 + 14.61 \times N - 0.17638 \times N^2$$

The fitted curves are displayed in Fig. 6.1.

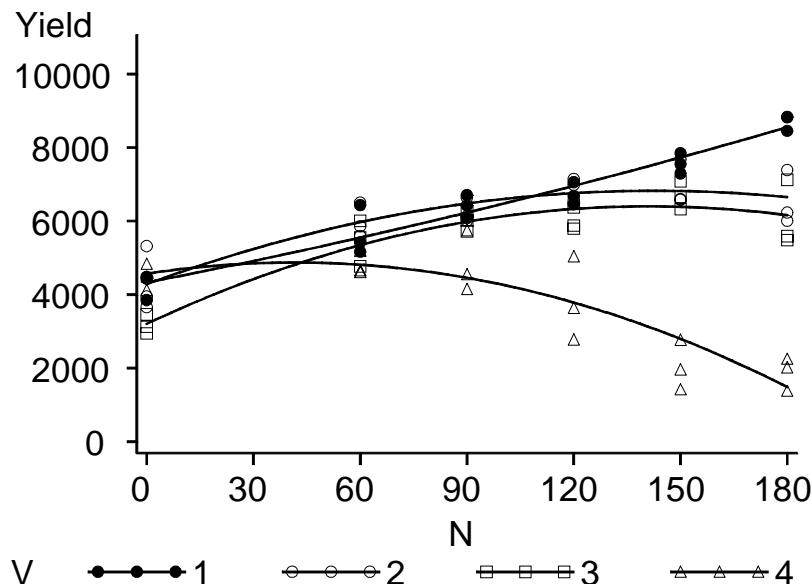


Fig. 6.3: Fitted regression curves for varieties V1 to V4.

Exercise 6.5: The two factors in the pigs data in Example 6.5, copper and phytase, are quantitative (**pigs.dat**). How would you fit polynomial terms for main effects and interaction? Hint: If x_i is the amount of copper at the i -th level and z_j the amount phytase at the j -th level, consider fitting terms $x_i z_j$ (linear-linear), $x_i^2 z_j$ (quadratic-linear), $x_i z_j^2$ (linear-quadratic), and $x_i^2 z_j^2$ (quadratic-quadratic) for interaction. Observe the marginality principle when fitting these terms.

Hint:

Fit the following model and consider different fitting sequences, adhering to the marginality principle:

$$y_{ijk} = \mu + \alpha_1 x_i + \alpha_2 x_i^2 + \beta_1 z_j + \beta_2 z_j^2 + \gamma_1 x_i z_j + \gamma_2 x_i z_j^2 + \gamma_3 x_i^2 z_j + \gamma_4 x_i^2 z_j^2 + e_{ijk}$$

The marginality principle dictates, that a term of the form $x_i^r z_j^s$ should be fitted only after all terms $x_i^t z_j^u$ with $t \leq r \leq 0$ and $u \leq s \leq 0$ have been fitted (Note that, e.g., when $t = 0$ then $x_i^t z_j^u = x_i^0 z_j^u = z_j^u$). Test the interaction first. The data for CuG are unbalanced, so you need to consider the following orders for the interaction terms (fitted after main effects α_1 , α_2 , β_1 and β_2):

- (I1) $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ (tests for $\gamma_1, \gamma_3, \gamma_4$)
- (I2) $\gamma_1, \gamma_3, \gamma_2, \gamma_4$ (tests for $\gamma_1, \gamma_2, \gamma_4$)

If interactions are not significant, the following fitting orders are appropriate for the main effects:

- (M1) $\alpha_1, \alpha_2, \beta_1, \beta_2$ (tests for β_1, β_2)
- (M2) $\beta_1, \beta_2, \alpha_1, \alpha_2$ (tests for α_1, α_2)

SAS hints:

Fitting order I1 for interaction and M2 for main effects:

```

data;
input
  Pig      Cu      Phyt      CuG       CuL      CuS;
datalines;
  1      20      0      5.16      8.07      1.63
  2      20      0      3.66     11.56      1.83
<more data>
  71     175     500     3.66     12.21      1.51
  72     175     500    10.10     11.82      2.50
;

proc glm;
model CuG=phyt phyt*phyt
  cu      cu*cu
  cu*phyt cu*cu*phyt cu*phyt*phyt cu*cu*phyt*phyt;
run;

```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Phyt	1	8.7979687	8.7979687	1.39	0.2426
Phyt*Phyt	1	0.9984188	0.9984188	0.16	0.6924
Cu	1	296.9727127	296.9727127	46.99	<.0001
Cu*Cu	1	19.2950549	19.2950549	3.05	0.0856
Phyt*Cu	1	0.7931739	0.7931739	0.13	0.7244
Phyt*Cu*Cu	1	0.6827636	0.6827636	0.11	0.7435
Phyt*Phyt*Cu	1	25.0197091	25.0197091	3.96	0.0511
Phyt*Phyt*Cu*Cu	1	2.7815361	2.7815361	0.44	0.5095

None of the interaction effects is significant here. This result is also obtained with the fitting order I2 (not shown). Only the linear main effect for copper is significant. Effects for phytase

are not significant with fitting order M1 (not shown). We remove non-significant terms fit the reduced model as follows:

```
proc glm;
model CuG=cu/solution;
run;
```

Output:

Parameter	Estimate	Standard	t Value	Pr > t
		Error		
Intercept	2.086019885	0.53488948	3.90	0.0002
Cu	0.032146290	0.00475667	6.76	<.0001

The fitted model is:

$$\text{CuG} = 2.08 + 0.0321 \times \text{Cu}$$

Additional remark: The sources of variation in the ANOVA table are commonly denoted as follows:

Source of variation (prose)	Model term	GLM code
Main effect copper		
Linear	$\alpha_1 x_i$	cu
Quadratic	$\alpha_2 x_i^2$	cu*cu
Main effect phytase		
Linear	$\beta_1 z_j$	phyt
Quadratic	$\beta_j z_j^2$	phyt*phyt
Interaction phytase \times copper		
Linear \times linear	$\gamma_1 x_i z_j$	cu*phyt
Linear \times quadratic	$\gamma_2 x_i z_j^2$	cu*phyt*phyt
Quadratic \times linear	$\gamma_3 x_i^2 z_j$	cu*cu*phyt
Quadratic \times quadratic	$\gamma_4 x_i^2 z_j^2$	cu*cu*phyt*phyt

Exercise 6.6 (Gomez and Gomez, p. 401): Nitrogen uptake of rice was studied in a two-factor experiment involving duration of water stress (x) and level of nitrogen application (z). The data are given Table 6.10 (**uptake.dat**). The experiment was a greenhouse experiment with four water-stress treatments (different duration in days) as main-plot treatments and four nitrogen rates (kg/ha) as sub-plot treatments, in four replications (blocks).

Table 6.10: Nitrogen uptake of the rice plants, grown with four degrees of water stress and four rates of nitrogen application.

Water Stress (days)	Nitrogen rate (kg/ha)	Nitrogen uptake (g/pot)			
		Block I	Block II	Block III	Block IV
0	0	0.250	0.321	0.373	0.327
	90	0.503	0.493	0.534	0.537
	180	0.595	0.836	0.739	0.974
	270	1.089	1.297	1.007	0.677
10	0	0.254	0.373	0.349	0.367
	90	0.506	0.613	0.588	0.625
	180	0.692	0.754	0.548	0.713
	270	1.033	0.757	1.034	0.831
20	0	0.248	0.234	0.267	0.305
	90	0.428	0.397	0.493	0.587
	180	0.484	0.453	0.457	0.372
	270	0.507	0.498	0.477	0.619
40	0	0.099	0.103	0.093	0.084
	90	0.154	0.142	0.133	0.129
	180	0.111	0.102	0.098	0.152
	270	0.089	0.142	0.138	0.141

Perform a factorial polynomial regression.

Hints:

Fit the following terms:

Source of variation (prose)	Model term
-----------------------------	------------

Main effect water

Linear	$\alpha_1 x_i$
Quadratic	$\alpha_2 x_i^2$
Cubic	$\alpha_3 x_i^3$

Main effect nitrogen

Linear	$\beta_1 z_j$
Quadratic	$\beta_2 z_j^2$
Cubic	$\beta_3 z_j^3$

Interaction water \times nitrogen

Linear \times linear	$\gamma_1 x_i z_j$
Linear \times quadratic	$\gamma_2 x_i z_j^2$
Linear \times cubic	$\gamma_3 x_i z_j^3$

Quadratic \times linear	$\gamma_4 x_i^2 z_j$
Quadratic \times quadratic	$\gamma_5 x_i^2 z_j^2$
Quadratic \times cubic	$\gamma_5 x_i^2 z_j^3$
Cubic \times linear	$\gamma_6 x_i^3 z_j$
Cubic \times quadratic	$\gamma_7 x_i^3 z_j^2$
Cubic \times cubic	$\gamma_9 x_i^3 z_j^3$

Note that the model is saturated in that the three d.f. for each main effect and the nine d.f. for interaction are used up by the polynomial terms.

(Note: Gomez and Gomez erroneously analyse the data with the last water stress level set equal to 30 instead of 40).

The model needs to have a main-plot error term. You need to define a class variable corresponding to water stress level in order to be able to fit the main plot error along with polynomial terms involving water stress. The data are balanced, so one fitting sequence is sufficient, observing the marginality principle.

SAS hints:

```

data;
input
  w      n      block    uptake   ;
  water_class=w;
datalines;
  0      0      1      0.250
  0      0      2      0.321
<more data>
  40     270    3      0.138
  40     270    4      0.141
;

proc glm;
class block w n;
model uptake=block w w*block n w*n;
random w*block/test;
run;

proc glm;
class block water_class;
model uptake=block w w*w w*w*w
          water_class*block
            n          n*n        n*n*n
            w*n       w*n*n      w*n*n*n
            w*w*n    w*w*n*n   w*w*n*n*n
            w*w*w*n w*w*w*n*n w*w*w*n*n*n/ss1 e1;
random water_class*block/test;
run;

```

Output:

Dependent Variable: uptake

Source	DF	Type I SS	Mean Square	F Value	Pr > F
block	3	0.008079	0.002693	0.57	0.6497
* w	1	2.854215	2.854215	602.22	<.0001
* w*w	1	0.052208	0.052208	11.02	0.0090
* w*w*w	1	0.052298	0.052298	11.03	0.0089
Error	9	0.042656	0.004740		
Error: MS(block*water_class)					
* This test assumes one or more other fixed effects are zero.					
Source	DF	Type I SS	Mean Square	F Value	Pr > F
block*water_class	9	0.042656	0.004740	0.44	0.9013
* n	1	1.260648	1.260648	118.21	<.0001
* n*n	1	0.004883	0.004883	0.46	0.5030
* n*n*n	1	0.021698	0.021698	2.03	0.1624
* w*n	1	0.608688	0.608688	57.08	<.0001
* w*n*n	1	0.002770	0.002770	0.26	0.6134
* w*n*n*n	1	0.001948	0.001948	0.18	0.6716
* w*w*n	1	0.017179	0.017179	1.61	0.2125
* w*w*n*n	1	0.008372	0.008372	0.79	0.3815
* w*w*n*n*n	1	0.024586	0.024586	2.31	0.1377
* w*w*w*n	1	0.017338	0.017338	1.63	0.2105
* w*w*w*n*n	1	0.002866	0.002866	0.27	0.6073
w*w*w*w*n*n*n	1	0.000668	0.000668	0.06	0.8038
Error: MS(Error)	36	0.383913	0.010664		
* This test assumes one or more other fixed effects are zero.					

The linear \times linear interaction is significant. Thus, by the marginality principle, we need to retain the linear main effects for both factors. Both of these are significant anyway. Also, the quadratic and cubic main effects for water stress are significant. Dropping non-significant terms, the model is estimated as follows (use MIXED to account for the fact that WATER_CLASS*BLOCK, the main plot error, is random):

```
proc mixed;
class block water_class;
model uptake=block w w*w w*w*w n w*n/solution;
estimate 'intercept' intercept 4 block 1 1 1 1/divisor=4;
random water_class*block;
run;
```

Output:

Solution for Fixed Effects

Effect	block	Estimate	Standard		t Value	Pr > t
			Error	DF		
Intercept		0.3232	0.04124	9	7.84	<.0001
block	1	-0.02487	0.03467	9	-0.72	0.4913
block	2	0.004688	0.03467	9	0.14	0.8954

block	3	-0.00700	0.03467	9	-0.20	0.8445
block	4	0
w		0.01905	0.008272	46	2.30	0.0259
w*w		-0.00159	0.000617	46	-2.58	0.0132
w*w*w		0.000025	0.000011	46	2.33	0.0241
n		0.002542	0.000189	46	13.47	<.0001
w*n		-0.00007	8.236E-6	46	-7.96	<.0001

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > t
intercept	0.3164	0.03536	9	8.95	<.0001

It turns out that the regression coefficients are rather small. To obtain more significant digits, water stress and nitrogen level are divided by 100, and the code is re-run:

```

data;
input
  w      n      block    uptake  ;
  water_class=w;
  w=w/100;
  n=n/100;
datalines;
  0      0      1      0.250
  0      0      2      0.321
<more data>
;

```

Output:

Solution for Fixed Effects

Effect	block	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		0.3232	0.04124	10	7.84	<.0001
block	1	-0.02487	0.03467	10	-0.72	0.4895
block	2	0.004688	0.03467	10	0.14	0.8951
block	3	-0.00700	0.03467	10	-0.20	0.8440
block	4	0
w		1.9050	0.8272	45	2.30	0.0260
w*w		-15.9203	6.1746	45	-2.58	0.0133
w*w*w		24.9844	10.7132	45	2.33	0.0242
n		0.2542	0.01887	45	13.47	<.0001
w*n		-0.6553	0.08236	45	-7.96	<.0001

Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr > t
intercept	0.3164	0.03536	10	8.95	<.0001

The estimated regression is:

$$\begin{aligned} N\text{-uptake} = & 0.3164 + 1.9050 \times W - 15.92 \times W^2 + 24.98 \times W^3 \\ & + 0.2542 \times N \\ & - 0.6553 \times W \times N \end{aligned}$$

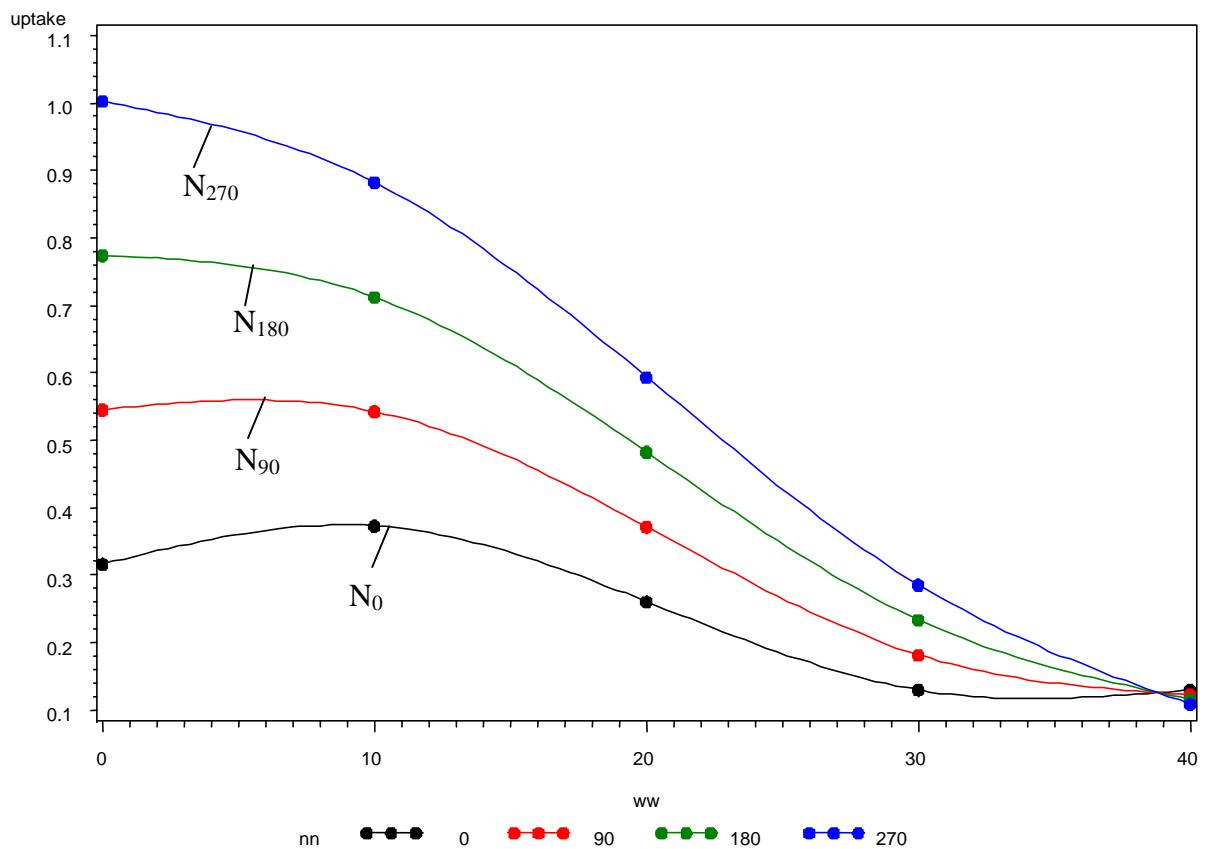
where W and N are water stress and nitrogen, divided by 100. The fitted model can be plotted as follows:

```

data plot;
do ww=0 to 40 by 10;
  do nn=0 to 270 by 90;
    w=ww/100;
    n=nn/100;
    uptake= 0.3164 + 1.9050 * w - 15.92 * w * w + 24.98 * w * w * w
           + 0.2542 * n
           - 0.6553 * w * n;
    output;
  end;
end;

symbol value=dot i=spline;
proc gplot;
plot uptake*ww=nn;
run;

```



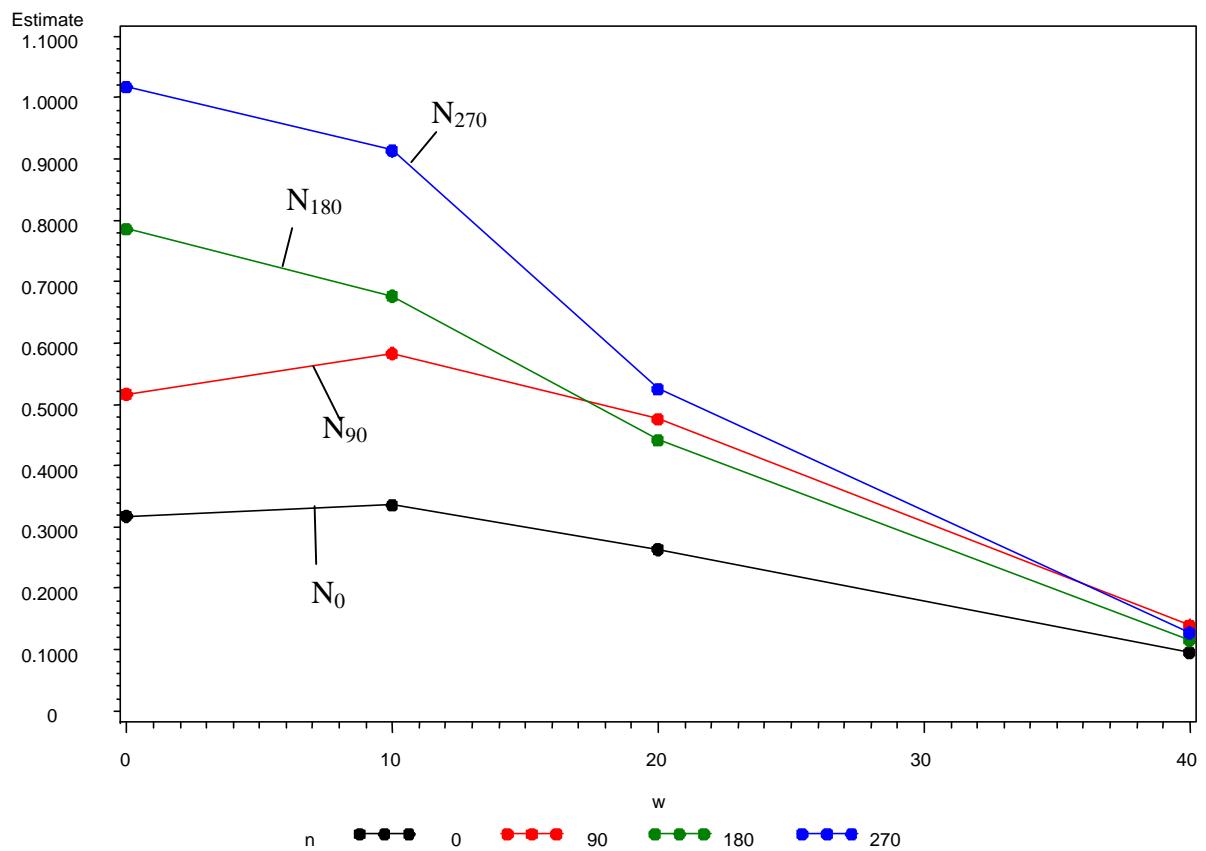
The ups and downs of the curves for N₀ and N₉₀, particularly at high water stress, should not be overinterpreted. They are probably just an artefact due to the polynomial regression. Nevertheless, the fitted curves agree reasonably well with the observed treatment means:

```

ods output lsmeans=lsmeans;
proc mixed;
class block w n;
model uptake=block w n w*n/solution;
random water_class*block;
lsmeans w*n;
run;

symbol value=dot i=join;
proc gplot data=lsmeans;
plot estimate*w=n;
run;

```



At high water stress, the effect of nitrogen is limited, while nitrogen can unfold its full potential when water stress is low or absent. This conclusion emerges, no matter if we analyse both factors as if they were qualitative (see last display of means) or by polynomial regression.

Exercise 6.7: Perform a factorial polynomial regression for CuL and CuS in the pig data ([pigs.dat](#)).

7. Repeated measures

The linear models and experimental designs considered so far have assumed that all observations are independent (exception: split-plot design, see below). The independence assumption is not appropriate when repeated measurements are taken on the same experimental unit (randomization unit). Repeated measures may be taken in time or in space or both. Here, we will give two examples. Clearly, repeated measurements on the same unit are correlated. The analysis of repeated measures calls for specialized methods, which account for the correlation. Basically, there are two approaches:

- (1) Compute a summary value, e.g., the mean, for each randomization unit. For each unit, a single summary value is obtained, which may be analysed by standard procedures.
- (2) Model and analyse replicate data directly.

The second approach is more powerful, but also much more difficult. Thus, one should always think about the summary value approach. In my experience, this approach is often sufficient.

We have already seen an example of repeated measurements: In a split-plot design, repeated measurements are taken on the same main-plot (randomization unit). The correlation among different observations on the same unit is modelled by the random main plot error term (f_{ik}). Note that all observations on the same main plot "carry" the same main-plot error term (f_{ik}). In this chapter, we will see further examples, where correlations of repeated measurements can be modeled by random effects, giving rise to mixed models.

When discussing an example, we will consider the summary value first and then think about a more refined mixed model approach.

7.1 The goats data

Example 7.1 (Ossama Dimassi, Tierzüchtung der Tropen und Subtropen, Universität Hohenheim): Three groups of five goats each were compared regarding the quality of milk for cheese production. The groups differed in two variables (treatment variables):

BREED [1=Dahlem Cashmere, 2=Bunte Deutsche Edelziege (alpine)]
PARITY (2 = second lamb, 3 = third lamb)

Breed 2 was tested only with goats for which PARITY = 2, so the combination BREED = 2 and PARITY = 3 was not observed. The GROUP variable is defined as follows:

GROUP	BREED	PARITY
1	1	3
2	1	2
3	2	2

Among others, the following response variables were assessed in biweekly intervals over a period of 28 weeks:

PROTEIN = protein content in percent weight of milk

CASEIN = casein content in percent weight of milk

CCV = Cheese conversion value = cheese weight (g) \times cheese dry matter (%) / milk weight (g)

Also, the following variables were recorded:

WEEK = number of week

GOAT = running number of goat

For each goat and week, there were two replicate measurements to assess measurement errors in the lab. Some measurements are missing, because the contracting lab failed to provide analysis results.

The objective of the study was to compare the groups with respect to milk quality. Data are stored in **cheese.dat**.

Cheese conversion value

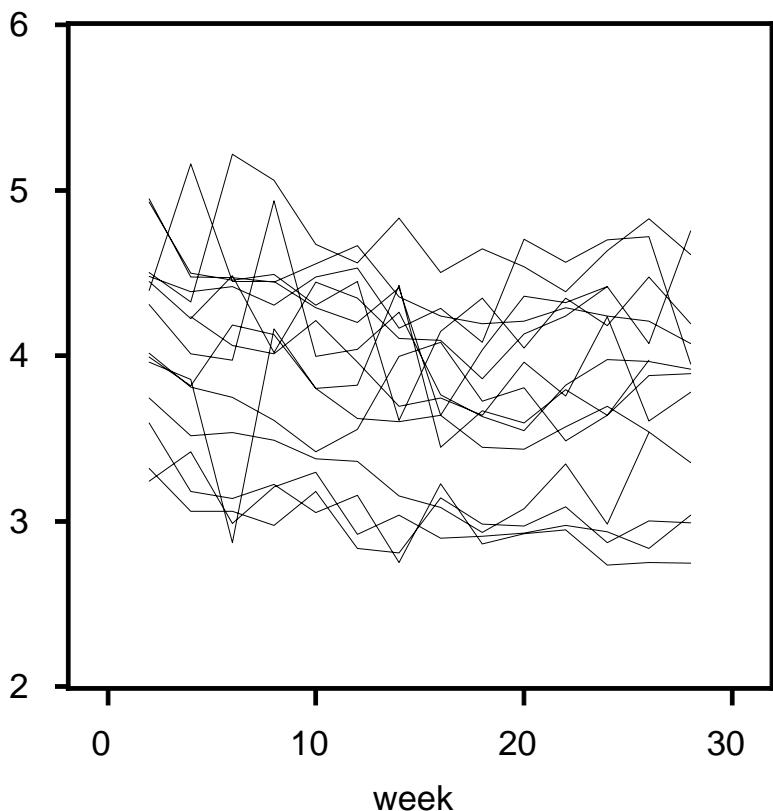


Fig. 7.1: Cheese conversion value profiles for fifteen goats.

The summary value approach

Weekly measurements were taken mainly to increase replication. Moreover, it was of interest to see if groups differ in their time profile. Fig. 7.1 shows the profiles for the three groups. There seems to be a slight downward trend for all goats, regardless of the group. Thus, one could just take the means across weeks for each goat, i.e., for each goat, one computes one mean value (summary statistic). This mean will reflect the average performance over time.

The fifteen means are then analysed by an appropriate linear model. The simplest type of analysis is by a one-way analysis for comparing the three groups according to the model

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

where

y_{ij} = mean for j -th goat in i -th group

α_i = effect of i -th group

Using means of all observations of a goat for analysis is the simplest way to tackle the repeated measures nature of the data. Clearly, observations on the same animal (subject) are not independent. Some form of correlation is to be expected. We cannot, therefore, analyse the replicate data as if they were independent. Note that **the usual linear model analysis assumes independence of all observations**. Thus, if we were to analyse the replicate data, the correlation would need to be modelled, as we will see later. The need to model the correlation is circumvented here by computing simple means. Note that means from different goats are independent, so standard procedures can be used for analysis.

A major disadvantage of the means approach is that time effects cannot be tested or studied. Also, if there are missing values, it is not obvious how to deal with the problem. In the case at hand, there are only few missing values, so the problem is not dramatic. Here we will just compute simple means across all available observations. It should be stressed, however, that this approach may cause biases in case there are many missing values. Thus, a more refined analysis is preferable, which can more properly deal with the missing data problem (see below).

SAS hints

Means for goats may be computed by PROC MEANS. The data need to be sorted by the variables GROUP and GOAT before running PROC MEANS, because means will be computed for goats within groups. Sorting may be done by PROC SORT.

```

data t;
input
Group Week Month Goat parity      breed kids   protein    casein    ccv;
if ccv=0 then ccv=.; → scrutiny of the data revealed that there were
                           some zeros for CCV. These were taken as missing values.
datalines;
1   2     1     1     3     1     2     3.56  2.64  4.45
1   2     1     1     3     1     2     3.55  2.64  .
1   4     1     1     3     1     2     3.58  2.68  4.25
1   4     1     1     3     1     2     3.58  2.67  4.2
<more data>
3   26    7     15    2     2     1     3.35  2.53  3.99
3   26    7     15    2     2     1     3.36  2.53  3.95
;
proc sort data=t out=t; → need to sort data before computing means for goats
by group goat;           within groups. Save sorted data under "T".

```

```

proc means data=t;
var protein casein ccv; → variables for which means are to be computed
by group goat; → compute mean for each goat within a group
output out=simple mean=; (use BY variables in same order as in PROC SORT!)
run; and save means in dataset "SIMPLE"
proc glm data=simple;
class group;
model ccv=group;
lsmeans group/pdiff;
means group/lsd;
run;

```

The data (means!) are balanced, so LSMEANS and MEANS yield the same result.

LSMEANS:

The GLM Procedure
Least Squares Means

Group	ccv	LSMEAN	
		LSMEAN	Number
1	4.25624266	1	
2	4.08233974	2	
3	3.23150000	3	

Least Squares Means for effect Group
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: ccv

i/j	1	2	3
1		0.4093	0.0003
2	0.4093		0.0013
3	0.0003	0.0013	

MEANS:

Alpha	0.05
Error Degrees of Freedom	12
Error Mean Square	0.103425
Critical Value of t	2.17881
Least Significant Difference	0.4432

Means with the same letter are not significantly different.

t Grouping	Mean	N	Group
A	4.2562	5	1
A			
A	4.0823	5	2
B	3.2315	5	3

Table 7.1: Group means computed from animal means over weeks and replicate measurements.

GROUP	BREED	PARITY	Mean [§]
1	1	3	4.25 ^a
2	1	2	4.08 ^a
3	2	2	3.23 ^b

§ Means followed by the same letter are not significantly different.

There is a significant difference among breeds, when parity is held constant (group 2 versus group 3), but there is no effect of parity, when breed is held constant (group 1 versus group 2). The comparison of group 1 versus group 3 is significant, but not interpretable, because groups differ in both breed and parity. Thus, the means table must be interpreted with the underlying factorial structure (factors breed and parity) in mind.

Exercise 7.1: Repeat the analysis for the cheese conversion value (**cheese.dat**). Do the analysis for protein and casein.

The groups have factorial structure. The two factors are breed and parity. Each factor has two levels, but not all factorial combinations are tested. As a result, interaction cannot be tested. This is easily seen by noting that, e.g., breeds can be compared only at PARITY=2. To test for interaction, we would need to have information on the difference among breeds at PARITY=3. This information is not available, however, because at PARITY=3 only breed 1 was tested.

We can do a two-way analysis based on a model without interaction:

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

y_{ijk} = mean for k -th goat for i -th breed and j -th parity

α_i = effect of i -th breed ($i = 1$: BREED=1; $i = 2$: BREED = 2)

β_j = effect of j -th parity ($j = 1$: PARITY = 2; $j = 2$: PARITY = 3)

It is very important to stress here that we need to make the strong assumption of no interaction to be able to do a two-way analysis. This assumption cannot be tested, so the analysis needs to be interpreted with care. Some might say, a two-way analysis is invalid because of unwarranted assumptions. I would not go so far. I would just say, that interpretation needs to account for the possibility of interaction. In fact, little is gained here by the two-way analysis

compared to the one-way analysis. All information is contained in the three group means. Despite of this, we will look at the two-way analysis to learn more about linear models!

We can compute marginal means for breed and parity. F-tests for α_i and β_j correspond to comparisons among marginal means for breed and parity. Basically, the least squares analysis estimates effects α_i and β_j from the three treatment combinations. It then estimates the cell means for all four treatments (!) and then computes marginal means in the complete estimated table of cell means. The least squares estimates of effects obtained by GLM are as follows:

Parameter	Estimate	
Intercept	3.405402913	B
breed 1	0.850839744	B
breed 2	0.000000000	B
parity 2	-0.173902913	B
parity 3	0.000000000	B

The expected values of the four treatments under the assumed model are given in Table 7.2.

Table 7.2: Expected values for four (!) treatments under two-way model without interaction

		PARITY		Marginal means
		2	3	
BREED	1	$\mu + \alpha_1 + \beta_1$	$\mu + \alpha_1 + \beta_2$	$\mu + \alpha_1 + \bar{\beta}_\bullet$
	2	$\mu + \alpha_2 + \beta_1$	$\mu + \alpha_2 + \beta_2$	$\mu + \alpha_2 + \bar{\beta}_\bullet$
Marginal means		$\mu + \bar{\alpha}_\bullet + \beta_1$	$\mu + \bar{\alpha}_\bullet + \beta_2$	

The expected values can be estimated by plugging in the least squares solutions for effects. For example, the expected value for BREED=1 and PARITY=1 is estimated as follows:

$$\hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_1 = 3.405 + 0.851 - 0.174 = 4.082$$

Table 7.3: Least squares estimates of expected values for four (!) treatments under two-way model without interaction.

		PARITY		Marginal mean
		2	3	
BREED	1	4.082	4.256	4.169
	2	3.232	3.405	3.318
Marginal mean		3.657	3.831	

Note that marginal means should be interpreted with care because the interaction cannot be tested and thus usefulness of marginal means is doubtful.

SAS hints

```
proc glm data=simple;
  class breed parity;
  model ccv=breed parity/solution;
  lsmeans breed parity/pdiff;
  estimate 'mu11' intercept 1 breed 1 0 parity 1 0;
  estimate 'mu12' intercept 1 breed 1 0 parity 0 1;
  estimate 'mu21' intercept 1 breed 0 1 parity 1 0;
  estimate 'mu22' intercept 1 breed 0 1 parity 0 1; } estimate simple means
run;                                         (for illustration only)
```

Output:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
breed	1	2.93150778	2.93150778	28.34	0.0002
parity	1	0.07560556	0.07560556	0.73	0.4093

Source	DF	Type III SS	Mean Square	F Value	Pr > F
breed	1	1.80982067	1.80982067	17.50	0.0013
parity	1	0.07560556	0.07560556	0.73	0.4093

Parameter	Standard				Pr > t
	Estimate	Error	t Value		
Intercept	3.405402913 B	0.24910880	13.67	<.0001	
breed 1	0.850839744 B	0.20339648	4.18	0.0013	
breed 2	0.000000000 B	.	.	.	
parity 2	-0.173902913 B	0.20339648	-0.85	0.4093	
parity 3	0.000000000 B	.	.	.	

The GLM Procedure
Least Squares Means

breed	HO:LSMean1=		
	ccv	LSMEAN	Pr > t
1	4.16929120		0.0013
2	3.31845146		

parity	HO:LSMean1=		
	ccv	LSMEAN	Pr > t
2	3.65691987		0.4093
3	3.83082278		

Dependent Variable: ccv

Parameter	Estimate	Standard Error	t Value	Pr > t
mu11	4.08233974	0.14382303	28.38	<.0001
mu12	4.25624266	0.14382303	29.59	<.0001
mu21	3.23150000	0.14382303	22.47	<.0001
mu22	3.40540291	0.24910880	13.67	<.0001

The correct Type I SS for BREED are obtained by fitting breed after PARITY:

```
proc glm data=simple;
class breed parity;
model ccv= parity breed;
run;
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
parity	1	1.19729267	1.19729267	11.58	0.0052
breed	1	1.80982067	1.80982067	17.50	0.0013

Source	DF	Type III SS	Mean Square	F Value	Pr > F
parity	1	0.07560556	0.07560556	0.73	0.4093
breed	1	1.80982067	1.80982067	17.50	0.0013

A number of things are worth mentioning:

- (1) The Type III SS yield the correct test for both effects. The SS are the same regardless of the order of fitting. In this case, the Type III SS for an effect are corrected for all other effects in the model. Note that Type III do not always yield correct F-tests, so I recommend to stick with Type I SS, unless you know exactly what you are doing.
- (2) The p-values for the correct F-tests are the same as the p-values for the marginal least squares means.
- (3) The p-values for marginal means are the same for the corresponding p-values of cell means in the one-way ANOVA. Also, the pairwise differences among both sets of means are the same. For example, the difference of the marginal means for breed is 0.85, which is identical to the difference of cell means for breeds 1 and 2 at parity=2 (groups 2 and 3). This result underlines the fact that no new information can be gained here by the factorial ANOVA compared to the one-way ANOVA. In fact, the factorial ANOVA cannot test the interaction due to lack of data.
- (4) Estimated cell means are the same as observed cell means for the three observed cells.

Exercise 7.2: Reproduce the two-way analysis based on animal means. Try fitting an interaction term for breed \times parity. Interpret the result!

Modeling replicated data

The major advantage of analysing replicated data is that the time main effect and the interaction group \times time can be tested. A further benefit is that the missing data problem can be dealt with more satisfactorily. Finally, there may be a gain in power to detect treatment effects.

For each animal there are 14 replicated times of measurement, and at each point in time, two replicates are measured. It is typical of repeated measures that observations are correlated. Ignoring correlation, we would model the replicate data as

$$y_{ijtk} = \mu + \alpha_i + \gamma_t + (\alpha\gamma)_{it} + g_{ijtk}$$

where

y_{ijtk} = k -th measurement on j -th animal in i -th group at t -th point in time

μ = general effect

α_i = main effect of i -th group

γ_t = main effect of t -th point in time

$(\alpha\gamma)_{it}$ = interaction of i -th group with t -th point in time

g_{ijtk} = residual error term corresponding to y_{ijtk}

To account for correlation among observations on the same animal, we may add a number of random effects. The simplest approach is to add a random animal effect

e_{ij} = effect of j -th animal in i -th group; $e_{ij} \sim N(0, \sigma_e^2)$

The refined model is

$$y_{ijtk} = \mu + \alpha_i + \gamma_t + (\alpha\gamma)_{it} + e_{ij} + g_{ijtk}$$

The refined model introduces a correlation among observations on the same animal through the animal effect e_{ij} . However, the model implies that observations made at the same point in time are no more correlated than observations from different points in time. This is unrealistic: it is more natural to assume that observations made at the same point in time are more highly correlated than observations made at different points in time. To account for this, we may add a random interaction of time and animal:

f_{ijt} = random deviation of j -th animal in i -th group at t -th point in time; $f_{ijt} \sim N(0, \sigma_f^2)$

The updated model reads

$$y_{ijtk} = \mu + \alpha_i + \gamma_t + (\alpha\gamma)_{it} + e_{ij} + f_{ijt} + g_{ijtk}$$

Note that the model contains three random terms. It is therefore a **mixed model** and needs to be analysed by mixed model procedures (e.g., by PROC MIXED). This lecture cannot cover mixed model theory in much detail (but see Appendix B), so we will mainly look at the implementation in MIXED, making a few remarks on methodology.

```

data t;
input
Group Week Month Goat parity      breed kids protein      casein      ccv;
if ccv=0 then delete;
datalines;
1   2    1    1    3    1    2    3.56  2.64  4.45
1   2    1    1    3    1    2    3.55  2.64  .
<more data>
3   26   7    15   2    2    1    3.35  2.53  3.99
3   26   7    15   2    2    1    3.36  2.53  3.95
;
proc mixed data=t;
class week goat group;
model ccv=group week group*week/ddfm=kr;
random goat*group goat*group*week;
run;

```

e_{ij} f_{ijt}

Kenward-Roger method to determine d.f. and to compute standard errors. Extension of Satterthwaite method. Best currently available method to find d.f. in mixed model analysis. Details cannot be explained here.

Output:

Cov Parm	Estimate
Goat*Group	0.09749
Week*Goat*Group	0.04517
Residual	0.01361

The variances of the animal effect and the animal-by-time effect are rather larger than the residual variance, indicating that our modelling is quite appropriate.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Group	2	12	15.23	0.0005
Week	13	151	4.63	<.0001
Week*Group	26	151	0.88	0.6311

There is no interaction of group with time, but the time main effect is significant. Thus, we drop the interaction term and estimate marginal means for group and time.

```

proc mixed data=t;
class week goat group;
model ccv=group week/ddfm=kr;
random goat*group goat*group*week;
lsmeans group/pdiff;
run;

```

Output:

Least Squares Means

Effect	Group	Estimate	Standard	DF	t Value	Pr > t
			Error			
Group	1	4.2657	0.1423	12	29.98	<.0001
Group	2	4.0917	0.1424	12	28.73	<.0001
Group	3	3.2279	0.1424	12	22.67	<.0001

Differences of Least Squares Means

Effect	Group	_Group	Standard	DF	t Value	Pr > t
			Error			
Group	1	2	0.1740	0.2013	0.86	0.4043
Group	1	3	1.0378	0.2013	5.16	0.0002
Group	2	3	0.8638	0.2014	4.29	0.0010

The group means are almost the same as those computed from the animal means (Table 7.1, p. 219; also see Table 7.4). Also, the p-values are almost the same compared to the animal means analysis. This indicates that the animal means analysis is quite valid.

Table 7.4: Group means from mixed model analysis based on replicate data and simple means from animal means data.

GROUP	BREED	PARITY	Mean	Mean
			Replicate data	Animal means
1	1	3	4.27 ^a	4.25 ^a
2	1	2	4.09 ^a	4.08 ^a
3	2	2	3.23 ^b	3.23 ^b

Means in a column followed by the same letter are not significantly different.

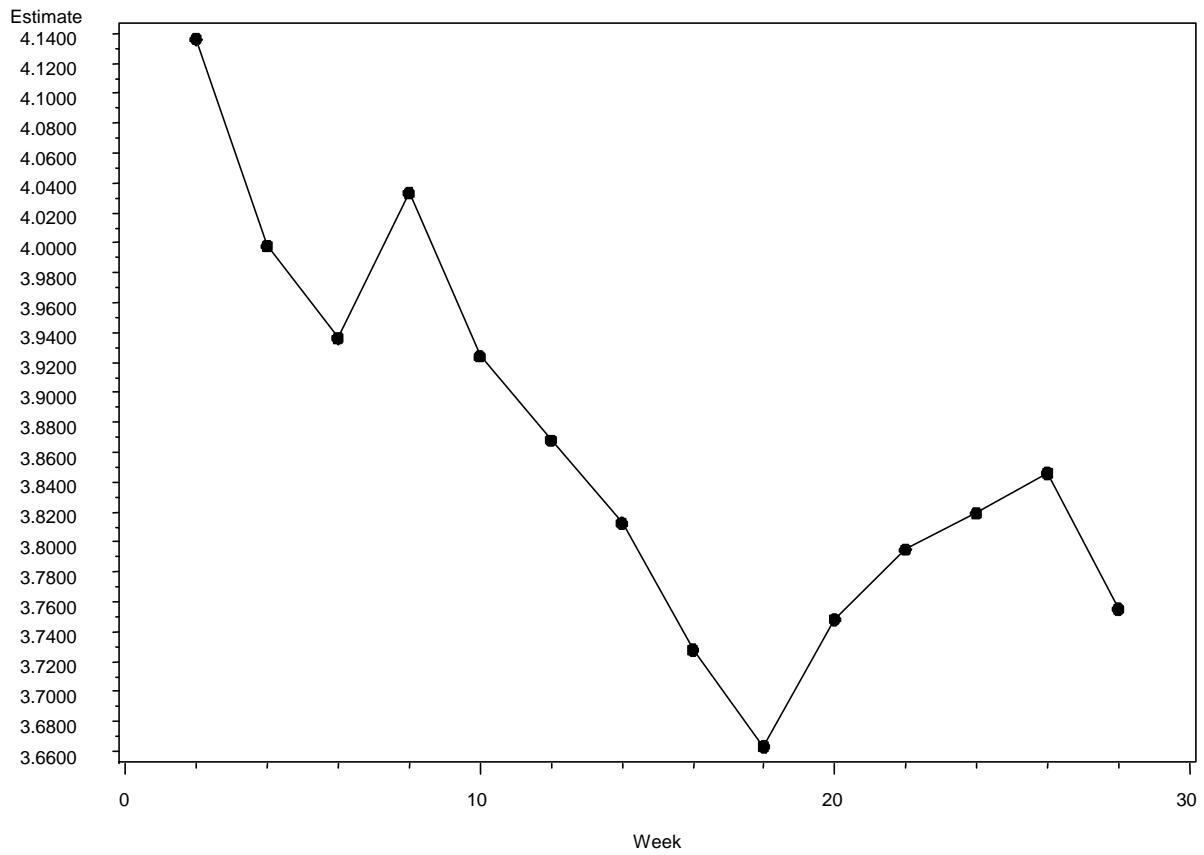
```

ods output lsmeans=lsmeans;
proc mixed data=t;
class week goat group;
model ccv=group week/ddfm=kr;
random goat*group goat*group*week;
lsmeans week/pdiff;
run;

symbol i=join value=dot;
proc gplot data=lsmeans;
plot estimate*week;
run;

```

Output:



Obviously, the time trend is negative for the first 18 weeks. After this, the trend is upwards. This pattern is in accordance with the expected lactation profile for goats (Dimassi, personal communication). We could try fitting a polynomial for the time trend, but this is not done here for brevity.

In summary: The mixed model analysis has allowed us the study the time effects. Moreover, the result for group means have been confirmed.

Exercise 7.3: Reproduce the analysis of replicate data. Try using PROC GLM to do the same analysis. Compare the results. Note: GLM is able to produce correct F-tests, but the standard errors for mean comparisons are not appropriate.

7.2 The Sorghum data

Example 7.2: In a greenhouse experiment, three different intensities of double use of Sorghum (grain for human consumption and leaves for fodder) were tested: (1) control (no leaves removed); (2) removal of all leaves except the top leaf; (3) Removal of all except the six top-most leaves (Piepho, 1997). Leaves were removed at four points in time within a period of three weeks. The experimental design was an RCBD with four replications. Sorghum plants were planted in pots, one plant per pot. A replication comprised 10 pots per treatment. The ten pots were placed on a tray. Thus, a complete block consisted of three trays, one for each treatment. Measurements were taken on a plant-basis. Here we look at the thousand kernel weight (tkw; see Table 7.5). A few plants died before any leaves were removed. Thus, the missing data pattern is independent of treatments, i.e., data are **missing completely at random (MCAR)**. Due to missing values, the data are unbalanced.

The objective of the analysis is to compare the means for the three treatments to see if double use is a useful alternative to grain production.

Table 7.5: Thousand kernel weight (TKW) of Sorghum in greenhouse experiment on double use (fodder and grain production; **sorghum.dat**).

Treatment	Block	Plant no.									
		1	2	3	4	5	6	7	8	9	10
1	1	41.89	29.97	27.82	29.68	27.84	33.93	34.77	30.00	27.71	32.66
	2	28.21	34.80	31.25	34.35	37.50	36.74	32.23	38.35	29.50	
	3	41.54	44.19	40.44	35.20	26.76	32.37	31.18	34.23	34.22	25.01
	4	35.26	26.65	35.47	29.84	34.38	26.79	41.39	40.60	28.16	28.11
2	1	41.28	32.86	35.15		21.76	34.60	19.36	46.22	41.64	32.91
	2	27.55	40.89	33.14	29.62		48.43	25.39	35.45	27.73	29.45
	3	32.27	38.62	24.18	29.34	22.77		23.87	28.01	24.70	23.70
	4	42.36	35.61	17.47	23.55	18.21	33.21	32.76	26.47	21.68	
3	1	27.40	32.66	24.15	27.98	35.43		31.94	32.24	29.87	
	2	44.37	30.01		36.49	40.40	29.48	24.69	35.03	27.48	
	3	35.98	32.82	27.16	29.81	33.91	28.28	29.67	33.66		29.50
	4	36.58	27.64	27.53	26.23		31.00	31.36	29.67	30.12	27.33

Fig. 7.1 shows the arrangement of plants within a block.

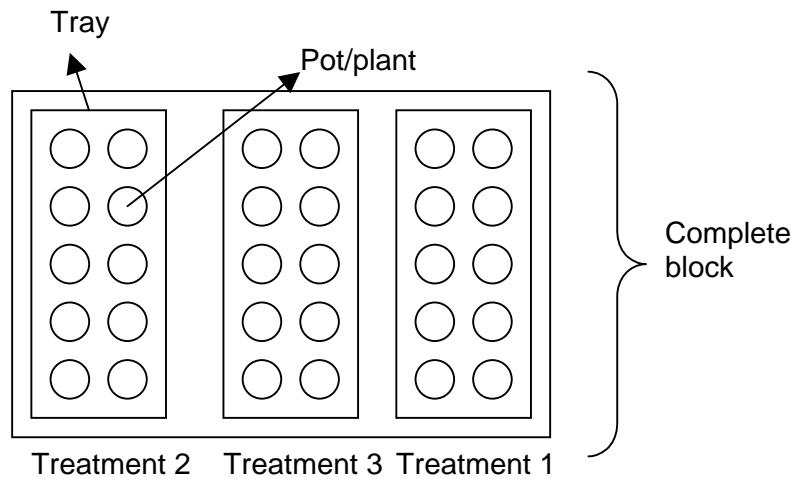


Fig. 7.1: Arrangement of plants (pots) within a block.

It is important to reflect that trays, not pots, are the randomization units. Clearly, pots are pseudo-replications, while trays are true replications. Another way of looking at the experiment is to consider plots (or measurements taken on pots) as repeated measurements on the same experimental unit (tray).

The summary value approach

We can compute means per tray and analyse tray means by standard procedures for RCBD. There are four tray means per treatment (one for each block), so the means data has a total of twelve observations. Ideally, the means would have been computed from ten plants on each tray. However, some observations are missing, so some means are based on only nine or eight observations. It is known the the standard error of a mean decreases with the number of observations, so some means will be more accurate than others. This difference in accuracy cannot be accounted for by a simple analysis of means. In fact, the analysis of means is not strictly valid because heterogeneity of variance is ignored. We will see later how a mixed model can be used for a more refined analysis.

SAS hints

Means can be computed by PROC MEANS. Means are then analysed by PROC GLM.

```
data t;
input
treat    block    plant     TKW;
datalines;
  1      1       1      41.89
  1      1       2      29.97
<more data>
  3      4       9      30.12
  3      4      10      27.33
;
proc sort data=t out=t;
by block treat;

proc means data=t;
var tkw;
output out=s mean=;
by block treat;

proc print data=s;
run;
```

Output:

block	treat	_TYPE_	_FREQ_	TKW
1	1	0	10	31.6270
1	2	0	10	33.9756
1	3	0	10	30.2088
2	1	0	10	33.6589
2	2	0	10	33.0722
2	3	0	10	33.4938
3	1	0	10	34.5140
3	2	0	10	27.4956
3	3	0	10	31.1989

```

4      1      0      10      32.6650
4      2      0      10      27.9244
4      3      0      10      29.7178
;
proc glm data=s;
class block treat;
model tkw=block treat;
means treat/lsd;
run;

```

Output:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
block	3	17.71300386	5.90433462	1.19	0.3890
treat	2	13.84383849	6.92191925	1.40	0.3172
Source	DF	Type III SS	Mean Square	F Value	Pr > F
block	3	17.71300386	5.90433462	1.19	0.3890
treat	2	13.84383849	6.92191925	1.40	0.3172

The GLM Procedure

t Tests (LSD) for TKW

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	6
Error Mean Square	4.948435
Critical Value of t	2.44691
Least Significant Difference	3.8489

Means with the same letter are not significantly different.

t Grouping	Mean	N	treat
A	33.116	4	1
A			
A	31.155	4	3
A			
A	30.617	4	2

There are no significant differences among the three treatments.

Modeling the replicate data

Ignoring correlation among observations on the same tray, we would base the analysis of replicate data on the following model:

$$y_{ijk} = \mu + b_j + \alpha_i + e_{ijk}$$

where

y_{ijk} = measurement on k -th plant for i -th treatment in j -th block

μ = general effect (intercept)

α_i = effect of i -th treatment

b_j = effect of j -th block

e_{ijk} = residual error of ijk -th plant; $e_{ijk} \sim N(0, \sigma_e^2)$

This model ignores the fact that plants from the same tray will be correlated due to common environmental conditions. Clearly, if the treatments were the same on each tray, plants on the same tray would be expected to be more similar than plants from different trays. The correlation of plants on the same tray can be modeled by adding a random tray effect,

u_{ij} = random effect of ij -th tray; $u_{ij} \sim N(0, \sigma_u^2)$

The refined model is

$$y_{ijk} = \mu + b_j + \alpha_i + u_{ij} + e_{ijk}$$

It is useful to consider the statistical properties of tray means under the mixed model. It can be shown that the tray means, $\bar{y}_{ij\bullet}$, have the following variance:

$$\text{var}(\bar{y}_{ij\bullet}) = \sigma_u^2 + \frac{\sigma_e^2}{n_{ij}}$$

where n_{ij} is the number of plants per tray. Obviously, the variance of a mean depends on the number of observations, n_{ij} . Thus, for unbalanced data, there will be heterogeneity of variance among the means, and, strictly speaking, an important assumption for ANOVA is violated. It transpires from the variance formula, however, that variance heterogeneity will be small when the tray variance, σ_u^2 , is much larger than the plant variance, σ_e^2 .

SAS hints:

```

data t;
input
treat    block    plant     TKW;
datalines;
1       1       1      41.89
1       1       2      29.97
<more data>
3       4       9      30.12
3       4      10     27.33
;
proc mixed data=t;
class treat block;
model tkw=treat block/ddfmethod=kr;
lsmeans treat/pdiff;
random block*treat;
run;

```

Kenward-Roger method for computing
approximate d.f. and standard errors.

Output:

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
treat	2	6.15	1.43	0.3097
block	3	6.13	1.11	0.4136

Least Squares Means

Effect	treat	Estimate	Standard		t Value	Pr > t
			Error	DF		
treat	1	33.1393	1.0965	5.53	30.22	<.0001
treat	2	30.6169	1.1288	6.23	27.12	<.0001
treat	3	31.1691	1.1546	6.78	27.00	<.0001

Differences of Least Squares Means

Effect	treat	_treat	Estimate	Standard		t Value	Pr > t
				Error	DF		
treat	1	2	2.5223	1.5737	5.87	1.60	0.1611
treat	1	3	1.9702	1.5922	6.14	1.24	0.2612
treat	2	3	-0.5521	1.6147	6.5	-0.34	0.7432

There are no significant treatment effects. The result agrees very well with the analysis of tray means. For example, the value of F_{exp} was 1.40 for the means data and is 1.43 for the replicate data.

Covariance Parameter Estimates

Cov Parm	Estimate
treat*block	1.2805
Residual	34.3450

The tray variance, σ_u^2 , is small (1.28) relative to the pot variance, σ_u^2 . This is one reason why results are so similar, despite the unbalancedness. Another reason is that unbalancedness is mild. Nevertheless, the mixed model analysis is more satisfactory here because it accounts for the missing data pattern and for the repeated measures nature of the data.

7.3 An addendum to the goats data

Having detected group differences in the cheese conversion value (ccv), the researcher wanted to know if variation in ccv could be at least partly explained by casein and/or protein level in the milk. Both casein and protein are quantitative variables, so association could be assessed

by simple correlation. The simple correlation, however, will not account for the complex design structure and the repeated measures nature of the data. Thus, simple correlation is not appropriate here. Instead, one may attack the problem by analysis of covariance. **It is a common mistake that simple correlation is used in situations were the design calls for a more complex analysis.**

For the cvv, we had selected a model with main effects for time and group (the interaction was n.s.):

$$y_{ijtk} = \mu + \alpha_i + \gamma_t + e_{ij} + f_{ijt} + g_{ijtk}$$

where

y_{ijtk} = k -th measurement on j -th animal in i -th group at t -th point in time

μ = general effect

α_i = main effect of i -th group

γ_t = main effect of t -th point in time

e_{ij} = effect of j -th animal in i -th group; $e_{ij} \sim N(0, \sigma_e^2)$

f_{ijt} = random deviation of j -th animal in i -th group at t -th point in time; $f_{ijt} \sim N(0, \sigma_f^2)$

g_{ijtk} = residual error term corresponding to y_{ijtk}

To assess the association of cvv with protein, we may simple add a linear regression term for protein:

$$\gamma_1 x_{ijtk}$$

where x_{ijtk} is the protein value associated with the $ijtk$ -th observation and γ_1 is a regression coefficient. We may add further polynomial terms. Furthermore, we can test the interaction of the regression with groups.

SAS hints

```
proc mixed data=t;
  class week goat group;
  model cvv=group week protein protein*protein/htype=1;
  random goat*group goat*group*week;
run;
```

Output:

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Group	2	12	18.60	0.0002
Week	13	169	4.76	<.0001
protein	1	189	8.08	0.0050
protein*protein	1	189	1.96	0.1636

The linear term is significant, while the quadratic is not. We continue by dropping the quadratic term and adding an interaction between group and protein.

```
proc mixed data=t;
class week goat group;
model ccv=group week protein group*protein/htype=1;
random goat*group goat*group*week;
run;
```

Output:

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Group	2	12	19.11	0.0002
Week	13	169	4.66	<.0001
protein	1	188	8.39	0.0042
protein*Group	2	188	1.16	0.3164

The interaction is not significant, so we drop the term from the model. The linear regression is estimated using the solution option:

```
proc mixed data=t;
class week goat group;
model ccv=group week protein/htype=1 solution;
random goat*group goat*group*week;
run;
```

Output:

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Group	2	12	18.11	0.0002
Week	13	169	4.73	<.0001
protein	1	190	6.77	0.0100

Solution for Fixed Effects

Effect	Week	Group	Estimate	Standard Error		DF	t Value	Pr > t
<stuff deleted>			0.1603	0.06158		190	2.60	0.0100

An increase of protein by one unit is associated with an increase of the cheese conversion value by 0.16 units. This association is **adjusted for both group and week effects**: it is an association "within groups and weeks". To study the association of protein and cvv between groups and between weeks, we could contrast the corresponding least squares means for protein and cvv (not done here).

Exercise 7.4: Analyse the association of ccv with casein (**cheese.dat**).

Exercise 7.5: Consider the example given in section 14.7. (p.449-460) given in D.C. Howell (Statistical methods for psychology) (**Stat Meth f Psychology.pdf**). Perform a mixed model analysis for this repeated measures dataset (**Statistik_Howell_Tab14_4.xls**) and compare the results to the more traditional analysis provided in Howell's book.

7.4 The colon data

Example 7.3 (J. Mentschel, FG Tierhaltung und Leistungsphysiologie, Institut Tierhaltung und Tierzüchtung, Universität Hohenheim - description of investigation partly verbatim from Claus et al., 2002): The colon of pigs has a fine microstructure, as depicted in Fig. 7.1. The colon epithel is covered by crypts, small cavities producing cells (enterocytes). The crypt cells have different functions, e.g., absorption of nutrients and production of mucus to facilitate the movement of faeces through the colon. The cells originate from mitoses of base cells at the bottom of the crypts (stem cell region), subsequently differentiate and migrate to the luminal end, and are secreted into the lumen of the colon after programmed cell death (apoptosis).

Evidence exists that butyrate inhibits apoptosis (death) of colon crypt cells in vivo so that less tryptophan from cell debris is available for skatol formation by microbes in the pig colon. Cell death tends to occur mainly at the luminal end of the crypts. One objective of the present study was to investigate the effect of two types of starch on the crypt cell survival. Two types of potatoe starch were compared: starch containing a high proportion of resistant starch and regular starch. The hypothesis to be tested was that increased butyrate formation will occur in the colon and contribute to reduced epithelial cell apoptosis thus leading to reduced skatole formation and absorption. Two groups of six barrows were provided with catheters into the jugular vein and fed either a ration with pre-gelatinized starch (high ileal digestibility; control) or potatoe starch (low ileal digestibility) as the main carbohydrate. After euthanizing barrows at the end of the feeding period, colon tissue for histological quantification of mitosis and apoptosis were obtained. For each animal, three tissue samples (segments) were taken from the proximal and the distal colon region. Cells undergoing apoptosis were identified by a modified TUNEL assay that leads to a staining of apoptosis-specific DNA-fragments. If possible, for each of the three regional segments 30 well oriented crypts (ideally those whose entire length could be completely visualised and those whose bases were in contact with the underlaying epithelium) were counted. For each animal, the mean value of the number of apoptotic cells per crypt was calculated for each region.

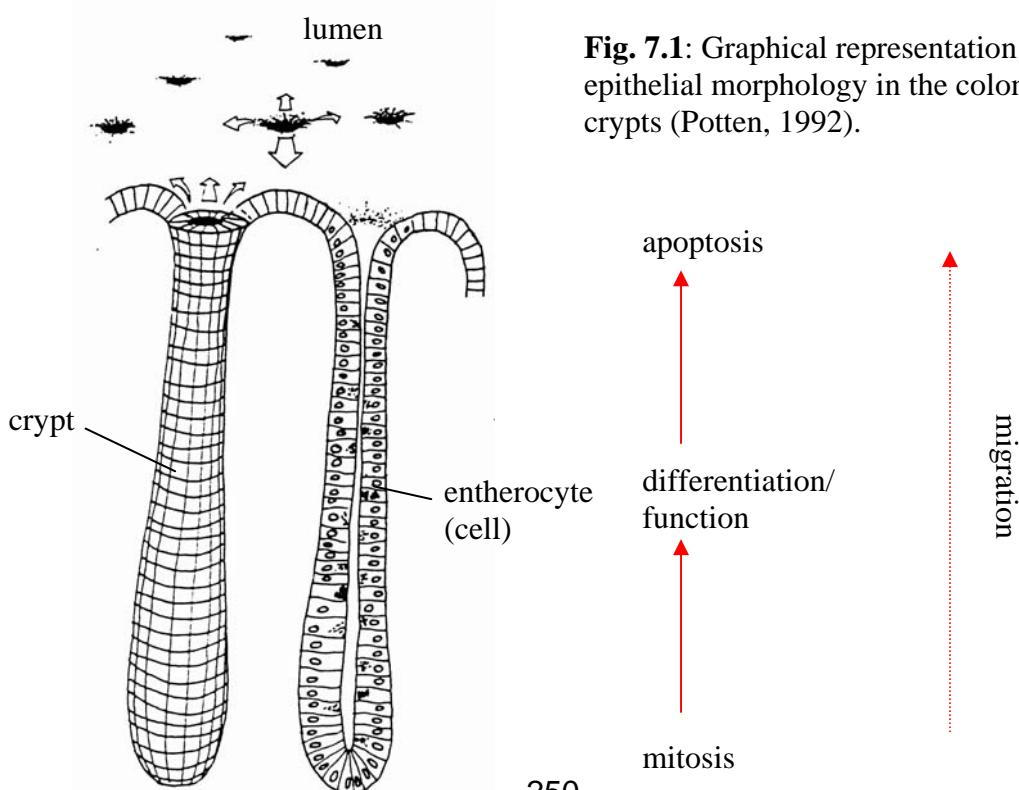


Fig. 7.1: Graphical representation of the epithelial morphology in the colon, with two crypts (Potten, 1992).

The following data were recorded (**colon.dat**):

group = group (1 = control, 3 = resistant starch = RS)
 region = region of colon (1 = proximal, 2 = distal)
 animal = animal number
 avcount = average count of apoptotic cells per crypt from at least 30 crypts per sample,
 averaged across three samples per animal and region

group	region	animal	avcount
1	1	1	0.93667
1	1	2	1.18000
1	1	3	0.85667
1	1	4	0.77667
1	1	5	0.49667
1	1	6	0.89000
1	2	1	1.34000
1	2	2	1.17000
1	2	3	1.26667
1	2	4	0.78000
1	2	5	0.71000
1	2	6	0.96333
3	1	12	0.37000
3	1	13	0.54000
3	1	14	0.36333
3	1	15	0.54333
3	1	16	0.40000
3	1	17	0.44333
3	2	12	0.36333
3	2	13	0.30667
3	2	14	0.46000
3	2	15	1.13000
3	2	16	0.61000
3	2	17	1.07667

The major objective here is to compare the two groups (control versus RS) to see if resistant starch has an effect on cell death. Secondly, it is of interest to compare the two colonial regions (proximal versus distal). Thus, we have a two-way factorial treatment structure, where each factor has two levels. Measurements in different regions on the same animal are correlated, so there is a repeated measures problem.

In what follows, we will analyse the data set in a step-by-step fashion, using different approaches of increasing complexity, starting from simple t-tests.

7.4.1 Groups - the unpaired t-test

We can compare groups separately for the proximal region and for the distal region. The simple t-test can be used for group comparisons. The t-test is equivalent to a one-way ANOVA for two groups (Sections 4.1 and 4.2). The t-statistic is computed as

$$t_{\text{exp}} = \frac{|\bar{y}_1 - \bar{y}_2|}{s \sqrt{\frac{2}{n}}}$$

where

n = number of replications (animals!) per group

\bar{y}_1, \bar{y}_2 = means of groups 1 and 2

$$s^2 = \frac{s_1^2 + s_2^2}{2}$$

s_1^2, s_2^2 = sample variances for groups 1 and 2

The observed t-value (t_{exp}) is compared to a tabular t (t_{tab}) with $2(n-1)$ d.f.

The above formula is valid for balanced data (for unbalanced data use formula in section 4.1).

Region	Group	Mean (\bar{y}_j)	Variance (s_j^2)	t_{exp}	t_{tab}
Proximal	Control	0.856	0.0496	4.26	2.228
	RS	0.443	0.0066		
Distal	Control	1.038	0.0681	2.09	2.228
	RS	0.658	0.1300		

The comparison is significant for the proximal region, and cell death is reduced for the RS treatment. There is no significant difference for the distal region, though the difference is about the same for both regions. The reason is the larger variance for the distal region.

7.4.2 Regions - the paired t-test

In a similar fashion, we may compare regions separately for each group. In a group, there are six animals, so there are six measurements for the proximal region and six measurements for the distal region. It is tempting to use the same t-test as the one used for comparing groups, but this would not be appropriate for an important reason: The twelve observations are not independent. Specifically, observations made on the same animal are correlated. In fact, there are six pairs of observations, one for each animal. This pairing of observations invalidates the t-test used in section 7.4.1. Instead, we need to use the "paired" t-test. This works as follows:

Paired t-test:

- Compute differences of both values (distal and proximal), d_i , for each subject (animal)
- Compute mean, \bar{d} , and standard deviation, s_d , of differences d_i
- Compute $t_{\text{exp}} = \frac{|\bar{d}|}{s_d} \sqrt{n}$, where n is the number of subjects (animals)
- Compare t_{exp} to a t-distribution with $n-1$ d.f.

This is exemplified for the first group (control):

Proximal	distal	difference (d_i)
0.93667	1.34000	-0.40333
1.18000	1.17000	0.01000
0.85667	1.26667	-0.41000
0.77667	0.78000	-0.00333
0.49667	0.71000	-0.21333
0.89000	0.96333	-0.07333

$$\begin{aligned}\bar{d} &= -0.1822 \\ s_d &= 0.1911 \\ t_{exp} &= 2.336 \\ t_{tab} &= 2.571 \quad (n = 6, d.f. = 5)\end{aligned}$$

Note that the mean of differences (\bar{d}) is the same as the difference computed from the two means of group 1, i.e.,

$$\bar{d} = 0.856 - 1.038 = -0.182$$

The two regions are not significantly different for the first group. For the second group, we find $t_{exp} = 1.55$, which is not significant.

7.4.3 Unpaired t-test for groups and one-way ANOVA

Instead of a t-test, we may perform an F-test by a one-way ANOVA according to the following model (compare sections 4.1 and 4.2):

$$y_{ik} = \mu + \alpha_i + e_{ik}$$

where

y_{ik} = average count for k -th animal in i -th group

μ = general effect

α_i = effect of i -th group

e_{ik} = error

F-test and t-test are equivalent because there are only two groups.

SAS hints

The ANOVA can be performed using GLM or MIXED. Here, we use MIXED. Means can also be compared by a t-test using LSMEANS.

```

data;
input
group    region    animal      avcount;
datalines;
  1        1         1       0.93667
  1        1         2       1.18000
  1        1         3       0.85667
  1        1         4       0.77667
  1        1         5       0.49667
  1        1         6       0.89000
  1        2         1       1.34000
  1        2         2       1.17000
  1        2         3       1.26667
  1        2         4       0.78000
  1        2         5       0.71000
  1        2         6       0.96333
  3        1         12      0.37000
  3        1         13      0.54000
  3        1         14      0.36333
  3        1         15      0.54333
  3        1         16      0.40000
  3        1         17      0.44333
  3        2         12      0.36333
  3        2         13      0.30667
  3        2         14      0.46000
  3        2         15      1.13000
  3        2         16      0.61000
  3        2         17      1.07667
;

proc sort;
by region group;

proc mixed;
class group;
model avcount=group;
lsmeans group/pdiff;
by region;
run;

```

Output:

----- region=1 -----

The Mixed Procedure

Type 3 Tests of Fixed Effects

Effect		Num	Den	F Value	Pr > F		
		DF	DF				
group		1	10	18.19	0.0017		
Least Squares Means							
Effect	group	Estimate	Standard Error	DF	t Value		
group	1	0.8561	0.06844	10	12.51		
group	3	0.4433	0.06844	10	6.48		
Differences of Least Squares Means							
Effect	group	_group	Estimate	Standard Error	DF	t Value	Pr > t
group	1	3	0.4128	0.09679	10	4.26	0.0017

----- region=2 -----

Type 3 Tests of Fixed Effects

Effect		Num	Den	F Value	Pr > F		
		DF	DF				
group		1	10	4.39	0.0627		
Least Squares Means							
Effect	group	Estimate	Standard Error	DF	t Value		
group	1	1.0383	0.1285	10	8.08		
group	3	0.6578	0.1285	10	5.12		
Differences of Least Squares Means							
Effect	group	_group	Estimate	Standard Error	DF	t Value	Pr > t
group	1	3	0.3806	0.1817	10	2.09	0.0627

Note that the p-values of the F-test and the t-test are exactly the same. Also note that $F_{exp} = (t_{exp})^2$, as expected. The t-values and resulting conclusions are the same as those computed in section 7.4.1.

7.4.4 Paired t-test for regions and ANOVA for block design

Instead of a t-test, we may perform an F-test by an ANOVA according to the model

$$y_{jk} = \mu + f_k + \beta_j + e_{jk}$$

where

y_{jk} = average count for j -th region on k -th animal

μ = general effect

β_j = effect of j -th region

f_k = effect of k -th animal

e_{jk} = error

Note that the model is equivalent to that for a randomized complete block design, if we identify animals with blocks. In fact, each animal constitutes a complete block with treatments "proximal" and "distal". A difference compared to the block design is that the regions cannot be randomized. Nevertheless, we may use the same form of model.

If animals can be regarded as a random sample, the animal effect is random with zero mean and variance σ_f^2 . It is important to recognise that the random animal effect models a correlation of observations on the same animal. It is reasonable to expect such correlation, and the model reflects this through the animal effect. In fact, the correlation is given by

$$\rho = \frac{\sigma_f^2}{\sigma_f^2 + \sigma^2},$$

where σ^2 is the residual error variance. The larger the variance of animal effects, the larger the correlation.

Even though animal effects are random by design, the ANOVA is unaltered, if the animal effect is formally regarded as fixed. The analysis for fixed animal effects is simpler and thus preferable. Also note that F-test and t-test are equivalent here because there are only two regions.

SAS hints

The ANOVA can be performed using GLM or MIXED. Here, we use MIXED. Means can also be compared by a t-test using LSMEANS.

```

proc sort;
by group region;

proc mixed;
class region animal;
model avcount=animal region;
lsmeans region/pdiff;
by group;
run;

```

Output:

----- group=1 -----

The Mixed Procedure

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
animal	5	5	5.45	0.0431
region	1	5	5.46	0.0667

Least Squares Means

Effect	region	Estimate	Standard		Pr > t
			Error	DF	
region	1	0.8561	0.05516	5	15.52 <.0001
region	2	1.0383	0.05516	5	18.82 <.0001

Differences of Least Squares Means

Effect	region	_region	Estimate	Standard		Pr > t
				Error	DF	
region	1	2	-0.1822	0.07801	5	-2.34 0.0667

----- group=3 -----

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
animal	5	5	1.37	0.3704
region	1	5	2.39	0.1828

Least Squares Means

Effect	region	Estimate	Standard		Pr > t
			Error	DF	
region	1	0.4433	0.09809	5	4.52 0.0063
region	2	0.6578	0.09809	5	6.71 0.0011

Differences of Least Squares Means

Effect	region	_region	Estimate	Standard		Pr > t
				Error	DF	
region	1	2	-0.2144	0.1387	5	-1.55 0.1828

Note that, again, the p-values of the F-test and the t-test are exactly the same. Also note that $F_{exp} = (t_{exp})^2$, as expected. The t-values and conclusions are the same as those computed in section 7.4.2.

7.4.5 A t-test for interaction

As we shall see in section 7.4.8, we can perform a two-way ANOVA to test the interaction between groups and region. In this section, it will be shown, that in the present case, we can test the interaction by a t-test. This is so because both factors just have two levels.

An interaction occurs when the difference between distal and proximal region is not the same for both groups. Thus, we may compute the difference between regions in both groups and then look at the **difference of differences**.

Region	Group	Mean (\bar{y}_j)	Variance (s_j^2)	t_{exp}	t_{tab}
Proximal	Control	0.856	0.0496	4.26	2.228
	RS	0.443	0.0066		
Distal	Control	1.038	0.0681	2.09	2.228
	RS	0.658	0.1300		

The difference among groups is 0.413 for the proximal region and 0.380 for the distal region. The two differences are quite similar. The difference of the two differences is 0.033, so there is little indication of interaction. To show that there is no interaction, we need to demonstrate that the observed difference of differences is not significantly different from zero.

We may look at this the other way round. The difference among regions is 0.182 for group 1 and 0.215 for group 3. Again, of course, the difference of differences is 0.033.

How can we test the null hypothesis of no interaction? First consider differences among both regions on the same animal.

Group (i)	Proximal	Distal	Difference (d_{ik})
Control	0.93667	1.34000	-0.40333
	1.18000	1.17000	0.01000
	0.85667	1.26667	-0.41000
	0.77667	0.78000	-0.00333
	0.49667	0.71000	-0.21333
	0.89000	0.96333	-0.07333
	$\bar{y}_1 = 0.8561$	$\bar{y}_2 = 1.0383$	$\bar{d}_1 = -0.1822$
RS	$s_{d1}^2 = 0.0365$		
	0.37000	0.36333	0.00667
	0.54000	0.30667	0.23333
	0.36333	0.46000	-0.09667
	0.54333	1.13000	-0.58667
	0.40000	0.61000	-0.21000
	0.44333	1.07667	-0.63334
	$\bar{y}_1 = 0.4433$	$\bar{y}_2 = 0.6578$	$\bar{d}_2 = -0.2144$
	$s_{d2}^2 = 0.1155$		

Note that the difference in cell means for proximal and distal region within a group equals the mean of differences d_{ik} within a group. The model for a difference d_{ik} from the k -th animal in the i -th group can be written

$$d_{ik} = \delta_i + e_{ik}$$

where

d_{ik} = difference between proximal and distal cell count on k -th animal in i -th group

δ_i = expected difference between proximal and distal region for the i -th group

e_{ik} = error corresponding to d_{ik}

If there is no interaction, the expected differences are the same in both groups, i.e.,

$$H_0: \delta_1 = \delta_2$$

This hypothesis may be tested by an unpaired t-test on the differences. We find

$$t_{\text{exp}} = \frac{|\bar{d}_1 - \bar{d}_2|}{s \sqrt{\frac{2}{n}}} = \frac{|-0.1822 + 0.2144|}{\sqrt{\frac{0.0365 + 0.1155}{2}} \sqrt{\frac{2}{6}}} = 0.20$$

on $2(n-1) = 10$ d.f. (The pooled variance was used to estimate s^2 , i.e., $s^2 = (s_{d1}^2 + s_{d2}^2)/2$).

This is not significant, since $t_{\text{tab}} = 2.228$ ($p = 0.8436$). Thus, the interaction is not significant.

7.4.6 Comparing marginal means for groups by an unpaired t-test (assuming no interaction)

Since there is no interaction, it is reasonable to compare marginal means both for groups as well as for regions. Let us consider the comparison of groups first. It makes sense to compute the average of proximal and distal counts per animal, i.e., to average across the regions. We thus obtain six averages for each group, one per animal. The group averages of these averages are **marginal means** for the group factor. The marginal means may be subjected to a simple **unpaired t-test**.

Group (i)	proximal	distal	mean (m_{ik})
Control	0.93667	1.34000	1.13834
	1.18000	1.17000	1.17500
	0.85667	1.26667	1.06167
	0.77667	0.78000	0.77834
	0.49667	0.71000	0.60334
	0.89000	0.96333	0.92667
			$\bar{m}_1 = 0.9472$
			$s_{m1}^2 = 0.04974$
Group (i)	proximal	distal	mean (m_{ik})
RS	0.37000	0.36333	0.36667
	0.54000	0.30667	0.42334
	0.36333	0.46000	0.41167
	0.54333	1.13000	0.83667
	0.40000	0.61000	0.50500
	0.44333	1.07667	0.76000
			$\bar{m}_2 = 0.5506$
			$s_{m2}^2 = 0.03942$

From this, we compute

$$t_{\text{exp}} = \frac{|\bar{m}_1 - \bar{m}_2|}{s \sqrt{\frac{2}{n}}} = \frac{|0.9472 - 0.5506|}{\sqrt{\frac{0.04974 + 0.03942}{2}} \sqrt{\frac{2}{6}}} = 3.25$$

which is compared against t_{tab} with $2(n-1) = 10$ d.f., i.e., $t_{\text{tab}} = 2.228$ (p-value = 0.0087) (The pooled variance was used to estimate s^2 , i.e., $s^2 = (s_{m1}^2 + s_{m2}^2)/2$). Thus, there is a significant difference among marginal group means. Cell death is reduced for RS. Note that the error term for the t-test depends on the between-animal variability. This is in contrast to the error term for the comparison of region means (see below).

7.4.7 Comparing marginal means for regions by a paired t-test (assuming no interaction)

We may compare proximal and distal region on every animal. Assuming there is no interaction, the observed difference d_{ik} for the k -th animal in the i -th group must have the same expected value in both groups. Thus, to estimate the marginal difference between regions, we may compute the average across all differences d_{ik} , i.e., across all twelve animals from the two groups. This average is identical to the difference of marginal means for regions. The mean difference may be tested by a paired t-test, using the twelve difference listed below, ignoring the group variable (this is okay because there is no interaction).

Group (i)	proximal	distal	difference (d_{ik})
Control	0.93667	1.34000	-0.40333
	1.18000	1.17000	0.01000
	0.85667	1.26667	-0.41000
	0.77667	0.78000	-0.00333
	0.49667	0.71000	-0.21333
	0.89000	0.96333	-0.07333
RS	0.37000	0.36333	0.00667
	0.54000	0.30667	0.23333
	0.36333	0.46000	-0.09667
	0.54333	1.13000	-0.58667
	0.40000	0.61000	-0.21000
	0.44333	1.07667	-0.63334
$\bar{y}_1 = 0.6497$		$\bar{y}_2 = 0.8481$	$\bar{d} = -0.1984$
			$s_d^2 = 0.06936$

Note that the estimated difference equals the difference of marginal means for proximal and distal region:

$$\bar{y}_1 - \bar{y}_2 = 0.6497 - 0.8481 = -0.1984 = \bar{d}$$

The t-statistic for the paired t-test of differences d_{ik} is computed as

$$t_{\text{exp}} = \frac{|-0.1984|}{\sqrt{0.06936}} \sqrt{2n-1} = 2.609$$

which is significant compared to $t_{\text{tab}} = 2.201$ on $2n-1 = 11$ d.f. (p-value = 0.0243). Note that this test assumes absence of interaction. Also, it should be pointed out that the error term for the t-test depends only on the variability within animals, not that between animals. In the extreme, when there is no within-animal variability, the difference is the same on each animal, even when there is considerable between-animal variation of average counts. What matters here is only how variable are the differences, not how variable are the observed average counts. This is in contrast to the error term for comparing group means, which depends on between-animal variability.

7.4.8 A linear model for two-way analysis

So far, we have done most analyses based on t-tests, which is intuitively appealing. The analysis based on t-tests is possible here because each factor (group, region) has just two levels. If any of the two factors has more than two levels, a fully efficient analysis is available only by two-way ANOVA based on a suitable linear model.

The standard two-way ANOVA model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

where

y_{ijk}	= mean count at j -th region on k -th animal within i -th group
μ	= general effect
α_i	= main effect of i -th group
β_j	= main effect of j -th region
$(\alpha\beta)_{ij}$	= group \times region interaction
e_{ijk}	= residual error

This model assumes that all observations are independent. This assumption is not valid here, because repeated measurements (distal, proximal) are made on each animal. The correlation may be modelled by an animal effect as follows (compare section 7.4.4):

$$y_{ijk} = \mu + \alpha_i + f_{ik} + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

where

$$f_{ik} = \text{effect of the } k\text{-th animal within the } i\text{-th group.}$$

The model has two random effects, f_{ik} and e_{ijk} . These are assumed to follow a normal distribution with zero mean, which may be expressed by

$$\begin{aligned} f_{ik} &\sim N(0, \sigma_f^2) \\ e_{ijk} &\sim N(0, \sigma_e^2) \end{aligned}$$

Before embarking on an ANOVA based on the above model, it is useful to consider the means and differences used in previous sections to perform various t-tests. The ingredients were the sums (means) and differences for proximal and distal region on every animal. The difference among regions on the ik -th animal is

$$\begin{aligned} d_{ik} &= y_{i1k} - y_{i2k} = \mu + \alpha_i + f_{ik} + \beta_1 + (\alpha\beta)_{i1} + e_{i1k} \\ &\quad - [\mu + \alpha_i + f_{ik} + \beta_2 + (\alpha\beta)_{i2} + e_{i2k}] \\ &= \beta_1 + (\alpha\beta)_{i1} + e_{i1k} - [\beta_2 + (\alpha\beta)_{i2} + e_{i2k}] \end{aligned}$$

An important feature is that the random animal effect, f_{ik} , cancels out. Thus, the variance of d_{ik} depends only on the random errors e_{ijk} , which reflects within-animal variation. The variance of d_{ik} is free of between-animal variation, modelled by f_{ik} . Note that the t-tests for interaction and test of marginal region means are computed from differences d_{ik} . Thus, the error term for testing interaction and marginal region means depends only on residual errors, i.e., they depend only on within-animal variation, modelled by e_{ijk} .

The mean across both regions on the ik -th animal is given by

$$\begin{aligned} m_{ik} &= \frac{y_{i1j} + y_{i2j}}{2} = [\mu + \alpha_i + f_{ik} + \beta_1 + (\alpha\beta)_{i1} + e_{i1k} \\ &\quad + \mu + \alpha_i + f_{ik} + \beta_2 + (\alpha\beta)_{i2} + e_{i2k}] / 2 \end{aligned}$$

The expression of the mean in terms of the model effects cannot be simplified, because no terms cancel out. In particular, the animal effects do not cancel out. Thus, the error term for the t-test of marginal group means, which is computed from the animal means m_{ik} , depends on animal effects, f_{ik} , as well as on residual error effects, e_{ijk} , i.e., it depends on both between-animal and within-animal variation.

7.4.9 Two-way ANOVA

The ANOVA can be based on the following model-building sequence:

Model	SS_{error}	Reduction in error SS (RSS) compared to preceding model
$y_{ijk} = \mu + e_{ijk}$	$SS(\mu)$	
$y_{ijk} = \mu + \alpha_i + e_{ijk}$	$SS(\alpha_i, \mu)$	$RSS(\alpha_i \mu)$
$y_{ijk} = \mu + \alpha_i + f_{ik} + e_{ijk}$	$SS(f_{ik}, \alpha_i, \mu)$	$RSS(f_{ik} \alpha_i, \mu)$
$y_{ijk} = \mu + \alpha_i + f_{ik} + \beta_j + e_{ijk}$	$SS(\beta_j, f_{ik}, \alpha_i, \mu)$	$RSS(\beta_j f_{ik}, \alpha_i, \mu)$
$y_{ijk} = \mu + \alpha_i + f_{ik} + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$	$SS[(\alpha\beta)_{ij}, \beta_j, f_{ik}, \alpha_i, \mu]$	$RSS[(\alpha\beta)_{ij} \beta_j, f_{ik}, \alpha_i, \mu]$

Reductions are computed as usual, e.g.,

$$RSS(\beta_j|f_{ik}, \alpha_i, \mu) = SS(f_{ik}, \alpha_i, \mu) - SS(\beta_j, f_{ik}, \alpha_i, \mu)$$

The reduction due to addition of a term appears as the SS of that term in the ANOVA table.

Source	$d.f.$	SS
Groups (main effect α_i)	($a-1$)	$RSS(\alpha_i \mu)$
Animals within groups (f_{ik})	($r-a$)	$RSS(f_{ik} \alpha_i, \mu)$
Regions (main effect β_j)	($b-1$)	$RSS(\beta_j f_{ik}, \alpha_i, \mu)$
Group \times region interaction $[(\alpha\beta)_{ij}]$	($a-1)(b-1$)	$RSS[(\alpha\beta)_{ij} \beta_j, f_{ik}, \alpha_i, \mu]$
Error (e_{ijk})	($N-r-ab+a$)	$SS[(\alpha\beta)_{ij}, \beta_j, f_{ik}, \alpha_i, \mu]$

a = number of groups

b = number of regions

r = total number of animals

N = total number of observations

Table 7.6: Expected mean squares in the split-plot ANOVA, assuming balanced data.

Source	d.f.	Expected $MS = E(MS)$	H_0 tested by MS	Expected MS under H_0
Groups	$(a-1)$	$\sigma^2 + b\sigma_f^2 + \frac{bn}{a-1} \sum_{i=1}^a (\alpha_i - \bar{\alpha}_\bullet)^2$ (*)	$\alpha_1 = \alpha_2 = \dots$	$\sigma^2 + b\sigma_f^2$
Animals	$a(n-1)$	$\sigma^2 + b\sigma_f^2$	None	$\sigma^2 + b\sigma_f^2$
Regions	$(b-1)$	$\sigma^2 + \frac{an}{(b-1)} \sum_{j=1}^b (\beta_j - \bar{\beta}_\bullet)^2$ (*)	$\beta_1 = \beta_2 = \dots$	σ^2
Interaction	$(a-1)(b-1)$	$\sigma^2 + \frac{n}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b [(\alpha\beta)_{ij} - (\bar{\alpha}\beta)_{i\bullet} - (\bar{\alpha}\beta)_{\bullet j} + (\bar{\alpha}\beta)_{\bullet\bullet}]^2$	$(\alpha\beta)_{11} = (\alpha\beta)_{12} = \dots$	σ^2
Error	$a(b-1)(n-1)$	σ^2	None	σ^2

(*) Assuming there is no interaction!

n = number of animals per group

Division of the SS of a term by its d.f. yields the mean square (MS). The d.f. given above assume that all factorial combinations (group \times region) were observed. The model is a mixed model, so it is not appropriate to test all terms against the error term. To obtain the correct F-tests, we need to look at the expected MS . These have a simple form only for balanced data (same number of animals per group, no missing observations). For unbalanced data, things are more complicated. The unbalanced case will be dealt with briefly in section 7.4.11 and 7.4.12. For balanced data the expected MS are as given in Table 7.6. The expected mean squares show that the main effects for groups needs to be tested against the MS for animals, while both interaction and main effect for regions are tested against the residual MS . This set up is basically the same as in the analysis of a split-plot experiment with main plots arranged according to a completely randomized design. In section 6.4, the split-plot design for main plots arranged in complete blocks was discussed, for which analysis very similar to the one presented here. The analogy with field experiments is as follows:

main plots = animals
 subplots = regions within animals

The only conceptual difference lies in the fact that regions (subplots) cannot be randomized within animals (main plots). This fact becomes particularly important with more than two levels of the subplot factor, when specialized models are needed to account for the lack of randomization. For only two levels of the subplot factor, we may always use the split-plot model. In fact, the split-plot model is often quite appropriate with more than two subplots, though alternatives are available (Verbeke and Molenberghs, 1997).

SAS hints

```
proc glm;
  class region animal group;
  model avcount=group animal*group region region*group;
  random animal*group/test;
  run;

proc mixed nobound;
  class region animal group;
  model avcount=region group region*group/ddfm=satterth;
  random animal*group;
  run;
```

GLM and MIXED yield the same ANOVA, except that MIXED does not provide an F-test for animals, because the corresponding effect is random (details of the reasons for this difference in output cannot be explained here; see Appendix B).

Output (GLM):

Dependent Variable: avcount

Source	DF	Type III SS	Mean Square	F Value	Pr > F
* group	1	0.944075	0.944075	10.59	0.0087
Error	10	0.891603	0.089160		

Error: MS(animal*group)

* This test assumes one or more other fixed effects are zero.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
animal*group	10	0.891603	0.089160	2.35	0.0973
* region	1	0.236017	0.236017	6.21	0.0319
region*group	1	0.001558	0.001558	0.04	0.8436
Error: MS(Error)	10	0.379929	0.037993		

* This test assumes one or more other fixed effects are zero.

The F-test for interaction is not significant. Note that the p-value of the F-statistic is the same as that of the t-test for interaction in section 7.4.5, and that $F_{exp} = 0.04 = (t_{exp})^2 = (0.20)^2$. Because of the non-significant interaction, we may test main effects. The p-value for the F-test group means is the same as that for the corresponding t-test in section 7.4.6, and it is easily checked that $F_{exp} = 10.59 = (t_{exp})^2 = (3.25)^2$. The F-test for marginal region means differs slightly from the t-test in section 7.4.7, though both are significant ($p = 0.243$ for t-test and $p = 0.0319$ for F-test). Also, the t-test had 11 error d.f., while the F-test has 10 d.f. The same result is obtained for the F-test and the t-test in section 7.4.7 if we drop the interaction term from the model. There is justification in doing so, because the interaction was not significant. Also note, that for the t-tests of marginal means in sections 7.4.6 and 7.4.7, we assumed absence of interaction. The ANOVA for the reduced model is obtained as follows:

Dependent Variable: avcount

Source	DF	Type III SS	Mean Square	F Value	Pr > F
group	1	0.944075	0.944075	10.59	0.0087
Error	10	0.891603	0.089160		
Error: MS(animal*group)					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
animal*group	10	0.891603	0.089160	2.57	0.0685
region	1	0.236017	0.236017	6.81	0.0243
Error: MS(Error)	11	0.381487	0.034681		

Now, the F-test for regions has 11 error d.f., and the p-value is the same as for the t-test in section 7.4.7. Also, $F_{exp} = 6.81 = (t_{exp})^2 = (2.609)^2$. Note that the F-test for groups has remained unchanged compared to the analysis based on the model with interaction.

Output (MIXED):

Covariance Parameter Estimates

Cov Parm	Estimate
animal*group	0.02558 $\longrightarrow \sigma_f^2$
Residual	0.03799 $\longrightarrow \sigma^2$

The output provides estimates of the variance components. The estimates are obtained by the so-called Restricted Maximum Likelihood (REML) method (cannot be explained here). The between-animal component (σ_f^2) is slightly smaller than the within-animal component (σ^2).

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
region	1	10	6.21	0.0319
group	1	10	10.59	0.0087
region*group	1	10	0.04	0.8436

The F-tests are the same as those obtained with GLM. No test for the ANIMAL*GROUP effect is obtained, because this effect is random, and a variance component is estimated in place of performing an F-test. The ANOVA for the reduced model is as follows:

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
region	1	11	6.81	0.0243
group	1	10	10.59	0.0087

Again, the result for GROUP and REGION are identical to that obtained with GLM.

7.4.10 Mean comparisons (balanced data)

Mean comparisons for balanced data work essentially the same as for the split-plot design with main plots arranged in complete blocks (section 6.4.3). The details are repeated here for clarity (Table 7.7).

Table 7.7: Standard errors of a difference (*s.e.d.*) and associated error d.f. for balanced split-plot design with main-plot factor A (groups) and sub-plot factor B (regions).

Comparisons	Standard error of a difference (<i>s.e.d.</i>)	Error d.f.
Marginal group means (A)	$s_a \sqrt{\frac{2}{nb}}$	$a(n-1)$
Marginal region means (B)	$s_b \sqrt{\frac{2}{na}}$	$a(b-1)(n-1)$
Cell means at same level of group (A)	$s_b \sqrt{\frac{2}{n}}$	$a(b-1)(n-1)$
Cell means at same level of region (B)	$\sqrt{\frac{2[(b-1)s_b^2 + s_a^2]}{nb}}$	$df_{satterth}$ (*)

a (b) = no. of levels of factor A (B); n = no. of animals = subjects = "main plots" per group; s_a^2 = main-plot error MS (animals within groups); s_b^2 = sub-plot error MS (residual); (*) see text.

The *s.e.d.* appropriate for mean comparisons depends in the type of comparison. Table 7.6 shows the *s.e.d.*'s and associated error d.f. for the balanced case. Note that in the balanced case cell means are identical to least squares means as for other balanced designs (Section 6.3). For either comparison, the least significant difference, *LSD*, is computed as

$$LSD = t_{tab} \times s.e.d.$$

in the usual way. The *s.e.d.* for comparing AB means at the same level of B are computed from a linear combination of ANOVA MS : s_a^2 and s_b^2 . Consequently, the error d.f. are a weighted mean of the d.f. associated with these two MS . According to the **Satterthwaite method** the d.f. associated with a linear combination of mean squares

$$c_1 MS_1 + c_2 MS_2 + \dots$$

is given by

$$df_{satterth} = \frac{(c_1 MS_1 + c_2 MS_2 + \dots)^2}{\frac{(c_1 MS_1)^2}{df_1} + \frac{(c_2 MS_2)^2}{df_2} + \dots}$$

where c_i are the coefficients of MS_i and df_i are the error d.f. associated with mean squares MS_i . For the comparison of AB means at the same level of B we find the following d.f.:

$$df_{satterth} = \frac{\left(\frac{2s_a^2}{nb} + \frac{2(b-1)s_b^2}{nb} \right)^2}{\frac{\left(2s_a^2 \right)^2}{nb} + \frac{\left(2(b-1)s_b^2 \right)^2}{nb}} \cdot \frac{a(n-1)}{a(b-1)(n-1)}$$

SAS hints

Warning: PROC GLM does not compute appropriate *s.e.d.*! Thus, PROC MIXED should be used to do all computations. For balanced data, the NOBOUND option needs to be used to produce the same F-tests as with GLM (The NOBOUND option allows negative variance component estimates. Details cannot be explained here).

In the present case, interactions were not significant, so we drop the interaction and just compare marginal means:

```
proc mixed nobound;
class region animal group;
model avcount=region group/noint solution ddfm=satterth;
random animal*group;
lsmeans group/pdiff;
lsmeans region/pdiff;
run;
```

Output:

Type 3 Tests of Fixed Effects

Effect	Num	Den	F Value	Pr > F
	DF	DF		
region	1	11	6.81	0.0243
group	1	10	10.59	0.0087

Least Squares Means

Effect	region	group	Standard			Pr > t
			Estimate	Error	DF	
group		1	0.9472	0.08620	10	10.99 <.0001
group		3	0.5506	0.08620	10	6.39 <.0001
region	1		0.6497	0.07183	17	9.04 <.0001
region	2		0.8481	0.07183	17	11.81 <.0001

Differences of Least Squares Means

Effect	region	group	_region	_group	Standard			Pr > t
					Estimate	Error	DF	
group		1		3	0.3967	0.1219	10	3.25 0.0087
region	1		2		-0.1983	0.07603	11	-2.61 0.0243

Group and region marginal means are significantly different, which is the same result as the one obtained before. For illustration, we also look at the code for comparing cell means (not actually needed here because interaction is not significant):

```
proc mixed nobound;
class region animal group;
model avcount=region group region*group/ddfmsatterth;
random animal*group;
lsmeans group*region/pdiff;
run;
```

Output:

Effect	region	group	Least Squares Means					Pr > t
			Estimate	Standard Error	DF	t Value		
region*group	1	1	0.8561	0.1029	17.2	8.32	<.0001	
region*group	1	3	0.4433	0.1029	17.2	4.31	0.0005	
region*group	2	1	1.0383	0.1029	17.2	10.09	<.0001	
region*group	2	3	0.6578	0.1029	17.2	6.39	<.0001	

Effect	region	group	Differences of Least Squares Means					Pr > t
			_region	_group	Estimate	Standard Error	DF	
region*group	1	1	1	3	0.4128	0.1456	17.2	0.0113
region*group	1	1	2	1	-0.1822	0.1125	10	-0.1365
region*group	1	1	2	3	0.1983	0.1456	17.2	0.1906
region*group	1	3	2	1	-0.5950	0.1456	17.2	0.0008
region*group	1	3	2	3	-0.2144	0.1125	10	0.0858
region*group	2	1	2	3	0.3806	0.1456	17.2	0.0180

The main advantage of cell mean comparisons based on error *MS* from a split-plot ANOVA compared to the simple t-tests presented in sections 7.4.1 and 7.4.2 is that all data are used to estimate the error term. Thus, the t-tests for comparing groups have more d.f. (17.2 Satterthwaite d.f., compared to 10 d.f. for the simple t-tests), so there is a potential gain in power to detect group differences for the different regions. These comparisons are relevant only when there is interaction, so that differences are expected to differ among regions. Here, interaction is not significant, so comparisons can serve for demonstration purposes only.

7.4.11 The t-test for interaction revisited

The t-test for interaction in 7.4.5 was based on a difference of differences in means and thus involved four means. The difference of differences is equivalent to a linear contrast involving four means:

\bar{y}_{11} = mean for group 1 (control), proximal region = 0.856

\bar{y}_{12} = mean for group 1 (control), distal region = 1.038

\bar{y}_{21} = mean for group 3 (RS), proximal region = 0.443

\bar{y}_{22} = mean for group 3 (RS), distal region = 0.658

The difference among groups for the proximal region is

$$\bar{y}_{11} - \bar{y}_{21} = 0.856 - 0.443 = 0.413$$

The difference among groups for the distal region is

$$\bar{y}_{12} - \bar{y}_{22} = 1.038 - 0.658 = 0.380$$

The difference of differences is

$$L = \bar{y}_{11} - \bar{y}_{21} - (\bar{y}_{12} - \bar{y}_{22}) = 0.413 - 0.380 = 0.033$$

The difference of difference is seen to be a contrast of the form

$$L = c_{11}\bar{y}_{11} - c_{12}\bar{y}_{12} - c_{21}\bar{y}_{21} + c_{22}\bar{y}_{22}$$

with

$$c_{11} = 1$$

$$c_{12} = -1$$

$$c_{21} = -1$$

$$c_{22} = 1$$

SAS hints

```
proc mixed;
  class region animal group;
  model avcount=region group region*group;
  random animal*group;
  estimate 'interaction' region*group 1 -1 -1 1;
run;
```

Output:

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
region	1	10	6.21	0.0319
group	1	10	10.59	0.0087
region*group	1	10	0.04	0.8436

Estimates					
Label	Estimate	Standard	DF	t Value	Pr > t
		Error			
interaction	0.03223	0.1591	10	0.20	0.8436

Note that the p-value for the interaction contrast is the same as that of the F-test in 7.4.9, and that the t-value is the same as that computed in section 7.4.5.

7.4.12 Analysis for three groups (unbalanced)

So far the analysis comprised two groups. The complete experiment was actually performed with three groups (**colon2.dat**). The second group was fed with the same resistant starch (RS) as group 3, but as opposed to group 3, the animals in group 2 went through a starving period before the feeding of RS started. The three groups may be denoted as follows:

Group	Treatment	Number of animals
1	Control	6
2	RS+ (with starving period)	5
3	RS- (without starving period)	6

The ANOVA is basically the same as for the two-group case. Note that now it is no longer possible to perform a complete analysis solely based on t-test. The data are unbalanced, because group 2 comprises only 5 animals. Thus, mean comparisons cannot be done using the *s.e.d.* provided in section 7.4.10. More general methods need to be used to compute means, *s.e.d.*, F-statistics, t-statistics, and degrees of freedom. Technical details will not be considered here. We only mention that the Kenward-Roger method, an extension of the Satterthwaite method, is the best currently available method to compute error d.f. as well as *s.e.d.*. It is available in MIXED with the DDFM = KR option.

SAS hints

```

data;
input
group region animal avcount;
datalines;
  1      1      1      0.93667
  1      1      2      1.18000
  1      1      3      0.85667
  1      1      4      0.77667
  1      1      5      0.49667
  1      1      6      0.89000
  1      2      1      1.34000
  1      2      2      1.17000
  1      2      3      1.26667
  1      2      4      0.78000

```

```

1      2      5      0.71000
1      2      6      0.96333
2      1      18     0.43000
2      1      19     0.66667
2      1      20     0.31667
2      1      21     0.53000
2      1      22     0.37333
2      2      18     0.36667
2      2      19     0.19667
2      2      20     0.53333
2      2      21     0.50333
2      2      22     0.66667
3      1      12     0.37000
3      1      13     0.54000
3      1      14     0.36333
3      1      15     0.54333
3      1      16     0.40000
3      1      17     0.44333
3      2      12     0.36333
3      2      13     0.30667
3      2      14     0.46000
3      2      15     1.13000
3      2      16     0.61000
3      2      17     1.07667
;

proc mixed;
class group region animal;
model avcount=group region group*region/ddf=kr;
random animal(group);
run;

```

Output:

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
group	2	14	11.76	0.0010
region	1	14	3.51	0.0820
group*region	2	14	0.98	0.4001

Interactions are not significant, as before, but the significance of region means has vanished. Thus, the significance for regions detected for the analysis of two groups may be a Type I error. Nevertheless, the p-value is not so far from the conventional 5% level, so there is some evidence of differences between regions. Marginal means (weighted least squares) are computed and compared as follows:

```

proc mixed;
class group region animal;
model avcount=group region/ddfm=kr;
random animal(group);
lsmeans group region/pdiff;
run;

```

Output:

Least Squares Means							
Effect	group	region	Estimate	Standard Error	DF	t Value	Pr > t
group	1		0.9472	0.07387	14	12.82	<.0001
group	2		0.4583	0.08093	14	5.66	<.0001
group	3		0.5506	0.07387	14	7.45	<.0001
region		1	0.5835	0.05578	27.3	10.46	<.0001
region		2	0.7206	0.05578	27.3	12.92	<.0001

Differences of Least Squares Means									
Effect	group	region	_group	_region	Estimate	Standard Error	DF	t Value	Pr > t
group	1		2		0.4889	0.1096	14	4.46	0.0005
group	1		3		0.3967	0.1045	14	3.80	0.0020
group	2		3		-0.09222	0.1096	14	-0.84	0.4141
region		1		2	-0.1371	0.06845	16	-2.00	0.0625

Groups 2 and 3 are not significantly different, so starving had no effect on cell counts. Both RS treatments are significantly different from the control. Note that the *s.e.d.* is not constant, so a common *LSD* cannot be computed. Regions are not significantly different at the 5% level, though nearly so.

7.4.13 Analysis of replicate data (unbalanced)

The mean cell count for a region is based on three different samples. Actually, for some animals, only one or two segments could be analysed per region (either the segment did not have any crypts or the crypts could not be satisfactorily visualized). The replicate data are given below (**colon3.dat**). The first six observations are for animals no. 1, 2 and 3. For animal 1, there are only two samples, and for animal 2 only one sample is available. For animal 3, three samples are recorded. This exemplifies the unbalancedness of the data.

slide group region grureg animal avcount						
632	1	2	11	1	1.57	}
633	1	2	11	1	1.11	}
644	1	2	11	2	1.17	→
655	1	2	11	3	0.89	}
656	1	2	11	3	1	
657	1	2	11	3	1.91	

just two samples

only one sample

three samples, as planned

667	1	2	11	4	0.38
668	1	2	11	4	1.1
669	1	2	11	4	0.86
679	1	2	11	5	0.67
680	1	2	11	5	0.6
681	1	2	11	5	0.86
703	1	2	11	6	0.7
704	1	2	11	6	0.6
705	1	2	11	6	1.59
634	1	1	12	1	1
635	1	1	12	1	1.31
636	1	1	12	1	0.5
646	1	1	12	2	1.32
647	1	1	12	2	0.75
648	1	1	12	2	1.47
658	1	1	12	3	1.07
659	1	1	12	3	0.5
660	1	1	12	3	1
670	1	1	12	4	1
671	1	1	12	4	0.33
672	1	1	12	4	1
682	1	1	12	5	0
683	1	1	12	5	0.82
684	1	1	12	5	0.67
706	1	1	12	6	0.67
707	1	1	12	6	0.6
708	1	1	12	6	1.4
607	2	2	41	18	0.42
608	2	2	41	18	0.43
609	2	2	41	18	0.25
619	2	2	41	19	0.46
620	2	2	41	19	0
621	2	2	41	19	0.13
437	2	2	41	20	0.6
438	2	2	41	20	0.29
439	2	2	41	20	0.71
449	2	2	41	21	0.58
450	2	2	41	21	0.5
451	2	2	41	21	0.43
461	2	2	41	22	1
462	2	2	41	22	0.5
463	2	2	41	22	0.5
610	2	1	42	18	0.5
611	2	1	42	18	0.5
612	2	1	42	18	0.29
622	2	1	42	19	1
623	2	1	42	19	0.5
624	2	1	42	19	0.5
440	2	1	42	20	0.5
441	2	1	42	20	0.36
442	2	1	42	20	0.09
452	2	1	42	21	0.5
453	2	1	42	21	0.52
454	2	1	42	21	0.57
464	2	1	42	22	0.25
465	2	1	42	22	0.67

```

466  2    1    42    22   0.2
571  3    2    51    12   0.71
572  3    2    51    12   0.25
573  3    2    51    12   0.13
583  3    2    51    13   0.42
584  3    2    51    13   0.33
585  3    2    51    13   0.17
595  3    2    51    14   0.56
596  3    2    51    14   0.44
597  3    2    51    14   0.38
401  3    2    51    15   1.56
403  3    2    51    15   0.7
413  3    2    51    16   0.83
414  3    2    51    16   0.67
415  3    2    51    16   0.33
425  3    2    51    17   0.83
426  3    2    51    17   0.9
427  3    2    51    17   1.5
575  3    1    52    12   0.55
576  3    1    52    12   0.19
587  3    1    52    13   0.33
588  3    1    52    13   0.75
598  3    1    52    14   0.67
599  3    1    52    14   0.22
600  3    1    52    14   0.2
404  3    1    52    15   0.6
405  3    1    52    15   0.43
406  3    1    52    15   0.6
416  3    1    52    16   0.2
417  3    1    52    16   1
418  3    1    52    16   0
428  3    1    52    17   0.4
429  3    1    52    17   0.43
430  3    1    52    17   0.5
;

```

slide = running number of microscope slide

grureg = crossed variable between group and region (not needed here)

The problem with unbalancedness here is that the accuracy of a mean depends on the number of samples. A mean cell count of three samples is three times more accurate than a cell count from a single sample. To account for this, a weighted analysis is in order, which gives means based on three samples a higher weight than means based on only two samples. The weighted analysis is automatically performed when a suitable mixed model is formulated in MIXED for the replicate data (see Appendix B). The appropriate model is

$$y_{ijkh} = \mu + \alpha_i + f_{ik} + \beta_j + (\alpha\beta)_{ij} + e_{ijk} + g_{ijkh}$$

where

y_{ijkh} = count of h -th sample at j -th region on k -th animal within i -th group

μ = general effect

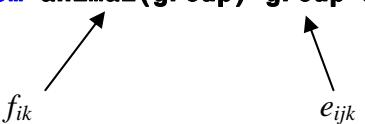
α_i = main effect of i -th group

- β_j = main effect of j -th region
 $(\alpha\beta)_{ij}$ = group \times region interaction
 f_{ik} = random effect of ik -th animal (between-animal effect)
 e_{ijk} = random residual effect of j -th region within ik -th animal (within-animal effect)
 g_{ijkh} = residual error (sampling error)

SAS hints

```

proc mixed data=t;
  class group region animal;
  model avcount=group region group*region/ddf=kr;
  random animal(group) group*animal*region;
run;
  
```



Output:

Covariance Parameter Estimates

Cov Parm	Estimate
animal(group)	0.01152 $\longrightarrow \sigma_f^2$
group*region*animal	0.004166 $\longrightarrow \sigma_e^2$
Residual	0.1005 $\longrightarrow \sigma_g^2$

Type 3 Tests of Fixed Effects

Effect	Num	Den	F Value	Pr > F
	DF	DF		
group	2	13	11.82	0.0012
region	1	13.2	2.91	0.1115
group*region	2	13.2	0.86	0.4439

Interaction and the main effects for region are not significant. This analysis is the most efficient analysis. Thus, I would conclude that there is only weak evidence of differences among regions (proximal, distal), while differences among groups are highly significant. The weighted least squares means for groups are:

```

proc mixed data=t;
  class group region animal;
  model avcount=group/ddf=kr;
  random animal(group) group*animal*region;
  lsmeans group/pdiff;
run;
  
```

Output:

Least Squares Means							
Effect	group	Standard					
		Estimate	Error	DF	t Value	Pr > t	
group	1	0.9281	0.07029	13.6	13.20	<.0001	
group	2	0.4583	0.07434	12.3	6.17	<.0001	
group	3	0.5404	0.07001	13.7	7.72	<.0001	

Differences of Least Squares Means							
Effect	group	group	Standard				
			Estimate	Error	DF	t Value	Pr > t
group	1	2	0.4697	0.1023	12.9	4.59	0.0005
group	1	3	0.3877	0.09920	13.7	3.91	0.0016
group	2	3	-0.08204	0.1021	12.9	-0.80	0.4363

7.4.14 Summary

We have shown how to analyse a two-factor experiment with repeated measures using different approaches. Simple unpaired and paired t-tests were the point of departure. Essentially the same conclusions were obtained by a full-fledged mixed model analysis, which is expected to be most efficient and which can accommodate unbalanced data. If a mixed model package is available, the analysis is straightforward, provided the mixed model is correctly specified.

Exercise 7.4: Reproduce the different analyses for the colon data. The data are stored in three different datasets:

colon.dat : two groups, mean counts per region
colon2.dat : three groups, mean counts per region
colon3.dat : three groups, mean counts per sample (up to three samples per region)

Exercise 7.5: The effect of lecithin against the Alzheimer disease was tested on two groups of patients (Hand und Crowder, 1996, p. 44 and p. 176). The first group received a placebo, while the lecithin medication was administered to the second group. Each person was read a number of words they were asked to memorize. The recorded response variable is the number of memorized words on a given occasion. The investigation was repeated five times with the same persons. The dates were codes as 0, 1, 2, 4 and 6 (**alzheimer.dat**). Fit a two-factorial model with factors group and medication. Use a split-plot model to allow for correlation of observations on the same person. Test if there are significant differences among the time profiles. Is there a significant effect of lecithin?

Group	0	1	2	4	6
1	20	19	20	20	18
1	14	15	16	9	6
1	7	5	8	8	5
1	6	10	9	10	10
1	9	7	9	4	6
1	9	10	9	11	11
1	7	3	7	6	3
1	18	20	20	23	21
1	6	10	10	13	14
1	10	15	15	15	14
1	5	9	7	3	12
1	11	11	8	10	9
1	10	2	9	3	2
1	17	12	14	15	13
1	16	15	13	7	9
1	7	10	4	10	5
1	5	0	5	0	0
1	16	7	7	6	10
1	5	6	9	5	6
1	2	1	1	2	2
1	7	11	7	5	11
1	9	16	17	10	6
1	2	5	6	7	6
1	7	3	5	5	5
1	19	13	19	17	17
1	7	5	8	8	6

Group	0	1	2	4	6
2	9	11	14	11	14
2	6	7	9	12	16
2	13	18	14	20	14
2	9	10	9	8	7
2	6	7	4	5	4
2	11	11	5	10	12
2	7	10	11	8	5
2	8	18	19	15	14
2	3	3	3	1	3
2	4	10	9	17	10
2	11	10	5	15	16
2	1	3	2	2	5
2	6	7	7	6	7
2	0	3	2	0	0
2	18	19	15	17	20
2	15	15	15	14	12
2	14	11	8	10	8
2	6	6	5	5	8
2	10	10	6	10	9
2	4	6	6	4	2
2	4	13	9	8	7
2	14	7	8	10	6

Number of memorized words

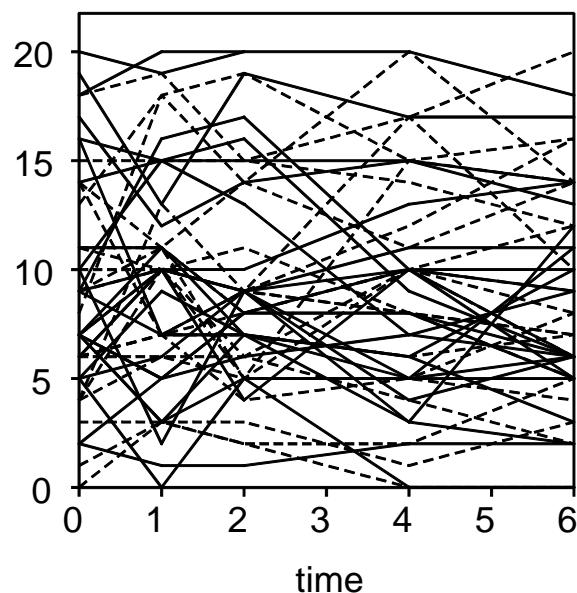


Fig. 7.2: Number of memorized words per person and date. Two groups: dotted = Lecithin; solid = Placebo). Five points in time.

References

- Claus, R., Lösel, D., Lacorn, M., Mentschel, J., Schenkel, H. (2002). Effects of butyrate on apoptosis in the pig colon and its consequences for skatole formation and tissue accumulation. *submitted.*
- Potten, C. S. (1992). The significance of spontaneous and induced apoptosis in the gastrointestinal tract of mice. *Cancer & Metastasis Reviews*, 11(2), 179-95.
- Verbeke, G., Molenberghs, G. (eds.) (1997). Linear mixed models in practice. Springer, Berlin.

8. Sample size and power for elementary procedures

When planning an experiment, one important task is to determine the sample size. Generally, planning the sample size requires some prior information on the variance, as well as some choice of the precision required or desired. There are two types of inferential procedures: Confidence intervals and statistical tests. For a confidence interval, required precision is usually specified by the half width of the interval one is aiming to attain. For a significance test, the Type I and Type II error rates (or power) need to be specified, as well as the minimal effect size that is to be detected with pre-specified power. Recall that the Type I error rate, also denoted as α , is defined as the probability of falsely (erroneously) rejecting the null hypothesis, when in fact the null hypothesis is true, while the Type II error rate, also denoted as β , is the probability of falsely rejecting the alternative, when in fact the alternative hypothesis is true. The power of a test is just the complement of β , i.e. the power equals $1 - \beta$. The following two tables summarize the key facts related to significance tests.

Decision of test	Reality	
	H_0 true	H_0 false
H_0 rejected	Type I error	correct decision
H_0 accepted	correct decision	Type II error

Probabilities:

Decision of test	Reality	
	H_0 true	H_0 false
H_0 rejected	α	$1 - \beta$
H_0 accepted	$1 - \alpha$	β

This section shows how to do simple sample size and power calculations in SAS using PROC POWER.

8.1. Theoretical background

Let μ be the parameter of interest (e.g. a mean or a mean difference) and $\hat{\mu}$ its estimator (e.g., sample mean or sample mean difference).

8.1.1 Confidence interval

Let $s.e.(\hat{\mu})$ be the estimated standard error of the parameter of interest. For example, when estimating a population mean (in a population of infinite size), the standard error of the mean is

$$s.e.(\hat{\mu}) = \sqrt{\frac{\sigma^2}{n}},$$

where σ^2 is the population variance, which is estimated by the sample variance

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1},$$

and n is the sample size. For the determination of sample size it is important to note that the standard error is inversely proportional to the square-root of n . Let v be the degrees of freedom for error. For example, $v = n - 1$ for a single sample mean.

An $(1 - \alpha)$ confidence interval is computed as

$$\hat{\mu} \pm t_{v;1-\alpha/2} s.e.(\hat{\mu}), \quad (8.1)$$

where $t_{v;1-\alpha/2}$ is the critical percentage point of a t-distribution with v degrees of freedom at a probability of $(1 - \alpha/2)$. Note that the critical t also depends on sample size through the degrees of freedom v .

According to formula (8.1) the half width of the confidence interval is given by

$$HW = t_{v;1-\alpha/2} s.e.(\hat{\mu}).$$

When planning sample size for a confidence interval, we can pre-specify the desired precision in terms of the desired half width HW . The half width HW depends on sample size through both the critical t-value $t_{v;1-\alpha/2}$ as well as the standard error $s.e.(\hat{\mu})$. Thus, to determine sample size, we can set the specified value of HW equal to $t_{v;1-\alpha/2} s.e.(\hat{\mu})$ and then solve for the sample size n . There is no explicit solution for this task, but a numerical solution can be obtained, and this is implemented in PROC POWER.

8.1.2 t-test

Suppose the true population mean is μ and you want to test the null hypothesis $H_0 : \mu = \mu_0$ versus the alternative hypothesis $H_A : \mu \neq \mu_0$. The test-statistic of the t-test for this problem is

$$t_{\text{exp}} = \frac{\hat{\mu} - \mu_0}{s.e.(\hat{\mu})}.$$

The null hypothesis is rejected when

$$|t_{\text{exp}}| = \frac{|\hat{\mu} - \mu_0|}{s.e.(\hat{\mu})} > t_{v;1-\alpha/2}.$$

This can be converted to

$$\begin{aligned}\hat{\mu} &> K_U = \mu_0 + t_{v;1-\alpha/2} s.e.(\hat{\mu}) \text{ or} \\ \hat{\mu} &< K_L = \mu_0 - t_{v;1-\alpha/2} s.e.(\hat{\mu}),\end{aligned}$$

where K_L and K_U are, respectively, the lower and upper critical limits for $\hat{\mu}$. Thus, the null hypothesis H_0 is rejected when $\hat{\mu}$ falls outside the interval $[K_L; K_U]$.

Fig. 1 gives the distributions of the sample mean $\hat{\mu}$ under H_0 and under the alternative hypothesis $H_A : \mu \neq \mu_0$ for a particular $\mu > \mu_0$. The Type I error rate, i.e., the probability of erroneously rejecting H_0 , is denoted as α . The Type II error rate, i.e., the probability of failing to detect a true difference, is denoted as β . The power of the test is given by the complement of this probability, i.e.

$$\text{Power} = 1 - \beta$$

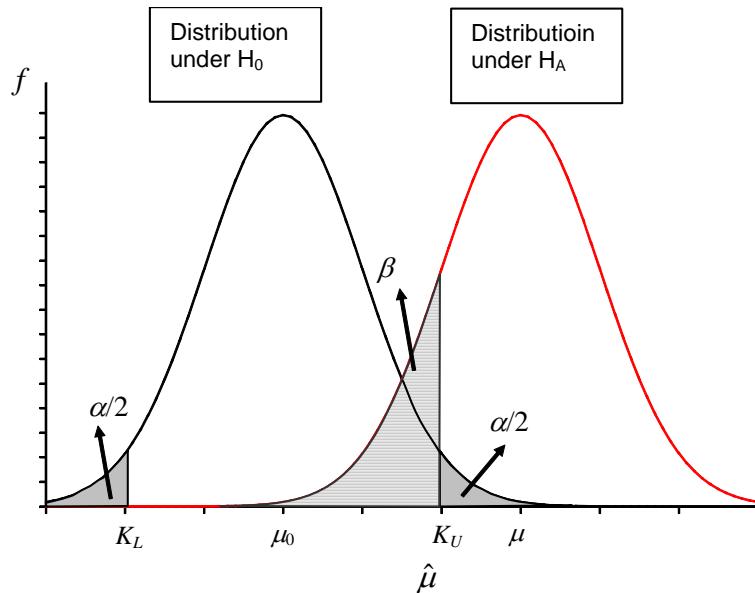


Fig.8.1: Two-sided t-Test - α - and β -error.

The key fact to observe here is that the critical limits K_L and K_U as well as the width of the distributions of $\hat{\mu}$ depend on sample size. This is depicted in Fig. 2, where the sample size is one forth of that in Fig. 1. The Type I error rate β is considerably reduced and hence the power is increased.

Also, it should be clear that when the difference of true mean and hypothesized mean, i.e. the effect size

$$\delta = \mu - \mu_0$$

gets larger in absolute value, the power will increase.

If we pre-specify the standard deviation σ , the two error rates α and β , and the effect size $\delta = \mu - \mu_0$, we can use the power curve to determine the necessary sample size. Again, there is no explicit formula for the exact solution, but a numerical solution can be found by the methods implemented in PROC POWER.

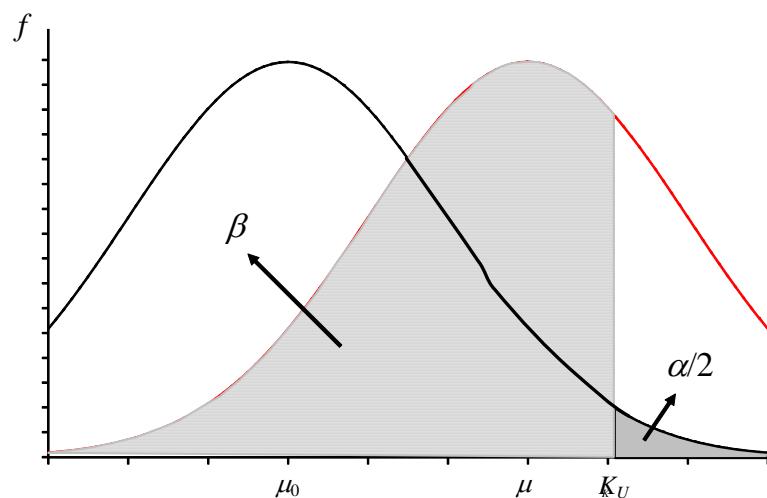


Fig.8.2: Two-sided t-Test - α - and β -error when sample size is one forth of that in Fig. 1.

In the following the implementation of sample size and power calculations based on the principles just outlined will be explained with a few examples for three elementary cases.

8.2 Single population mean

8.2.1 Confidence interval

Assume you want to estimate a population mean μ such that the half width of a 95% confidence interval is no more than 1.5 units. The standard deviation is estimated to be 4.5. You want to determine the sample size for this problem.

```
proc power;
onesamplemeans ci = t      /*compute confidence interval using t-distrib.*/
halfwidth = 1.5
stddev = 4.5
ntotal = .
probwidth = 0.95; /*1 - alpha*/
run;
```

Output:

The POWER Procedure

Confidence Interval for Mean

Fixed Scenario Elements

Distribution	Normal
Method	Exact
CI Half-Width	1.5
Standard Deviation	4.5
Nominal Prob(Width)	0.95
Number of Sides	2
Alpha	0.05
Prob Type	Conditional

Computed N Total

Actual	
Prob	N
(Width)	Total

0.957	50
-------	----

The required sample size is 50.

8.2.2 t-test

Suppose you want to detect a difference of a sample mean from a hypothetical population mean of $\mu_0 = 50$ at a significance level of $\alpha=5\%$ by a two-sided t-test. Prior to the experiment, the standard deviation is estimated to be $\sigma=7.5$. You want to determine the sample size n for this problem. You want to realize a power of 90% when the actual population mean is $\mu=52$. Note that the effect size for this problem is $\delta = 52 - 50 = 2$, and this is the relevant quantity as regards power. Thus, e.g., you would obtain the same answer, e.g., with means $\mu_0 = 15$ and $\mu = 17$.

```
proc power;
onesamplemeans test=t
  nullmean=50
  mean = 52
  stddev = 4.5
  ntotal = .
  power = 0.95;
run;
```

Output:

```
The POWER Procedure
One-sample t Test for Mean
```

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Null Mean	50
Mean	52
Standard Deviation	4.5
Nominal Power	0.95
Number of Sides	2
Alpha	0.05

Computed N Total	
Actual Power	N Total
0.951	68

The required sample size is 68.

Suppose that for the same problem your budget restricts the sample size to $n = 4$ and you want to know the power $(1 - \beta)$ of this design.

```
proc power;
onesamplemeans test=t
  nullmean=50
  mean = 52
  stddev = 4.5
  ntotal = 4
  power = .;
run;
```

Output:

```
The POWER Procedure
One-sample t Test for Mean

Fixed Scenario Elements

Distribution          Normal
Method               Exact
Null Mean            50
Mean                 52
Standard Deviation   4.5
Total Sample Size     4
Number of Sides       2
Alpha                0.05
```

```
Computed Power

Power
0.098
```

The power is only 9.8%, so this is not a sufficient sample size.

8.3 Two unpaired samples

8.3.1 t-test

Assume you want to compare two fertilizer treatments. The t-test is to detect a difference of $\delta = 0.3$ t/ha with a power of 80%. From previous experiments, the standard deviation is expected to be $\sigma = 0.2$ t/ha. What is the sample size to achieve the desired power?

```
proc power;
```

```

twosamplemeans test=diff
  meandiff = 0.3
  stddev = 0.2
  npergroup = .
  power = .80;
run;

```

Output:

```

The POWER Procedure
Two-sample t Test for Mean Difference

```

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Mean Difference	0.3
Standard Deviation	0.2
Nominal Power	0.8
Number of Sides	2
Null Difference	0
Alpha	0.05

Computed N Per Group

Actual Power	N Per Group
0.848	9

You conducted an experiment and afterwards you would like to assess the power of your experiment. Suppose the two treatment means were 5.6 and 6.1 t/ha and the standard deviation was 1.5 t/ha. The sample sizes were 3 and 4.

```

proc power;
twosamplemeans test=diff
  groupmeans = 5.6 | 6.1
  stddev = 1.5
  groupns = (3 4)
  power = .;
run;

```

Output:

```

The POWER Procedure
Two-sample t Test for Mean Difference

```

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Group 1 Mean	5.6
Group 2 Mean	6.1
Standard Deviation	1.5
Group 1 Sample Size	3
Group 2 Sample Size	4
Number of Sides	2
Null Difference	0
Alpha	0.05

```

Computed Power

Power

0.065

```

The power of this design was only 6.5%, so the experiment was poorly designed.

8.3.2 Confidence interval

Suppose you want the half width of a confidence interval for a mean difference in an unpaired sample to be no larger than $HW = 2.5$ units at a coverage probability of 95%. The standard deviation is $\sigma = 12$ units.

```

proc power;
twosamplemeans ci = diff
  halfwidth = 2.5
  stddev = 12
  npergroup = .
  probwidth = 0.95;
run;

```

Output:

```

The POWER Procedure
Confidence Interval for Mean Difference

Fixed Scenario Elements

Distribution          Normal
Method               Exact
CI Half-Width        2.5
Standard Deviation   12
Nominal Prob(Width)  0.95
Number of Sides       2
Alpha                0.05
Prob Type            Conditional

Computed N Per Group

Actual
Prob      N Per
(Width)    Group

0.954     200

```

The required sample size is 200 observations per group.

8.4. Two paired samples

There is a special option in the POWER procedure for the paired sample case, but this requires specification of the correlation in addition to the standard deviation. A simpler approach exploits the fact that paired differences can be analysed in the same way as a sample from a single population. In other words, the paired t-test is equivalent to a single-sample t-

test applied to paired differences. The same applies to confidence limits for the mean difference.

8.4.1 t-test

Assume you want to detect a difference of $\delta = 7.0$ units for two paired means. The standard deviation of paired differences is $\sigma = 5.8$ units.

```
proc power;
onesamplemeans test=t
  nullmean=0
  mean = 7
  stddev = 5.8
  ntotal = .
  power = 0.95;
run;
```

Output:

```
The POWER Procedure
One-sample t Test for Mean

Fixed Scenario Elements

Distribution          Normal
Method               Exact
Null Mean            0
Mean                 7
Standard Deviation   5.8
Nominal Power        0.95
Number of Sides      2
Alpha                0.05

Computed N Total

Actual      N
Power     Total

0.967      12
```

The required sample size is 12 pairs.

8.4.2 Confidence interval

Suppose you want to determine a confidence interval with 95% coverage probability and the standard deviation is expected to be $\sigma = 12$ units. The half width is to be no larger than $HW = 2.5$ units.

```
proc power;
onesamplemeans ci = t    /*compute confidence interval using t-distrib.*/
  halfwidth = 2.5
  stddev = 1.2
  ntotal = .
  probwidth = 0.95; /*1 - alpha*/
run;
```

Output:

```
The POWER Procedure
Confidence Interval for Mean

Fixed Scenario Elements

Distribution          Normal
Method               Exact
CI Half-Width        2.5
Standard Deviation   1.2
Nominal Prob(Width)  0.95
Number of Sides       2
Alpha                0.05
Prob Type            Conditional

Computed N Total

      Actual
      Prob      N
      (Width)    Total

      0.975      5
```

Exercises

1. You want to detect a difference of 2 units from the true population mean. The standard deviation is 1.2 units. What is the required sample size, when testing at a significance level of 5% with a desired power of 95%?
2. You are comparing two means in an unpaired design. Assuming that the standard deviation is 1.7 units, what is the power if the sample size in each group is $n = 5$ and the true means differ by 0.5 units?
3. You are running a paired samples design and want to make sure the half width of a 95% confidence interval will be no larger than 7.2 units. Assuming the expected standard deviation of paired differences is 35 units, what is the required number of pairs (sample size)?

Appendices

A. Some linear model theory

This section briefly gives a few key results for the general linear model (GLM), which are used by linear model procedures such as GLM of the SAS System. For brevity, these results will be given with little explanation. A full understanding of these results requires some familiarity with matrix algebra. If you don't have a background in matrix algebra, leafing through these pages will just give you a glimpse of what linear model packages do. The virtue of a matrix formulation of "the" linear model outlined in this Appendix is its complete generality, which allows one to analyse any linear model within a single framework based on the same principles. For simple settings (balanced one-way layout, linear regression), matrix expressions have simple scalar counterparts, some of which have been given occasionally in these lecture notes, but this is not generally true for more complex designs (e.g., for unbalanced data). The generality of the matrix formulation of the GLM is the key to the flexibility of PROC GLM and similar procedures. For details see, e.g., Searle, S.R., 1971, Linear models. Wiley, New York, and Searle, S. R., 1987, Linear models for unbalanced data. Wiley, New York. Also see "SAS/STAT User's Guide Version 8 (SAS Institute, Cary: NC).

(1) Any linear model can be written in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where

- \mathbf{y} = column vector of observations
- \mathbf{X} = design matrix
- $\boldsymbol{\beta}$ = parameter vector
- \mathbf{e} = error vector corresponding to \mathbf{y}

The errors are assumed to be independently identically distributed with variance σ^2 , i.e.,

$$\text{var}(\mathbf{y}) = \mathbf{I}\sigma^2$$

where \mathbf{I} is an identity matrix (square matrix with ones everywhere)

Example: Linear regression for crab data by

$$y_i = \alpha + \beta x_i + e_i$$

where y_i = premolt size of i -th crab, x_i = postmolt size of i -th crab.

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{pmatrix}$$

(2) The ordinary least squares solution (OLSE) for the parameter vector is obtained by

$$OLSE(\beta) = \hat{\beta} = (X'X)^{-1} X'y$$

where X' is the transpose of X and $(X'X)^{-1}$ is a generalized inverse (g-inverse) of $X'X$. Among others, a g-inverse of $(X'X)^{-1}$ has the following properties:

- (i) $X'X(X'X)^{-1}X'X = X'X$
- (ii) $X'X(X'X)^{-1}X' = X'$

(3) Fitted values for observations:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y$$

Example: Linear regression for crab data

$$X\hat{\beta} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \hat{\alpha} + \hat{\beta}x_1 \\ \hat{\alpha} + \hat{\beta}x_2 \\ \hat{\alpha} + \hat{\beta}x_3 \\ \cdot \\ \cdot \\ \hat{\alpha} + \hat{\beta}x_n \end{pmatrix}$$

(4) Residuals:

$$r = y - X\hat{\beta} = [I - X(X'X)^{-1}X']y$$

Note that $M = I - X(X'X)^{-1}X'$ is idempotent, i.e., $MM = M$.

(4) Error SS:

$$SS = r'r = y'MM'y = y'My = y[I - X(X'X)^{-1}X']y$$

(5) Estimate of error variance:

$$s^2 = \frac{SS}{df}$$

where

$$df = N - \text{rank}(X),$$

N is the number of observations in y and $\text{rank}(X)$ is the rank of the matrix X .

(6) The ordinary least squares estimator of an estimable linear function $\mathbf{k}'\boldsymbol{\beta}$ is given by

$$\mathbf{k}'\hat{\boldsymbol{\beta}}$$

The estimated standard error of this estimator is

$$s.e.(\mathbf{k}'\hat{\boldsymbol{\beta}}) = s\sqrt{\mathbf{k}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{k}}$$

where s is the square root of s^2 in (5).

Example: Least square means, as obtained by the LSMEANS statement of GLM, are of this form. A block design with four treatments and three blocks is analysed according to the model

$$y_{ij} = \mu + b_j + \alpha_i + e_{ij} ,$$

where b_j are block effects and α_i are treatment effects. The mean of, e.g., the 1st treatment is defined as

$$\eta_1 = \mu + \frac{b_1 + b_2 + b_3}{3} + \alpha_1 .$$

Plugging in least squares estimators for parameters, this is known as the least square mean for the 1st treatment. With

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

this is of the form $\eta_1 = \mathbf{k}'\boldsymbol{\beta}$, where

$$\mathbf{k}' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

The difference of the means of the first two treatments is

$$\begin{aligned} \eta_1 - \eta_2 &= \mu + \frac{b_1 + b_2 + b_3}{3} + \alpha_1 - \left(\mu + \frac{b_1 + b_2 + b_3}{3} + \alpha_2 \right) = \alpha_1 - \alpha_2 \\ \mathbf{k}' &= (0 \ 1 \ -1 \ 0 \ 0 \ 0 \ 0 \ 0) \end{aligned}$$

The least squares estimate of this difference appears in the column labeled "ESTIMATE" in the differences of least squares means (PDIFF option to the LSMEANS statement).

Example: The ESTIMATE statement of GLM also uses the results given above.

(7) t-test of $H_0: \mathbf{k}'\boldsymbol{\beta} = 0$:

$$t_{\text{exp}} = \frac{|\mathbf{k}'\hat{\boldsymbol{\beta}}|}{s.e.(\mathbf{k}'\hat{\boldsymbol{\beta}})}$$

This is compared against t_{tab} with $df = N - \text{rank}(\mathbf{X})$ degrees of freedom.

Example: p-values of pairwise comparisons among least squares means obtained by LSMEANS (GLM) are computed based on t_{exp} . The same is true of the t-test produced by the ESTIMATE statement.

(8) Confidence interval for $\mathbf{k}'\boldsymbol{\beta}$:

$$\mathbf{k}'\hat{\boldsymbol{\beta}} \pm t_{\text{tab}} s.e.(\mathbf{k}'\hat{\boldsymbol{\beta}})$$

Example: The CL option to the LSMEANS and ESTIMATE statements uses this result.

(9) This item and the one that follows have the major purpose to show that the procedures GLM and MIXED produce equivalent F-tests for the general linear model, specifically when Type I hypotheses are tested, which correspond to sequential SS (reductions in SS). The reader is advised that the two items are a bit lengthy and involve several matrix results not explained here.

An F-test based on sequential SS can be derived from the partitioned model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

Note that this is of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X} = (\mathbf{X}_1 \mathbf{X}_2)$ and $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$.

Example: One-way ANOVA

$$\beta_1 = \mu, \quad \beta_2 = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_I \end{pmatrix}$$

The reduction in error SS due to fitting β_2 , after having fitted β_1 is (Searle, 1987, p. 275)

$$RSS(\beta_2 | \beta_1) = \mathbf{y}' A_{12} \mathbf{y}$$

where

$$A_{12} = \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{M}_1$$

with

$$\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1$$

The matrix A_{12} is idempotent, i.e.,

$$A_{12} = A_{12} A_{12}$$

Example: $RSS(\beta_2 | \beta_1) = \mathbf{y}' A_{12} \mathbf{y}$ corresponds to Type I SS in PROC GLM (sequential SS).

Example: For simple linear regression (e.g., crab data), we have

$$\mathbf{X}_1 = \mathbf{1}_N \quad (\text{an } N\text{-vector of ones})$$

$$\mathbf{X}_2 = (x_1, x_2, \dots, x_N)'$$

$$\beta_1 = \alpha$$

$$\beta_2 = \beta$$

$$\mathbf{M}_1 = \mathbf{I} - \bar{\mathbf{J}}_N, \quad \text{where } \bar{\mathbf{J}}_N \text{ is a square matrix with elements } 1/N \text{ everywhere}$$

(Searle, 1987, p. 282),

$$\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2 = \sum_{i=1}^N (x_i - \bar{x})^2 = SS_x$$

$$\mathbf{y}' \mathbf{M}_1 \mathbf{X}_2 = \sum_{i=1}^N y_i (x_i - \bar{x}) = \sum_{i=1}^N (y_i - \bar{y})(x_i - \bar{x}) = CP_{xy}$$

$$RSS(\beta_2 | \beta_1) = RSS(\beta | \alpha) = \mathbf{y}' A_{12} \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{M}_1 = CP_{xy} SS_x^{-1} CP_{xy} = \hat{\beta}^2 SS_x$$

where

$$\hat{\beta} = \frac{CP_{xy}}{SS_x}$$

Thus, in this case the reduction in SS can be cast into a very simple algebraic form. Note that the reduction involves only $\hat{\beta}_2 = \hat{\beta}$.

The reduction is a quadratic form, which may be used to test

$$H_{01}: \mathbf{T}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

where \mathbf{T}' has full row rank and its rows are eigenvectors corresponding to the non-zero eigenvalues of \mathbf{A}_{12} , i.e.,

$$\mathbf{A}_{12} = \mathbf{T}\mathbf{T}' \text{ and } \mathbf{T}'\mathbf{T} = \mathbf{I}. \quad (\text{Searle, 1987, p. 235})$$

Due to the idempotency of \mathbf{A}_{12} , the null hypothesis is equivalent to

$$H_{01}: \mathbf{A}_{12}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

From $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$ (Searle, 1987, p. 224), and hence $\mathbf{A}_{12}\mathbf{X}_1 = \mathbf{0}$, this is seen to be equal to

$$H_{01}: \mathbf{A}_{12}\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbf{0}$$

Finally, since

$$\mathbf{A}_{12}\mathbf{X}_2 = \mathbf{M}_1\mathbf{X}_2$$

(Theorem 7.1 in Searle, 1987, p. 218), the hypothesis is equivalent to

$$H_{01}: \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbf{0}$$

This shows that the null hypothesis tested by $RSS(\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1)$ involves only $\boldsymbol{\beta}_2$.

Example: Simple linear regression

$$\mathbf{X}_1 = \mathbf{1}_N \quad (\text{an N-vector of ones})$$

$$\mathbf{X}_2 = (x_1, x_2, \dots, x_N)'$$

$$\boldsymbol{\beta}_1 = \alpha$$

$$\boldsymbol{\beta}_2 = \beta$$

$$H_{01}: \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbf{M}_1\mathbf{X}_2\beta = \mathbf{0}$$

For arbitrary \mathbf{X}_2 , this hypothesis can be true only if $\beta = 0$, i.e., a regression slope of zero.

Also, equivalently, we find:

$$A_{12} = \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \cdot \\ \cdot \\ x_N - \bar{x} \end{pmatrix} SS_x^{-1} (x_1 - \bar{x} \quad x_2 - \bar{x} \quad \cdot \quad \cdot \quad x_N - \bar{x})$$

$$\mathbf{T} = \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \cdot \\ \cdot \\ x_N - \bar{x} \end{pmatrix} SS_x^{-0.5}$$

$$\mathbf{T}'\mathbf{T} = 1$$

$$\mathbf{T}'\mathbf{X} = \begin{pmatrix} 0 & \frac{1}{\sqrt{SS_x}} \end{pmatrix}$$

Thus,

$$H_{01}: \mathbf{T}'\mathbf{X}\beta = \frac{\beta}{\sqrt{SS_x}} = 0 \Leftrightarrow H_{01}: \beta = 0$$

The F-statistic is given by

$$F_{exp} = \frac{RSS(\beta_2 | \beta_1)}{(r_{12} - r_1)s^2} \quad (\text{A1})$$

where r_{12} is the rank of $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2]$ and r_1 is the rank of \mathbf{X}_1 . This is compared against an F -distribution with $(r_{12} - r_1)$ numerator d.f. and $(N - r_{12})$ denominator d.f.

Example: The last F-test produced by PROC GLM under "Type I SS" is of this form. For simple linear regression, we find $r_{12} = 2$, $r_1 = 1$, and

$$F_{exp} = \frac{\hat{\beta}^2 SS_x}{s^2}$$

(10) The null hypothesis in (9) is seen to be of the form

$$H_{02}: \mathbf{K}'\beta = \mathbf{0}$$

with

$$\mathbf{K}' = \mathbf{T}'\mathbf{X}$$

where \mathbf{T}' has full row rank and its rows are eigenvectors corresponding to the non-zero eigenvalues of A_{12} (see item 9). Equivalently, the rows of \mathbf{K}' are eigenvectors corresponding to $\mathbf{X}'A_{12}\mathbf{X} = \mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2$ (compare SAS/STAT User's guide, p. 167).

Generally, a linear hypothesis of the form $\mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$ with any matrix \mathbf{K}' of full row rank and $\mathbf{K}'\boldsymbol{\beta}$ estimable can be tested by

$$F_{exp} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{K}\left[\mathbf{K}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}}}{s^2 \text{rank}(\mathbf{K})} , \quad (\text{A2})$$

which is compared against an F -distribution with $\text{rank}(\mathbf{K})$ numerator d.f. and denominator d.f. associated with s^2 (Searle, 1987, p. 291). The test-statistic in (A2) is sometimes referred to as Wald-type F-statistic.

Example: For linear models with a single error term, PROC MIXED computes Wald-type F-statistics according to (A2). When $\mathbf{K}' = \mathbf{T}'\mathbf{X}$, as corresponds to reductions in SS (Type I SS in GLM), the hypothesis is equivalent to a "Type I hypothesis" in MIXED. Type I hypotheses are invoked by the HTYPE = 1 option.

Choosing $\mathbf{K}' = \mathbf{T}'\mathbf{X}$, we find

$$\mathbf{K}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{K} = \mathbf{T}'\mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}'\mathbf{T} = \mathbf{T}'\mathbf{T}\mathbf{T}'\mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}'\mathbf{T}\mathbf{T}'\mathbf{T} = \mathbf{T}'\mathbf{A}_{12}\mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}_{12}\mathbf{T}$$

Using the g-inverse (Searle, 1987, p. 224)

$$(\mathbf{X}\mathbf{X})^{-1} = \left\{ \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -(\mathbf{X}_1'\mathbf{X}_1)^{-1}(\mathbf{X}_1'\mathbf{X}_2) \\ \mathbf{I} \end{bmatrix} (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1} \begin{bmatrix} -(\mathbf{X}_2'\mathbf{X}_1)(\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{I} \end{bmatrix} \right\} \quad (\text{A3})$$

where \mathbf{I} is an identity matrix, it can be shown (Searle, 1987, p. 225) that

$$\mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' + \mathbf{A}_{12} .$$

Also, since $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$ (Searle, 1987, p. 224), we have $\mathbf{A}_{12}\mathbf{X}_1 = \mathbf{0}$, and, noting that \mathbf{A}_{12} is idempotent,

$$\mathbf{K}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{K} = \mathbf{T}'\mathbf{A}_{12}\mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}_{12}\mathbf{T} = \mathbf{T}'\mathbf{A}_{12}\mathbf{T} = \mathbf{T}'\mathbf{T}\mathbf{T}'\mathbf{T} = \mathbf{I}$$

From this

$$\hat{\boldsymbol{\beta}}'\mathbf{K}\left[\mathbf{K}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{K}\mathbf{K}'\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{T}\mathbf{T}'\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{A}_{12}\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\mathbf{A}_{12}\mathbf{X}_2\hat{\boldsymbol{\beta}}_2$$

The last equality follows since $\mathbf{A}_{12}\mathbf{X}_1 = \mathbf{0}$. Now with the choice of g-inverse in (A3), we have the following solution for $\boldsymbol{\beta}_2$ (Searle, 1987, p.263):

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y} ,$$

so that

$$\hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\mathbf{A}_{12}\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{y}'\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{A}_{12}\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}$$

We can now use the fact that \mathbf{M}_1 is idempotent, i.e., $\mathbf{M}_1 = \mathbf{M}_1\mathbf{M}_1$, and hence,

$$\mathbf{A}_{12} = \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{M}_1 = \mathbf{M}_1 \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{M}_1 \mathbf{M}_1 = \mathbf{M}_1 \mathbf{A}_{12} \mathbf{M}_1$$

from which

$$\hat{\boldsymbol{\beta}}'_2 \mathbf{X}'_2 \mathbf{A}_{12} \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 = \mathbf{y}' \mathbf{A}_{12} \mathbf{A}_{12} \mathbf{y} = \mathbf{y}' \mathbf{A}_{12} \mathbf{y} = \text{RSS}(\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1)$$

This shows, that the F -statistics in (A1) and (A2) are equivalent, when $\mathbf{K}' = \mathbf{T}'\mathbf{X}$.

Example: F-tests based on Type I SS obtained by PROC GLM, which are computed according to (A1), yield the same results as Wald-type F-tests in PROC MIXED, computed according to (A2), when the HTYPE=1 option is used. The HTYPE=1 option invokes computation of \mathbf{K} so that essentially $\mathbf{K}' = \mathbf{T}'\mathbf{X}$ (SAS/STAT User's Guide, p. 167).

(11) Consider the partitioned model

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \mathbf{X}_3 \boldsymbol{\beta}_3 + \boldsymbol{\epsilon}$$

The reductions in SS from sequential fitting test the following hypotheses:

Reduction	Hypothesis [§] (Searle, 1987, p. 280)
-----------	--

$\text{RSS}(\boldsymbol{\beta}_3 \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$	$\mathbf{M}_1 \mathbf{X}_2 \boldsymbol{\beta}_2 + \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_3 \boldsymbol{\beta}_3$
$\text{RSS}(\boldsymbol{\beta}_2 \boldsymbol{\beta}_2, \boldsymbol{\beta}_1)$	$[\mathbf{I} - \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{M}_1] \mathbf{X}_3 \boldsymbol{\beta}_3$

The important point here is that the hypothesis tested by $\text{RSS}(\boldsymbol{\beta}_3 | \boldsymbol{\beta}_2, \boldsymbol{\beta}_1)$ involves only $\boldsymbol{\beta}_3$, while the hypothesis corresponding to $\text{RSS}(\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1)$ involves both $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_2$. Thus, to test a hypothesis about just one term (e.g., $\boldsymbol{\beta}_3$), that term needs to be **fitted last** in the sequence. Conversely, tests for **terms not fitted** last generally involve more terms than just the term in question and thus **do not provide adequate tests**, except in special circumstances (e.g., some balanced designs).

Example: A design with incomplete blocks (**unbalanced data**) can be analysed according to (Section 5.6.2)

$$y_{ij} = \mu + b_j + \alpha_i + e_{ij}$$

where

y_{ij} = measurement of j -th replicate of i -th treatment

μ = general term

α_i = effect of i -th treatment

b_j = effect of j -th block

We may set $\boldsymbol{\beta}_1 = \mu$, $\boldsymbol{\beta}_2 = (b_1, b_2, \dots)'$, and $\boldsymbol{\beta}_3 = (\alpha_1, \alpha_2, \dots)'$. Fitting the general term, μ , first, block effects, b_j , second, and the treatment effect, α_i , last, the reductions are as follows:

Reduction	Hypothesis [§] (Searle, 1987, p. 121)
$RSS(b_j \mu)$	$b_j + \sum_{i=1}^I n_{ij} \alpha_i / n_{i\bullet}$ equal for all j
$RSS(\alpha_i b_j, \mu)$	α_i equal for all i

§ $n_{ij} = 1$ when y_{ij} is observed; $n_{ij} = 0$ when y_{ij} is missing. $n_{i\bullet} = \sum_{j=1}^J n_{ij}$.

Thus, $RSS(\alpha_i | b_j, \mu)$ is adequate for testing the equality of treatment effects. However, the hypothesis tested by $RSS(b_j | \mu)$ involves not only block effects, but also a linear combination of treatment effects with weights depending on n_{ij} , which depend on both treatments and blocks. The crux here is that the term

$$\sum_{i=1}^I n_{ij} \alpha_i / n_{i\bullet}$$

generally depends on j . Thus, $RSS(b_j | \mu)$ is not generally adequate for testing the equality of block effects b_j . To obtain a test for equality of blocks, the order of fitting must be changed, i.e., blocks need to be fitted second and treatments last.

Note that for **balanced data**, $n_{ij} = 1$ for all i and j , and thus

$$H_0: \sum_{i=1}^I n_{ij} \alpha_i / n_{i\bullet} = \sum_{i=1}^I \alpha_i / I = \bar{\alpha}_\bullet$$

which does **not** depend on j . The hypothesis tested by $RSS(b_j | \mu)$ now becomes

$$H_0: b_j + \bar{\alpha}_\bullet \text{ equal for all } j$$

which implies

$$H_0: b_j \text{ equal for all } j$$

since $\bar{\alpha}_\bullet$ does not depend on j . Thus, for balanced data, $RSS(b_j | \mu)$ does provide a valid test for the equality of block effects! In fact, for balanced data,

$$RSS(b_j | \mu) = RSS(\alpha_i | b_j, \mu)$$

and

$$RSS(\alpha_i | \mu) = RSS(b_j | \alpha_i, \mu)$$

As a result, the order of fitting block and treatment effects is immaterial for balanced data.

B. Some mixed linear model theory

This section gives a few key results on mixed model analysis as implemented, e.g., in the MIXED procedure of the SAS System. More details may be found in Searle, S.R., Casella, G., McCulloch, C. E. (1992). Variance components, Wiley, New York. Also see: SAS/STAT User's Guide Version 8.

(1) Any mixed linear model can be written in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta}$ is a parameter vector of random effects and \mathbf{u} is a vector of random effects.

Example: For the colon data (Section 7.4), we have

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} \quad (i = 1, 2 \text{ groups}; j = 1, \dots, 2 \text{ regions}; k = 1, \dots, 6 \text{ animals})$$

where

y_{ijk}	= mean count at j -th region on k -th animal within i -th group
μ	= general effect
α_i	= main effect of i -th group
β_j	= main effect of j -th region
$(\alpha\beta)_{ij}$	= group \times region interaction
f_{ik}	= effect of the k -th animal within the i -th group; $f_{ik} \sim N(0, \sigma_f^2)$
e_{ijk}	= residual error; $e_{ijk} \sim N(0, \sigma^2)$

In matrix form this is

(2) The variance-covariance matrix of \mathbf{y} is

$$\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{ZGZ} + \mathbf{R}$$

where

$\mathbf{G} = \text{var}(\mathbf{u})$ and $\mathbf{R} = \text{var}(\mathbf{e})$.

Example: For the colon data we have

$$\text{var}(\mathbf{u}) = \text{var} \begin{pmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{14} \\ f_{15} \\ f_{16} \\ f_{21} \\ f_{22} \\ f_{23} \\ f_{24} \\ f_{25} \\ f_{26} \end{pmatrix} = \begin{pmatrix} \sigma_f^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_f^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_f^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_f^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_f^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_f^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_f^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_f^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_f^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_f^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_f^2 & 0 \end{pmatrix} = \mathbf{I}\sigma_f^2$$

and $\mathbf{R} = \mathbf{I}\sigma^2$. The zero off-diagonals in \mathbf{G} indicate that the effects are uncorrelated, i.e., the covariance is zero.

(3) The generalized least squares estimator of $\boldsymbol{\beta}$ is computed as

$$GLSE(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}_V = [\mathbf{X}^T V^{-1} \mathbf{X}]^{-1} \mathbf{X}^T V^{-1} \mathbf{y}$$

This estimator is the BLUE, i.e., the best linear unbiased estimator of $\boldsymbol{\beta}$. For V known, it is also the Maximum Likelihood Estimator (MLE) of $\boldsymbol{\beta}$, assuming multivariate normality.

(4) In practice, variance components in V need to be replaced by estimators, so that the "empirical" BLUE is

$$\hat{\boldsymbol{\beta}}_{\hat{V}} = [\mathbf{X}^T \hat{V}^{-1} \mathbf{X}]^{-1} \mathbf{X}^T \hat{V}^{-1} \mathbf{y}$$

The variance components may be estimated by the restricted maximum likelihood method (REML), details of which cannot be explained here.

Example: REML is the default method of MIXED for estimating variance components. Estimable functions are estimated by empirical BLUE.

(5) The null hypothesis

$$H_0: \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$$

is tested by a Wald-type F-statistic of the form

$$F_{\text{exp}} = \frac{\hat{\boldsymbol{\beta}}_{\hat{V}}' \mathbf{K} \left[\mathbf{K}' (\mathbf{X} \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{K} \right]^{-1} \mathbf{K}' \hat{\boldsymbol{\beta}}_{\hat{V}}}{\text{rank}(\mathbf{K})} \quad (\text{A5})$$

The numerator d.f. is equal to $\text{rank}(\mathbf{K})$. Determination of denominator d.f. is generally a difficult problem, and various approximations have been suggested. The best currently available option is the Kenward-Roger procedure, which is not explained here for brevity.

Example: The DDFM = KR option in PROC MIXED invokes the Kenward-Roger method.

(6) The generalized least squares estimator of an estimable linear function $\mathbf{k}'\boldsymbol{\beta}$ is given by

$$\mathbf{k}'\hat{\boldsymbol{\beta}}_{\hat{V}}$$

The estimated standard error of this estimator is

$$s.e.(\mathbf{k}'\hat{\boldsymbol{\beta}}_{\hat{V}}) = \sqrt{\mathbf{k}' (\mathbf{X} \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{k}}$$

An improved estimator, which accounts for errors in the weight matrix $\hat{\mathbf{V}}$, was suggested by Kenward and Roger (1997) and is implemented in PROC MIXED.

(7) t-test of $H_0: \mathbf{k}'\boldsymbol{\beta} = 0$:

$$t_{\text{exp}} = \frac{|\mathbf{k}'\hat{\boldsymbol{\beta}}_{\hat{V}}|}{s.e.(\mathbf{k}'\hat{\boldsymbol{\beta}}_{\hat{V}})}$$

This is compared against t_{tab} with approximated d.f., e.g. by the Kenward-Roger method.

Example: p-values of pairwise comparisons among least squares means obtained by LSMEANS (MIXED) are based on t_{exp} . The d.f. are computed by the Kenward-Roger method, if the DDFM = KR option is invoked.

Example: The ESTIMATE statement of MIXED uses the results given above.

(8) Confidence interval for $\mathbf{k}'\boldsymbol{\beta}$:

$$\mathbf{k}'\hat{\boldsymbol{\beta}}_{\hat{V}} \pm t_{tab} s.e.(\mathbf{k}'\hat{\boldsymbol{\beta}}_{\hat{V}})$$

Degrees of freedom can be approximated by the Kenward-Roger method.

(9) Note that a fixed effects linear model as considered in Appendix A corresponds to

$$V = I\sigma^2$$

with which

$$\text{GLSE}(\beta) = \text{OLSE}(\beta).$$

Example: When PROC MIXED is used to analyse a fixed effects model, the LSMEANS and ESTIMATE statement produce the same means as with PROC GLM. Also, the F-tests are identical.

Also, with $\hat{\sigma}^2 = s^2$, the Wald-type F-statistic in (A5) is equivalent to the one in (A2).

Example: With the HTYPE=1 option, PROC MIXED produces the same F-tests as the Type I SS in GLM.

All of this reflects the fact that a fixed effects model (Appendix A) is just a special case of a mixed model.

(10) For balanced data and for simple variance-covariance structures V , $\text{GLSE}(\beta) = \text{OLSE}(\beta)$.

Example: For the balanced split plot design, GLSE are the same as OLSE, and both are just simple arithmetic means. Thus, GLM and MIXED yield the same estimates with the LSMEANS and ESTIMATE statements. Note, however, that GLM does not compute correct standard errors and t-tests. For unbalanced data, the two estimators yield different results, and so the output of LSMEANS and ESTIMATE generally differs between GLM and MIXED.

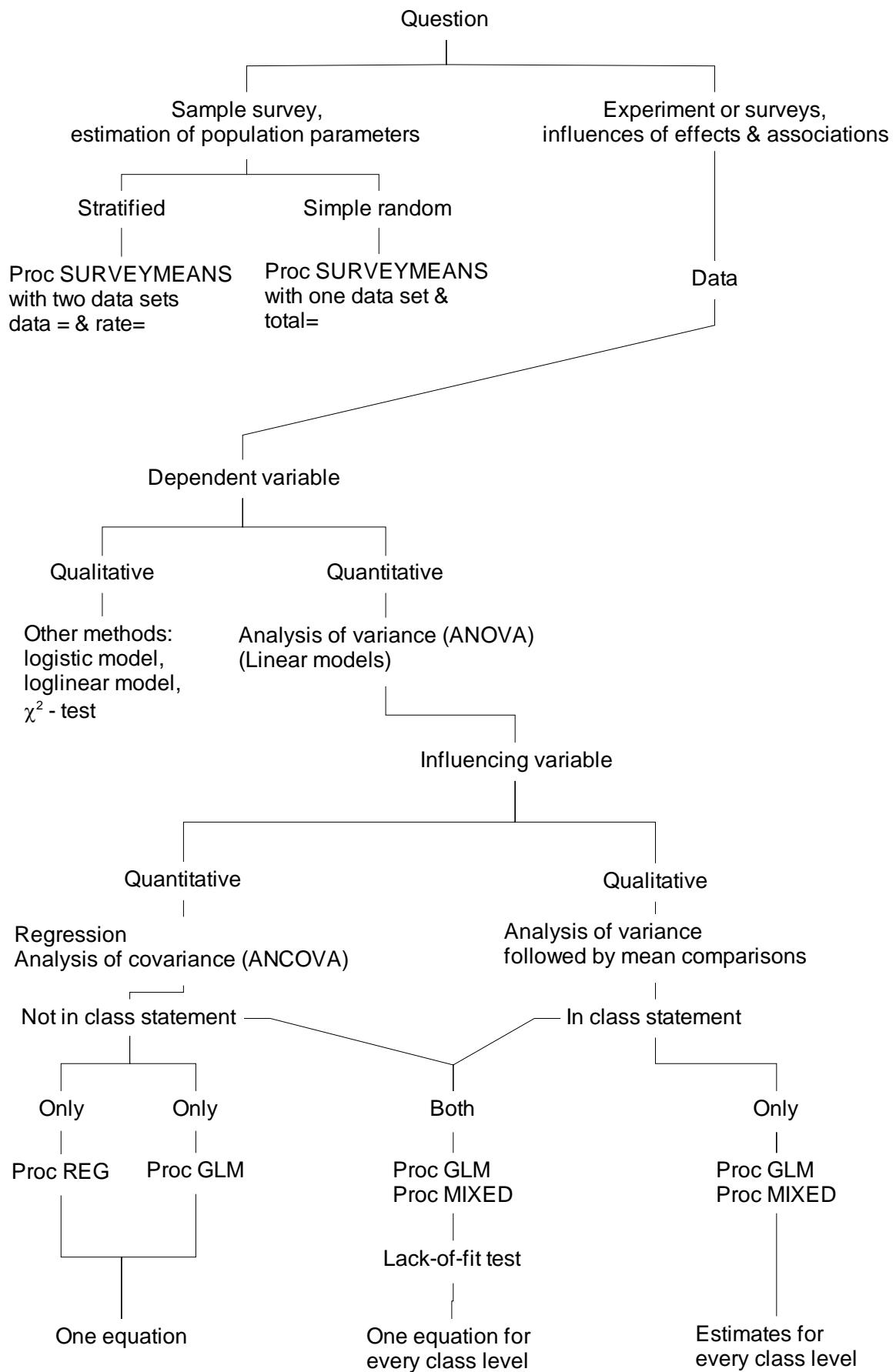


Fig. A1: Decision tree to select an analysis method based on methods covered in this lecture.

These terms mean the same:

Response = dependent variable = outcome variable

Predictor = explanatory variable = influencing variable

Which procedure when?

Type of data	Suitable statistics	Procedure	Class statement	Solution
Quantitative	Regression	REG, GLM or MIXED	No	Equation
Qualitative	ANOVA	GLM or MIXED	Yes	Least-square means estimates
Qualitative & quantitative	ANOVA & Regression (ANCOVA)	GLM or MIXED	Yes, for qualitative variables	One equation per level of class-variable (class-combination if more than one)

How to build a model

Always put

- **design** (e.g. block) variables in model **before treatment effects**
- **main effects before simple interactions** and than **complex interactions** and/or
- **linear regression terms before higher order** regression terms (x^2, x^3, \dots) and
- **lack-of-fit is always the last** effect in the model.

So the model looks like this:

Response = design variables + treatment variables + (lack-of-fit)