Modelling Biology

Basic Applications of Mathematics and Statistics in the Biological Sciences

Part II: Data Analysis and Statistics

Script B

Introductory Course for Students of Biology, Biotechnology and Environmental Protection

Werner Ulrich

UMK Toruń
2007
Contents

Introduction.................................................................................................................................................. 3
1: Planning a biological study..................................................................................................................... 4
2: Bayesian inference and maximum likelihood .......................................................................................... 9
3: The general linear model and analysis of variance ................................................................................. 18
4: Linear multiple regression ..................................................................................................................... 29
5: Logistic and other regression techniques ............................................................................................ 40
6: More on variance analysis..................................................................................................................... 48
7: Cluster analysis ..................................................................................................................................... 54
  7.1. K-means clustering ........................................................................................................................... 59
  7.2. Neighbour joining ............................................................................................................................. 60
8. Rank order and model II regression ....................................................................................................... 62
Literature................................................................................................................................................... 67
Statistics and the internet ........................................................................................................................... 68
Links to multivariate techniques .............................................................................................................. 70
Introduction

This is the second part of the two term lecture about biological modelling and the second part of the statistics lecture. It deals with Bayesian inference, model II regression, and basic multivariate techniques.

Many such techniques are now implemented in powerful statistic packages. The aim of any statistical education is to understand what these programs do, which errors can be made by applying them, and how to interpret their results. Today statistics is a holistic approach to our data and any basic lecture has also to deal with techniques of study planning, variable choice, power analysis, or data mining.

Again I tried to show the basic techniques by applying them to every day problems in biology. Special attention is paid to the technical side of data handling. The mathematical background, on the other hand, is discussed as short as possible and only as detailed as necessary to understand the process of data analysis.

Again this script is not a textbook. It serves as accompanying material to the lecture. Many very good online textbooks on basic and applied statistic are now available in the internet. Additionally, the internet provides for most of the every day problems of data analysis worked examples and ready to use solutions.

A one term basic lecture cannot deal with everything. Important multivariate techniques like principal component and discriminate analysis are left out. Missing are also recent developments in meta analysis. The wide field of null model building is only shortly discussed. Many statistical approximation techniques and manual techniques for simple pairwise comparisons (for instance the use of distribution tables) are not longer necessary. Instead the lecture contains many examples of how to use Excel and statistic packages for data analysis.
1. Planning a scientific study

Any sound statistical analysis needs appropriate data. **Any application of statistics to inappropriate or even wrong data is as unscientific as sampling data without prior theoretical work and hypothesis building.** Nevertheless both types of making 'science' have quite a long history in biology. One the one hand there is a strong mainly European tradition of natural history (faunistic, systematic, morphological) where data acquisition and simple description dominate. There is an untranslatable German phrase 'Datenhuberei' for this. Datenhuberei resulted, for instance, in large collections of morphological and faunistic data, manifested in mountains of insects in alcohol tubes or insect boxes. In many cases, the scientific value of these data is zero.

Another, mainly American tradition is called (also untranslatable) 'Sinnhuberei', the exaggeration of theoretical speculations and sophisticated statistical and mathematical analysis. The result was a huge number of theoretical papers, full of statistics, speculation, and discussion of other papers of the same type, but without foundation in reality.

Science but has to be founded both in good data and in good theory. **The first thing however must be theory.** Because this is still not always clear to students and even scientists I shall give a short introduction to the planning of scientific work. The schemes below and on the next side show a very general model of how to plan a scientific study. The latter scheme is based on a more specific one of Jürgen Bortz (Statistik für Sozialwissenschaftler, Springer 1999) designed for the social sciences.

We can identify six main phases of a scientific study. The first phase must always be a detailed analysis of the existing literature, where the problem to be solved is clearly stated. Fig. 1.1 shows this phase in a very simple diagram. You have to state your motivation by answering three main questions: **What is interesting? Why is it interesting? Cui bono?** To whom is it interesting and who might benefit from it?

The second theory building phase consists in the formulation or adaptation of a sound and logical theory. A **theory is a set of axioms and hypotheses that are formalized in a technical language.** What we need is a test whether our theory is **logically sound.** We also need **models for testing it and criteria for accepting or modifying it.** Hence, we must formulate specific **hypotheses** that follow from our theory. These hypotheses have to be tested by experiments or observations. And we must formulate exactly when to reject our hypotheses. Hence this phase has already to include possible methods for data analysis.

Next come the planning and data acquisition phases. Planning has to be done with the way of data analysis in mind. Often **pilot studies** are necessary. Data sizes have to be estimated by **power analysis.** The **analytical phase** consists of setting appropriate significance levels for rejecting a hypothesis and of choosing appropriate test statistics. In
Data analysis and statistics

1. **Searching phase**
   - Definition of the problem, study of literature

2. **Theoretical phase**
   - Formulating a general theory
     - Logical verification
     - Formulating criteria for accepting hypotheses
     - Deducing hypotheses
     - Deducing appropriate null models

3. **Planning phase**
   - Planning of the study, Experimental design

4. **Observational phase**
   - Getting data
   - Setting significance values $\alpha$ for acceptance
   - Choosing the appropriate statistical test
     - $P < \alpha$

5. **Analytical phase**
   - Criterion for accepting hypotheses fulfilled
   - Theory useful
     - Formulating further tests of the theory

6. **Decision phase**
   - Study ill designed
     - Modifying theory
     - Theory exhausted
     - Theory useful
   - Theory inappropriate
   - Theory exhausted
this phase many errors occur when choosing tests for which the data set is inappropriate. The data have to be checked whether they met the prerequisites of each test. Therefore, be aware of the possibilities and limitations of the statistical tests you envision prior to data analysis.

The last phase is the decision phase. The researcher has to interpret the results and to decide, whether the hypotheses have passed the tests and the theory seems worth further testing or whether all or some hypotheses must be rejected. In the latter case the theory requires revision or must also be rejected.

### Preparing the experimental or data collecting phase

Let’s look a bit closer to data collecting. Before you start any data collecting you have to have a clear vision of what you want to do with these data. Hence you have to answer some important questions:

1. For what purpose do I collect data?
2. Did I read the relevant literature?
3. Have similar data already been collected by others?
4. Is the experimental or observational design appropriate for the statistical data analytical tests I want to apply?
5. Are the data representative?
6. How many data do I need for the statistical data analytical tests I want to apply?
7. Does the data structure fit into the hypotheses I want to test?
8. Can I compare my data and results with other work?
9. How large are the errors in measuring? Do these errors prevent clear final results?
10. How large might the errors be for the data being still meaningful?

Statistics can answer some of these questions. This is the field of **power analysis**, with which we will deal later in detail.

### Data banks

How to process your experimental or observational data? Most popular is the data storage in special data bank programs like Access. But in most cases Excel data banks do the same and are easier to handle. Besides an Access data bank is shown. Important is that you have to prepare your data bank structure and data filter procedures before you start data collecting. By this you can immediately fit your experimental and data collecting protocols to the requirements of your data bank and the analytical techniques envisioned. Otherwise it might be that you are unable to incorporate data into your data bank.

However, in many cases it will be most time saving to write your data immediately into the spreadsheet of a statistical package. All advanced packages like **Statistica**, **SPSS**, **SAS**, **Systat**, **BMDP**, **MatLab**
Data analysis and statistics

(Matrix laboratory), or the **R-package** allow for data storing and processing. In this lecture we will mainly work with *Past* and *Statistica*. However, recently *R* got popularity among biologists because of its free distribution, its open structure and its wide analytical possibilities. Additionally, many specialized software solutions for non-standard statistical analyses are available (in part in the internet). The appendix lists some of them. Excel provides basic statistical distributions and tests and free add ins are available that allow for elementary data analysis. However, for any deeper analysis common statistical packages are indispensable.

**Publications**

An essential part of any scientific study is the publication, may be in form of a **paper** in a scientific journal, or in form of a **technical report** in the case of industrial research (R & D, research and development), or as an **opinion** for private or public organisations. Unpublished data are worthless. Writing a scientific paper is quite difficult and a lot of experience is necessary to publish successfully. Fortunately, there is now a series of books available that give advices. Many good examples and advices for all sort of biological publishing can be found at [http://classweb.gmu.edu/biologyresources/writingguide/index.htm](http://classweb.gmu.edu/biologyresources/writingguide/index.htm) or [http://abacus.bates.edu/~ganderso/biology/resources/writing/HTWtoc.html](http://abacus.bates.edu/~ganderso/biology/resources/writing/HTWtoc.html)

Scientific publications of any type are classically divided into 6 major parts:

1. **Title, affiliations and abstract**
   In this part you give a short and meaningful title that may contain already an essential result. The abstract is a short text containing the major hypothesis and results. The abstract should make clear why a study has been undertaken

2. **The introduction**
   The introduction should shortly discuss the state of art and the theories the study is based on, describe the motivation for the present study, and explain the hypotheses to be tested. Do not review the literature extensively but discuss all of the relevant literature necessary to put the present paper in a broader context. Explain who might be interested in the study and why this study is worth reading!

3. **Materials and methods**
   A short description of the study area (if necessary), the experimental or observational techniques used for data collection, and the techniques of data analysis used. Indicate the limits of the techniques used.

4. **Results**
   This section should contain a description of the results of your study. Here the majority of tables and figures should be placed. Do not double data in tables and figures.
5. Discussion
This part should be the longest part of the paper. Discuss your results in the light of current theories and scientific belief. Compare the results with the results of other comparable studies. Again discuss why your study has been undertaken and what is new. Discuss also possible problems with your data and misconceptions. Give hints for further work.

6. Acknowledgments
Short acknowledgments, mentioning of people who contributed material but did not figure as co-authors. Mentioning of fund giving institutions

7. Literature
A general rule in all writing is that fuzzy writing speaks of fuzzy thinking. Hence, try to think hard. Particularly important is that you clearly cite all sources you used. Otherwise you are in danger of making plagiarisms. How the scientific community understands correct citing can be found under http://abacus.bates.edu/pubs/Plagiarism/plagiarism.html

The example besides shows how a typical manuscript for an international journal has to look like. Of course every journal has its own requirements but in general these are quite similar.

Werner Ulrich
Nicolaus Copernicus University Toruń
Department of Animal Ecology
Gagarina 9, 87-100 Toruń; Poland
e-mail: ulrichw @ uni.torun.pl
Internet: www.uni.torun.pl/~ulrichw
Phone: 0048 – 56 - 611 4469

Species co-occurrences and neutral models: reassessing J. M. Diamond's assembly rules

Running title: Assembly rule and neutral models
Number of words: 5563

Abstract
The question whether....
Key words: Neutral model, assembly rules...

1. Introduction
One of the fundamental questions...

2. Methods
I used the zero-sum multinomial neutral model approach...

3. Results
The simple placement...

4. Discussion
Recently, Gotelli and McCabe (2002)...

Acknowledgements I thank Nicholas Gotelli for helpful and encouraging comments on the manuscript. Miss H. Pearson kindly improved my English. This work was in part supported by a grant of the Polish Scientific Committee, KBN No. P04 -034-22.

5. References


Table 1: Summary results...

Figure 1: Standardized effect sizes...
2. Bayesian inference and maximum likelihood

On the Bayesian side

In the Polish television Zygmunt Chajzer moderated a game show (Idź na całość) where candidates had to choose among three gates. Behind one was a car and behind the two others small evils ("zonks"). To enhance tension the game master showed the candidate after his first choice one of the two remaining gates behind which was the evil and asked whether the candidate now changes his choice. Looking from a probabilistic point of view (Fig. 2.1): should the candidate change his choice or doesn’t it matter? Of course, the probability that the candidate chooses the right gate at the first trial is 1/3 and at the second trial where only two gates remain 1/2. So, at first sight, it doesn’t matter whether the candidate changes. The probability that it is behind one or another is the same. However, things are not so easy. In probabilistic theory this problem is known as the “goats problem”, was formulated by the French mathematician Joseph Bertrand in 1889 and has been solved as late as 1990. The fact that the game master showed the second “Zonk” gate brings no further information. The candidate already knew that one of the two remaining gates after the first choice contained a “Zonk”. Hence changing the gate and opening the remaining opens in fact two gates, the “Zonk” gate and the unknown. The probability to get the car by changing is therefore not 1/2 it is 2/3! Fig. 2.2 makes this point clear. The green boxes denote the first choice, the red boxes are opened by the game master. Remaining at the first choice will win in one out of the three possibilities. Changing the gate will win in two out of the three possibilities. Hence the candidate should change. What has happened? We changed our perspective. One time we looked without prior knowledge to the problem (p = 1/2). We performed two independent choices. The second times we reconsidered our choice in the light of new information (the moderator opened a gate). This changes the probability distribution while including the history of decision.

In the probability part of the Math lecture we calculated dependent probabilities a priori that means without prior knowledge. The probability that events A and B occur is the product of the probability of A under the condition B and the probability of B.

\[
p(A \land B) = p(A | B) p(B)
\]

\[
p(A | B) = \frac{p(A \land B)}{p(B)}
\]

(2.1)

Likewise
Because \( p(A \land B) = p(B \land A) \) we can combine both equations and get

\[
 p(A \land B) = p(A | B)p(B) = p(B | A)p(A) \Rightarrow \\
 p(A | B) = \frac{p(B | A)p(A)}{p(B)}
\]  

(2.3)

This is the simplest formulation of the **theorem of Bayes** (after Thomas Bayes, British reverend, 1702-1761). It states that we can compute the conditional probability \( A | B \) from the inverse probabilities \( B | A \). Hence, if it is impossible to give \( p(A | B) \) we try to estimate this from \( p(B | A) \). If we deal with two events of which \( B \) is prior to \( A \) we estimate \( p(A | B) \) from the probability of the occurrence of \( B \) after \( A \) has occurred. \( p(A | B) \) is called the **posterior or conditional probability** of \( A \). \( p(A) \) is called the **prior or marginal probability** of \( A \) without any knowledge of the occurrences of \( B \). In other words the Bayesian theorem includes a priori knowledge to calculate probabilities. Within this terminology we can restate the Bayes theorem

\[
 p(A | B) = \frac{\text{conditional } \times \text{prior}(A)}{\text{prior}(B)}
\]

(2.4)

Now assume that the event \( B \) can have several states \( B_i \). In this case we can try to estimate \( p(A) \) (the total probability of \( A \) under all \( B_i \)) from an **a priori** knowledge of the occurrences of the various possibilities of \( A \). Eq. 2.1 tells that \( p(A \land B_i) = p(A | B_i)p(B_i) \). The sum of all \( p(A \land B_i) \) gives the total probability \( p(A) \) of \( A \)

\[
 p(A) = \sum_{i=1}^{n} p(A | B_i)p(B_i)
\]

(2.5)

This is known as the **theorem of total probability**.

We introduce eq. 2.5 into eq. 2.3 and get

\[
 p(B_i | A) = \frac{p(B_i)p(A | B_i)}{\sum_{i=1}^{n} p(A | B_i)p(B_i)}
\]

(2.6)

This is the important reformulation of the theorem of Bayes. It tells us how to estimate a probability of an event \( B_i \) in combination with \( A \) when the probability of \( A \) cannot be estimated directly. Then we estimate indirectly and need for this the **a priori** probabilities \( p(A | B_i) \).

Now we can solve the “Idż na całość” problem from a Bayesian perspective (Fig. 2.2). Assume we choose gate 1 \((G1)\) at the first choice. We are looking for the probability \( p(G1 | M3) \) that the car is behind gate 1 if we know that the moderator opened gate 3 \((M3)\). Hence

\[
 p(G1 | M3) = \frac{p(M3 \land G1)}{p(M3)} = \frac{p(M3 | G1)p(G1)}{p(M3 | G1)p(G1) + p(M3 | G2)p(G2) + p(M3 | G3)p(G3)}
\]

\[
 = \frac{1/2 \times 1/3}{1/2 \times 1/3 + 1 \times 1/3 + 0 \times 1/3} = \frac{1}{3}
\]
The probability that the car is behind gate 1 if the candidate chose gate 1 in the first trial is only 1/3. We should change the gate to get a probability of 1-1/3 = 2/3 that the car is behind gate 2.

Another example how to apply the Bayes theorem. Suppose we undertake a faunistic survey. We study the occurrence of the damselfly *Calopteryx splendens* at small rivers and want to use this data for bioindication. We know from the literature that *Calopteryx* occurs at about 10% of all rivers. Occurrence depends on water quality. We take data of water quality from environmental studies and compare them with occurrence data. From this we compute the probability of occurrence in certain quality classes. Suppose we have five quality classes that occur in 10% (class I), 15% (class II), 27% (class III), 43% (class IV), and 5% (class V) of all rivers. The probability to find *Calopteryx* in these five classes is 1% (class I), 7% (class II), 14% (class III), 31% (class IV), and 47% (class V). To which class belongs probably a river if we find *Calopteryx*? We compute

\[
p(\text{class I}|A) = \frac{0.1 \times 0.01}{0.1 \times 0.01 + 0.15 \times 0.07 + 0.27 \times 0.14 + 0.43 \times 0.31 + 0.05 \times 0.47} = 0.0048
\]

We compute for the other classes \(p(\text{class II}|A) = 0.051, p(\text{class III}|A) = 0.183, p(\text{class IV}|A) = 0.647, p(\text{class V}|A) = 0.114\). The sum of all probabilities is of course again 1. We conclude that finding *Calopteryx* indicates with probability of ca. 65% that we deal with a river of quality class IV. It is very improbable that we deal with a river of water classes I or II. Now assume that we study the whole damselfly fauna and use the tabulated abundance data. By combining all probabilities the Baysian theorem can serve as the basis to establish **indicator values** for these species. Taking all species together we would be able to asses water quality alone from occurrence data of these indicator species. Indeed for flatworms and caddis flies respective indicator values have been worked out to use them as bioindicators for water quality.

Another example how to apply the Bayes theorem. In a behavioural experiment crows have to choose earthworms from two boxes. Box 1 contains 5 earthworms and 5 plastic imitations, box 2 contains 5 earthworms and 10 plastic imitations. How fast does the crow lean to pick from box 1. In the first trial the crow picks an earthworm. What is the probability that it picked it from box 1. This is our probability \(p(A|B)\). We need three other probabilities:

1. \(P(A)\) is the a priori probability of picking from box 1. This probability is \(p(A) = 0.5\)
2. \(P(B)\) is the probability of picking an earthworm regardless of the box. This probability comes from eq. 2.6 \(p(A|B)\) is the probability to get an earthworm out of each box. Hence \(p(B) = 5/10 \times 0.5 + 5/15 \times 0.5 = 0.417\). This is not the same as the total probability to get an earthworm irrespective of the box \(p(B_{total}) = 15/25 = 0.4\!
3. \(P(B|A)\) is the probability to get an earthworm out of box A. Hence \(p(B|A) = 0.5\).

The probability that the earthworm came from box 1 is therefore

\[
p(A|B_i) = \frac{0.5 \times 0.5}{0.417} = 0.6
\]

This result seems intuitively clear because the proportion of real earthworms in box 1 was greater.

Bayes theorem is extensively discussed in forensic. Today’s DNA or drug tests are able to identify persons with high accuracy. However, if we have small probabilities of occurrence in a populations these tests have severe limitations. This is called the **false positive fallacy**. Let’s take a standard DNA test (Fig. 2.3). The test has a precision of more than 99%. Hence of 500 suspects we expect about 5 false positives. That means
that the probability that a DNA test points to the correct person in Fig. 2.3 is not 1/500. It is only 1/(1+4) = 0.20. The seemingly high precision of the test turns out to be fallacious. This phenomenon is called the “error of the prosecutor”.

Mathematically speaking

\[
p(c|+) = \frac{p(+|c)p(c)}{p(+)} = \frac{1 \times 1/500}{5/500} = \frac{1}{5}
\]

where + stands for a positive test and c stands for being the criminal. In this case the probability 5/500 can be viewed as stemming from correct and false positives (two exclusive possibilities). The Bayes theorem (eq. 2.6) modifies to

\[
p(c|+) = \frac{p(+|c)p(c)}{p(+|c)p(c) + p(+|\neg c)p(\neg c)}
\]

(4.7)

In this form the Bayes theorem is used in the juristic literature. This gives for the previous example

\[
p(c|+) = \frac{1 \times 1}{500 \times \left(1 \times \frac{1}{500} + 4 \times \frac{1}{499}\right)} = \frac{1}{5}
\]

An example for the complementary “error of the advocate”: In the process against the basketball star E. O. Simpson, one of his advocates (a Harvard professor) argued that Simpson sometimes has beaten his wife. However, only very few men who beat their wives later murder them (about 0.1%). The statistician I. J. Good however, (in a letter to nature) showed that from a Bayesian perspective the case is not as clear. Assume 10000 beaten wives (Fig. 2.4). In the mean 10 will be murdered. For simplicity we assume a mean duration of marriage of ten years. Hence the probability of a beaten women to be murdered in a given year is 1/10000. In the USA about 25000 people per year fall victims of murder (of a total of 250 Mill.). The probability to be murdered is also \(p = 1/10000\). Hence of 10000 beaten women two per year will be murdered: one by her husband and one by another person. The probability that a beaten woman was murdered by her husband is therefore 1/2 and not 1/1000! Again the difference comes from the perspective. Once \((p = 1/10000)\) we look from a priori and once \((p = 1/2)\) we look from a posteriori. Thus

\[
p(m|h) = \frac{p(m|h)\cdot p(h)}{p(m|h)\cdot p(h) + p(m|\neg h)\cdot p(\neg h)} = \frac{1}{2}
\]

Another example of Bayesian logic is the functioning of Anti-SPAM filters.
They use the following reasoning:

\[
p(\text{SPAM}|\text{specific words}) = \frac{p(\text{specific words}|\text{SPAM}) \cdot p(\text{SPAM})}{p(\text{specific words})}
\]

The advantage of such filters is that they can be trained. They use a posteriori knowledge about frequencies of words in SPAMs = \(p(\text{specific words}|\text{SPAM})\). Recent spammer tactics try to counterstroke these learning capabilities by loading the mails with lots of “innocent” words to reduce the frequency (probability) of the SPAM sensitive words.

This latter example directs us to the theory of searching. Assume a hawk searches a terrain for prey. Let \(p(\text{suc})\) be the probability to find a prey. Assume for simplicity he spots a number of pixel at a time that can be modelled by squares. Hence he surveys a grid. Of course he knows where to search. That means he knows a priori probabilities for each square of the virtual grid \(p(\text{prey})\). However there is another probability for each square, the probability \(p(\text{suc}|\text{prey})\) to find the prey. Our hawk will now systematically modify his a priori probabilities \(p(\text{prey})\) depending on his successive failures and search where he assumes the highest probability of success. Hence (eq. 2.7)

\[
p(\text{suc} | \text{prey}) = \frac{\frac{p(\text{prey} | \text{suc}) \cdot p(\text{suc})}{p(\text{prey} | \text{suc}) \cdot p(\text{suc}) + p(\text{prey} | \neg\text{suc}) \cdot p(\neg\text{suc})}}{p(\text{suc} | \text{prey}) \cdot (1 - p(\text{suc}))}
\]

This gives the modified probability of being prey under the experience of having no success. Now let \(A_T\) be the total area of search and \(A_E\) the empty part without prey. \(k\) denotes the number of successful hunts within the part of the area with prey. Hence \(p(\text{suc}|\text{prey}) = \frac{k}{(A_T-A_E)}\). We get

\[
p(\text{prey} | \neg\text{suc}) = \frac{\frac{k}{A_T} - \frac{k}{A_T} \cdot \frac{k}{A_T - A_E}}{\frac{k}{A_T - A_E} \cdot \frac{k}{A_T} - \frac{k}{A_T - A_E}} = \frac{A_T - A_E - k}{A_T(\frac{A_T - A_E}{A_T - A_E})} = \frac{A_T - A_E - k}{A_T - k}
\]

The hawk needs a priori knowledge on three parameters, the total search area, the proportion that might be empty and the number of successful attacks. If the total area is empty the probability that there is prey is of course zero, if there is no empty cell the probability is one in accordance with our expectation.

Another problem that includes a posteriori probabilities deals with the best choice problem. Assume a parasitic wasps that attacks clutches of aphids. These clutches are of different quality (size, exposition). The wasp visits one clutch after another. However, because it has of course competitors it has to choose after a certain time. The question is how long should the wasp search to make the best choice that means to attack the best clutch in the given situation? Assume the wasp “knows” that it can visit 20 clutches a day but cannot return to a previous. Such a problem can be solved if you can define gain and cost functions. The point where the costs exceed the gains should define the moment to attack. In the present case we apply the odds strategy. Odds are a possibility to presents probabilities. They are defined as

\[
\text{Odds} = o = \frac{p}{q} = \frac{p}{1-p}
\]
For instance, given that the probability to be 180 cm high is 0.7 for a man and only 0.1 for a woman the ratio of probabilities (the odds) for a man of being above 180 cm is 70:30 = 2.33 and for a woman 10:90. The chance of having above 180 cm is for a man 21 times higher than for a woman. Odds can be viewed as a measure of dependence of probabilities.

Using odds we can now define the cost function as a stopping rule. This was first done by the German–Belgian mathematician Thomas Bruss in 2005. The stopping rule comes from the sum of all odds
\[
O = \sum_{k=1}^{n} o_k
\]  
You should stop if this value $O > 1$. The probability of success is given the product
\[
p(S) = \prod_{k=n}^{1} (1-p_k) \sum_{k=n}^{1} \frac{p_k}{1-p_k} = \prod_{k=n}^{1} q_k \sum_{k=n}^{1} o_k
\]
(2.10)

The method is shown in the Excel example beside (Fig. 2.5). The probability that patch $k$ is the best is $p = 1/k$. The odds are added and at $k = 8$ this sum > 1. The respective probability function (eq. 2.10) shows that the probability to make the best choice is highest at $k = 8$. The wasp should attack the eight clutch it finds.

Tackling the problem from a mathematical point of view we consider a series of $n$ trials. The wasp should attack that clutch (at position $r$) that is better than the best of the previous $r-1$. The probability that the second best clutch is within the first $r$ trials is $p_2 = r/(a-1)$ where $a$ is the position of the best clutch. Hence the probability distribution that we stop at the best clutch number is given by the sum of the single distributions to make the best choice divided by $n$. For large $n$ we can approximate by an integral
\[
p(S) = \frac{1}{n} \sum_{r=1}^{n} \frac{r}{a-1} \approx \frac{1}{n} \int_0^n \frac{r}{a} \, da = -\ln \left( -\frac{r}{n} \right)
\]
For obtaining the maximum of $p(S)$ we have to set the derivative of $p(S) = 0$. Hence
\[
\frac{dS}{dr} = -\ln \left( -\frac{r}{n} \right) + \frac{r}{n} \frac{1}{n} = 0 \rightarrow \ln \left( -\frac{r}{n} \right) = -1 \rightarrow r = \frac{n}{e}
\]
For our clutch example $r = 7.358$. This called the \(1/e\)-stopping rule.

There is a long and ongoing discussion about the interpretation of the Bayesian theorem. Bayesianists...
and **frequentists** differ in the question how to interpret the probabilities involved. For a frequentist probability is the frequency of being observed. A Bayesionist sees probability as a measure of the assumed probability of being observed (the “degree of belief”). Bayesian inference tries therefore to estimate probabilities of our hypothesis by subsequent experience within a single model whereas frequentists seek to verify models from probabilities of alternative hypotheses.

**Maximum likelihoods**

Closely related to Bayesian inference is the use of **maximum likelihoods**. Suppose you studied 50 patients in a clinical trial and detected at 30 of them the presence of a certain bacterial disease. What is the most probable frequency of this disease in the population? Our first guess would be 60%. The maximum likelihoods method provides us with a systematic way to solve such problems to infer moments of statistical distributions.

The basic idea is to infer a function that describes the probabilities for a moment. The maximum of this function would be the desired value. This function is obtained from the distribution of known events. It is therefore an a posteriori method. Sir Ronald Fisher first introduced this method. Using observed (or assumed) frequencies p one may calculate the density function for a given process \( f_p(x_1,...,x_i) \). The maximum likelihood function \( L_p \) is now defined as

\[
L_p = f_p(x_1,...,x_i)
\]

Likelihood means probability. However probabilities always sum up to 1 when summed over all possible states. Likelihoods do not sum to one. In the above example we have to calculate the probabilities for 30 positives out of 50 trials for different probabilities p.

\[
p_{0.5}(30|50) = \binom{50}{30} \left( \frac{1}{2} \right)^{50} = 0.042
\]

\[
p_{0.6}(30|50) = \binom{50}{30} \left( \frac{3}{5} \right)^{30} \left( \frac{2}{5} \right)^{20} = 0.115
\]

\[
p_{0.8}(30|50) = \binom{50}{30} \left( \frac{4}{5} \right)^{30} \left( \frac{1}{5} \right)^{20} = 0.001
\]

Doing this for all probabilities p you get the likelihood function (Fig. 2.6). The maximum of this function is at \( p = 0.6 \), the value of our intuitive guess. Analytically we can solve the problem by setting the first derivative of \( L_p \) to zero.

\[
L_p = \binom{50}{30} p^{30} (1-p)^{20} \rightarrow \frac{dL_p}{dp} = \binom{50}{30} 30p^{29}(1-p)^{20} - \binom{50}{30} p^{30} 20(1-p)^{19} = 0
\]

\[3(1-p) = 2p \rightarrow p = \frac{3}{5}\]

\( p = 3/5 \) gives with \( \mu = np \) also the mean of the distribution \( \mu = 30 \).
Because it is often easier to reduce powers prior to calculation we can apply the \textit{log likelihood estimator} $\ln(L_p)$. This gives
\[
\ln(L_p) = \ln\left(\frac{50}{30}\right) + 30 \ln(p) + 20 \ln(1-p) \rightarrow \\
\frac{d \ln L_p}{dp} = 0 = \frac{30}{p} - \frac{20}{1-p} \rightarrow p = \frac{3}{5}
\]

The use of the log likelihood is possible because the logarithm of $L$ is a monotonically raising function and therefore proportional to $L$.

If we have consecutive trials and these trials are independent the probability to get exactly the observed events comes from the product of the single probabilities. In this case the maximum likelihood function is defined as the product

\[
L(p_1...p_n | O_1...O_m) = \prod_{i=1}^{n} f(p_i, O_i)
\]

For instance you study morphological variability in butterflies. You take a sample of 5 and find an anomaly in one of the cases. The normal variant occurs in $p = \frac{4}{5}$ cases. Your likelihood estimator is then
\[
L_p = p^4(1-p) \\
\frac{dL_p}{dp} = 4p^3(1-p) - p^4 = 0 \rightarrow p = \frac{4}{5}
\]

You expect the anomaly to occur in 20% of the population. Of course these are trivial examples and the overall result was that the best estimator for a population frequency is the sample frequency.

A typical biological example for the use of likelihoods is the construction of phylogenetic trees from DNA sequence data. Such a construction proceeds in a two step way. The first step includes the definition and measurement of the distances between the sequences. The second step uses one of the clustering algorithms to construct most parsimonious trees. Cluster analysis will be dealt with later. We look at the first step. Assume you have two sequences of length $n = 12$. At $x = 5$ sites these sequences differ (Fig. 2.7). The \textit{Jukes Cantor model} now assumes that the probabilities $\lambda$ of any transition within these 4 nucleotides is the same. However, there can be single, multiple, parallel or even backward transitions. Comparing the sequences has to consider this. Assuming that transition probability is time independent (every period has the same transition probability) the probability distribution follows an \textit{Arrhenius model}.

\[
T \rightarrow C \\
T \rightarrow C \\
C \rightarrow G \\
A \rightarrow G \\
A \rightarrow G \\
G \rightarrow C \\
G \rightarrow C \\
G \rightarrow C \\
A \rightarrow C \\
A \rightarrow C \\
C \rightarrow C \\
G \rightarrow T \\
\]

Single substitution
Parallel substitution
Back substitution
Multiple substitution

\[
p_{\text{trans}} \propto 1 - e^{-\lambda t}
\]

In our case $\lambda = 1/3$. Hence all nucleotide combinations have the same probability to occur. We can now calculate the probability that the nucleotides of a site differ. In order to do this we apply a Marcov chain model. Assume you start with A at time $t_0$. Because of equal transition probabilities you will end after time $t_i$ in an equal probability distribution of the four nucleotides. This is shown in the next \textit{Mathematica} example. The largest eigenvec-
The transition matrix $U$ has the eigenvector $\{1,1,1,1\}$ that means all states have the same probability of occurrence irrespective of the initial state. In vector notation: you might start from $\{1,0,0,0\}$ and end at $\{1/4,1/4,1/4,1/4\}$. We are now interested in the probability to get from the sequence at $t_0$ to the sequence at $t_1$. The **Theorem of Chapman-Kolmogorov** tells that this probability is the sum of all possible states in the Markov chain. Hence, the probability for a single difference is

$$p = \frac{3}{4} \left(1 - e^{-4t}\right) = \frac{3}{4} \left(1 - e^{-4/3t}\right)$$  \hspace{1cm} (2.13)

For estimating the probability of $x$ differences among $n$ sites we apply the binomial distribution.

$$L(x; d) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\ln(L(x; d)) = \binom{n}{x} + x \ln\left(\frac{3}{4} - \frac{3}{4} e^{-4/3d}\right) + (n-x) \ln\left(1 - \left(\frac{3}{4} - \frac{3}{4} e^{-4/3d}\right)\right)$$

Setting the first derivative to zero gives

$$t = -\frac{3}{4} \ln\left(1 - \frac{4x}{3n}\right)$$  \hspace{1cm} (2.14)

This is the mean time to get $x$ different sites. It is also a measure of distance that depends only on the number of substitutions.

Likelihoods can also be used to test hypotheses. This is done by the likelihood ratio test. In the previous example we might wish to test the hypothesis $p = 0.6$ against the alternative hypothesis $p = 0.5$. The respective likelihoods are $L(0.5) = 0.041$ and $L(0.6) = 0.115$. The ratio test calculates now

$$\text{LLR} = 2 \ln \left(\frac{L_1}{L_2}\right) = 2(\ln L_1 - \ln L_2)$$

In our case this gives 5.47. In the next chapter we will deal with the $\chi^2$ distribution. The likelihood ratio is $\chi^2$ distributed with 1 degree of freedom.
3. The general linear model and analysis of variance

Up to now we only dealt with statistics regarding one or two variables. But in biological research or applications we mostly deal with a manifold of factors that influence certain dependent variables. For instance, we design an experiment in which we manipulate more than one factor and are interested in the effects of all factors combined and of single effects. These are typical questions for multivariate statistics, statistics that deal with the relationships of two (or even more) sets of variables. One variable set contains all so-called independent variables and one set contains those variables that result from the independent variables, the dependent variables. The variables of both groups might, of course, also be interdependent. Assume for instance a forest in which a group of plants occur. The growth of these plants depends on a set of variables, light, for instance, humidity, temperature, mineral availability, soil structure, or pollination. We might treat these variables as independent variables. The second group of variables are the plant species. Variables that influence this group are for instance population density, biomass, assimilation rate, evaporation and others. But the variables of the second group are also interrelated. They compete for resources, have mutualistic interactions and/or are characterised by parasite - host relations. The aim of all multivariate techniques is now to describe the behaviour of these dependent variables in terms of all or of subsets of the independent variables.

A typical data sheet might look as below. We got data from 30 sites (or experiments, or observational stands, or, or, or). We measured 13 variables. Of course, the choice of these variables was directed by an underlying hypothesis about factors that influence plant growth and species richness. Building up an appropriate data bank is the next step. Now we have to apply our theoretical background. We have to decide which variables...
depend on others and which are independent. We might also decide that all these variables are independent of each other. How to test for this will be explained in the next lecture.

In our case we decided that the plant species densities depend on abiotic factors and on earthworm densities. These are clearly stated hypotheses and we have to discuss this choice.

In general a typical data matrix for further statistical analysis looks as follows. We have columns of variables and rows of cases.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
</tr>
</tbody>
</table>

There are six major groups of standard multivariate techniques:

The **analysis of variance** compares means of dependent variables under the influence of interacting independent variables.

**Multiple regression** tries to describe dependent variables from one or more sets of linear algebraic functions made up of independent variables.

**Cluster analysis** groups a set of entities according to the expression of a set of variables defining these entities.

**Factor analysis** is designed to condense a larger set of variables or observations that share some qualities into a new and smaller set of artificial variables that are easier to handle and to interpret.

**Discriminant analysis** is a regression analysis that tries to classify elements of a grouping variables into predefined groups.

**Multidimensional scaling** arranges a set of variables in a n-dimensional space so that the distances of these variables to the axes become minimal.

Except of cluster analysis all of these techniques (at least in their basic form) depend on the same set of prerequisites. These are called the **general linear model**.

**The general linear model**

Assume you have one dependent variable Y and 3 independent variables $X_1$, $X_2$, and $X_3$. The general linear model assumes now that all relationships between these variables can be described by simple linear algebraic functions of the type

$$Y = mX_i + b$$

$$X_i = mX_{j\neq i} + b$$

A second assumption is that

$$\sigma^2_{total} = \sum \sigma^2_i + \sum \text{cov}_{i,j}$$

If we now define one of the variables Y as a dependent variable we would be able to express the variance of Y from the variances of the other independent variables $X_i$

$$\sigma^2_Y = \sum \sigma^2_{i\neq Y} + \sum \text{cov}_{i,j\neq Y} + (\sigma^2_{total} - \sum \text{cov}_{i,j}) = \sum \sigma^2_{i\neq Y} + \sum \text{cov}_{i,j\neq Y} + \sigma^2_{error}$$

(3.1)

In other words, the variance of the dependent variable Y can be expressed by the sums of variances and covariances of the independent variables and a term that contains the variances introduced or co-introduced by
Y. Additionally, we know already that if we can write the variances in such an additive form, we have the same relation for the means.

\[ \mu_Y = \sum \mu_i + \mu_{Error} \]  

(3.2)

This last result can be interpreted as follows. For multivariate statistic to be computed the mean of our dependent variable has to be a linear combination of all independent variables plus an error term. Equations 3.1 and 3.2 are very important. The general linear model (GLM) relates the applicability of a multivariate technique to these basic assumptions. The general linear model assumes that the dependent variables can be constructed by linear combinations of the set of independent variables.

Now, we can state the question about the applicability of any multivariate technique that depends on the GLM more precisely. We have to ask what structure our data set has to have to be describable by the general linear model:

- First of all, there have to be linear relationships. If our variables would be connected in a nonlinear manner (this is most often the case in biological sciences) we would first have to linearize them by appropriate data transformations.
- A next assumption is that the errors around the means have to have a normal distribution. Be careful. A logarithmic transformation of data distorts the normal distribution or errors. The errors are then log-normally distributed.
- Additionally, this normal distribution of errors has to have a similar variance for all data points. A violation of this assumption is shown in the Figure beside (Fig. 3.1). It shows the relation between the dependent variable Y and one of the independent variables X. X is of course a discrete variable with five expressions. Blue are the data points, the mean values are given in red and the error bars show one standard deviation. The within error raises systematically. This is a typical violation of the variance homogeneity. It is often termed heteroscedasticity.

Later we will deal with some other restrictions of the general linear model. In general, the number of observations has to be large enough so that errors can vary normally around the mean. If the number of data points would be too small outliners or single deviating observations would have an overproportional influence on the result. A general rule is that the number of cases in our data matrix must never be smaller than the number of variables in the model, for most techniques the number of cases should be at least five times the number of variables. The relative effect of single data points decreases with the sample size. Too small samples lead therefore to small-size effects, so-called Poisson errors. They should be minimized.
Consider an experiment designed to infer the influence if light intensity on plant growth. Four experimental plots (A to D) are installed each with five replicates. The four plots differ in total light energy supply. After the experiment we measure total plant biomass in each of the 25 experimental plots. The raw data are given in the Table beside. The question is now: does light intensity influence total biomass? Of course we could take our t-test and compare all five plants separately (giving 10 single comparisons). But it is possible that none of the pairwise comparisons gives a significant result although the treatments differ due to a trend in the data set. It would be better to treat all groups simultaneously. This is the aim of an analysis of variance. The idea behind the ANOVA is very simple. It was developed by the British biostatistician and genetic Sir Ronald Fisher (1890-1962) especially for the biological sciences.

The total variance of the whole data set is given by

\[
S^2_{\text{total}} = \frac{k}{N-1} \sum_{k=1}^{N} (x_k - \bar{x})^2 = \frac{k}{N-1} \sum_{j=1}^{n} \sum_{j=1}^{n} (x_{i,j} - \bar{x})^2
\]

k denotes the number of groups (treatments). N is the total number of data points (the sample size) and ni denote the numbers of data points within each treatment. For simplicity the ANOVA uses sums of squares \( SS = s^2 \times (n-1) \). Hence

\[
SS_{\text{total}} = \sum_{i=1}^{k} \left( \sum_{j=1}^{n} (x_{i,j} - \overline{x_{\text{total}}})^2 \right)
\]

(3.3)

Now we divide the total variance of all 25 plots into two groups. One group contains the variance within the 4 groups, the second group the variance between the groups (computed from the grand mean) (Fig. 3.2).

Under the assumption that the within group variability is caused by the same set of unknown variables (that operate within the groups in a similar way) any difference in the between group variability should be caused by the treatment. In a next step we compare both variances. We know already a way to compare variances. It is the F-test of Fisher.

\[
F = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}}
\]

(3.4)

How to compute within and between group variances. We compute

\[
SS_{\text{between}} = \sum_{i=1}^{k} n_i (\overline{x_i} - \overline{x_{\text{total}}})^2
\]

(3.5)
We sum over the $k$ groups. Why multiplying with $n_i$, the number of data points in each treatment? To compute the between variance we assume that the total variance comes solely from the differences between the treatments. Hence, we replace the within values through the within mean. The new total sum of squares is then according to eq. 3.3

$$SS_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{total})^2 = n_i (\bar{x}_{j} - \bar{x}_{total})^2$$

Now we need the degrees of freedom. If we would in our example divide $SS_{between}$ through the group number ($4$) we would underestimate the total variance. Why? Because if we have means of three groups, the variance computed by including the fourth group is no longer a random variable. It is already determined by the other four values. We have instead to divide through $3$, the number of variables that can fluctuate at random. In other words, we have to divide through the number of degrees of freedom. This is for the between group variance $df_{between} = k - 1$ if $k$ denotes the number of groups.

The within group variance is given by

$$SS_{within} = \sum_{i=1}^{k} \left( \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \right)$$

(3.6)

$SS_{within}$ is therefore the sum of the variances computed within the groups. This is the part of variance not influenced by the treatment. It is introduced by other (often unknown or not measured) factors and is often also called $SS_{error}$. If $SS_{within}$ would equal $SS_{between}$ the treatment would not influence the dependent variable. Otherwise $SS_{between}$ would introduce an additional part of variance. Variances of independent random variates are additive (eq. 5.21). We get a fundamental equation on which the analysis of variance is based.

$$SS_{total} = SS_{between} + SS_{within}$$

(3.7)

For the degrees of freedom a similar equation holds. We have $df_{within} = N-k$ degrees of freedom. Hence.

$$df_{total} = df_{between} + df_{within}$$

(3.8)

In our case equation 3.7 gives $(25-1) = (4-1) + (25-4)$. This is immediately clear. If we would compute the total variance (computed from all cases) according to eq. 4.8 we would have to divide through $n-1 = 24$. $SS_{total}$ has 24 degrees of freedom.

Our F-test looks therefore as follows

$$F = \frac{SS_{between}}{SS_{within}} \cdot \frac{k-1}{N-k} = \frac{N-k}{k-1} \frac{SS_{between}}{SS_{within}}$$

(3.9)

A numerical example how to compute an ANOVA is shown in the next table. A standard matrix that contained data from four treatments with five observations each gave four group means. Do these group means differ? $SS_{between}$ and $SS_{within}$ are computed separately and compared with $SS_{total}$. You see that the sum of $SS_{between}$ and $SS_{within}$ is indeed identical to $SS_{total}$. The statistic package Statistica (and all others too) give the following output (in this case it would even be faster to compute $F$ by hand). The variable light intensity (between
variability) has three degrees of freedom, the within variability has $25 - 4 = 21$ degrees of freedom. *Statistica* denotes MS (mean sums of squares) instead of SS. $MS = SS / df$.

$MS_{\text{effect}}$ is the between group variability, $MS_{\text{error}}$ the within group variability. $F$ is the quotient of $MS_{\text{effect}} / MS_{\text{error}} = 22.79 / 6.34 = 3.597$. The corresponding probability level $p(F)$ is 0.031. In other words, we conclude that light intensity has a statistically significant effect on plant growth. Our error level for making a type I error is about 3%. But be careful. This error level is only valid if the prerequisites of the GLM are met. In particular, the errors around the mean have to be at least approximately normally distributed.

In the previous case we dealt with only one independent variable (or predictor variable), the light intensity, plant biomass was the dependent variable. The ANOVA can easily be extended to deal with more than one predictor variable. In this case we speak of multifactorial designs. With the same reasoning as above we can divide the total variance into parts where each part represents the fraction of variance introduced by a certain variable. The next Table gives our plant data with an additional variable, nutrient availability. Now, we have a two factorial design. Equation 3.6 modifies to...

<table>
<thead>
<tr>
<th>Observations</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>SSbetween</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.08</td>
<td>0.19</td>
<td>0.83</td>
<td>2.80</td>
<td>0.404</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>1.21</td>
<td>0.71</td>
<td>2.69</td>
<td>0.404</td>
</tr>
<tr>
<td>3</td>
<td>0.19</td>
<td>1.97</td>
<td>1.10</td>
<td>1.93</td>
<td>0.404</td>
</tr>
<tr>
<td>4</td>
<td>0.51</td>
<td>0.19</td>
<td>0.11</td>
<td>2.57</td>
<td>0.404</td>
</tr>
<tr>
<td>5</td>
<td>0.73</td>
<td>0.19</td>
<td>0.30</td>
<td>2.58</td>
<td>0.404</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group mean</th>
<th>0.448</th>
<th>0.750</th>
<th>0.811</th>
<th>2.515</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSwithin</td>
<td>0.131</td>
<td>0.319</td>
<td>0.046</td>
<td>0.081</td>
</tr>
<tr>
<td>SSError</td>
<td>0.085</td>
<td>1.484</td>
<td>0.244</td>
<td>0.342</td>
</tr>
<tr>
<td>df</td>
<td>0.004</td>
<td>0.314</td>
<td>0.250</td>
<td>0.004</td>
</tr>
<tr>
<td>F</td>
<td>0.082</td>
<td>0.312</td>
<td>0.096</td>
<td>0.004</td>
</tr>
</tbody>
</table>

| Total SSwithin | 4.11 |
| Total SSbetween| 13.96 |
| Grand mean     | 1.05 |
| Grand SS       | 0.14 |
| SSbetween+SSwithin | 18.07 |
| F              | 18.14 |
| F-test         | 2.118E-05 |

<table>
<thead>
<tr>
<th>Treatments</th>
<th>SSwithin</th>
<th>SSbetween</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4.11</td>
<td>13.96</td>
</tr>
<tr>
<td>B</td>
<td>9.00</td>
<td>5.00</td>
</tr>
<tr>
<td>C</td>
<td>22.79</td>
<td>6.34</td>
</tr>
<tr>
<td>D</td>
<td>22.79</td>
<td>6.34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observations</th>
<th>Light intensity</th>
<th>Nutrients</th>
<th>Biomass</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>A</td>
<td>11</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>A</td>
<td>7</td>
</tr>
<tr>
<td>D</td>
<td>A</td>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>E</td>
<td>B</td>
<td>A</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>B</td>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>G</td>
<td>B</td>
<td>A</td>
<td>5</td>
</tr>
<tr>
<td>H</td>
<td>B</td>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>I</td>
<td>B</td>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>J</td>
<td>B</td>
<td>A</td>
<td>8</td>
</tr>
<tr>
<td>K</td>
<td>B</td>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>L</td>
<td>B</td>
<td>A</td>
<td>5</td>
</tr>
<tr>
<td>M</td>
<td>B</td>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>N</td>
<td>B</td>
<td>A</td>
<td>15</td>
</tr>
<tr>
<td>O</td>
<td>B</td>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>P</td>
<td>B</td>
<td>A</td>
<td>4</td>
</tr>
<tr>
<td>Q</td>
<td>B</td>
<td>A</td>
<td>11</td>
</tr>
<tr>
<td>R</td>
<td>B</td>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>S</td>
<td>B</td>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>T</td>
<td>D</td>
<td>A</td>
<td>7</td>
</tr>
<tr>
<td>U</td>
<td>D</td>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>V</td>
<td>D</td>
<td>A</td>
<td>5</td>
</tr>
<tr>
<td>W</td>
<td>D</td>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>X</td>
<td>D</td>
<td>A</td>
<td>7</td>
</tr>
<tr>
<td>Y</td>
<td>D</td>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>Z</td>
<td>D</td>
<td>A</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>D</td>
<td>A</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>D</td>
<td>A</td>
<td>5</td>
</tr>
<tr>
<td>E</td>
<td>D</td>
<td>A</td>
<td>6</td>
</tr>
</tbody>
</table>
The degrees of freedom come from

\[ df_{\text{total}} = df_A + df_B + df_{AxB} + df_{\text{error}} \]

and

\[ kmn = (k - 1) + (m - 1) + (k - 1)(m - 1) + km(n - 1) \]

where \( k \) and \( m \) are the numbers of Categories in A and B and \( n \) is the number of cases.

We have the variance part introduced by the factor light (SS\(_A\)) only, by the additional factor nutrients (SS\(_B\)) only, by the combined effect of light and nutrients (SS\(_{AxB}\)), and of course by the within variance (SS\(_{\text{error}}\)).

SS\(_A\) is computed separately for both groups of nutrients to exclude the influence of nutrients. SS\(_B\) is computed separately for the four groups of nutrients to exclude the influence of light intensity. Additionally we need the SS\(_{AxB}\). Under the hypothesis that light and nutrients are not independent the combined action of both variables might influence the plant biomass.

Be careful. For every treatment of light intensity you have to have at least two treatment cases of nutrients. Otherwise it would be impossible to compute the within variability of nutrients inside the treatment light. However in reality two cases are much too less. Even for an approximated estimate of the within variance you need at least 6 treatments for each combination of light and nutrients. Hence, you need at least 4*2*6 = 48 cases. The number of experimental data necessary for multifactorial designs raises very fast.

We computed the total sums of squares and the sums of squares between light intensity and between nutrient availability. This is already quite time consuming to do it by hand. But our statistic package gives immediately the output shown above. The Statistica output shows separate results for the variables (treatments) light (effect 1) and nutrients (effects 2). These are called the main effects. Additionally the combined effect of light and nutrients is given, the secondary or combined effects. Again, we have the within variance MS error, which is the same for all three combinations of effects. The latter are computed over all groups of combinations of the main effects. The quotients of MS\(_{\text{effect}}\) (MS\(_{\text{between}}\)) and MS\(_{\text{error}}\) (MS\(_{\text{within}}\)) is again our F-value. For light I used the same data as in the monofactorial design. We see that including a new treatment variable changed the significance levels for light too. Now light appears to be less significant.

Nutrient availability and the combined effect of light and nutrient do not significantly contribute to total variance. Of course, we have always to take care that sums of squares can be computed. So, for each combina-
tion of variables at least two data points must be available. Otherwise our experimental design would be incomplete. The table beside shows such a typical *incomplete design*. We have one metrically scaled independent variable and three predictors. However, a full ANOVA including all three predictors is impossible to compute because not all predictor combinations occur. We will deal with incomplete designs in chapter 6.

Now comes one example of the category how to lie with statistics. The next *Statistica* table contains the result of an ANOVA with three effects (VAR2, VAR3, and VAR4). We see that VAR2 and the combined effect of all predictors are statistically significant at the 5% error level. However, as independent variables I used simple random numbers. There shouldn’t be any significant results. The probability for a significant result should be 1 out of 20 = 5%). What has happened? We used an error level $\alpha$ of 5% for the whole ANOVA result. The probability to make no error is $1 - \alpha$. The probability to make no error in n experiments is therefore $(1 - \alpha)^n$. What is then the probability that in n experiments we reject erroneous H0, that is that we make a type II error? This probability is $\beta = 1 - (1 - \alpha)^n$. In our above example we have 3 independent variables and their combinations. Our probability level is therefore not $\alpha = 0.05$. It should be between $\beta = 1 - (1 - \alpha)^3 = 0.14$ and $\beta = 1 - (1 - \alpha)^7 = 0.30$. Hence for single contrasts we get a new error level of 14 to 30% and we expect to make an erroneous decision in 1 out of 6 to 1 out of 3 cases. Our level of $\alpha = 0.05$ is often termed the *test wise error rate*. But if we use a test several times with the same set of data or compute a multivariate test with many variables simultaneously we have to deal with the *experiment wise error rate*. We have to correct our initial significance level by the following equation

$$\alpha^* = 1 - (1 - \alpha)^{1/n} \approx \frac{\alpha}{n}$$

(3.10)

The latter expression $\alpha^* \approx \alpha / n$ results from the binomial function if $\alpha$ is small (from a Taylor expansion). This so-called *Bonferroni correction* reduces the test wise error level to reach an acceptable experiment wise error level. You divide your initial error level through the number of independent variables, or, if you want to be on the sure side, through the number of combinations of these variables. In our above case we should therefore only accept significance levels $p(F)$ below $0.05 / 7 = 0.007$. 

<table>
<thead>
<tr>
<th>Dep</th>
<th>Pred A</th>
<th>Pred B</th>
<th>Pred C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.166325</td>
<td>A</td>
<td>1</td>
<td>H</td>
</tr>
<tr>
<td>0.259949</td>
<td>A</td>
<td>1</td>
<td>H</td>
</tr>
<tr>
<td>0.594445</td>
<td>A</td>
<td>1</td>
<td>H</td>
</tr>
<tr>
<td>0.425759</td>
<td>A</td>
<td>1</td>
<td>H</td>
</tr>
<tr>
<td>0.05286</td>
<td>B</td>
<td>1</td>
<td>H</td>
</tr>
<tr>
<td>0.005181</td>
<td>B</td>
<td>1</td>
<td>H</td>
</tr>
<tr>
<td>0.41623</td>
<td>B</td>
<td>2</td>
<td>H</td>
</tr>
<tr>
<td>0.111211</td>
<td>B</td>
<td>2</td>
<td>H</td>
</tr>
<tr>
<td>0.960788</td>
<td>B</td>
<td>2</td>
<td>H</td>
</tr>
<tr>
<td>0.819748</td>
<td>C</td>
<td>2</td>
<td>H</td>
</tr>
<tr>
<td>0.649385</td>
<td>C</td>
<td>2</td>
<td>H</td>
</tr>
<tr>
<td>0.937118</td>
<td>C</td>
<td>2</td>
<td>H</td>
</tr>
<tr>
<td>0.219705</td>
<td>C</td>
<td>2</td>
<td>H</td>
</tr>
<tr>
<td>0.379865</td>
<td>C</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.266792</td>
<td>C</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.34661</td>
<td>D</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.356833</td>
<td>D</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.767452</td>
<td>D</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.894365</td>
<td>E</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.202289</td>
<td>E</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.230626</td>
<td>E</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.722637</td>
<td>E</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.438195</td>
<td>E</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.561168</td>
<td>E</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>0.786699</td>
<td>E</td>
<td>2</td>
<td>I</td>
</tr>
</tbody>
</table>
What is the necessary sample size for the within sums of squares $SS_{\text{error}}$? In other words what is the equivalent of a power analysis in the case of an ANOVA. We can compute this by a similar t-test model as above. We define the effect size by the maximum difference of means of the groups. We get

$$t \approx \sqrt{\frac{N_{\text{effect size}}}{2k}} \sigma_{\text{within}}$$

Hence

$$N = 2k \left( \frac{t \sigma_{\text{within}}}{\text{effect size}} \right)^2$$

where $k$ is the number of groups and $N$ is the total sample size (the total number of observations). Again, we need information about the within group variability.

Lastly, deal with two non-parametric alternatives to the ANOVA. Assume we have an observational series with a single discrete independent variable, a monofactorial design. Then, we can use the so-called Kruskal-Wallis test or Kruskal-Wallis ANOVA by ranks. In principle, the Kruskal-Wallis test is computed...
in the same way as the U-test. We compute the total rank sums as the base for a test statistic. Our test statistic is

\[ KW = \frac{12}{N(N+1)} \sum_{i=1}^{r} \frac{R_i^2}{n_i} - 3(N+1) \]  

(3.13)

where \( N \) is the total number of observations, \( n_i \) the number of observations of group \( i \), and \( k \) the number of groups. The Table shows an example how to compute a Kruskal-Wallis test with Excel. We have a series of 53 observations divided into 4 groups or treatments. The KW-value of 9.15 has again to be compared with tabulated ones. For values of \( r > 5 \) or larger sample sizes KW is approximately \( \chi^2 \) distributed and values can be taken from a \( \chi^2 \) table with \( r-1 \) degrees of freedom. In our case this value for 3 degrees of freedom and \( \alpha = 0.95 \) is \( \chi^2 = 7.81 \). Our value of 9.15 is larger than 7.81 and we accept the hypothesis that there is a difference between the treatments at an error level of 5% (the exact error level given by Statistica is 2.73%). The Kruskal-Wallis test has in monofactorial designs nearly the same power as an ordinary ANOVA and should be used in all cases where we are not sure about the underlying frequency distributions of the populations.

A second alternative is to use a rank order ANOVA. In this case we apply an ordinary ANOVA but use for the dependent metrically scaled variable ranked data. Such a rank ANOVA is applicable in monofactorial and multifactorial designs. An example is shown in the next table. We have 29 data sets grouped into four effects. Now we rank these data from the largest (0.929 = rank 1) to the smallest (0.082 = rank 29) using the Excel build in function POZYCJA( .. ). We compare the results of ordinary ANOVA, ANOVA with ranked data and KW-test. The ANOVA gives a probability level for \( H_0 \) of \( p(F) = 0.007 \), the ANOVA with ranked data proposes \( p(F) = 0.014 \), and the Kruskal-Wallis test returns \( p = 0.024 \). All three tests point to significant differences between the four groups. The non-parametric alternative returns a higher significance level. This is caused by the loss of information due to the ranking of the data. But the probability to make a type I error (to accept the wrong hypothesis) is also reduced.

<table>
<thead>
<tr>
<th>Dep</th>
<th>Ranks</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.489864</td>
<td>13</td>
<td>A</td>
</tr>
<tr>
<td>0.366997</td>
<td>17</td>
<td>A</td>
</tr>
<tr>
<td>0.253452</td>
<td>22</td>
<td>A</td>
</tr>
<tr>
<td>0.777929</td>
<td>11</td>
<td>A</td>
</tr>
<tr>
<td>0.321003</td>
<td>19</td>
<td>A</td>
</tr>
<tr>
<td>0.126373</td>
<td>24</td>
<td>A</td>
</tr>
<tr>
<td>0.111894</td>
<td>26</td>
<td>A</td>
</tr>
<tr>
<td>0.055178</td>
<td>27</td>
<td>A</td>
</tr>
<tr>
<td>0.982267</td>
<td>8</td>
<td>B</td>
</tr>
<tr>
<td>0.946089</td>
<td>9</td>
<td>B</td>
</tr>
<tr>
<td>1.763621</td>
<td>3</td>
<td>B</td>
</tr>
<tr>
<td>0.361607</td>
<td>18</td>
<td>B</td>
</tr>
<tr>
<td>1.620806</td>
<td>5</td>
<td>B</td>
</tr>
<tr>
<td>1.854719</td>
<td>2</td>
<td>B</td>
</tr>
<tr>
<td>1.206337</td>
<td>7</td>
<td>B</td>
</tr>
<tr>
<td>0.627104</td>
<td>12</td>
<td>C</td>
</tr>
<tr>
<td>0.236488</td>
<td>23</td>
<td>C</td>
</tr>
<tr>
<td>0.380474</td>
<td>15</td>
<td>C</td>
</tr>
<tr>
<td>0.311656</td>
<td>20</td>
<td>C</td>
</tr>
<tr>
<td>0.370961</td>
<td>16</td>
<td>C</td>
</tr>
<tr>
<td>0.404568</td>
<td>14</td>
<td>C</td>
</tr>
<tr>
<td>0.042592</td>
<td>29</td>
<td>C</td>
</tr>
<tr>
<td>0.12325</td>
<td>25</td>
<td>D</td>
</tr>
<tr>
<td>0.256909</td>
<td>21</td>
<td>D</td>
</tr>
<tr>
<td>1.535243</td>
<td>6</td>
<td>D</td>
</tr>
<tr>
<td>2.754946</td>
<td>1</td>
<td>D</td>
</tr>
<tr>
<td>0.945028</td>
<td>10</td>
<td>D</td>
</tr>
<tr>
<td>0.048601</td>
<td>28</td>
<td>D</td>
</tr>
<tr>
<td>1.63511</td>
<td>4</td>
<td>D</td>
</tr>
</tbody>
</table>
Lastly, the ANOVA is of course also an alternative to a simple t-test in pairwise comparisons. If you have only two groups your ANOVA should give results similar to a t-test and the Kruskal-Wallis test should be comparable to the U-test. This is indeed the case, as the last example shows. The following data set was compared by a t-test and by the F-test. The t-test returns:

\[ \text{TEST.T(A2:A9,D2:D8,2,2)} = 0.000623. \] The ANOVA gives an identical result.

<table>
<thead>
<tr>
<th>Dep</th>
<th>Ranks</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.488964</td>
<td>8</td>
<td>A</td>
</tr>
<tr>
<td>0.36997</td>
<td>9</td>
<td>A</td>
</tr>
<tr>
<td>0.253452</td>
<td>12</td>
<td>A</td>
</tr>
<tr>
<td>0.777929</td>
<td>7</td>
<td>A</td>
</tr>
<tr>
<td>0.321003</td>
<td>11</td>
<td>A</td>
</tr>
<tr>
<td>0.126373</td>
<td>13</td>
<td>A</td>
</tr>
<tr>
<td>0.111894</td>
<td>14</td>
<td>A</td>
</tr>
<tr>
<td>0.055178</td>
<td>15</td>
<td>A</td>
</tr>
<tr>
<td>0.982267</td>
<td>5</td>
<td>B</td>
</tr>
<tr>
<td>0.946089</td>
<td>6</td>
<td>B</td>
</tr>
<tr>
<td>1.763621</td>
<td>2</td>
<td>B</td>
</tr>
<tr>
<td>0.361607</td>
<td>10</td>
<td>B</td>
</tr>
<tr>
<td>1.620806</td>
<td>3</td>
<td>B</td>
</tr>
<tr>
<td>1.854719</td>
<td>1</td>
<td>B</td>
</tr>
<tr>
<td>1.206337</td>
<td>4</td>
<td>B</td>
</tr>
</tbody>
</table>

Advices for using ANOVA

- You need a specific hypothesis about your variables. In particular, designs with more than one predictor level (multifactorial designs) have to be stated clearly.
- ANOVA is a **hypothesis testing method**. Pattern seeking will in many cases lead to erroneous results.
- Predictor variables should really measure different things, they should not correlate too highly with each other.
- The general assumptions of the GLM should be fulfilled. In particular predictors should be additive. The distribution of errors should be normal.
- It is often better to use log-transformed values.
- In monofactorial designs where only one predictor variable is tested it is often preferable to use the non-parametric alternative to ANOVA, the **Kruskal-Wallis test**. The latter test does not rely on the GLM assumptions but is nearly as powerful as the classical ANOVA.
- Another **non-parametric alternative** for multifactorial designs is to use ranked dependent variables. You loose information but become less dependent on the GLM assumptions.
- ANOVA as the simplest multivariate technique is quite robust against violations of its assumptions.
4. Linear multiple regression

Surely, the most often used statistical tool of biologists is regression analysis. Multiple regression is the generalization of simple linear regression where you try to predict one dependent variable from a second independent variable. In multiple regression you have again one dependent but now a set of n independent variables. In mathematical terms

\[ Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n = b_0 + \sum_{i=1}^{n} b_iX_i \]  

(4.1)

Recall that X and Y are series of observations. Hence, we can write them in vector notation

\[
\begin{bmatrix}
    Y_1 \\
    Y_2 \\
    \vdots \\
    Y_n
\end{bmatrix}
= 
\begin{bmatrix}
    1 & x_{1,1} & \ldots & x_{1,k} \\
    1 & x_{2,1} & \ldots & x_{2,k} \\
    \vdots \\
    1 & x_{n,1} & \ldots & x_{n,k}
\end{bmatrix}
\begin{bmatrix}
    b_0 \\
    b_1 \\
    \vdots \\
    b_n
\end{bmatrix}
\]

(4.2)

We solve this equation as before and take care of the ordering of elements.

\[ Y = (X'X)^{-1}X'Y \]

(4.3)

Let’s take a simple example. Three predictor (independent) variables X are assumed to influence the dependent variable Y. The complete multiple regression needs two lines of programming and the output gives us the parameters \( b_0 = 0.54, b_1 = -0.21, b_2 = 0.23, \) and \( b_3 = -0.006 \). Our regression model looks as follows

\[ Y = 0.54 - 0.21X_1 + 0.23X_2 - 0.006X_3 \]

The previous model used raw data. Now we take a slightly different approach. We standardize our variables and use the common Z-transformation. Recall that \( Z_i = (X_i - \mu)/\sigma \) with \( Z \) having a mean of 0 and a standard deviation of 1. The regression model looks as follows

\[ Z = b_0 + b_1Z_{X_1} + b_2Z_{X_2} + b_3Z_{X_3} + \ldots + b_nZ_{X_n} = b_0 + \sum_{i=1}^{n} b_iZ_{X_i} \]  

(4.4)
Data analysis and statistics

Applying equation 3 we get

\[ Z_Y = Z_X b \rightarrow Z_X 'Y = (Z_X 'Z_X)b \rightarrow b = (Z_X 'Z_X)^{-1}(Z_X 'Z_Y) \]  

(4.5)

How to interpret \(Z_X Z_X^\prime\) and \(Z_X 'Z_Y\)? Look at the definition of the coefficient of correlation:

\[ r = \frac{1}{n-1} \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \]

(4.6)

This is a very important equation that tells that a coefficient of correlation is the sum of all pairwise \(Z\)-values of \(X\) and \(Y\). Now look at the \(Z_X 'Z_Y\) matrix. This is nothing else that this sum calculated for all \(X\) variables. For the \(Z_X 'Z_X\) matrix holds the same. It is identical to the correlation matrix for all \(X\). Hence we can write

\[
\begin{bmatrix}
\frac{1}{n-1} \sum_{i=1}^{n} Z_{X1} Z_{X1} & \cdots & \frac{1}{n-1} \sum_{i=1}^{n} Z_{X1} Z_{Xk} \\
\vdots & \ddots & \vdots \\
\frac{1}{n-1} \sum_{i=1}^{n} Z_{Xk} Z_{X1} & \cdots & \frac{1}{n-1} \sum_{i=1}^{n} Z_{Xk} Z_{Xk}
\end{bmatrix} = \begin{bmatrix} r_{X1X1} & \cdots & r_{X1Xk} \\
\vdots & \ddots & \vdots \\
r_{XkX1} & \cdots & r_{XkXk}\end{bmatrix}
\]

(4.7)

From this identity we get a new simple equation for the multiple regression.

\[ \beta = R_{XX}^{-1}R_{XY} \]

(4.8)

where \(R\) denotes the respective correlation matrices. The Mathematica solution for the example in the previous table is shown below. Our model is

\[ Y = -0.71 - 1.08X_1 + 1.20X_2 - 0.03X_3 \]

This result based on Z-transformed data differs from the above that was based on raw data. In the latter case we got standardized correlation coefficients, so-called beta values.

Our multiple regression tries to predict the values of the dependent variable on the basis of the independent variables. However, the model does not tell anything about the goodness of fit. In ordinary regression this measure is the coefficient of determination \(R^2\). \(R^2\) is defined as the proportion of variance in \(Y\) that is explained by the model. Using the model with Z-transformed data we get
where \( k \) is the number of independent variables. Hence \( R^2 \) equals the sum of all Z-transformed predicted values divided through the number of cases \( n \). In a last step we need information whether the grand \( R^2 \), the model wide coefficient of variation and the beta values of each model variable explains a significant part of total variance. In order to do this we have two possibilities. First we can use standard F and t tests. Hence

\[
(4.9)
\]

\[
t(f) = \frac{\beta}{\text{Standard error of } \beta} \tag{4.10}
\]

with \( f \) being the degree of freedom. The standard error of beta is computed from

\[
\text{SE}(\beta_i) = \sqrt{\frac{r_{ii}(\Sigma^{-1})(1-R^2)}{n-k-1}} \tag{4.11}
\]

with \( r_{ii}(\Sigma^{-1}) \) being the element \( ii \) of the inverted covariance matrix \( \Sigma \) (as shown above) and \( n \) the number observations and \( k \) the number of independent variables in the model. \( n-k-1 \) denote the degrees of freedom of the t-test. The significance of the whole model (the significance of \( R^2 \)) is tested by an F-test.

\[
F = \frac{R^2}{(1-R^2)} \frac{n-k-1}{k} \tag{4.12}
\]

Both tests assume \textbf{normally distributed, homoscedastic errors} and are only approximations. Today standard errors are most often obtained from permutation tests. Hence the values \( X_i \) of the predictor variables are randomly reshuffled and after each such reshuffling a new multiple correlation is computed. This method gives a range of different beta values for each predictor variable. The distribution of values gives then the required standard error of beta.

The Table below presents a typical problem biologists are faced with. We got data about abundances of the gall inducing hymenopteran insect \textit{Andricus curvator}, a small cynipid wasp, which induces large galls in \textit{Torymus sp.} and \textit{Charips sp.} The table shows the correlation coefficients between the predictor variables and the response variable.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.572</td>
<td>0.000</td>
<td>0.000</td>
<td>0.041</td>
<td>0.690</td>
<td>0.842</td>
<td>1.841</td>
<td>2.227</td>
<td></td>
</tr>
<tr>
<td>0.983</td>
<td>0.304</td>
<td>0.038</td>
<td>0.646</td>
<td>1.211</td>
<td>0.443</td>
<td>1.143</td>
<td>3.245</td>
<td></td>
</tr>
<tr>
<td>0.654</td>
<td>0.961</td>
<td>0.875</td>
<td>1.948</td>
<td>1.294</td>
<td>0.585</td>
<td>1.456</td>
<td>2.181</td>
<td></td>
</tr>
<tr>
<td>0.579</td>
<td>0.066</td>
<td>0.023</td>
<td>0.230</td>
<td>1.439</td>
<td>0.680</td>
<td>0</td>
<td>2.074</td>
<td></td>
</tr>
<tr>
<td>0.329</td>
<td>0.291</td>
<td>0.266</td>
<td>0.623</td>
<td>1.263</td>
<td>0.471</td>
<td>1.901</td>
<td>1.908</td>
<td></td>
</tr>
<tr>
<td>0.335</td>
<td>0.097</td>
<td>0.061</td>
<td>0.292</td>
<td>0.836</td>
<td>0.619</td>
<td>1.257</td>
<td>2.321</td>
<td></td>
</tr>
<tr>
<td>0.295</td>
<td>0.139</td>
<td>0.019</td>
<td>0.278</td>
<td>0.476</td>
<td>0.015</td>
<td>1.245</td>
<td>1.476</td>
<td></td>
</tr>
<tr>
<td>0.404</td>
<td>0.761</td>
<td>0.433</td>
<td>1.583</td>
<td>1.240</td>
<td>0.556</td>
<td>1.669</td>
<td>1.715</td>
<td></td>
</tr>
<tr>
<td>0.506</td>
<td>0.129</td>
<td>0.100</td>
<td>0.307</td>
<td>0.885</td>
<td>0.444</td>
<td>1.728</td>
<td>2.527</td>
<td></td>
</tr>
<tr>
<td>0.127</td>
<td>0.606</td>
<td>0.452</td>
<td>6.147</td>
<td>0.860</td>
<td>0.932</td>
<td>0.192</td>
<td>0.727</td>
<td></td>
</tr>
</tbody>
</table>
the leaves of oaks. We are interested to see on which factors numbers of gall wasps depend. We got data about two competing wasp species, *Neuroterus albipes* and *Cynips quercusfolii*, and the parasitic wasps *Torymus* sp. and *Charips* sp. Additionally, we have data about precipitation and temperature during the sampling period. At last we speculate that the height of the galls (in m) and the abundance of *Andricus* in the previous generation might influence the actual gall densities. From this we make a model that contains one dependent variable, (*Andricus*) and seven independent variables. In total we have 10 observational data sets (cases) (Tab. 1).

Now we take a common statistical package and copy our data to the package spreadsheet and run the multiple regression option with all default settings. We get Tab. 2 and are lucky. The multiple regression points to highly significant predictor variables. Only temperature seems to be unimportant.

The output contains a number of important metrics. First it gives a multiple regression coefficient of correlation $R$, in analogy to the simple Pearson coefficient. It denotes the fraction of total variance explained by the model $R^2$, as well as a corrected level of variance that corrects for small sample size corr $R^2$ (correction for shrinkage after Carter). The significance level (the probability to make a type I error) is computed by an ordinary $F$-test. Next we have the predicted model parameters $B$ and the beta-values. Note that beta and $b$-values are connected by the following equation.

$$
\text{beta}(X_i) = B(X_i) \frac{s(X_i)}{s(Y)}
$$

(4.13)

where $s$ again denotes the standard deviation of variables $X_i$ and $Y$.

In the case of only two variables beta-values are identical with the Pearson coefficient of correlation. The output contains also standard errors of the parameters. Be careful, standard errors not standard deviations. All multivariate statistics give standard errors. For $t$-tests or other pairwise comparisons instead you need standard deviations. Don’t use the errors of regression analysis unconsciously for further analysis, for instance for tests whether coefficients of correlations or slopes of two models differ. Because standard errors are always lower than standard deviations you would get highly significant results even where no such are. The $t$-test at the end of the table tests whether the beta-values differ significantly from zero.

We should undertake a stepwise regression analysis. That means we should step by step eliminate all variables from our regression that are not significant at a given $p$-level (this is in most cases the 5% error level). Stepwise regression throws out Temperature and gives us a model where all independent (predictor) variables...
Data analysis and statistics

are significant at the 5% error level. It seems that we can prepare a paper.

Really? In this short example we made a whole set of errors. We applied a method without prior checking whether the prerequisites of the method are met. In fact, the data of Tab. 1 are nothing more than simple random numbers or combinations of random numbers. How is it possible to get highly significant results with them?

A multiple regression relies on the general linear model (GLM). Hence it must be possible to formulate the whole regression model as a linear combination of variables. The variables must be connected by linear regressions of the type $Y = aX + b$. We check this and don’t detect any significant deviation from linearity. In fact, however, I computed the previous generation data of *A. curvator* as a power function with random

---

**How to interpret beta-values**

An important but also difficult problem in multiple regression is the interpretation of beta-values. Remember that

$$beta(X_i) = B(X_i) \frac{s(X_i)}{s(Y)}$$

Remember also that the coefficient of correlation for a simple correlation between two variables is defined as

$$r = b \frac{s_x}{s_y}$$

It is immediately evident that beta-values are generalisations of simple coefficients of correlation. However, there is an important difference. The higher the correlation between two or more predicator variables (multicollinearity) is, the less will $r$ depend on the correlation between $X$ and $Y$. Hence other variables might have more and more influence on $r$ and $b$. For high levels of multicollinearity it might therefore become more and more difficult to interpret beta-values in terms of correlations. Because beta-values are standardized $b$-values they should allow comparisons to be make about the relative influence of predicator variables. High levels of multicollinearity might let to misinterpretations.

Hence high levels of multicollinearity might

- reduce the exactness of beta-weight estimates
- change the probabilities of making type I and type II errors
- make it more difficult to interpret beta-values.

There is an additional parameter, the so-called **coefficient of structure** that we might apply. The coefficient of structure $c_i$ is defined as

$$c_i = \frac{r_{iy}}{\sqrt{R^2}}$$

where $r_{iy}$ denotes the simple correlation between predicator variable $i$ and the dependent variable $Y$ and $R^2$ the coefficient of determination of the multiple regression. Coefficients of structure measure therefore the fraction of total variability a given predicator variable explains. Again, the interpretation of $c_i$ is not always unequivocal at high levels of multicollinearity.
offset from the *A. curvator* data. We did not detect this because we have only ten observations and the range of values is too small to detect deviations from linearity. This leads us immediately to a second major error. We have eight variables but only ten observations. The number of cases is much too low. A general rule is that the number of data sets should be at least 2 times the number of variables included in the original model. In our case we must reformulate our initial model to include at most four variables.

There is a simple equation to compute the optimal sample size \( n \) (although in reality we will seldom have the opportunity to have such large sample sizes. Instead, we will most often only deal with minimal sample sizes.

\[
n = \frac{L(1 - R^2)}{R^2}
\]

(4.14)

where \( R \) is the desired experiment wise coefficient of determination (explained variance). This value depends on what we intend to accept as a significant difference, the effect size. We know effect sizes already from bivariate comparisons and the discussion of the t-test. \( R^2 \) and effect size \( \varepsilon^2 \) are related by the following equation

\[
\varepsilon^2 = \frac{R^2}{1 - R^2}
\]

(4.15)

For \( \varepsilon^2 = 0.02 \) (week effect) \( R^2 = 0.02; \) \( \varepsilon^2 = 0.15 \) (medium effect) \( R^2 = 0.13; \) for \( \varepsilon^2 = 0.35 \) (strong effect) \( R^2 = 0.26 \) The \( L \) values that can be obtained from the following Table

<table>
<thead>
<tr>
<th>Variables</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L )</td>
<td>7.8</td>
<td>9.7</td>
<td>11.1</td>
<td>12.3</td>
<td>13.3</td>
<td>14.3</td>
<td>15.1</td>
<td>15.9</td>
<td>16.7</td>
<td>17.4</td>
<td>18.1</td>
<td>18.8</td>
<td>19.5</td>
<td>20.1</td>
<td>20.7</td>
<td>22.5</td>
<td>23.1</td>
<td>23.7</td>
<td>24.3</td>
<td>25.9</td>
</tr>
</tbody>
</table>

If we have for instance 5 variables in the model and want to have a multiple \( R^2 \) value of 0.95 we need 35 data sets to see a strong effect.

If we find non-linear relationships between variables we have to linearize them, for instance by appropriate logarithmization. But be careful. Logarithmization changes the distribution of errors around the mean. The GLM but relies on the assumption that the errors are distributed normally around the mean, and that they are not correlated with the dependent variable. They have to be homoscedastic. Otherwise we speak of heteroscedasticity.

A good example is Taylor’s power law. Insects fluctuate in abundance. Assume we study ten species with different upper limits of abundance, different carrying capacities. Insight these boundary abundances fluctuate at random. This is a typical statistical process called proportional rescaling. For such a process it is easy to show that mean and variance are connected by a simple equation

\[
\text{Variance} \propto \text{mean}^2
\]

(4.16)

Read: The variance is proportional to the square of the mean. Hence the variance, the distribution of errors around the mean, rises linearly with the squared mean. Mean and variance are correlated. This violates the assumptions of the GLM. Fortunately, the model is very robust against this type of correlation. The B– and beta values are nearly unaffected but the error levels \( p \) might be distorted.

Next the errors around the mean must not be correlated with each others. They must not be autocorrelated. In non-linear relationships variances and means are often correlated. Be careful at high levels of autocor-
relation. All statistical packages contain a simple test for autocorrelation, the **Durbin-Watson test for serial correlation** $D$. Without going into the details I tell that $D$-values around 2 indicate low levels of autocorrelation. In our example $D$ is 2.93. This high value points to severe autocorrelation and we have to check our data which of the variables causes this problem. We have to eliminate the variable from the initial model. Another method to detect violations of the GLM assumptions is a visual inspection of the residuals. **Residuals** are the deviation of the data from the regression line and are often used for further analysis. Residuals should be normally distributed around the mean (the value of the regression line). Any regularities in the distribution of residuals hence point to **heteroscedasticity** or **autocorrelation** and therefore to an incorrect use of multiple regression. However, as already stressed, multiple regression is a quite robust method and only severe distortions of the residual distribution make the results unreliable.

It’s still unclear why we got highly significant results from random data. The first answer is that some of the predictor variables were highly correlated. Height and *Torymus* were highly correlated ($R^2 = 0.99$). This violates the assumption of the GLM. Although multiple regression is a technique designed to detect interrelations between sets of predictor variables **pairwise correlations must not be too high.** In other words they must not explain the same part of variance. If they were perfectly correlated a multiple regression would be impossible. Technically speaking the matrix of correlation coefficients would be singular and the inverse would not be defined. Including such variables would also give evidence of an ill designed model. That something is wrong with the variables Height and *Torymus* is also indicated by the values of beta. As coefficients of correlation, beta-values should range between $-1$ and $1$. The above equation however tells that absolute values larger than one are possible. However, **very high or low values of beta often indicate some violation of the GLM assumptions.**

A next answer lies in the distinction between **test wise** and **experiment wise error rates.** What does this mean? If you test whether single variables are significant or compare two means you perform a test and have a probability to make a type I error (your significance level). However, if you test a whole model (the outcome of an experiment) you have to perform a series of single tests and you get a type I error level that refers to the whole experiment. In our case we got 8 test wise error levels and one experiment wise multiple $R^2$ with $p(t) < 0.051$. If you accept a single test at $p < 0.05$ and perform $n$ tests your probability to get at least one significant result is

$$p = 1 - (1 - 0.05)^n$$

(4.17)

Hence, for $n = 7$ you get $p = 0.30$, a value much higher than 0.05. This means for our example that the probability to get at least one significant result out of seven predictor variables is nearly 30%. In a multiple regression this value is even too low because you have to consider all combinations of variables. Additionally, the lower the number of observations is, the higher is the probability for a significant result (why?). Hence to be sure to make no error you have to reduce your test wise error rate to obtain an experiment wise error rate of less than 0.05. The most easiest way to do this is to use a **Bonferroni correction** obtained from the first element of the Taylor expansion of eq. 4.17. You simply divide your test wise error rate $\alpha$ by the number of variables in the model

$$\alpha^* = \frac{\alpha}{n}$$

(4.18)
In our example of the first table you should accept only variables with \( p < 0.05 / 8 = 0.006 \). Hence, only the previous generation of \( A. \text{curvator} \) remains in the model. This correction should always be used in multivariate statistics.

Let us at the end assume that the data of Tab. 1 were real. We can then redefine our initial model to perform a correct multiple regression. First, we have to reduce the number of variables. By this we also define a logical model that we want to test with the regression analysis. Our data contain three groups of variables. \( \text{Torymus} \) and \( \text{Charips} \) are parasitoids or inquilines of \( \text{Andricus} \). We simply combine both variables and use a new variable 'parasitoid abundance' as the sum of both. By this we also reduce the degree of multicollinearity in the data set because \( \text{Torymus} \) and \( \text{Charips} \) abundances are highly correlated. We also have two variables that describe weather conditions. They are not correlated. We define the precipitation / temperature ratio \( P/R \), the hydrothermic coefficient, as a new variable that describes moisture. Low values indicate rather dry periods, high values moist conditions. Height was highly correlated with \( \text{Torymus} \). At the moment we have no good explanation for this pattern and skip the variable height. We would have to undertake further experiments to clear this point.

With this variables we undertake a new regression analysis and find only the competitors to be significant. The Durbin-Watson statistics gives a value of \( D = 2.07 \). We do not expect higher levels of autocorrelation. We drop the other variables and get at the end a simple linear regression between \( \text{Andricus} \) and its competitors \( (\text{Andricus} = (0.39 \pm 0.09)\text{Competitors} - (0.33 \pm 0.19)) \); \( R^2_{\text{corr.}} = 0.69; p < 0.002 \). Note that the errors are standard errors not standard deviations. The p-value is smaller than our Bonferroni corrected acceptable error level of 0.006 and we might accept the hypothesis about a close relation between \( \text{Andricus} \) and \( \text{Cynips} \) abundances. But be careful. Both data sets consist of pure random numbers. I needed only five reshuffles to get the data matrix for this result. The example shows how easy it is to get highly significant results from nothing.

With this variables we undertake a new regression analysis and find only the competitors to be significant. The Durbin-Watson statistics gives a value of \( D = 2.07 \). We do not expect higher levels of autocorrelation. We drop the other variables and get at the end a simple linear regression between \( \text{Andricus} \) and its competitors \( (\text{Andricus} = (0.39 \pm 0.09)\text{Competitors} - (0.33 \pm 0.19)) \); \( R^2_{\text{corr.}} = 0.69; p < 0.002 \). Note that the errors are standard errors not standard deviations. The p-value is smaller than our Bonferroni corrected acceptable error level of 0.006 and we might accept the hypothesis about a close relation between \( \text{Andricus} \) and \( \text{Cynips} \) abundances. But be careful. Both data sets consist of pure random numbers. I needed only five reshuffles to get the data matrix for this result. The example shows how easy it is to get highly significant results from nothing.

Multiple regression uses ordinary least squares and relies on coefficients of correlation as distance measure. But recall the basic model to calculate the regression

\[
\beta = R^{-1}_{XY} R_{XX}
\]

(4.19)

The model contains two association matrices. We might use other measures of distance to get different regression models. For instance look at the next table. It gives the occurrences of 12 species at five sites. This is a presence-absence table. Does the occurrence at sites two to five predict the occurrence at site one? We use the Soerensen index and get the respective association matrix. With this matrix we run a multiple regression and get the respective beta values and the inverse of \( R \). This is enough to calculate the coefficients of determination and significance levels using equations 8 to 11. The matrix below shows the results and the calculations. We see that none of the single predictors is statistically significant. The reason for the low t-values is that the numbers of species (cases) used is too low. The regression used in this way has a week power that means it too often fa-
vours the null hypothesis of no influence. For a sound regression model we need at least 50 species. The whole model, in turn, appears to be significant before Bonferroni correction. The Bonferroni corrected significance level would be \( p(F) = 0.03 \times 5 = 0.15 \). Hence despite the high \( R^2 \) value of 0.82 (calculated from the beta values and the respective association coefficients according to eq. 4.9) we infer that the predictor variables do not allow for a prediction of site occurrences at site one. Again the reason for this discrepancy is the too low number of cases.

We used a multiple regression approach of presence absence data associated with the Soerensen index. Important is that we run a multiple regression of nominally scaled data. Indeed, we can even used mixed models with nominally, ordinary and metrically scaled variables. Important is that we have to be able to use the same association measure for all of these data types. Often we have to apply a previous classification. Important is also that we do not get regression equations to predict occurrences. We only infer significant influences.

A second way to assess single dependencies in a system of interdependent variables is the computation of so-called partial correlations. These are linear correlations between the \( X \) and \( Y \) variables where the influence of a third or more variables has been eliminated. Assume you have 3 variable \( X, Y, \) and \( Z \) (Fig. 4.1). The relations between them are represented by the following path diagram. \( Y \) depends on \( X \) and \( Z \) and \( X \) depends on \( Y \). If we compute a simple regression to infer the influence of \( X \) on \( Y \) we would ignore the common influence of \( Z \) on both variables. A partial correlation eliminates this influence. It is calculated from the co-variance of the residuals of the variables invoked. Suppose we want to infer the relationship between \( X \) and \( Y \) without the effect of the third variable \( Z \). We start with the following model

\[
X = X (Y) + X (Z) \\
Y = Y (X) + Y (Z)
\]

hence we assume that the variability in \( X \) results from an additive process where both variables \( Y \) and \( Z \) contribute.

If we now regress \( X \) on \( Z \) and \( Y \) on \( Z \) we get the part of the variability explained by \( Z \) and parts (the residuals) that are not explained. The correlation between these residuals is the partial correlation we are look-
Hence the partial correlation is a correlation of residuals (Fig. 4.2). The partial correlation between X any Y after eliminating the influence of Z is

$$r_{X/Y|Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r_{YZ}^2}}$$

(4.20)

This procedure can be extended step by step to eliminate the influence of other variables. At the end we get a set of correlation coefficients that describe the influence of each of the predictor variables Xᵢ on the dependent variable Y. A semipartial correlation correlates a variable with one residual only. Hence the semipartial correlation between X and Y might use the residuals of the regression between X and Z and correlates them with Y. It is calculated from

$$r_{(X|Y)Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r_{YZ}^2}}$$

(4.21)

The semipartial correlation asks whether a third variable influences the dependence of a variable Y on X via X. Semipartial correlations are therefore used to infer the importance of the single predictor variables in multiple regression. The residuals of X after eliminating the of predictors are correlated with the dependent variable Y. The higher this semipartial correlation is, the higher is the influence of X on Y.

To test whether a partial correlation is statistically significant or whether two coefficients differ we transform the coefficients via Fisher’s Z transformation for correlation coefficients. The value

$$Z = (Z_{\text{Fisher},1} - Z_{\text{Fisher},2})\sqrt{n - k + 1}$$

(4.22)

is the approximately normally distributed and can be compared with the standard normal.
Advices for using multiple regression

- First of all, multiple regression is a tool for testing predefined hypotheses. It is not designed for hypothesis generating. Hence the first thing you must do is to formulate a sound model of how your variables might be related. A blind pattern seeking will often lead to erroneous results because with a sufficient number of variables and small sample sizes you will very often if not always find some ‘significant’ dependencies.
- Think hard about cause and effect. In many cases it is not obvious which variable is the dependent and which the independent. In many cases a path diagram helps in formulating logical hypotheses.
- All variables must have a metric scale.
- Prior to computation check the correlation matrix of all single correlations. Eliminate independent variables that will not explain any part of total variance because they fully or nearly fully depend on another variable (multicollinearity). As a rule eliminate them if they are correlated with another independent variable with $r > 0.9$.
- Check for non-linear dependencies and try to eliminate them. Often logarithmic rescaling of some or all variables gives better results.
- In cases of heavy violations of the assumptions of the GLM use non-parametric regression. In its simplest form use ranked variables.
- The number of observations $N$ must be sufficiently large. $N$ should be at least twice the initial number of variables in the model.
- After computing the regression check for autocorrelation (using the Durbin-Watson static; $D$ should be between 1 and 3) and inspect visually the distribution of residuals. These should not show any regularities. Patterns in the distribution of residuals indicate violations of the general linear model.
- At small data set size use always Bonferroni corrected significance levels
- Use a stepwise analysis and eliminate in a stepwise manner all variable that do not significantly explain parts of total variance.
- Check the resulting multiple regression equation whether it is sound and logical.
- Try to check the resulting equation using other data. This leads often to an improvement of the original model.
- If you have a large number of observations (data sets) divide the data set prior to computation at random into a part with which you generate your multiple regression equation and a part with which to test this equation. If your test leads to unsatisfactory results you have to reformulate the whole model and start again. This means often that you have to include other previously not considered variables.
- A last step should be the verification of the model by other independent data sets.
5. Logistic and other regression techniques

Bivariate or multiple regression needs continuous variables to be run. But often our data are of a discrete type or we have mixed data. Is it possible to run a regression with such data types? Look at the next Figure 5.1. Four species of closely related plants were studied at five stands (the colored bars) and their aboveground biomasses compared. An analysis of variance resulted in a highly significant difference ($F = 58$, $p(F) < 0.001$). But we can also try a simple linear regression between species (coded from 1 to 4) and biomass. We get a sufficiently good regression. But this regression model would not be able to predict a species from its body weight. A body weight of 2.5 would point either to species B or C. We might try to solve the above regression equation for $x$ and get $x = (2.0 + 1.05)/1.42 = 2.5$. This is exactly between 2 and 3. From the regression we are not able to decide unequivocally whether certain biomasses point to one or another species.

But we can solve our problem. This statistical solution is the so-called **logistic** or **logit regression**. The logic behind this type of regression is very simple and best explained by the next example. In archaeological research it is sometimes difficult to separate the sexes from cranial or postcranial skeleton rudiments because for most traits sexes highly overlap. Assume, for instance, that we have a series of skeleton fragments and want to infer the sex. To do this we first have to have a model. We generate such a model from a series of skeleton samples with known sex and perform a multiple regression. But our dependent variable, the sex, has only two values, male and female. Multiple regression, however, assumes the dependent variable to be continuous. We have to transform it and we do this via the already known logistic function. The logistic function has the form

$$Z = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$  \hspace{1cm} (5.1)

Our multiple regression has the general form

$$Y = a_0 + \sum_{i=1}^{n} a_i x_i$$

To derive an appropriate model we start with odds, that means the quotient of the probabilities $O = p/(1-p)$. Odds have values from 0 to $\infty$. $\ln(O)$ in turn goes from $-\infty$ to $\infty$. Now we use a simple multiple regression to estimate $\ln(O)$ from or predictors. Hence

$$\ln \left( \frac{p}{1-p} \right) = a_0 + \sum_{i=1}^{n} a_i x_i \rightarrow \frac{p}{1-p} = e^{a_0 + \sum_{i=1}^{n} a_i x_i} \rightarrow p = \frac{e^{a_0 + \sum_{i=1}^{n} a_i x_i}}{1 + e^{a_0 + \sum_{i=1}^{n} a_i x_i}}$$

We got the general **logistic regression** equation where logit stands for log odds.

$$Z = \frac{e^{a_0 + \sum_{i=1}^{n} a_i x_i}}{1 + e^{a_0 + \sum_{i=1}^{n} a_i x_i}}$$  \hspace{1cm} (5.2)
The Figure 5.2 shows how the logit regression transforms the y value into a range between 0 and 1. We see a threshold where an unequivocal decision is not possible but for most regression values we should be able to decide whether a given set of morphological traits stems from a male or a female skeleton. The next table shows values of morphological traits together with the sex of the person from whom these values were measured. Statistica computes for this the following regression function

\[ Z = \frac{e^{0.19 + 0.2A - 6.36B + 1.77C}}{1 + e^{0.19 + 0.2A - 6.36B + 1.77C}} \]

The result is shown in Fig 5.3. Of 20 samples 14 allow an unequivocal decision to be make. In five cases a clear decision is impossible and there is one obviously wrong result. From this we conclude that the power of our regression model is about 85%. If we now have a set of samples from a skeleton of unknown sex, we might use our regression model to establish the sex. The probability that we choose the right sex is about 85% (17 / 20).

The logit transformation transforms the data into the range between 0 and 1. This is the same range a probability can take. We might therefore interpret the result of a logit regression as a probability that a certain set of data belongs to one of both types of Y.

Fig. 5.4 shows a comparison of our logit result with an ordinary multiple regression. We see that the multiple regression to a certain extent also finds out males and females. But the result is much less decisive than before. There are only seven unequivocal decisions, four results are erroneous. We also see that the errors in both models differs.

An alternative to logit regression is probit regression (= probability unit). Again a categorized dependent variable Y is transformed in such a way to be appropriate for multiple regression. However in this case the inverse cumulative standard normal is used. Hence the probit model fits the following function

\[ p(x) = \text{Norm}(0,1) = \text{Norm}(a_0 + \sum a_i x_i) \]

In other words we use the cumulative standard normal to fit the regression model. Logit and probit transformations are very similar and the respective regression models give also very similar results. Due to its reference to odds and its ease calculate logit regression has become the standard.

Logit and probit regression can be viewed as two cases of a general logistic model. This can be stated as
Data analysis and statistics

This model contains apart from the independent variable coefficients two additional parameters \( b_0 \) that sets the upper limit and \( b_1 \) that defines the shape. Such a logistic regression model is often fit to growth curves.

Another special regression model that is used in pharmacology to describe drug responsiveness is a modification of the logistic

\[
Y = b_0 - \frac{b_0}{1 + \left(\frac{X}{b_1}\right)^{b_2}}
\]

This model is quite flexible (Fig. 5.5) and its parameters have specific interpretations. \( b_0 \) is the maximum response at dose saturation, \( b_1 \) is the concentration that produces a half maximum response and \( b_2 \) determines the slope of the function, that means it is a measure how fast the response increases with increasing drug dose.

The previous models are **intrinsically non-linear regression models**. Other models are **intrinsically linear**. These are model that can be linearized.

An intrinsically linear model is for instance a polynomial model of the form

\[
Y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \ldots + a_k y + a_{k+1} y^2 + a_{k+2} y^3 \ldots
\]

Such a model can easily be solved using matrix notation

\[
Y = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & y_1 & y_1^2 \\ 1 & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & y_n & y_n^2 \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_k \end{bmatrix}
\]

Other intrinsically linear models are logarithmic, exponential, or power function models. These types of models can be solved using either the linearized model an linear regression or non-linear estimation using the non-linear function.

For instance a power function can be fitted from the linearized version. In this case we fit the model

\[
y = a X^z \rightarrow \ln y = \ln(a) + z \ln(x) \\Rightarrow \\
\ln y = \ln(a) + z \ln(x) + \varepsilon \rightarrow y = ax^z e^\varepsilon
\]

In this model the error \( \varepsilon \) is multiplicative. Using non-linear estimation we fit

\[
y = ax^z + \varepsilon
\]

In this model the error term is additive. Both regression model result in different parameter values.

The solution of intrinsically non-linear models needs two steps. A search algorithm for parameter values and a method to assess the goodness of fit of a particular step. Hence non-linear estimation starts from the initial parameter settings and modifies step by step these parameter values until the goodness of fit measure
reaches a minimum. In any case we need some hypotheses about the underlying relationships. We need a precise model to be tested. We then test whether our data fit to this model or not. This can be done by least square statistics, by a Kolmogorov-Smirnov test or by some other tests. Most often we will have some hypotheses about the model and we are interested in the parameters that define this model. Let’s look at the next example (Fig.5.5). We studied a population of bacteria in the presence of a limiting resource. Therefore, we expect that there is an upper boundary of population size and our null model is that the process follows a Lotka - Volterra or logistic growth model. We know already that the general equation for such a process is

\[ N = \frac{a}{1 + be^{-t}} \]  

(5.6)

Our question is now whether the observed data follow such a model. We do not know the parameter values of \(a\) and \(b\). A Kolomogov - Smirnov test is not applicable. We shall ask our statistic program. For instance, the *Statistica* module *Nonlinear Estimation* produces after the input \(y = a/(1+b\exp(-x))\) and using the quasi Newton approximation a function that is shown in Fig. 5.6. To fit a model to given data, the program starts with an initial set of parameter values and changes them according to a predefined algorithm (the best known algorithm is probably the Newton approximation). The fitting process stops when the least square test variable reaches a minimum. In our case I did not use real data but randomized data generated with the above logistic model and the parameters \(a = b = 100\). The *Statistica* approximation is therefore very good and the program predicts a coefficient of determination \(R^2 = 0.84\).

However, be careful. Our fit is so good because we used the correct model. If we do not know what model might apply, data fitting might be highly misleading. But even if we know the correct model we may go wrong if we start with wrong initial parameter settings. Look at Fig. 5.6. The blue data points were computed via a randomization from a general Michaelis - Menten model of the form

\[ N = \frac{at}{b + t} \]  

(5.7)

A very simple equation. *Statistica* fits with the initial settings \(a = 1000\) and \(b = 0.001\) and the quasi Newton method the curve shown in Fig. 5.7. This is in essence no fit. What has happened? We simply started with the wrong initial settings. Starting with reasonable settings \((a = 1, b = 1)\) gives after a few iterations a very good
data fit (Fig. 5.8). The initial settings to generate the data points were \( a = 10 \) and \( b = 5 \). So, data fitting is a sensitive tool that has to be used with caution and with a good deal of a priori knowledge about realistic model types and parameter values.

**Path analysis and linear structure models**

Multiple regression requires a predefined model that shows relationships between variables. At least it must be possible to define dependent and independent variables. How do we know which of the variables is dependent? We need some prior knowledge or assumptions about causal relations. With four predictor variables our model looks as in Fig. 5.9. \( E \) denotes the unexplained part of total variance.

Path analysis tries to generalize our assumptions about causal relationships in our model. Now we also try to establish the causal relations between the predictor variables. We define a whole model and try to separate correlations into direct and indirect effects. A path analytical model might look as follows: \( X_1 \) to \( X_4 \) are the predictor variables, \( Y \) is the predicant (the dependent variable) and \( e \) denote the errors (the unexplained variance) of each single variable (Fig. 5.10). The aim of path analysis is now to estimate the correlation coefficients between all variables and the amounts of unexplained variances \( e \). By modifying the path analytical model the method tries to infer an optimal solution where a maximum of total variance is explained. **Hence, path analysis tries to do something that is logically impossible, to derive causal relationships from sets of observations.**

Path analysis might be a powerful tool. However, it demands very thorough model development and large data sets. The smaller a data set is the higher is the probability to get a significant global solution together with significant partial correlations. Path analysis is also not unequivocal. Very often you find more than one solution that fit well to the data.

Path analysis is largely based on the computation of partial coefficients of correlation. But you have to know which variables have to be eliminated. This requires thorough prior modelling efforts. Assume the following case with 4 variables. We model the interrelationships between these variables through the following path diagram. The aim of the path analysis is now to infer the relative strength of correlations between these variables. This could be done by computing the appropriate partial correlations. Path analysis instead uses standardized variables instead of raw data. That means all variables are Z-transformed having means of zero and variances of 1. The fundamental theorem of path analysis tells now that it is always possible to compute the correlation coefficients of standardized variables from the total set of simple correlation coefficients. In path analysis these correlation coefficients of standardized variables are called **path coefficients** \( p \) (Fig. 5.11).

We see that the results of path analysis can’t be better than the underlying model. **Path analysis is a model confirmatory tool. It should not be used to generate models or even to seek for**
models that fit the data set.

How to compute a path analysis? The model above contains four variables that are presumably connected. We get three regression functions

\[
\begin{align*}
W &= p_{xw} X + e \\
X &= p_{xy} Y + e \\
Z &= p_{zx} X + p_{zy} Y + e
\end{align*}
\]

\[-p_{xw} X + W - e &= 0 \\
X - p_{xy} Y - e &= 0 \\
-p_{zx} X - p_{zy} Y + Z - e &= 0
\]

This is a system of linear equations. We need a set of data for each variable and the equations are then solved by linear regression. However, we can make things easier. The input values in path analysis are already Z-transformed variables. Hence

\[
\begin{align*}
Z_w &= p_{xw} Z_X + e \\
Z_x &= p_{xy} Z_Y + e \\
Z_z &= p_{zx} Z_X + p_{zy} Z_Y + e \\
\downarrow \\
Z_w Z_v &= p_{xw} Z_X Z_Y + e Z_Y \\
Z_x Z_w &= p_{xy} Z_X Z_Y + e Z_w \\
Z_z Z_w &= p_{zx} Z_X Z_w + p_{zy} Z_Y Z_w + e Z_w \\
Z_x Z_z &= p_{xy} Z_X Z_Y + e Z_X \\
Z_x Z_Y &= p_{xy} Z_X Z_Y + e Z_Y \\
Z_z Z_Y &= p_{zx} Z_X Z_Y + p_{zy} Z_Y Z_Y + e Z_Y \\
\downarrow \\
r_w &= p_{xw} r_{xy} \\
r_{zw} &= p_{xyz} r_{xy} \\
r_z &= p_{zx} r_{xw} + p_{zy} r_{yw} \\
r_{xz} &= p_{xy} r_{yx} \\
r_{xy} &= p_{xy} \\
r_{zy} &= p_{zx} r_{xy} + p_{zy}
\end{align*}
\]

In a first step we multiplied the three equations to get all combinations of \(Z_{nm}\). Now remember equation 11.6. The sum of all \(Z_{nm} Z_{om}\) gives \(r_{mn}\), the coefficient of correlation. Summing over all cases gives a relation between path coefficients and coefficients of correlation. Because a simple sum of all Z-values is zero the error terms vanish and we have systems of linear equations that can be solved using ordinary matrix algebra. At the end we get a path diagram as shown in Fig. 5.12.

Linear path analysis cannot only handle with observable or observed variables. It can also be used to deal with complex and unobservable, so-called endogen, variables (Fig. 5.13). Look at the next example. Six exogenous measured variables can be divided into two groups. \(X_1\) to \(X_4\) are related and contain a common source of variance. This variance can be identified and we define a model with the intern endogen variable \(Y_1\). In the same way define \(X_5\) and \(X_6\) a variable \(Y_2\). \(Y_1\) and \(Y_2\) explain a part of the variance of the variable \(Z\). All
relationships can be defined by path coefficients. In this case we use path analysis for a whole model with measurable and latent variables. The latter type of variable can be identified by factor analysis, with which we deal later. To do so we need linear structure models and this technique is best known under the name LISREL (Linear structure models).

In the biological sciences this technique is seldom used, mainly due to the many difficulties in interpreting the results. It needs even more thorough a priori modelling effort. In biology the number of variables is in most cases rather limited (less than 10). For \( n = 10 \) we need \( 10 \times (10-1)/2 = 45 \) correlation coefficients. For estimating these at low error levels we need at least 10 independent observations for each variable pair. In praxis for even moderate large models hundreds of data sets are necessary. Therefore biologists more often rely on multiple regression in combination with factor analysis than on LISREL. In **Statistica** the SEPATH module is a powerful tool for developing structured models.

**Assessing the goodness of fit**

Look at the following plot. Our problem is to fit a model to a set of 15 observations. The question is what model fits best. Spreadsheet programs offer some simple build in functions (a power, a logarithmic, an exponential, and linear and higher order algebraic functions (up to the 5th order). Often we need other models and then we have to use the non-linear regression modules of statistic packages. They use least square algorithms to fit self-defined models. However, how to assess what model fits best? Of course we might use the fraction of explained variance, the **coefficient of determination** \( R^2 \). In the example of Fig. 5.14 we get: linear \( R^2 = 0.71 \); quadratic \( R^2 = 0.83 \); power function \( R^2 = 0.82 \). Which model fits best? We can’t decide. Judged only by the \( R^2 \) values the quadratic might look best. In fact, the data points were generated by a linear model that fitted worst. This is an often met problem and in fact there is no general solution to the problem of finding the best model. \( R^2 \) is not a very sensitive measure of goodness of fit because it is easy to get high values with different models. If possible we have to enlarge the number of data points.

Even worse is the situation if we have to compare multivariate models. For instance we have to compare results of multiple regressions that contain different numbers of predictor variables. Of course the higher the number of predictors the higher will be the variance explained. Thus, the problem is to find the optimum number of predictors. One criterion in multiple regression is that we should eliminate all variables that do not significantly contribute to the overall \( R^2 \) value. In our case we can only leave the quadratic or the linear term
in the quadratic model. To decide how many parameters to leave we can use the so-called Akaike information criterion (AIC). It gives the amount of information explained by a given model and is defined as

$$AIC = -2 \ln(L) + 2K$$

(5.8)

where $L$ is the maximum likelihood estimator of goodness of fit and $K$ the number of free parameters of the model. In the example above the linear and the power function models have two free parameters. The quadratic model has three free parameters. $L$ can be any appropriate measure of fit. Often $R^2$ is used. $K$ is often identical with the number of free variables (predictors) of the model. The lower AIC is, the better fits a given model to the data. For comparison of two models we can take the difference $|AIC_1 - AIC_2|$. This difference is normally distributed and we assume therefore that differences are significant at $p < 0.05$ if $\Delta AIC$ is larger than 1.96. For the example above we get $\Delta AIC$ (quadratic - power) = 1.96. $\Delta AIC$ (linear - power) = 0.27. Therefore we interpret that the power function with only two free parameters fits not significantly better than the linear model.

Another important criterion is the Bayesian information criterion (BIC) (sometimes termed Schwarz criterion)

$$SC = N \ln(L) + K \ln(N)$$

(5.9)

where $N$ is the number of data points. This models points in our case unequivocally against the quadratic function. $\Delta BIC$ (quadratic - power) = 2.96. $\Delta BIC$ (linear - power) = 0.27.

An important point is that AIC or BIC must not be used in parallel with $R^2$ as the measure of goodness of fit. Both methods rely on different philosophies. For instance in multiple regression variable reduction via AIC and via $R^2$ will result in different final models. $R^2$ of the AIC model is now not longer a measure of explained variance.

For detailed information about these two criteria and other less often used see http://www.aps.uoguelph.ca/~lrs/ANSC637/LRS16/.
6. More on variance analysis

A simple ANOVA is easy to compute. However, variance analytical techniques can be greatly extended. A first extension regards the inclusion of metric and categorical variables. In this case we perform a **covariance analysis** (ANCOVA). In an ANCOVA we use the residuals of a multiple regression of a dependent variable on independent metrically scaled variables (the **covariates**) as the test variable in a ordinary ANOVA. This is shown in the next plant example. We studied plant growth on different soils in dependence on light intensity (low/high). A simple ANOVA pointed indeed to Light as a factor that influenced plant growth. We further assume that growth is also dependent on the amount of nutrients in the soil. Fig. 6.1 shows indeed a general regression of growth on nutrients (in black). However for each light intensity class there is a different regression (the colored lines). Now we calculate the residuals of the single regression and the different regressions of each treatment. However using the different regressions of each treatment to calculate the residuals would include the effect of the treatment through the backdoor. Indeed the respective ANOVA gives $F = 0$ in the *Statistica* example on the next page.

The ANCOVA with nutrients as covariate now tells that light has only a marginal influence on growth after correcting for the effect of nutrients. Using the residuals of the grand regression between nutrients and growth as the dependent variable in a simple ANOVA gives nearly the same result. The different significance levels result from the different degrees of freedom. The simple ANOVA doesn't know anything about the variable nutrient and $df_{\text{error}} = 27$. In the ANCOVA one df goes to nutrients and $df_{\text{error}} = 26$.

The next example shows a bifactorial ANCOVA of the same plant example. After regressing growth against nutrients light intensity remains a significant predictor of plant growth. If we first regress growth on nutrients and use the residuals (ZMN4 in the *Statistica* example) in a simple ANOVA we...
get nearly the same result. Again the different significance levels stem from
the different degrees of freedom. Covariance analysis has the same limitations
and prerequisites than multiple regression and ANOVA. Both rely on GLM.
The covariates have to be metrically scaled.

Another special case of ANOVA regards repetitive designs. For instance in
medical research we test patients before and after medical treatment to infer
the influence of the therapy. In this case we have to divide the total variance
($SS_{\text{total}}$) in a part that contains the variance between patients ($SS_{\text{between}}$) and
within the patient ($SS_{\text{within}}$). The latter can be divided in a part that comes
from the treatment ($SS_{\text{treat}}$) and the error ($SS_{\text{error}}$) (Fig. 6.2). Hence we have
one variance more to consider, the a priori differences between the patients
that are not influenced by the medical treatment. Assuming a monofactorial
design we calculate the necessary sums of squares from

\[
\begin{align*}
\text{SS}_{\text{total}} &= \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{x})^2 \\
\text{SS}_{\text{between}} &= k \sum_{i=1}^{n} (\bar{P}_{i} - \bar{x})^2 \\
\text{SS}_{\text{within}} &= \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{P}_{j})^2 \\
\text{SS}_{\text{treat}} &= n \sum_{j=1}^{k} (\bar{T}_{j} - \bar{x})^2 \\
\text{SS}_{\text{error}} &= \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{P}_{i} - \bar{T}_{j} + \bar{x})^2
\end{align*}
\]  

(6.1)

where \(x_{ij}\) is the measurement of patient \(i\) within treatment \(j\), \(\bar{x}\) is the grand mean, \(\bar{T}\) is the mean of the \(k\) treatments, and \(\bar{P}\) the mean of the \(n\) patients. \text{SS}_{\text{total}}\) is of course the sum of these sums of squares.

The degrees of freedom come from

\[
\begin{align*}
df_{\text{total}} &= df_{\text{between}} + df_{\text{within}} = df_{\text{between}} + df_{\text{treat}} + df_{\text{error}} \\
k \cdot n - 1 &= n - 1 + n(k - 1) = n - 1 + k - 1 + (n - 1)(k - 1)
\end{align*}
\]

To test for significance we apply again the F test with \((k-1)\) and \((n-1)(k-1)\) degrees of freedom

\[
F = \frac{\text{SS}_{\text{treat}} \cdot df_{\text{treat}}}{\text{SS}_{\text{error}} \cdot df_{\text{treat}}} = \frac{n \sum_{j=1}^{k} (\bar{T}_{j} - \bar{x})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{P}_{i} - \bar{T}_{j} + \bar{x})^2} \frac{(n-1)(k-1)}{k - 1}
\]

(6.2)

The Excel example at the next page shows how to calculate a repeated measures ANOVA. The blood pressure of ten test persons were measured in the morning, in the afternoon, and in the night. The question is whether blood pressures depend on the time of day. We first calculate \text{SS}_{\text{within}}\) from the person means and the grand mean. Next we calculate \text{SS}_{\text{within}}\) from the single measurements and the person means. \text{SS}_{\text{treat}}\) comes from the mean of the three day times and the grand mean. Lastly, \text{SS}_{\text{error}}\) comes from the measurement and all three means. We see that indeed \text{SS}_{\text{total}} = \text{SS}_{\text{within}} + \text{SS}_{\text{between}}\) and \text{SS}_{\text{within}} = \text{SS}_{\text{treat}} + \text{SS}_{\text{error}}\).
Used in this way makes the repeated measure ANOVA very easy to calculate even with Excel. The significance test reduces to

\[
F = \frac{SS_{\text{treat}}}{SS_{\text{error}}} \frac{df_{\text{error}}}{df_{\text{treat}}} = \frac{n(n - 1)\sum_{j=1}^{k} \overline{T}_j^2}{\sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \overline{T}_j)^2}
\]

Of course, the above ANOVA scheme can be extended to many treatment variables.

The validity of the F test for the previous ANOVA depends on a series of prerequisites. Particularly the within variances have to be similar (homogeneous). Further, in repeated designs treatments are often correlated. In the tables above morning might correlate with evening and night because the test persons react in the same way on the time of day. These correlations have to be similar for all combinations of treatment. Otherwise the treatment would be inhomogeneous. Inhomogeneities result often in inflated type I error probabilities. To cor-
rect for this possibility we might **correct the degrees of freedom** to obtain lower F-values and significance levels. In eq. 6.2 we have (k-1) degrees of freedom for the effect and (n-1)(k1) degrees of freedom for the error. To correct for violations of the AOVa prerequisites we multiply these degrees of freedom by a factor \( \varepsilon \). At maximum heterogeneity

\[
\varepsilon = \frac{1}{k-1}
\]

(6.5)

In the above example we have to run the F-test with 1/2*18 degrees of freedom for the error and 1/2*2 df for the treatment. This gives \( p(F=14.324;9,1) = 0.202 \). There is no significant difference between the treatments. Another possibility is of course to use a Bonferroni correction and to divide the test wise error level through the number of treatments: 0.05 = 0.05/3 = 0.017. A third possibility is to use the **conservative AOVa** with (n-1) \( df_{\text{error}} \) and only one \( df_{\text{treat}} \).

Up to know we dealt with such experimental designs where for all combinations of variables at least 2 cases were available (the variances could be calculated). In this case we speak of **complete designs**. However, often we have only **incomplete designs** where not for all factor combinations within variances exist. For instance, we study our already known plants within the following design. We have three light categories (low, medium, high) and four nutrient categories (P, N, Fe, Mn) We do not expect that light and nutrients interact. That means we ignore the interaction term light x nutrients. Therefore, instead of at least 24 parallels \((3\times4\times2)\) we need much less. In the best case only eight. You don’t need all treatment combinations to be realized. In other words ignoring irrelevant interactions reduces the experimental effort. In the example beside we considered only the main effects and found only light to be significant.

Essentially, the incomplete design AOVa proceeds similar to the AOVa for complete designs. \( SS_{\text{error}} \) are calculated in the same way as above. \( SS_{\text{treat}} \) contains the between treatment variances of the factors.
A special case of incomplete designs are nested designs. In nested designs occur treatments of one factor only with certain treatments of the other factor. An example gives the next table. Phosphorus occurs only in the soils of the experiments that got much light, nitrogen is associated with the medium light treatment and Fe and MN with the low light treatments. Within the variance components module of Statistica we use the nested design option and get the output shown above. Again only light appeared to be significant.

Lastly we look at the F-test from the perspective of a multiple regression. A a simple ANOVA F is defined as

\[
F = \frac{SS_{\text{treat}}(n - k)}{SS_{\text{error}}(k - 1)}
\]

\(R^2\) of the multiple regression is defined as

\[
R^2 = \frac{\sigma^2_{\text{treat}}}{\sigma^2_{\text{total}}} = \frac{SS_{\text{treat}}}{SS_{\text{total}}}
\]

where \(\sigma^2_{\text{treat}}\) is of course the variance explained by the predictor variables. We further know that \(SS_{\text{total}} = SS_{\text{treat}} + SS_{\text{error}}\). Plugging this equation into the second gives

\[
SS_{\text{total}} = SS_{\text{treat}} + SS_{\text{error}} \rightarrow SS_{\text{total}}R^2 = SS_{\text{total}} - SS_{\text{error}}
\]

\[
SS_{\text{error}} = (1 - R^2)SS_{\text{total}}
\]

Now we go back to the first equation and get

\[
F = \frac{R^2SS_{\text{total}}(n - k)}{(1 - R^2)SS_{\text{total}}(k - 1)} = \frac{R^2(n - k)}{(1 - R^2)(k - 1)}
\]

(6.6)

This equation is identical to eq. 12.12 (with \(m = k - 1\) and enables us to test the significance of a multiple and a bivariate regression.
7. Cluster analysis

Cluster analysis is one of the multivariate techniques that is not based on the assumptions of the general linear model. Cluster analysis is a method that does not test hypotheses. It is an intrinsically hypotheses generating method. Simply speaking a cluster analysis tries to classify a set of objects into groups (cluster or classes) according to a predefined measure of distance. However, there is no single technique. Cluster analysis is a whole set of very different methods and solutions.

Consider the following example (Fig. 7.1). We take a matrix containing ten variables (objects) and 20 observations (cases). Then we start Statistica, copy the data into the cluster analysis module and perform a default analysis. The result (the Figure beside) is a dendrogram that visualizes the distances of the variables from each other according to a linkage algorithm. We are ready at hand with an interpretation. We develop an a posteriori hypothesis.

But what did Statistica really do? First of all it applied a measure of distance. The program measures case wise differences and computes from them a common distance value between two variables. It then tries to link variables according to these distances. The Jaccard measure of similarity is one measure of distance between two objects. It is defined as

\[
J = \frac{a}{b + c}
\]  

(7.1)

where \(a\) is the number of common elements and \(b\) and \(c\) the total numbers of elements in object 1 and 2.

It is possible to generalize the Jaccard index to get an equation that contains nearly all measures of similarity as special cases

\[
J = \frac{a + \varepsilon e}{a + \varepsilon e + \delta (b + c)}
\]  

(7.2)

where \(a\) is the number of common elements, \(b\) and \(c\) are the number of elements that occur only in object 1 or 2 and \(e\) is the number of elements that neither occur in 1 nor in 2.

The best known measure of distance for metrically scaled variables is of course the Euclidean distance.

\[
d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}
\]  

(7.3)

where \(p\) is the number of data sets and \(x_{ik}\) and \(x_{jk}\) the values of \(i,j\) at data point \(k\).

The most important alternative to Euclidean distance is the so-called city block or Taxi driver distance that measures the shortest way to get from point \(i\) to point \(j\). It is defined from the general Minkowski metric

\[
d_{ij} = \left( \sum_{k=1}^{p} |x_{ik} - x_{jk}|^{r} \right)^{1/r}
\]  

(7.4)
r = 1 defines the Taxi driver distance, r = 2 the Euclidean distance. Other measures of distance for metrically scaled variables are pairwise correlation coefficients or the Bray-Curtis measure dealt with in the math part.

More important for the outcome of a cluster analysis is the choice of the linkage algorithm. The basic procedures of such algorithms can be exemplified from the single linkage algorithm (Fig. 7.2). Starting with an association matrix we order the pairs of sites according to the degree of association. Then step by step objects are including into cluster that have a higher association than predefined. The first group is made of BD, the second group includes C because D and C are the next with respect to association. Lastly we include A. Single linkage has some undesired properties. It tends to produce a large number of clusters and has problems if three or more objects have the same distances to each other.

The Figure 7.3 shows the most important linkage algorithms. The exact form of the algorithms are given in most statistical textbooks. Important is that different algorithms frequently result in different clusters. The result of a cluster analysis relies therefore to a good deal on the choice of the linkage algorithm. In general we have four types of algorithms
Sequential versus simultaneous algorithms. In simultaneous algorithms the final solution is obtained in a single step and not stepwise as in the single linkage above.

Agglomeration versus division algorithms. Agglomerative procedures operate bottom up, division procedures top down.

Monothetic versus polythetic algorithms. Polythetic procedures use several descriptors of linkage, monothetic use the same at each step (for instance maximum association).

Hierarchical versus non-hierarchical algorithms. Hierarchical methods proceed in a non-overlapping way. During the linkage process all members of lower clusters are members of the next higher cluster. Non hierarchical methods proceed by optimization within group homogeneity. Hence they might include members not contained in higher order cluster.

There are some general properties of the linkage algorithms.

1. The **single linkage algorithm** uses the minimum distance between the members of two clusters as the measure of cluster distance. It favours chains of small clusters.

2. The **average linkage** uses average distances between clusters. It gives frequently larger clusters. The most often used average linkage algorithm is the **Unweighted Pair-Groups Method Average (UPGMA)**.

3. The **Ward algorithm** calculates the total sum of squared deviations from the mean of a cluster and assigns members as to minimize this sum. The method gives often clusters of rather equal size.

4. **Median clustering** tries to minimize within cluster variance.

The Fig. 7.4 shows the same set of eight variables clustered by four different linkage algorithms. The results look different. Only two constant patterns are detectable. Variables 1, 2, 3, and 6 cluster always together.
and variable 8 seems to form a single cluster. It is then a question of our interpretation of the results. Most popular are UPGMA and the Ward algorithm that give in most cases rather equally sized partitions and therefore clusters.

The Ward algorithm differs somewhat from the others by using a minimalization criterion. Using Euclidean distance the Ward algorithm tries to minimize for each cluster \( r \) the distance \( d_r^2 \) of all elements of \( r \) from the centroid (the middle) \( c \) of the cluster.

\[
d_r^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - c_{ij})^2 \rightarrow \min
\]  

(7.5)

One method to assess the ‘right’ number of clusters is the so-called elbow test (Fig. 7.5). The next Table shows the agglomeration process of ten variables. The final clustering indicates seven singles cluster. However, some of them seem to be poorly separated. We plot the Euclidean distances against the agglomeration stage as shown below. From stage seven to eight a step (an elbow) occurs, where the dissimilarity suddenly raises. At lower stages dissimilarities are rather low. The elbow criterion tells now that we should accept only those clusters above the elbow. Hence we should accept a three cluster solution with cluster one containing Var2, Var4, Var5, and Var9, cluster two containing Var6 and Var8, and cluster three containing Var1, Var3, Var7, and Var10 (Fig. 7.6).

However, there is no ‘correct’ clustering. Each method favours different aspects of distance. Hence, although cluster analysis is a genuine hypothesis generating technique we should again start with a sound and realistic hypothesis.
Advices for using a cluster analysis

- The first point is the number of data sets. Especially in the case of samples where you want to get information about the whole population sample sizes (data sets) must not be too small.
- Screen the data for outliers. They might have overproportional influence on the outcome and should be eliminated.
- As in other multivariate techniques an initial sound hypothesis building process helps to interpret the results.
- It is a good praxis to undertake in parallel a factor analysis and to compare the results with those of the cluster analysis. Sometimes (in the case of many variables) it is better to undertake first a factor analysis and use the factor loadings instead of the raw variables for the cluster analysis. Highly correlated variables are thus combined and the total number of variables is reduced.
- Even without a priori factor analysis variables that correlate highly with others ($r > 0.9$) should be eliminated. It is of course our interpretation which of the two or more correlated variables to throw out.
- If all data are metrically scaled I advise to standardize them by Z-transformation. Large differences in absolute values might bias the outcome. Try also log-transformations.
- Try out several linkage algorithm. Compare the results and accept only clusters that have been found by all or at least by the majority of algorithms.
- Do the results make sense? Compare them with your initial hypotheses.
- If you have a large data set of independent data you might divide them into two randomly chosen parts. Use one part in the cluster analysis for hypothesis building and the second part for verifying the result by using the same algorithm.
- Cluster analysis is a method that is open for all types of data manipulation. Therefore, you must always give basic information about how you received a given result:
  1. Distance measure and linkage algorithm
  2. Explain your method choice. Tell especially, why you used a given linkage algorithm
  3. Tell whether your results are stable.
7.2 K-means cluster analysis

Ordinary cluster analysis groups a set of variables into a number of clusters. This number is the result of the clustering process. A different technique is the clustering of variables into a predefined number of clusters. This is done by the k-means clustering. K-means clustering starts with a predefined number of clusters and puts the elements step by step into these clusters. It seeks for elements that have distances from the mean of the own cluster that are larger than those to another cluster. The element is then shifted to this cluster. The process stops if all elements have found ‘their’ cluster.

The k-means method does not result in a dendrogram. Instead the output contains the distances of each variable from the cluster mean. The results but can be visualized by distances of elements from cluster means. This is shown in the diagram beside. 20 Elements have been grouped into five clusters. They are defined by distances of cluster centres (the red arrows) and these are computed from the elements they contain (black elements).

The k-means method is a good alternative to the previous hierarchical approach if the number of clusters is a priori known. This is often not the case. Then you can first run an ordinary cluster analysis. This gives a hypothesis about the number of clusters. You then run a k-means analysis to reach in a better cluster model.

The Figure beside shows a plot of cluster means for each data set of a k-means cluster analysis of eight variables and ten data sets divided into three clusters. Such a plot helps to assess how well the clusters are separated. In our case the red and the green cluster are weakly separated.

A shortcoming of this method is that the final clustering might depend on the entry order of the elements into the clustering process. I advise therefore to run several analyses in which the order of variables is changed. Accept the solution that was found most often.
### 7.3 Neighbour joining

The most often used agglomerative cluster algorithm in phylogenetic analysis is **neighbour joining** that was particularly developed for phylogenetic analysis.

At the beginning the method puts all elements into different clusters (Fig. 7.3.1). It then proceeds stepwise. It starts with a dissimilarity matrix based on a certain distance measure. (shown as different branch lengths in the Fig.).

First, two clusters (elements) are chosen and combined in one cluster to get a new node in the tree (X in the Fig.). From this a new dissimilarity matrix (containing the elements A, X, D, E, F) is calculated and the method proceeds at step one. The third dissimilarity matrix contains the elements (A, X, Y, F).

The method depends therefore on three criteria: the algorithm to choose the pair of elements to join, the algorithm to calculate the dissimilarity matrix and an algorithm to estimate branch lengths. In the simplest case dissimilarities for any element are calculated from

\[
\delta_{(X,Y)} = \sum_{n} \delta(X,Y)
\]

(7.3.1)

\(\delta_{(X,Y)}\) are the dissimilarities for the elements X and Y calculated from the branch lengths. To select a pair of elements to be joined the method proceeds in calculating

\[
Q_{\delta_{(X,Y)}} = (n-2)\delta_{(X,Y)} - \Delta_{(X)} - \Delta_{(Y)}
\]

(7.3.2)

for all element pairs (X,Y). The pair with the lowest value of Q is selected for joining. Given that two elements A and B have been selected and joined to a new element U_{AB} the new dissimilarities are calculated from

\[
\delta_{(X, U_{AB})} = \frac{\delta_{(X, A)} + \delta_{(X, B)} - \delta_{(A, B)}}{2}
\]

(7.3.3)

This is called the **reduction step**. After this step the method proceeds in a recursive way starting again from the dissimilarities. Beside I show a simple example for four species. From an initial distance matrix dissimilarities (eq. 7.3.1) were calculated. The next step involves the selection of elements (eq. 7.3.2) and at last the reduction to a new distance matrix (eq. 7.3.3). This procedure is then repeated for the new matrix. At the end the method joins vertebrates and Protostomia.

We also need the branch lengths of U to A and B. These
are given by

\[
\delta(A, U) = \frac{(n - 2)\delta(A, B) + \Delta(A) - \Delta(B)}{2(n - 2)} \\
\delta(B, U) = \frac{(n - 2)\delta(A, B) - \Delta(A) + \Delta(B)}{2(n - 2)}
\]

(7.3.4)

Above a more complicated example with the program Past is shown. Starting with raw data the program first calculates the distances between the species (in this case from correlation coefficients) and then proceeds as shown above.

A very good and simple introduction to neighbour joining gives http://artedi.ebc.uu.se/course/sommar/njoin/index2.html.
8. Rank order and model II regression

In the first part of the statistics lecture we heard about linear regression and correlation. The correlation between two variables $x$ and $y$ was defined as

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$  \hspace{1cm} (8.1)

The value

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2}$$  \hspace{1cm} (8.2)

was defined as the coefficient of determination. No correlation means $r^2 = 0$; maximum correlation means $r^2 = 1$. In the latter case, there is no variance of the $y$-value for any given $x$-value. In other words, the total variance of $y$ had been explained in terms of $x$.

Now look at Fig. 8.1. From an experiment we got two sets of data. There is a highly significant correlation between both data sets. But be careful, the regression depends solely on two data points at the beginning and at the end of the spectrum of values. Without them there would be no significant correlation ($p(t) > 0.05$).

What to do? We have three possibilities to solve our problem. The first is that we leave out these data points. But this is of course in a certain sense a manipulation. A second possibility would be to give the data weights. Weighting data of which we think that they are of different reliability, influence, or importance is an often used method in statistics. In our case, we could, for instance, give the data between the two outliers weights that are two times the weights of the outliers. This is nothing more than to double each data point (we would not have ten but 18 data points in total) and to regress again. After this procedure our regression is not longer significant. But weighing data is of course a very subjective procedure and we have to be very careful with such a manipulation.

But we have a third possibility. We do not use the raw data; we use only the ranks of our data. Look at the next Table. Here the raw data are given together with the ranking of these data. In a next step we simply correlate only the ranks. By this way we loose information (the absolute values of our data) but we gain a better and non-subjective correlation. Ranking data eliminates particularly outliers. It also linearizes the data (Fig. 8.2). And it is independent of the above assumptions of the normality of errors and other restrictions of the general linear model. This method
of correlating ranks is called the **Spearman rank correlation** after the British psychologist Ch. Spearman (1863-1945). It is the most widely used alternative to the **Pearson parametric correlation** and should always be used if we are not sure about the true data structure. All statistic programs and Excel perform such a ranking and a rank correlation. In our case Spearman’s rank does not point to a significant correlation.

Up to now we used the Gaussian least square method to derive linear regressions. This means that we used only the deviations in the dependent variable \( y \). The method treats \( y \) as having error terms, the residuals, which introduce additional components of variance. These error terms might be caused by the influence of other variables or by simple measurement errors. But what about measurement errors of the independent variable \( x \)? For instance in Fig. 8.3 we tried to explain mammalian brain weight on the basis of the respective body weight. We treated the body weight as being measured exactly, without any measurement error. But in reality we measured both variables and both measurements will have a certain measurement error. If the errors in measuring \( x \) would be large our least square regression might give a quite misleading result. This problem has long been known as one method to overcome this problem is surely to use Spearman’s rank order correlation. However we may also use a different method for minimizing the deviations of data points from the assumed regression line. Look at Fig. 8.13. Ordinary least square regression depends on the variability of \( y \) but assumes \( x \) to be without error (OLR\(_y\)). But why not using OLR based on the variability of \( x \) only (OLR\(_x\))? The slope \( a \) of OLR\(_x\) is (why?)

\[
a_{OLRx} = \frac{s_y^2}{s_{xy}}
\]  

(8.3)

Now suppose we compute OLE\(_y\) and OLE\(_x\) separately. Both slopes differ. One depends solely on the errors of \( y \), the other solely on the errors of \( x \). We are tempted to use the mean value of both slopes as a better approximation of the ‘real’ slope. Now we use the geometric mean of both slopes.

\[
a_{RMA} = \sqrt{a \cdot a} = \sqrt{\frac{s_{xy}}{s_y^2}, \frac{s_x^2}{s_{xy}}} = \frac{s_y}{s_x}
\]

(8.4)

In effect this regression reduces the diagonal distances of the rectangle made up by \( \Delta y \) and \( \Delta x \) as shown in Fig. 8.4. Due to the use of the geometric mean it is called the **geometric mean regression** or the **reduced major axis regression** (**standard major axis regression**). Due to the use of standard deviations the method uses standardized data instead of raw data. Eq. 8.4 can be written in a
slightly different form. From eq. 8.8 of part A of the lecture we get

$$a_{RMA} = \frac{s_y}{s_x} = \frac{a_{OLR}}{r}$$

(8.5)

The RMA regression slope is therefore always larger than the ordinary least square regression slope \(a_{OLR}\). Hence, if an ordinary least square regression is significant the RMA regression is significant too.

Lastly, the most natural way of distance and minimizing distances to reach in a regression slope is to use the Euclidean distance (Fig. 8.4). If we use this concept we again consider the errors of \(x\) and those of \(y\). This technique is called \textbf{major axis regression} (MAR). We know already how to calculate principal (or major) axis from the matrix algebra part. These are the eigenvectors of a given matrix. Hence we need to solve

$$(\Sigma - \lambda I) \cdot u = 0$$

(8.6)

with \(\Sigma\) being the dispersion matrix and \(u = [x-\mu] \) the matrix of eigenvectors. We get the dispersion matrix from eq. 8.13 of part A (cf. the Excel example) as the dot product of the matrix of centralized X and Y column vectors with it’s transpose.

In a simple example we now take two variables \(x\) and \(y\) and want to calculate the principle axes. We need the dispersion matrix that contains the covariances. The variance of \(X\) is \(s_x^2 = 20\), of \(y\) \(s_y^2 = 25.03\), and the covariance is \(s_{xy} = 16.73\). We compute the eigenvalues and eigenvectors of \(S\) and get the first eigenvector \(E = \{-0.65, -0.76\}\). The principal axis slope is therefore \(m = -0.76/0.65 = 1.16\). From \(y = mx + b\) we get the intercept \(b = -3.64\) (Fig. 8.5). We performed a major axis regression.

The general solution is given below. We need the dispersion matrix \(\Sigma\) and compute the eigenvectors and eigenvalues. In general we get the slope \(m\) of the major axis regression from the values \(y/x\) of the first major axis, that means the axis of the largest eigenvalue.

$$m = \frac{-2s_{xy}}{s_x^2 - s_y^2 + \sqrt{(s_x^2 - s_y^2)^2 - 4s_{xy}^2}} \quad \lambda = \frac{1}{2}(s_x^2 + s_y^2 + \sqrt{(s_x^2 - s_y^2)^2 - 4s_{xy}^2}) - 2\lambda - s_y = s_x^2 + \sqrt{(s_x^2 - s_y^2)^2 - 4s_{xy}^2}$$

$$m = \frac{s_{xy}}{\lambda - s_y}$$

(8.7)
Data analysis and statistics

Shortly

\[ a_{\text{MAR}} = \frac{s_{xy}}{\lambda - s_y^2} \]  

(8.8)

with \( \lambda \) being

\[ \lambda = \frac{s_x^2 + s_y^2 + \sqrt{(s_x^2 + s_y^2)^2 - 4(s_x^2 s_y^2 - s_{xy}^2)}}{2} \]  

(8.9)

We get in our example \( m = 16.73/(39.42 - 25) = 1.16 \) (Fig. 8.5).

OLRx and OLRy are termed the model I regression. RMA and MAR are also termed model II regression.

Of course, the intercept of all these regression techniques is further computed from eq. 8.2 of part A

\[ b = \bar{y} - a \bar{x} \]  

(8.10)

Figure 8.6 shows a comparison of all four regression techniques. We see that at a high correlation coefficient of 0.77 the differences in slope values between all four techniques are quite small. Larger differences but would occur at lower correlation coefficients. The Excel model shows also how to compute a model II regression. They are not implemented into Excel and major statistic packages although they are recently very popular among biologists.

This leads us to the question of model choice (Fig. 8.7). When to use which type of regression? In general model II regression should be used when we have no clear hypothesis what is the dependent and what the independent variable. The reason is clear. If we don’t know what is x and y we also can’t decide which errors to leaf out. Hence we have to consider both errors, those of x and those of y. If we clearly suspect one variable to be independent we also should use model II regression if this variable has large measurement errors. As rule of thumb for model I regression the errors of x should be smaller than about 20% of the errors in y. Hence \( s_y < 0.2s_x \). Lastly, if the we intend to use a model II regression, we should use MAR if our data are of the same dimension (of the same type, for instance weight and weight, mol and mol, length and length and so on). Otherwise a RMA regression is indicated.

However, all these advice have only importance if we deal with loosely correlated data. For variables having a correlation coefficient above 0.9 the differences in slope become quite small and we can safely use an ordinary least square regression OLRy. Additionally, there is an ongoing discussion about the interpretation of RMA. RMA deals only with variances. The slope term does not contain any interaction of the variables (it
Data analysis and statistics

lacks the covariance). Hence, how can RMA describe a regression of $y$ on $x$? Nevertheless especially RMA and MAR have become very popular when dealing with variables for which common trends have to be estimated.

**Regression and causality**

At the end of this chapter we have to discuss shortly the relation between regression and causality. What is if we find a significant correlation between two sets of data? In theory, we have four possibilities

1. A change in $x$ causes respective changes in $y$
2. A change in $y$ causes respective changes in $x$ and we have to invert our regression
3. $x$ and $y$ are mutually related
4. $x$ and $y$ are not directly related and the observed pattern is caused by a third or more hidden factors that influence both, $x$ and $y$. In this case there is no causal relationship between $x$ and $y$.
5. $x$ and $y$ are neither directly nor indirectly related. The observed pattern is produced accidentally.

There has been a long and controversial debate among philosophers and statisticians whether regression or more generally statistics as a whole can prove causal relations. Today, such a possibility is generally excluded.

**Statistics as an empirical science cannot prove causality. It only ‘proves’ that certain observed patterns are more or less probable.** Mathematically speaking, a correlation between two variables $x$ and $y$ is a necessary, but not a sufficient prerequisite for a causal relationship.

But what is if we have a correlation with a significance level of 0.9999 in favour of $H_1$, that there is a connection between $x$ and $y$ of the form that $x$ implies $y$ ($x \rightarrow y$)? Does this mean that with 99.99% probability there is a causal relationship? No. This case only excludes possibility 5 of the above list. To point to causality we have to have additional knowledge. Was $x$ prior to $y$? Is $x$ the more basal pattern? Contradicts the relation $y \rightarrow x$ logic? What other variables have to be considered? Are mutual relations between $x$ and $y$ possible? By answering these questions we can step by step exclude (with a predefined probability level) the points 2 to 4 of our list until we decide that there should be a causal relation between $x$ and $y$ of the form $x \rightarrow y$. But again, we cannot prove it.
Literature

Bortz J. 1999—Statistik für Sozialwissenschaftler—Heidelberg (5. Aufl.).
Ennos R. 1999—Statistical and Data Handling Skills in Biology—Longman.
Kreyszig E. 1977—Statistische Methoden und ihre Anwendungen—Vandenhoeck und Ruprecht (7th Ed.).
Statistics and the Internet

Here are several useful links to internet pages that contain information about statistics

**Handbooks**

Hyperstat (large online statistic handbook and glossary): http://www.davidmlane.com/hyperstat/glossary.html


Statistica online handbook (the complete Statistica handbook and glossary for download) http://www.statsoft.com/textbook/stathome.html

StatPrimer (an online statistic handbook by Bud Gerstman) http://www.sjsu.edu/faculty/gerstman/StatPrimer/

Introductory statistics and multivariate statistics (Two very extensive online statistical textbook covering all sorts of useful techniques with many animations) http://www.psychstat.smsu.edu/multibook/mt00.htm; http://www.psychstat.smsu.edu/sbk00.htm

Statistik für Psychologinnen und Psychologen (a very well designed online script of statistics with many tables and examples) http://www.psychologie.uni-freiburg.de/signatures/leonhart/skript/


How to write a paper in scientific journal style and format? http://abacus.bates.edu/~ganderso/biology/resources/writing/HTWtoc.html


**Software sites**

Past (a very good freeware statistic package that contains all the necessary tools for basal statistical analysis) http://folk.uio.no/ohammer/past

Statpac (freeware programs for basic calculations) http://www.davidmlane.com/hyperstat/glossary.html

XLStat (statistic module for Excel, free trials) http://www.davidmlane.com/hyperstat/glossary.html

Virtual Library in Statistics (Many links to interesting web pages and programs) http://www.stat.ufl.edu/vlib/statistics.html

UCLA distribution calculator (very good probability and table calculator for many basic statistical distributions) http://calculators.stat.ucla.edu/cdf/

UCLA statistics calculators (a large collections of basic statistics calculators and tables, easy to use) http://calculators.stat.ucla.edu/

UCLA statistics http://ebook.stat.ucla.edu/textbook/singles/describe_single/probmodels/calc.html

Neville Hunt’s homepage (if you want to produce statistic tables by yourself, here you find a very good instruc-
Data analysis and statistics

- Excel tables for important statistical distributions: [http://www.mis.coventry.ac.uk/~nhunt/tables.htm](http://www.mis.coventry.ac.uk/~nhunt/tables.htm)
- Nellie. Ecological null models for the PC (a program for randomization tests): [http://www.uwec.edu/Academic/Curric/weiher/nullmodel.htm](http://www.uwec.edu/Academic/Curric/weiher/nullmodel.htm)
- EcoSim (a standard program for randomization of data matrices): [http://homepages.together.net/~gentsmin/ecosim.htm](http://homepages.together.net/~gentsmin/ecosim.htm)

**Libraries**

- Interactive statistical calculation pages (a browser to search all sorts of statistic tools, many links to only textbooks, glossaries and scripts): [http://members.aol.com/johnp71/javastat.html](http://members.aol.com/johnp71/javastat.html)
Links to multivariate techniques

**GLM:** A nice introductory description at http://trochim.human.cornell.edu/kb/genlin.htm

**Analysis of variance:** All about at http://faculty.vassar.edu/lowry/vsanova.html and http://www.statsoftinc.com/textbook/stanman.html

**Multiple regression:** all about at http://www.statsoftinc.com/textbook/stmulreg.html and http://www2.chass.ncsu.edu/garson/pa765/regress.htm

**Path analysis:** All about at http://www2.chass.ncsu.edu/garson/pa765/path.htm and http://luna.cas.usf.edu/~mbrannic/files/regression/Pathan.html
Short introduction at http://users.rcn.com/dakenny/pathanal.htm

**Cluster analysis:** All about at http://cne.gmu.edu/modules/dau/stat/regression/multregsn/multregsn_frm.html and http://www.clustan.com/what_is_cluster_analysis.html
A nice introduction at http://149.170.199.144/multivar/hc.htm

**Discriminant analysis:** All about at http://www.statsoftinc.com/textbook/stdiscan.html and http://www2.chass.ncsu.edu/garson/pa765/discrim.htm
Look also at http://www.doe-mbi.ucla.edu/~parag/multivar/da.htm

**Factor analysis:** All about at http://www2.chass.ncsu.edu/garson/pa765/factor.htm and http://www.statsoftinc.com/textbook/stfacan.html
Look also at http://www.doe-mbi.ucla.edu/~parag/multivar/pca.htm and http://www.chem.duke.edu/~reese/tutor1/factuemp.html


Of course *Wikipedia* gives many links to other sites.