LECTURE NOTES FOR

APPLIED STATISTICS

ANDERS TOLVER JENSEN AND HELLE SØRENSEN

DEPARTMENT OF NATURAL SCIENCES

LIFE SCIENCE FACULTY, UNIVERSITY OF COPENHAGEN

FREDERIKSBERG, NOVEMBER 2007

Preface

These lecture notes have been written for the course *Applied Statistics* at the Life Science Faculty at the University of Copenhagen (until 2007 known as KVL). The course is offered by the Department of Natural Sciences. The course consists of five weeks with "statistical themes" followed by a project period of three weeks. The material in these notes will be the basis for the first five weeks in the 2007 version of the course.

In short, the notes cover various kinds of Gaussian linear and non-linear models, Gaussian models with random effects, logistic regression and the proportional odds model. In particular the ANOVA models are assumed to be known and are only dealt with implicitly.

The notes are, to a great extend, inspired by earlier lecture notes by Bo Martin Bibby (*Noter til Praktisk Statistisk Dataanalyse* and *Noter til Regressionsanalyse*). Parts of the text are almost merely translations of these notes, while other parts have been prepared from scratch. An important difference, however, is that the present notes use `R` as well as `SAS`.

The intention never was to give a comprehensixe exposition of the subjects. In particular, problems like variance heterogeneity, multiple testing and numerical problems in `R` and `SAS` are dealt with as they occur. Instead, the intention was to give introductions to the areas which are (hopefully) useful for the project work in the course as well as for later use.

In all chapters one or a few data examples form the basis, and statistical analyses are carried out for these examples. The focus is on the statistical models, their interpretation and, not the least, the conclusions obtained from the analysis.

As an important part of the analyses, computer programs are provided. All analyses are carried out with both `R` and `SAS`, but the reader is supposed to choose one or the other. The notes contain no introduction to `SAS` and `R`, so the reader should be familiar with one of them already (see the webpage of the course, given below, for suggestions on introductory material).

Datasets and computer programs for examples and exercises are made available at the webpage of the course,

```
http://www.matfys.kvl.dk/stat/kurser/appliedstatistics/
```

Here you can also find other information about the course.

Frederiksberg, November 2007

Anders Tolver Jensen

Helle Sørensen

Contents

Chapter 1

Multiple linear regression

In this chapter we discuss models applicable in situations where the outcome of a (response) variable is naturally explained by the outcome of a number of qualitative variables, also called covariates.

In `SAS`, `PROC REG` and `PROC GLM` can both be used for (multiple) regression analysis. We will use `PROC GLM` since it works for general linear models. For example, it allows for quantitative explanatory variables (factors). In `R`, we will use the function `lm`. See the computer sections for computational details, and also for some more general comments about graphics in `R`.

1.1 Simple linear regression

A *simple linear regression analysis* is natural when, for each experimental unit, we have measured two quantities and we wish to (partly) explain the value of one of them by the other one.

Assume, for example, that we have registered the intake of some feed supplement and the increment in weight over some period for n pigs. For the i'th pig the measurements are x_i and y_i. We wish to explain the increment in weight from the intake of feed supplement, and a plot of y against x is natural to illustrate the relation (see Figure 1.1).

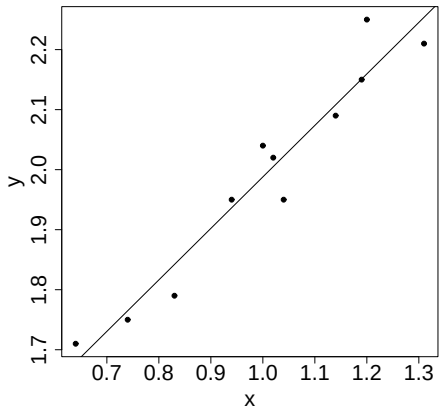


Figure 1.1: Simple linear regression.

Tree	Diameter	Height	Volume	Tree	Diameter	Height	Volume
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	16.9	66	64.3
11	11.3	79	24.2	27	17.3	81	55.4
12	11.4	76	21.0	28	17.5	82	55.7
13	11.4	76	21.4	29	17.9	80	58.3
14	11.7	69	21.3	30	18.0	80	51.5
15	12.0	75	19.1	31	18.0	80	51.0
16	12.9	74	22.2	32	20.6	87	77.0

Table 1.1: Diameter, height and volume of 32 cherry trees.

The corresponding simple linear regression model is given by

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n,$$

where the e_i's are $N(0, \sigma^2)$-distributed and independent.

The variable y is called the *response variable*, and x is called the *explanatory variable*, or a *covariate* (or a regressor). The parameter α is interpreted as the expected weight increment without feed supplement ($x = 0$) whereas β is the expexted *extra* weight increment obtained by increasing the amount of feed supplement by one unit.

1.2 Multiple linear regression

Assume now that for each experimental unit there are measurements of a response variable, y, as well as of *a number of* explanatory variables (covariates), $x_1, \ldots, x_p$. That is, for the i'th experimental unit there are measurements y_i and $x_{1i}, \ldots, x_{pi}$.

The purpose is to (partly) explain the variation of the response variable by means of the explanatory variables. The *multiple linear regression model* is given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + e_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where the e_i's are $N(0, \sigma^2)$-distribued and independent. The parameter β_1 is the expected difference between two experimental units for which the variable x_1 differs by one, but *all other explanatory variables are the same*. Similarly for $\beta_2, \ldots, \beta_p$.

Let us consider an example.

Example 1.1 (Volume of cherry trees) For each of 32 cherry trees one has measurements of the diameter (inches), the height (feet) and the volume (cubic feet). The dataset is given in Table 1.1.

The main interest is on prediction of the volume. The (economic) value of a tree is represented by the volume of the tree which is hard to measure without felling the tree. As opposed to this one can

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x_{1i}	2	6	9	14	16	4	10	8	11	3	1	7	13	12	5	15
x_{2i}	17	10	12	25	21	18	14	22	20	11	13	15	24	16	23	19
y_i	29	3	-2	9	-5	25	-1	21	8	14	24	10	10	-3	32	-6

Table 1.2: Fictious numbers used for Figure 1.2.

measure the diameter and the height without felling the tree. Hence, if the volume is well described from the diameter and the height, then the forester can estimate the economic value of the forest only from measurements of diameter and height.

Moreover, since it is much easier to measure the diameter than the height it is relevant to investigate whether measurements of the height holds extra information about the volume, in excess of the information from the diameter. In other words: does height explain parts of the variation in volume which is is not explained by diameter?

A multiple regression model can help us answer such questions. Let d be the diameter, h the height and v the volume, then the multiple regression of volume on diameter and height is given by

$$v_i = \beta_0 + \beta_1 d_i + \beta_2 h_i + e_i, \quad i = 1, \ldots, n,$$

where the e_i's are independent and $N(0, \sigma^2)$-distributed. The parameter β_1 is the expected difference in volume between two trees with a one inch difference in diameter and the same height. Similarly, β_2 is the expexted difference in volume for two trees with same diameter but a difference in height of one foot. □

Now, how do we check if a multiple linear regression is a reasonable model for the data we have collected? In the simple case with only one covariate a plot of the response against the covariate most often reveals whether a simple linear regression makes sense at all (although the plot cannot stand alone as model validation since it checks only the systematic part of the model).

In the case with several covariates the response may be plotted against each of the covariates but one has to be very careful with the interpretation of these plots. It is clear that if each of the plots shows a linear relationship, then the fixed/systematic part of the multiple linear regression model is indeed reasonable. There are, however, situations where the plots do not show a linear relationship but a multiple linear regression model is appropriate nonetheless.

Consider for example the fictious numbers in Table 1.2. Plots of y against x_1 and x_2, respectively, are shown in Figure 1.2. There seems to be a relationship between x_1 and y, but is it linear? There does not really seem to be a relationship between x_2 and y. Actually, the numbers satisfy

$$y = 1 - 3x_1 + 2x_2$$

(exactly, no error terms), so there *is* a linear relationship between x_1 and y as well as between x_2 and y.

The point is that even if there is only very small measurement errors then it is *not* possible to assess the appropriateness of the multiple regression model from the plots of the response against each covariate. The picture may be blurred due to the variation of the other variable.

Rather, model validation should be based on investigation of the residuals. In order to get the residuals, the model parameters should be estimated. The parameters in model (1.1) are $\beta_0, \beta_1, \ldots, \beta_p$ and σ^2. The mean parameters (the β_i's) are estimated by least squares and we denote the estimates by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$. The *predicted values* and *residuals* are then defined by

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi},$$

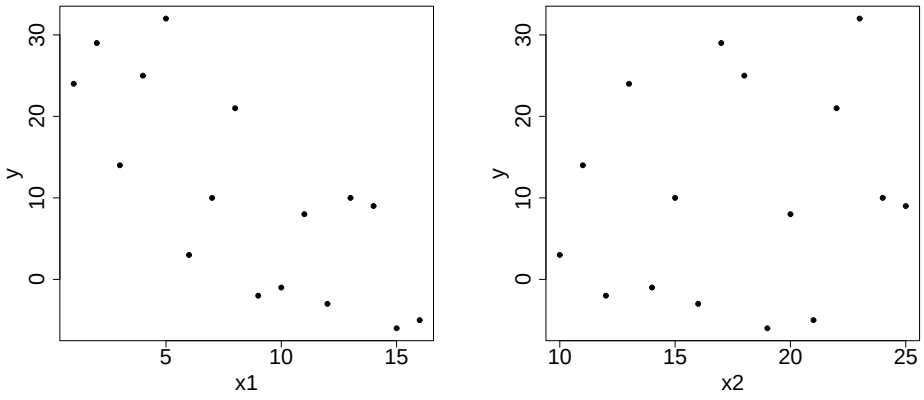


Figure 1.2: Scatter plots of y against x_1 and x_2 for the numbers in Table 1.2.

and

$$\hat{e}_i = y_i - \hat{\mu}_i.$$

The estimate for σ^2 is the so-called mean square error (MSE),

$$s^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{e}_i^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2.$$

If the model describe the data well, then the predicted values and the residuals are (almost) independent, so there will be no patterns in a scatter plot of the residuals against the predicted values. Moreover, *if* the model describe the data well, then the residuals do not depend on the covariates, so plots of the resisuals against each of the covariates will show no patterns. Hence, we can validate the model by investigating these plots, looking for systematic patterns in the way the points are spread out. No pattern is an indication that the model fits the data well, whereas a pattern in one of the plots indicates that the model does not describe the data well.

Usually we use the *standardized residuals* instead of the "raw" residuals (the $\hat{e}_i$'s above). The standardized residuals are defined as the raw residuals divide by their estimated standard deviation, and they are therefore all on the same scale. If the model describe the data well then the standardized residuals are approximately $N(0,1)$-distributed, so we sometimes make a QQ-plot of the standardized residuals to check if this is true.

If the model is not appropriate for the data, we cannot trust the results from the statistical analysis, that is, the estimates, confidence intervals and p-values etc. Hence, model validation is very important! There are many potential reasons that a model is not appropriate: the mean structure may be misspecified (ie. a missing covariate or a non-linear structure), there may be variance heterogeneity (all e_i's do not have the same variance), the observations may not be independent; the e_i's may not be normally distributed.

Often, but not always, the problems regarding variance heterogeneity and normality can be remedied by transformation of the response and/or the covariates. Once we are satisfied with the model validation part of the analysis, we can start model reduction (hypothesis testing), analysis of parameter estimates, prediction for other values of the covariates, etc.

Let us consider the example with cherry trees again for details on these issues.

Example 1.1 *(continued)* Recall the multiple regression model

$$v_i = \beta_0 + \beta_1 d_i + \beta_2 h_i + e_i, \quad i = 1, \ldots, n,$$

where the e_i's are independent and $N(0, \sigma^2)$-distributed, v is the volume, d is the diamater and h is the height.

Figure 1.3 shows the volume plotted against the diameter (to the left) and the volume plotted against the height (to the right). There is a clear relationship between diameter and volume as well as between height and volume, but recall that these figures are not appropriate to judge whether the multiple regression model describes the data well or not. We need to consider the residuals, which we will do in a moment.

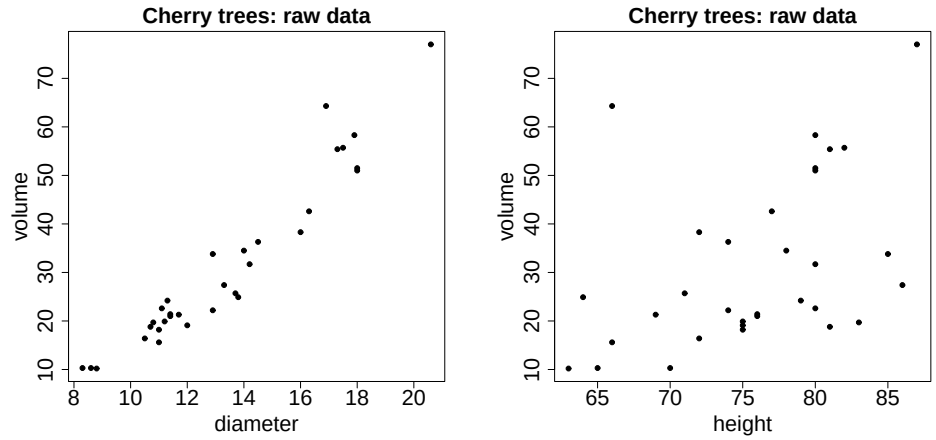


Figure 1.3: Volume af cherry trees plotted against diameter and height.

From the computer output (`R` or `SAS`) we find the estimates (with standard errors in parenthesis):

$$\hat{\beta}_0 = -44.8706\ (10.6234), \qquad \hat{\beta}_1 = 5.1606\ (0.3196), \qquad \hat{\beta}_2 = 0.0945\ (0.1548), \qquad s^2 = 25.57$$

The computer programs also carry out t-tests for the hypotheses $H_0 : \beta_i = 0$. The test statistic is given by

$$\frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \sim t(n - p - 1),$$

under H_0, and the p-values are given in the output.

In this example we see that the β_0 and β_1 are significantly different from zero ($p = 0.0002$ and $p < 0.0001$, respectively), whereas β_2 is not ($p = 0.55$). We have to be very careful with the conclusions: The interpretation is *not* that that there is no relationship between height and volume. Rather, we conclude that height does not contribute to the explanation of volume *when volume has been adjusted for diameter*.

This is an example of *multi-collinearity* which means that there is a (linear) relationship between some of the covariates. Here, height and diameter are positively correlated. Loosely speaking, they are both measurements of the same, namely the size of the tree. Had we used age of the tree as yet another covariate we would probably have reached a similar conclusion: age and volume are related but when we adjust for diameter (and height), there is no significant effect of age. Sometimes multi-collinearity is easily detected (as in this case); sometimes it is not so obvious (in particular when it involves three or more variables). In any case we must be careful with the conclusions from multiple regression analyses.

Recall that the above conclusions are valid only if the model describes the data well, and we need to validate the model. In fact, usually we would not be interested in the estimates and tests before having investigated the residuals. Figure 1.4 shows the standardized residuals plotted against the two covariates (top) and against the predicted values (bottom left). In two of the three figures we see a clear pattern: larger standardized residuals for small and large value than for medium values of the variable on the x-axis. The bottom right plot is a QQ-plot of the standardized residuals. A QQ-plot plots the quantiles of the standardized residuals to those of the standard normal distribution. Hence, if the assumption of Gaussian error terms e_i holds, then the points will be scattered around a straight line. In this case, the QQ-plot looks akay, except for a very large residual. However, the residual plots make us doubt that the model gives a reasonable description of the data, and hence make us doubt the conclusion from the model!

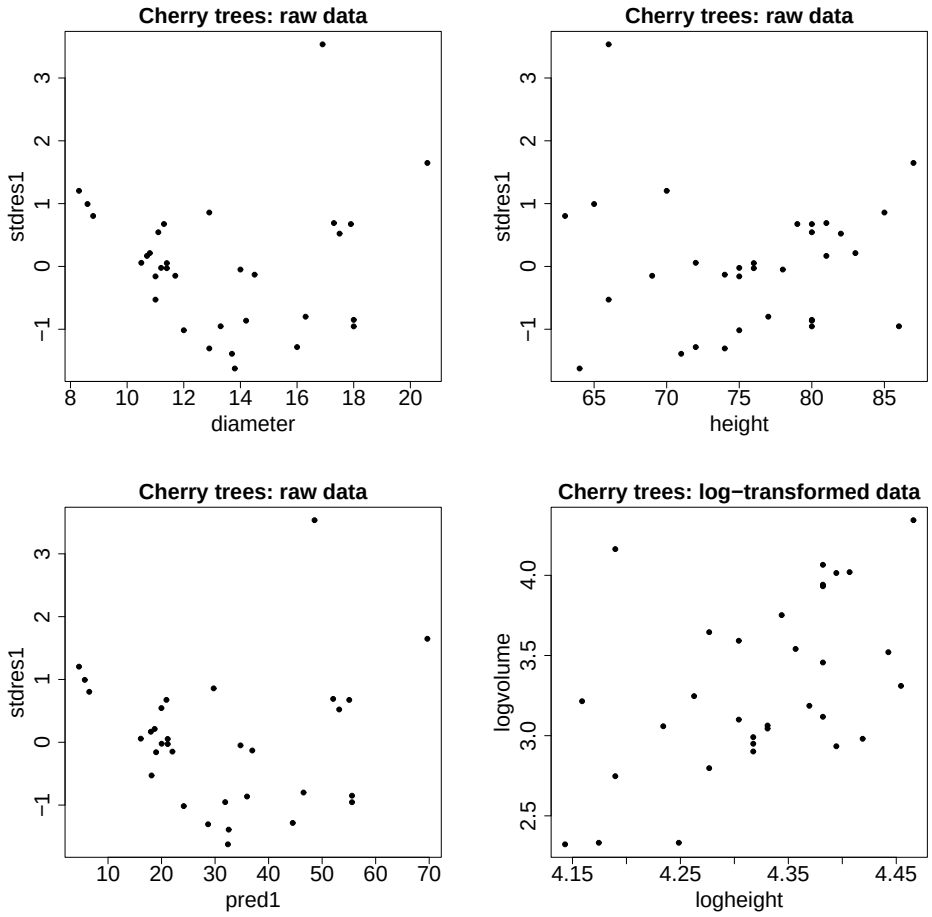
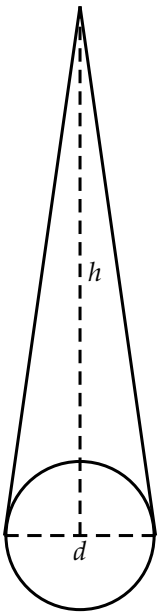


Figure 1.4: Residual plots and QQ-plot for the multiple linear regression model applied to the raw data.

Then what? Often a better fit is obtained by transformation of the response and/or the covariates. But there are many possible transformations so which one should we choose? In this example we can get help from the case story and from geometry.



If we think of a cherry tree as a cone with diameter d and height h, then the volume of that cone is given by

$$v = \frac{\pi}{12} \cdot h \cdot d^2.$$

A cherry tree is of course only approximately described by a cone so we extend the model by allowing the constant and the powers to be different from those above:

$$v = c \cdot d^{\beta_1} \cdot h^{\beta_2}.$$

Taking logarithms on both sides yields:

$$\log v = \beta_0 + \beta_1 \log d + \beta_2 \log h,$$

hvor $\beta_0 = \log c$ and log is the natural logarithm.

At best we can hope for the relation to hold on average, and hence only approximately for individual trees. We add Gaussian error terms e_i and get

$$\log v_i = \beta_0 + \beta_1 \log d_i + \beta_2 \log h_i + e_i, \quad i = 1, \ldots, n.$$

This is a multiple linear regression of $\log v$ on $\log d$ and $\log h$. The difference between this new model and the model analyzed above is that both the response variable and both covariates have been transformed.

We analyze the model as before. First we plot the response $\log v$ against each of the covariates, $\log d$ and $\log h$. These plots are shown in Figure 1.5. Again we see a clear relationship between the response and each of the covariates, and the relationships even seem approximately linear, so there is reason to believe that the multiple regression model fits the data well.

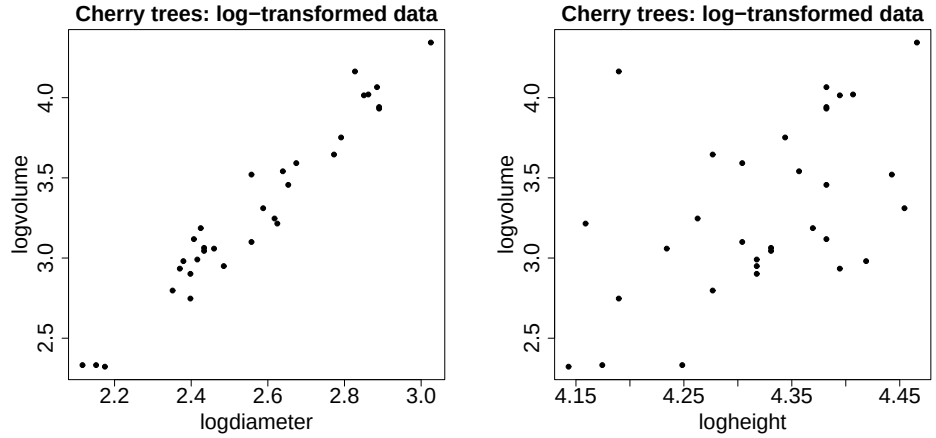


Figure 1.5: The logarithmic volume against the logarithmic diameter and the logarithmic height for the 32 cherry trees.

We also have to investigate the residuals in order to trust the analysis. Residual plots and a QQ-plot are

shown in Figure 1.6. There seems to be no systematic pattern in any of the reidual plots; the residuals are nicely spread out.

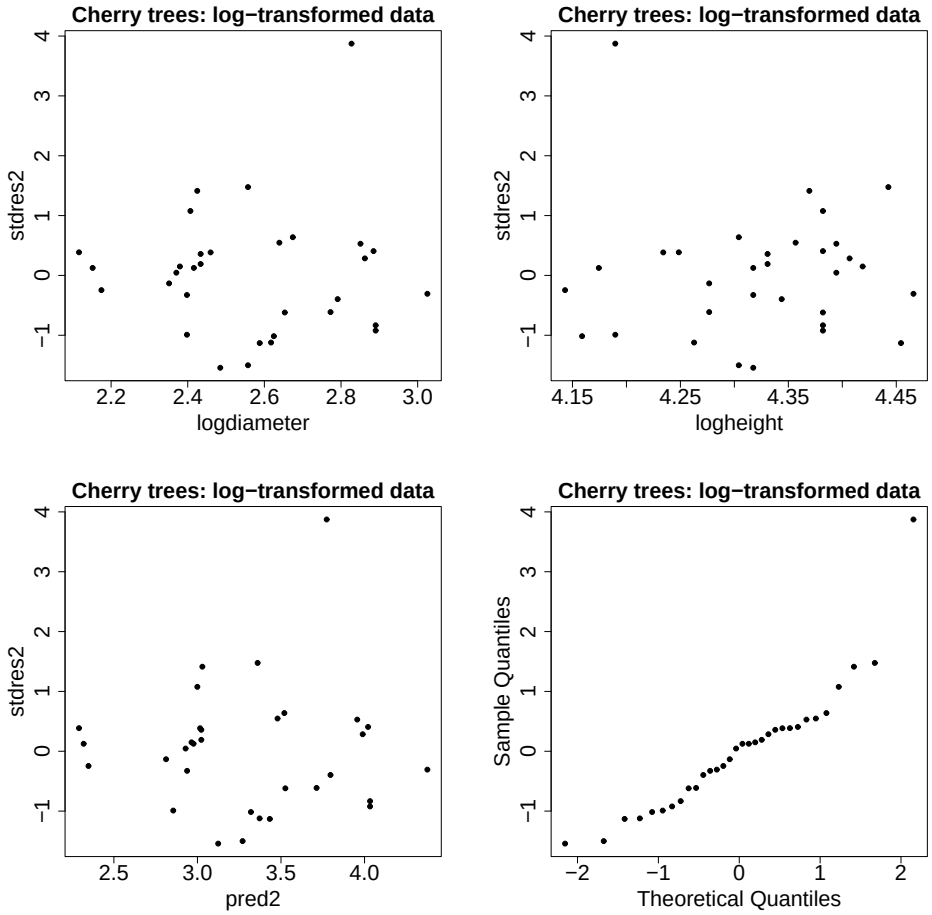


Figure 1.6: Residual plots and QQ-plot for the multiple linear regression model applied to the transformed data (all three variables log-transformed).

There is, however, an observation with a very large standardized residual, around 4. It comes from observation no. 26 which has a diameter of 16.9, a height of 66 and a volume of 64.3. Except for this observation the QQ-plot looks quite nice. The standardized residuals are approximately independent and standard normal so we would expect only 1 in 10000 standardized residuals to be as large numerically as 4. To have one with only 32 observations is quite suspicious.

The so-called *Cook's distance* is another way to to detect outliers or highly influential observations. Cook's distance is computed for each observation. Loosely speaking, D_j measures how much the predicted values change when the j'th observations is left out of the analysis. Formally, D_j is defined as

$$D_j = \frac{1}{(p+1)s^2}\sum_{i=1}^{n}(\hat{\mu}_i - \hat{\mu}_i^{(j)})^2, \quad j = 1, \ldots, n,$$

where $\hat{\mu}_i^{(j)}$ is the predicted value of y_i (the response) when the j'th observation is left out of the analysis. In other words, D_j is large for observations which has great influence on the estimates.

Figure 1.7 shows the Cook's distances against the observation number for the 32 cherry trees. One of the observations — no. 26, the same as before — is very different from the other ones.

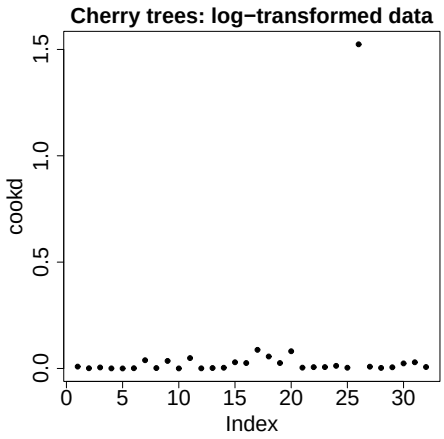


Figure 1.7: Cook's distances for the multiple regression model applied to the log-transformed data.

Now, what should we do about such an observation? On the one hand it is not a good idea to leave it out if there are no reasons why it should deviate. On the other hand it is not apt that the conclusions rely too strogly on one observation; then the conclusion is not very strong. Often, the best solution is to analyze the data both with and without the suspicious observation and state the conclusions from both analyses.

First, the analysis of the full dataset. From the computer output we find the following estimates (and standard errors):

$$\hat{\beta}_0 = -4.9907\ (1.0488), \qquad \hat{\beta}_1 = 2.1381\ (0.0982), \qquad \hat{\beta}_2 = 0.6490\ (0.2627), \qquad s^2 = 0.0132$$

Moreover, we find that the hypotheses $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$ are both convincingly rejected ($p < 0.0001$ for both hypotheses). The parameter β_2 is also significantly different from zero ($p = 0.02$), so the logarithmic height does indeed contribute to the explanation of the logarithmic volume, even after it has been adjusted for the logarithmic diameter. Note that this is contrary to the conclusion from the analysis of the original (untransformed) dataset.

Next, the analysis of the dataset without the influential observation, no. 26. The residual plots are not shown here but look fine (except a slight tendency that the variance increases with height; check it yourself!). The estimates are

$$\hat{\beta}_0 = -6.6316\ (0.7998), \qquad \hat{\beta}_1 = 1.9827\ (0.0750), \qquad \hat{\beta}_2 = 1.1171\ (0.2044), \qquad s^2 = 0.0066.$$

We see that the estimates are quite different from the estimates from the analysis of the full dataset. The variance estimates is halved when observation no. 26 is left out. Furthermore, the estimate of β_2 is almost doubled and is now highly significant ($p < 0.0001$). In the following we will use the estimates without observation no. 26.

Recall that we ended up using the log-transformed data. What does the analysis imply on the original scale? Consider two trees with same height, h, and diameters d_1 and d_2, respectively. The difference between the expected logarithmic volumes for these two trees is

$$\begin{aligned} \mathrm{E}(\log v_2 - \log v_1) &= \beta_0 + \beta_1 \log d_2 + \beta_2 \log h - (\beta_0 + \beta_1 \log d_1 + \beta_2 \log h) \\ &= \beta_1 (\log d_2 - \log d_1) \end{aligned}$$

Taking the exponential function on both sides yields (with somewhat sloppy notation)

$$\frac{v_2}{v_1} \approx \left(\frac{d_2}{d_1}\right)^{\beta_1}$$

This is of course not surprising, cf. the cone model. We have estimated β_1 to 1.98 (without observation no. 26). Hence, for example, an increment in diameter of 25% corresponds to a volume increment of 56% since $1.25^{1.98} = 1.56$. Moreover, the 95%-confidence interval for β_1 is $(1.83, 2.14)$, so the 95%-confidence interval for the volume increment is $(1.25^{1.83}, 1.25^{2.14}) = (1.50, 1.61)$, corresponding to (50%,61%).

Remember that prediction was an important issue in this example. Consider for example a tree with a diameter of 12 inches and a height of 75 feet. What is the predicted volume of this tree? Well, if we use the parameter estimates we get the predicted value

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \log(12) + \hat{\beta}_2 \cdot \log(75) = -6.6316 + 1.9827 \cdot \log(12) + 1.1171 \cdot \log(75) = 3.118$$

of the log-volume, and hence the a predicted volume of $\exp(3.118) = 22.61$. The prediction interval for the log-volume is computed to $(2.949, 3.288)$ by `R` and `SAS`, hence the prediction interval for the volume is $(19.08, 26.79)$. That is, with 95% probability a 12 feet high tree with diameter 12 inches will have a volume between 19 and 27 cubic feet.

Finally, a comment about the interpretation of the parameters, β_1 and β_2. Consider for a moment the raw residuals (`rvolume`) obtained by the simple linear regression of logarithmic volume on logarithmic diameter, and the raw residuals (`rheight`) obtained by the simple linear regression of logarithmic height on logarithmic diameter. The residuals represent the part of the logarithmic volume and logarithmic height, respectively, which is not explained by the logarithmic diameter. The two set of residuals are plotted against each other in Figure 1.8.

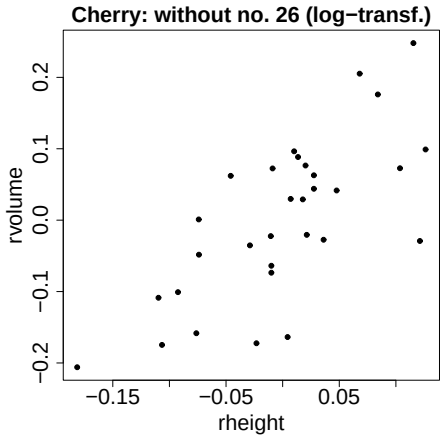


Figure 1.8: Added variance plot.

The estimated slope parameter in the corresponding simple linear regression of `rvolume` on `rheight` is 1.117, that is, exactly $\hat{\beta}_2$ from above. This is another way to say that β_2 measures how much of the logarithmic volume that can be explained by the logarithmic height, in excess of what has been explained by the logarithmic diameter. □

1.2.1 `R` programs and output

We use `lm` for analysis of multiple regression models.

Example 1.1 *(continued)*

Reading the data into R

Suppose that the dataset is saved in the ASCII-file `cherry.txt` as follows:

```
diameter height volume
  8.3            70           10.3
  8.6            65           10.3
  8.8            63           10.2
   .        .
   .        .            [more datalines here]
   .        .
 20.6            87           77.0
```

The dataset is read into R and attached, so we can use the variable names `volume`, `dimater` and `height`. The path to the file should of course be specified with the file name.

```
> cherry = read.table("cherry.txt",header=T)
> attach(cherry)
```

Scatter plots of response against covariates and a few general comments about graphical options

Simple scatter plots of `volume` against `diameter` and `height` are made by the `plot`-commands:

```
> plot(diameter,volume)
> plot(height,volume)
```

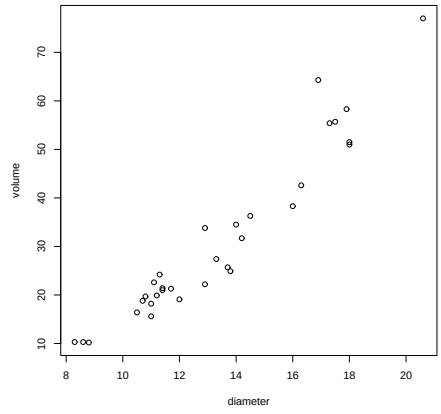


Figure 1.9: R-plot with default setting (no options), compare with the left part of Figure 1.3.

The first `plot`-command gives the scatter plot in Figure 1.9, which is not quite identical to the left plot in Figure 1.3:

- The plots in Figure 1.3 have solid bullets rather than open circles. This is obtained by the option `pch=16`; there are many other possibilities.

- The text and numbers on the axes are enlarged in Figure 1.3. This is obatined by the options `cex.axis=2` and `cex.lab=2`. Replace the "2" by another number if you want the text larger or smaller (the default is 1).
- There is a headline in Figure 1.3. This is obtained by the options `cex.main=2` and `main="Cherry trees: raw data"`.
- In order to make enough room for the enlarged labels and the title you may write something like `par(mar=c(5,4,2,1)+0.5)` before the `plot`-command.
- Moreover, the screen plot can be saved to a file with `dev.print`. Below is shown how to copy the screen plot to a pdf-file called `ex1_1.pdf`. If you prefer to use the eps-format instead, then write `dev.print(device=postscript, file="ex1_1.eps")`.

In summary, the plot in Figure 1.3 is made and written to a file with the commands:

```
> par(mar=c(5,4,2,1)+0.5)
> plot(diameter,volume,cex.axis=2,cex.lab=2,pch=16,
     cex.main=2,main="Cherry trees: raw data")
> dev.print(device=pdf, file="ex1_1.pdf")
```

In the following we write the simplest possible version of the `plot`-commands, without extra options. If you want to change the layout, then you may use some of the options above — or some of the numerous other options that exist in R, write `?par` for help.

Multiple linear regression on the original numbers

The multiple regression model of volume on diameter and height is analyzed with the `lm`-function: the response variable is written on the left hand side of a "tilde", the explanatory variables on the right hand side. Estimates from the fit are obtained by `summary`. A slightly edited version of the run looks as follows:

```
> model1 = lm(volume ~ diameter + height)
> summary(model1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -44.87058   10.62344  -4.224 0.000217 ***
diameter      5.16058    0.31958  16.148 4.94e-16 ***
height        0.09446    0.15476   0.610 0.546355

Residual standard error: 5.057 on 29 degrees of freedom
```

From the output we read off the estimates, $\hat{\beta}_0 = -44.87$ (intercept), $\hat{\beta}_1 = 5.16$ (the effect of diameter), $\hat{\beta}_2 = 0.09$ (the effect of height). We also find the corresponding standard errors as well as the t-tests for the hypotheses that the parameters are equal to zero. The residual standard deviation is estimated to $s = 5.057$ corresponding to $s^2 = 25.57$.

We get the predicted values by `predict(model1)` and the raw residuals by `residuals(model1)`. In order to get the standardized residuals the `MASS`-package should be loaded; then `stdres` gives the standardized residuals. Hence, the following commands result in three residual plots and a QQ-plot similar to those of Figure 1.4.

```
> library(MASS)
> pred1 = predict(model1)
> stdres1 = stdres(model1)

> plot(pred1,stdres1)
> plot(diameter,stdres1)
> plot(height,stdres1)
> qqnorm(stdres1)
```

Multiple linear regression on the log-transformed numbers

First the three variables are transformed with the natural logarithm, and the simple versions of Figure 1.5 are constructed:

```
> logvolume = log(volume)
> logdiameter = log(diameter)
> logheight = log(height)

> plot(logdiameter,logvolume)
> plot(logheight,logvolume)
```

Next, the multiple regression model of the logarithmic volume on the logarithmic diameter and the logarithmic height is fitted, and the residual plots and the QQ-plot are constructed (Figure 1.6). Moreover, the plot of Cook's distances is constructed (Figure 1.7).

```
> model2 = lm(logvolume ~ logdiameter + logheight)

> pred2 = predict(model2)
> stdres2 = stdres(model2)

> plot(pred2,stdres2)
> plot(logdiameter,stdres2)
> plot(logheight,stdres2)
> qqnorm(stdres2)

> cookd = cooks.distance(model2)
> plot(logvolume,cookd)
```

Finally, the parameter estimates are obtained with `summary`, which gives the following output:

```
> summary(model2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.9907     1.0488  -4.759 4.96e-05 ***
logdiameter   2.1381     0.0982  21.771  < 2e-16 ***
logheight     0.6490     0.2627   2.471   0.0196 *

Residual standard error: 0.1151 on 29 degrees of freedom
```

Multiple linear regression without observation no. 26

Finally, the analysis without the suspicious observation, no. 26. The variables without this observation are easily constructed as follows:

```
> logvolume3 = logvolume[-26]
> logdiameter3 = logdiameter[-26]
> logheight3 = logheight[-26]
```

The multiple linear regression model if fitted with only 31 observations, parameter estimates are obtained with `summary` as before, and confidence limits are found by `confint` (the residual plots, the QQ-plot and the Cook's plot are constructed exactly as before so the commands are left out):

```
> model3 = lm(logvolume3 ~ logdiameter3 + logheight3)
> summary(model3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.63162    0.79979  -8.292 5.06e-09 ***
logdiameter3  1.98265    0.07501  26.432  < 2e-16 ***
logheight3    1.11712    0.20444   5.464 7.81e-06 ***

Residual standard error: 0.08139 on 28 degrees of freedom

> confint(model3)
               2.5 %    97.5 %
(Intercept)  -8.269912 -4.993322
logdiameter3  1.828998  2.136302
logheight3    0.698353  1.535894
```

Prediction

In order to compute the prediction of log-volume for a tree with diameter 12 and height 75, we first create a new data set with the relevant values of the explanatory variables corresponding to `model3`. We use `predict` to compute the prediction and the prediction interval. Had we wanted the confidence interval instead (that is, an interval estimate for the expected volume of such a tree), we should write `interval="prediction"`.

```
> new = data.frame(logdiameter3=log(12), logheight3=log(75))
> pred = predict(model3,  newdata= new, interval="prediction")
> pred
          fit      lwr      upr
[1,] 3.118250 2.948598 3.287901
> exp(pred)
          fit      lwr      upr
[1,] 22.60677 19.07918 26.78659
```

1.2.2 SAS programs and output

We use `proc glm` for the analysis of (multiple) regression models since it works for linear models in general. Note that `proc reg` could also be used (and has special features for regression which are sometimes useful).

Example 1.1 *(continued)*

Reading the data into SAS

Suppose that the dataset is saved in the ASCII-file `cherry.txt` as follows:

```
diameter height volume
  8.3            70           10.3
  8.6            65           10.3
  8.8            63           10.2
  .        .
  .        .            [more datalines here]
  .        .
 20.6            87           77.0
```

Then a SAS-dataset `cherry` is constructed (and printed) in the following data-step. The path to the file should of course be specified with the file name. The option `firstobs=2` is needed because the first line is not a data line.

```
data cherry;
  infile 'c:\cherry.txt' firstobs=2;
  input diameter height volume;
proc print;
run;
```

Scatter plots of response against covariates

The scatter plots from Figure 1.3 are created with `proc gplot` as follows:

```
proc gplot data=cherry;
  plot volume*diameter=1 volume*height=1;
run;
```

Multiple linear regression on the original numbers

The multiple regression of volume on diameter and height is analyzed with `proc glm`. In the `model`-statement the response is on the right side of the equation sign while the explanatory variables are on the right side. As defualt `proc glm` makes both type I tests (successive tests) and type III tests (parallell tests); with the option `ss3` we ask SAS to only give the type III versions.

In this case we also make a new dataset `outvol` where the predicted values are saved as `p` and the standardized residuals (called `student` in SAS) are saved in the variable `sres`. The dataset also contains the original variables from the `cherry`-dataset. The residual plots similar to those in Figure 1.4 are constructed as the plots above, using the dataset `outvol`. The QQ-plot of the standardized residuals is made with `qqplot` in `proc univariate`. A straight line corresponding to the mean and standard deviation is added due to the option `normal(mu=est sigma=est)`.

```
proc glm data=cherry;
  model volume = diameter height / ss3;
  output out=outvol student=sres predicted=p;
run;
```

```
proc gplot data=outvol;
  plot sres*p=1 sres*diameter=1 sres*height=1;
run;

proc univariate data=outvol;
  qqplot sres / normal(mu=est sigma=est);
run;
```

The output from `proc glm` (slightly edited) looks like this:

```
                                The GLM Procedure

Dependent Variable: volume

                                        Sum of
   Source                     DF         Squares     Mean Square    F Value    Pr > F
   Model                       2     8492.936974     4246.468487     166.07    <.0001
   Error                      29      741.538026       25.570277
   Corrected Total            31     9234.475000

                 R-Square     Coeff Var      Root MSE    volume Mean
                 0.919699      16.18793      5.056706       31.23750


   Source                     DF     Type III SS     Mean Square    F Value    Pr > F
   diameter                    1     6667.711625     6667.711625     260.76    <.0001
   height                      1        9.526974        9.526974       0.37    0.5464

                                        Standard
          Parameter         Estimate           Error    t Value    Pr > |t|
          Intercept     -44.87057864     10.62343728      -4.22      0.0002
          diameter        5.16058067      0.31957874      16.15      <.0001
          height          0.09446500      0.15476086       0.61      0.5464
```

In the top we find the residual mean square error, $s^2 = 25.57$. The other parameter estimates are found in the bottom together with their standard errors and the t-tests for the hypothesis that the parameter equals zero. We see that $\hat{\beta}_0 = -44.87$ (the intercept), $\hat{\beta}_1 = 5.16$ (the effect of diameter), $\hat{\beta}_2 = 0.09$ (the effect of height). Note that the tests are also carried out as F-tests above the parameter estimates (parallell tests).

Multiple linear regression on the log-transformed numbers

Next, to the analysis of the log-transformed data. First, a new dataset `logcherry` with the log-transformed is constructed. Then the multiple regression analysis of the logarithmic volume on the logarithmic diamater and the logarithmic height is carried out as above. Note that the plot of Cook's distances (Figure 1.7) is also made; they are called `cookd` in `SAS`.

```
data logcherry;
  set cherry;
  logvolum = log(volume);
  logheig = log(height);
  logdiam = log(diameter);
```

```
run;

proc gplot data=logcherry;
  plot logvolum*logdiam=1 logvolum*logheig=1;
run;

proc glm data=logcherry;
  model logvolum = logdiam logheig;
  output out=outvol student=sres predicted=p cookd=cook;
run;

proc univariate data=outvol;
  qqplot sres / normal(mu=est sigma=est);
run;
```

An edited version of the output:

```
                                The GLM Procedure

Dependent Variable: logvolum

                                        Sum of
   Source                     DF         Squares     Mean Square    F Value    Pr > F
   Model                       2      8.69325063      4.34662531     328.08    <.0001
   Error                      29      0.38421406      0.01324876
   Corrected Total            31      9.07746469


                    R-Square     Coeff Var      Root MSE    logvolum Mean
                    0.957674      3.487375      0.115103         3.300570


   Source                     DF     Type III SS     Mean Square    F Value    Pr > F
   logdiam                     1      6.27988388      6.27988388     474.00    <.0001
   logheig                     1      0.08088327      0.08088327       6.10    0.0196


                                       Standard
         Parameter         Estimate         Error    t Value    Pr > |t|
         Intercept     -4.990707768    1.04879552      -4.76      <.0001
         logdiam        2.138067353    0.09820491      21.77      <.0001
         logheig        0.648977960    0.26265654       2.47      0.0196
```

Multiple linear regression without observation no. 26

The same analysis is now carried out without observation no. 26 which has diameter equal to 16.9 (as the only observation). A dataset, `logcherry2` is constructed without this observation and the multiple regression model is fitted as above. We also ask for the confidence limits for the mean parameters. This is done with the option `clparm` in the `model`-statement.

```
data logcherry2;
  set logcherry;
  if diameter = 16.9 then delete;
run;
```

```
proc glm data=logcherry2;
  model logvolum = logdiam logheig / clparm;
  output out=outvol student=sres predicted=p cookd=cook;
run;
```

The corresponding output:

```
                                The GLM Procedure

Dependent Variable: logvolum

                                     Sum of
      Source                DF       Squares     Mean Square    F Value    Pr > F
      Model                  2    8.12322615      4.06161308     613.19    <.0001
      Error                 28    0.18546337      0.00662369
      Corrected Total       30    8.30868953

                  R-Square     Coeff Var      Root MSE    logvolum Mean
                  0.977678      2.486793      0.081386         3.272732

      Source                DF     Type III SS     Mean Square    F Value    Pr > F
      logdiam                1      4.62750362      4.62750362     698.63    <.0001
      logheig                1      0.19778034      0.19778034      29.86    <.0001


                                Standard
Parameter         Estimate         Error    t Value    Pr > |t|      95% Confidence Limits
Intercept     -6.631617126    0.79978973      -8.29      <.0001    -8.269912123 -4.993322129
logdiam        1.982649910    0.07501061      26.43      <.0001     1.828997636  2.136302185
logheig        1.117123333    0.20443706       5.46      <.0001     0.698352998  1.535893668
```

Prediction

In order to compute the prediction of log-volume for a tree with diameter 12 and height 75, we first create a dataset with these values on the log-scale (the data set `tmp`). Then the values are appended to the dataset `logcherry2` with the log-values. The new dataset `pred` includes the 31 data points (not no. 26) as well as the values for prediction.

```
data tmp;
  logdiam = log(12);
  logheig = log(75);
run;

data pred;
  set logcherry2 tmp;
run;
```

Then the new dataset is used in a `proc glm`. The output with predicted values and lower and upper limits of the 95% prediction intervals are written to the dataset `predout` which is printed. Had we

wanted the confidence interval instead (that is, an interval estimate for the expected volume of such a tree), we should write `l95m` and `u95m` instead of `l95` and `l95`.

```
proc glm data=pred;
  model logvolum = logdiam logheig;
  output out=predout predicted=p l95=lowerpi u95=upperpi;
run;

proc print data = predout;
run;
```

The output from the `proc print` is like this:

Obs	diameter	height	volume	logvolum	logheig	logdiam	p	lowerpi	upperpi
1	8.3	70	10.3	2.33214	4.24850	2.11626	2.31027	2.13138	2.48916
2	8.6	65	10.3	2.33214	4.17439	2.15176	2.29788	2.11777	2.47799
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
31	20.6	87	77.0	4.34381	4.46591	3.02529	4.35545	4.17433	4.53657
32	.	.	.	.	4.31749	2.48491	3.11825	2.94860	3.28790

1.3 Polynomial regression

A special case of the multiple regression models occurs when powers of a covariate are used as explanatory variables. Consider the case with a response, y, and a single covariate, x. The polynomial regression of order p is given by

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + e_i, \quad i = 1, \ldots, n,$$

where as usual the e_i's are assumed to independent and $N(0, \sigma^2)$-distributed. The mean structure of the model describes y as a polynomium in x of order p, and the β's are the coefficients in the polynomium.

For example, a scatter plot of y against x might show "curvature", in the sense that the points "bend off" compared to a straight line. The curvature will be even more clear in the residual plot for the simple linear regression of y on x: negative residuals for small and large values of x, positive for medium values of x. A quadratic model, that is, a polynomial of order two, might then be appropriate:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, \ldots, n, \tag{1.2}$$

The polynomial model is a multiple linear regression model with the k'th covariate equal to the k'th power of the x, and the analysis hence follows the scheme from Section 1.2.

Example 1.2 (Optimal supply of nitrogen) An experiment of the effect of nitrogen supply on the yield of winter wheat has been carried out. The purpose is to be able to give advice to farmers about the optimal supply. The yield for six different amounts of supplied nitrogen are given in Table 1.3. This is of course an extremely small data material, and one should be very careful not to "overfit" such data, but let us analyze the data, anyway.

The data are plotted in Figure 1.10. The relationship between nitrogen supply and yield is obviously not linear — there is a clear curvature. We fit the quadratic model, see (1.2), with the yield as response

Nitrogen (kg/ha)	Yield (hkg/ha)
0	23.1590
50	38.6801
100	54.7080
150	57.4650
200	62.7166
250	62.3278

Table 1.3: The nitrogen data.

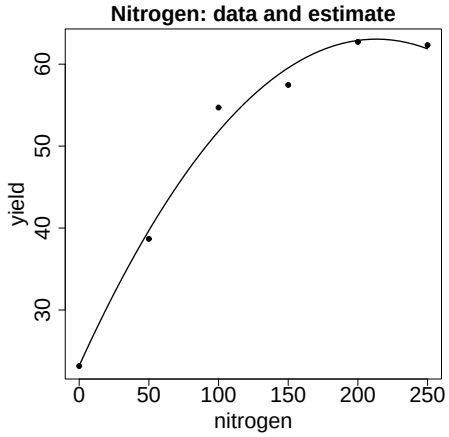


Figure 1.10: The nitrogen data: the data points with the estimated second order polymial.

and nitrogen supply and squared nitrogen supply as covariates. The estimated coefficients (and their standard errors) are

$$\hat{\beta}_0 = 23.2054\ (1.9575), \qquad \hat{\beta}_1 = 0.3737\ (0.03683), \qquad \hat{\beta}_2 = -0.0008761\ (0.0001414)$$

and the residual variance is estimated to $s^2 = 4.66$. The estimated (solid) curve in Figure 1.10 is thus given by

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 = 23.2054 + 0.3737 \cdot x - 0.0008761 \cdot x^2.$$

The experimenter was particularly interested in an optimal supply of nitrogen, that is, the amount of nitrogen that gives the highest yield. We easily get an estimate of this optimal supply from the quadratic model: if β_2 is negative then the largest value of $\beta_0 + \beta_1 x + \beta_2 x^2$ is obtained for $x = -\beta_1/(2\beta_2)$. Hence the optimal nitrogen supply is estimated to

$$-\frac{\hat{\beta}_1}{2\hat{\beta}_2} = -\frac{0.3737}{2 \cdot (-0.0008761)} = 213.27.$$

Computation of the standard error and the confidence interval for the optimal nitrogen supply are not straight-forward because the function $-\beta_1/(2\beta_2)$ is not linear in (β_1, β_2). However, an approximate 95%-confidence interval turns out to be $(164.20, 262.34)$. This is extremely wide (and probably useless in practice), due to the small number of observations. □

1.3.1 R-programs and output

Example 1.2 *(continued)*

Reading the data

Since the dataset has only six observations we enter them manually. Furthermore, the squared nitrogen supply is constructed and called `nitrogen2`.

```
> nitrogen = c(0, 50, 100, 150, 200, 250)
> yield = c(23.1590, 38.6801, 54.7080, 57.4650, 62.7166, 62.3278)
> nitrogen2 = nitrogen*nitrogen
```

The quadratic model

The quadratic model is fitted with `lm` with `yield` as response and `nitrogen` and `nitrogen2` as covariates. The parameter estimates can be read off from the `summary`-output.

```
> model1 = lm(yield ~ nitrogen + nitrogen2)
> summary(model1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.2054107  1.9575403  11.854  0.00129 **
nitrogen     0.3737180  0.0368264  10.148  0.00204 **
nitrogen2   -0.0008761  0.0001414  -6.196  0.00847 **

Residual standard error: 2.16 on 3 degrees of freedom
```

Now to the construction of Figure 1.10 where the data points are plotted with the estimated second order polynomium. First, the data points are plotted with `plot` as usual. Next, the estimated function is superimposed with the `points`-command. `points` works similar to `plot`, except that it superimposes the points/curve onto the present plot, instead of making a new plot. In order to make a smooth curve we construct a vector `x` of length 50 with equidistant values from zero to 250. For each of the 50 `x`-values we compute the value of the estimated polynomium,

$$\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 = 23.2054 + 0.3737 \cdot x - 0.0008761 \cdot x^2$$

and store the results in the vector `y`. Finally `y` is plotted against `x` with `points`. The option `type="l"` has the effect that the 50 points are joined (by linear interpolation), so a curve rather than 50 points are plotted.

```
> plot(nitrogen,yield)
> x = seq(0,250,length=50)
> y = 23.2054 + 0.3737 * x - 0.0008761 * x^2
> points(x,y,type="l",lwd=2)
```

1.3.2 SAS-programs and output

Example 1.2 *(continued)*

Reading the data

The data are read into the dataset `nitrogen` by `cards`. Moreover, the variable `nitro2` with squared values is constructed.

```
data nitrogen;
  input nitro yield;
  nitro2 = nitro*nitro;
  cards;
  0   23.1590
  50  38.6801
  100 54.7080
  150 57.4650
  200 62.7166
  250 62.3278
  ;
run;
```

The quadratic model

The quadratic model with `nitro` and `nitro2` is fitted with `proc glm` in the usual way:

```
proc glm data=nitrogen;
  model yield = nitro nitro2 / ss3;
run;
```

The output is as follows:

```
                                The GLM Procedure

Dependent Variable: yield

                                        Sum of
        Source                     DF       Squares     Mean Square    F Value    Pr > F
        Model                       2    1226.014110      613.007055     131.41    0.0012
        Error                       3      13.994999        4.665000
        Corrected Total             5    1240.009109


                      R-Square     Coeff Var      Root MSE    yield Mean
                      0.988714      4.333351      2.159861      49.84275


        Source                     DF     Type III SS     Mean Square    F Value    Pr > F
        nitro                       1     480.4205493     480.4205493     102.98    0.0020
        nitro2                      1     179.0973266     179.0973266      38.39    0.0085

                                              Standard
                Parameter        Estimate          Error    t Value    Pr > |t|
```

```
Intercept      23.20541071      1.95754033      11.85      0.0013
nitro           0.37371796      0.03682638      10.15      0.0020
nitro2         -0.00087611      0.00014140      -6.20      0.0085
```

A plot corresponding to that in Figure 1.10 is made as follows. First pairs of fictious nitrogen values and the corresponding estimated yield values are constructed and saved in the dataset `fit`. The `x`-values are $0, 5, \ldots, 250$. The estimated values, `y`, are computed from the parameter estimates. Then the two datasets, `nitrogen` and `fit` are merged because we want to plot variables from both in the same plot. Then the six data points and the (`x`,`y`)-values are plotted in the same plot. This is obtained by `overlay`. Note that the data points are points whereas the (`x`,`y`)-values are joined to a curve; this is obtained with `symbol1` and `symbol2`.

```
data fit;
  do x = 0 to 250 by 5;
    y = 23.2054 + 0.3737 * x - 0.0008761 * x*x;
   output;
  end;
run;

data fit;
  merge fit nitrogen;
run;

symbol1 i=none v=dot c=black;
symbol2 i=join v=none c=black;

proc gplot data = fit;
  plot yield*nitro=1 y*x = 2 / overlay;
run;
```

Chapter 2

Models with both factors and covariates

Usually the term *regression models* is used for models where all the explanatory variables are quantitative and are used as covariates. If all the explanatory variables are qualitative, that is, each variable groups the experimental units into certain groups we usually speak about *analysis of variance* (ANOVA). In this case the explanatory variables are also called factors. In this chapter we consider models where there are both factors and covariates. In that case we sometimes use the term *analysis of covariance* (ANCOVA). Models of all these types are *linear models*.

Actually, all three model types can be thought of as multiple regression models if certain so-called dummy-variables are introduced. This is "just" a matter of parameterization, and we will not discuss this issue any further. Rather, we will think of the models in what we believe is a more natural way: as models with both covariates and factors as explanatory variables.

2.1 A few general comments on computational aspects

Gaussian linear models are fitted with `lm` in `R` and with `proc gml` in `SAS`. In both programs it is important to specify if (numeric) variables are to be treated as factors or covariates. In `R`, we write `factor(x)` if a numerical variable `x` is to be treated as a factor. In `SAS` the variable should be included in a `CLASS` statement in order to be treated as a factor.

`R` and `SAS` use different default parameterizations which sometimes makes it a litte hard to compare output from the two programs directly. But, indeed, they come out with the same results. *It is very important to be able to read off the parameter estimates and test summaries correctly*, so make sure you are able to do so in `R` or `SAS`, whatever program you use!

Note that it is often useful to specify the same model in different ways (with different parameterizations). One specification may be useful for the model reduction part of the analysis whereas another may be more adequate for reporting the results. Again, make sure you understand the different specifications of the models. We will pay quite some attention to these matters in the computer sections.

2.2 An example

Let us consider a simple example with one factor and one covariate.

Example 2.1 (Cauliflowers) In a growth study the number of leaves on cauliflower plants were counted at seven dates in 1956 and at seven dates at 1957. At each date 10 plants were examined, and the average number of leaves was computed. Moreover, the accumulated day degrees over $32°F$ was registered and divided by 100.

Year 1956		Year 1957	
day degrees	leaves	day degrees	leaves
4.5	3.8	4.5	6.0
7.5	6.2	8.0	8.5
9.5	7.2	9.5	9.1
10.5	8.7	11.5	12.0
13.0	10.2	13.0	12.6
16.0	13.5	14.0	13.3
18.0	15.0	16.5	15.2

Table 2.1: Cauliflowers: the data.

The data are listed in Table 2.1 and plotted in the left part of Figure 2.1. The two lines are the estimated lines from the simple linear regressions for each year separately. We see that the number of cauliflower leaves increases as the number of day degrees increases, and that the relationships seem to be approximately linear for both years. Also, the plot indicates that the two lines are parallell, but we will test that hypothesis in a statistical model in a moment.

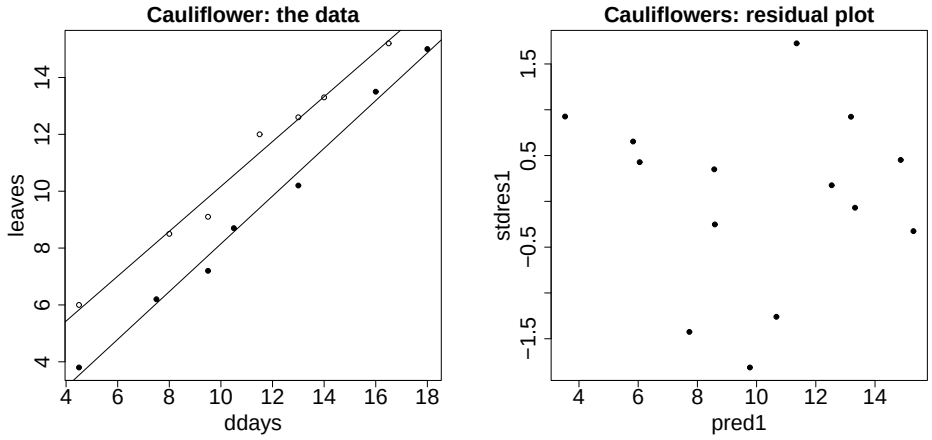


Figure 2.1: Caulifolowers: The data with fitted lines for each year (left) and residual plot for model with different slopes.

As a starting point we take the model with two different regression lines for 1956 and 1957 (different intercepts and different slopes). For the i'th observation, let y_i be the average number of cauliflower leaves, x_i the day degrees and $\texttt{year}_i$ be either 1956 or 1957. The model may then be written as

$$y_i = \alpha(\texttt{year}_i) + \beta(\texttt{year}_i) \cdot x_i + e_i$$

where the e_i's are independent $N(0, \sigma^2)$. The intercepts are $\alpha(1956)$ and $\alpha(1957)$, and the slopes are $\beta(1956)$ and $\beta(1957)$. The residual plot for the model is shown in the right part of Figure 2.1. It does not give rise to any objections. Neither does the QQ-plot (not shown).

It is important that `year` is used correctly, as a factor (qualitative variable). The variable is numeric with values 1956 and 1957, but the actual values do not matter — they could as well have been named 1 and 2 (or A and B, or whatever). In `R`, remember to use `factor(year)`; in `SAS`, remember to put `year` in a `CLASS` statement. The dependence on slope of year is specified to `R` and `SAS` as an interaction between the year factor and the day degrees covariate.

From the computer-output we get the estimates (see the details in the computers sections)

$$\begin{aligned}\hat{\alpha}(1956) &= -0.249; \qquad \hat{\beta}(1956) = 0.840\\ \hat{\alpha}(1957) &= \quad 2.276; \qquad \hat{\beta}(1957) = 0.789\\ s^2 &= 0.1653.\end{aligned}$$

One of the interesting hypotheses is that of a common slope (parallell lines) for the two years, H_0 : $\beta(1956) = \beta(1957)$. From the estimates we find that $\hat{\beta}(1957) - \hat{\beta}(1956) = -0.051$ which turns out to have a standard error of 0.054. The hypothesis is thus tested with a t-test on

$$t = \frac{-0.051}{0.054} = -0.934$$

which is $t(10)$-distributed under H_0. The corresponding p-value is 0.37, so the hypothesis of parallell lines is not rejected.

We therefore fit the model with common slope

$$y_i = \alpha(\texttt{year}_i) + \beta \cdot x_i + e_i \tag{2.1}$$

From the computer output we find that $\hat{\alpha}(1957) - \hat{\alpha}(1956) = 1.9624$ with a standard error of 0.2162. Hence, the number of leaves is significantly larger in 1957 than in 1956, even after adjustment for day degrees ($p < 0.0001$). The hypothesis $H_0 : \beta = 0$ is also rejected ($p < 0.0001$) so the day degrees has significant effect on the number of leaves.

In other words, model (2.1) cannot be reduced and is the final model. We find the following estimates of the mean parameters (with standard errors in parenthesis):

$$\hat{\alpha}(1956) = -0.0097\ (0.3365); \quad \hat{\alpha}(1957) = 1.9528\ (0.3298); \quad \hat{\beta} = 0.8186\ (0.0266)$$

Moreover, the variance estimate is $s^2 = 0.1634$. □

In the example we compared *two* regression lines. In particular we tested the hypothesis that the two slopes were the same with a t-test. A t-test can be used because the hypothesis reduces the number of parameters in the model with one. Had there been three regression lines (three years) the hypothesis of common slopes would have been $\beta(1) = \beta(2) = \beta(3)$, reducing the number of parameters by two (from three intercepts and three slopes to three intercepts and one slope). In that case the hypothesis should have been tested with an F-test.

Let us remind ourselves of the general F-test for linear hypothesis testing in linear models. Consider two models, A and B, where model B is a sub-model of model A. Let SS_e^A and SS_e^B be the corresponding residual sums of squares, and DF_e^A and DF_e^B be residual degrees of freedom. Then the test for model B against model A can be carried out as a test on

$$F_{AB} = \frac{(\mathrm{SS}_e^B - \mathrm{SS}_e^A)/(\mathrm{DF}_e^B - \mathrm{DF}_e^A)}{\mathrm{SS}_e^A/\mathrm{DF}_e^A} \tag{2.2}$$

which is F-distributed with $(\mathrm{DF}_e^B - \mathrm{DF}_e^A, \mathrm{DF}_e^A)$ degrees of freedom if model B is true. The test is one-sided with large values critical (large values indicate that model B is not true).

2.2.1 R programs and output

Example 2.1 *(continued)*

Reading the data into R *and construction the plot of the data*

Suppose that the dataset is saved in the ASCII-file `cauli.txt` as follows:

```
year ddays leaves
1956  4.5  3.8
1956  7.5  6.2
  .     .
  .     .     [more datalines here]
  .     .
1957 16.5 15.2
```

The data is read into R and attached. Moreover, a year-factor named `yearfac` is constructed which should be used in the analysis.

```
> cauli = read.table("cauli.txt", header=T)
> attach(cauli)
> yearfac = factor(year)
```

The left plot in Figure 2.1 is constructed as follows. First an "empty" plot with only axes and labels is constructed (`type="n"`) in order to get reasonable axes. Next, the data points for 1956 and 1957 are added with `points` and two different symbols. Then the simple linear regressions for each year are fitted and the parameters are used for the fitted lines which are added to the plot with `abline`.

```
> plot(ddays,leaves)
> points(ddays[year==1956],leaves[year==1956],pch=16)
> points(ddays[year==1957],leaves[year==1957],pch=1)

> model1956 = lm(leaves[year==1956] ~ ddays[year==1956])
### The estimated line is:  -0.24915 + 0.83980 * x for 1956
> abline(-0.24915,0.83980)

> model1957 = lm(leaves[year==1957] ~ ddays[year==1957])
### The estimated line is: 2.27622 + 0.78918 * x for 1957
> abline(2.27622,0.78918)
```

Analysis with test for parallell lines (common slope)

The model with different slopes is specified with an interaction between `yearfac` and `ddays`. The model can be specified in various ways; `model1` is appropriate for testing for parallell lines.

```
> model1 = lm(leaves ~ yearfac + ddays + yearfac:ddays)
> summary(model1)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.24915    0.42452  -0.587  0.57028
yearfac1957        2.52536    0.64032   3.944  0.00276 **
```

```
ddays              0.83980    0.03506  23.950 3.66e-10 ***
yearfac1957:ddays -0.05062    0.05416  -0.935  0.37199

Residual standard error: 0.4066 on 10 degrees of freedom
```

The reported estimates are

$$\hat{\alpha}(1956) = -0.249; \qquad \hat{\beta}(1956) = \quad 0.840;$$
$$\hat{\alpha}(1957) - \hat{\alpha}(1956) = 2.525; \qquad \hat{\beta}(1957) - \hat{\beta}(1956) = -0.051$$

so we get

$$\hat{\alpha}(1956) = -0.249; \qquad \hat{\beta}(1956) = 0.840$$
$$\hat{\alpha}(1957) = -0.249 + 2.525 = 2.276; \qquad \hat{\beta}(1957) = 0.840 - 0.051 = 0.789$$

These estimates would have come out immediately (supplied with standard errors) had we specified the model without the main effect of `ddays` and with no intercepts as follows:

```
> model1a = lm(leaves ~ yearfac + yearfac:ddays - 1)
> summary(model1a)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
yearfac1956       -0.24915    0.42452  -0.587 0.570283
yearfac1957        2.27622    0.47937   4.748 0.000782 ***
yearfac1956:ddays  0.83980    0.03506  23.950 3.66e-10 ***
yearfac1957:ddays  0.78918    0.04128  19.118 3.33e-09 ***

Residual standard error: 0.4066 on 10 degrees of freedom
```

It is important to realize that the two models, `model1` and `model1a`, are identital — only the parameterizations differ. Very often, just as here, one specification of the model is the most useful for hypothesis testing whereas another a most useful for reporting the estimates and confidence intervals.

Already from the `summary(model1)`-output above we see that the interaction between `yearfac` and `dday` is not significant ($p = 0.37$), so there is no indication of different slopes. We can see this from the `summary`-output only because the `year`-factor has no more than two levels. Had there been three years, say, the interaction would have been described by two parameters, and the hypothesis should be tested with the general F-test (2.2). In `R` this is easily carried out by fitting the model under the hypotheses and comparing the two with `anova`. In fact, this will be our usual way of testing hypotheses in linear models. In this case we thus fit the model with common slope and use `anova`:

```
> model2 = lm(leaves ~ yearfac + ddays)
> anova(model2,model1)
Analysis of Variance Table

Model 1: leaves ~ yearfac + ddays
Model 2: leaves ~ yearfac + ddays + yearfac:ddays
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     11 1.79725
2     10 1.65286  1   0.14439 0.8736  0.372
```

We get the same p-value (of course), 0.37. We thus accept `model2` and use `summary` to see the results. We also fit the model without intercept:

```
> summary(model2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.009691   0.336543  -0.029    0.978
yearfac1957  1.962451   0.216193   9.077 1.93e-06 ***
ddays        0.818580   0.026570  30.808 4.99e-12 ***

Residual standard error: 0.4042 on 11 degrees of freedom

> model2a = lm(leaves ~ yearfac + ddays - 1)
> summary(model2a)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
yearfac1956 -0.009691   0.336543  -0.029   0.9775
yearfac1957  1.952760   0.329796   5.921   0.0001 ***
ddays        0.818580   0.026570  30.808 4.99e-12 ***

Residual standard error: 0.4042 on 11 degrees of freedom
```

From `model2` we easily find the relavant t-values and p-values for the hypotheses $H_0 : \alpha(1956) = \alpha(1957)$ and $H_0 : \beta = 0$. From `model2a` we read off the estimates.

2.2.2 SAS programs and output

Example 2.1 *(continued)*

Reading the data into SAS

Suppose that the dataset is saved in the ASCII-file `cauli.txt` as follows:

```
year ddays leaves
1956  4.5  3.8
1956  7.5  6.2
  .     .
  .     .     [more datalines here]
  .      .
1957 16.5 15.2
```

The the dataset saved in the SAS dataset `cauli` by the following program lines:

```
data cauli;
  infile 'c:\cauli.txt' firstobs=2;
  input year ddays leaves;
run;
```

The plot of the data

The left plot of Figure 2.1 is constructed as follows. First, a dataset with only the observations from 1956 is made (`cauli1`). This is used for the simple linear regression of leaves on day degrees for 1956. The predicted values are needed for the estimates line and are saved in the dataset `out1`. Similarly for the

1957 observations. Then the two output datasets are merged, in order to make one plot with both years. Finally, the plots are made with gplot as usual.

```
data cauli1;
  set cauli;
  ddays1 = ddays;
  leaves1 = leaves;
  where year=1956;
proc glm data = cauli1;
  model leaves1 = ddays1;
  output out = out1 predicted = p1;
run;

data cauli2;
  set cauli;
  ddays2 = ddays;
  leaves2 = leaves;
  where year=1957;
proc glm data = cauli2;
  model leaves2 = ddays2;
  output out = out2 predicted = p2;
run;

data forplot;
  merge out1 out2;
run;

symbol1 i=none v=dot c=black;
symbol2 i=none v=circle c=black;
symbol3 i=join v=none c=black;

proc gplot data = forplot;
  plot leaves1*ddays1=1 leaves2*ddays2=2 p1*ddays1=3 p2*ddays=3 / overlay;
run;
```

Analysis with test for parallell lines

The model with different slopes is specified with an interaction between year and day degrees. The model can be specified in various ways corresponding to different parameterizations. some are more appropriate for model reduction (hypothesis testing); some are more appropriate for picking out the parameter estimates.

The first one is suitable for testing:

```
proc glm data = cauli;
  class year;
  model leaves = year ddays year*ddays / ss3 solution;
  output out = cauliout predicted = pred student = stdres;
run;

proc gplot data = cauliout;
  plot stdres*pred;
run;
```

The output is the following (slightly edited as usual):

```
                                The GLM Procedure

                            Class Level Information

                          Class         Levels    Values
                          year               2    1956 1957

Dependent Variable: leaves

                                          Sum of
    Source                      DF         Squares     Mean Square    F Value    Pr > F
    Model                        3     165.6764286      55.2254762     334.12    <.0001
    Error                       10       1.6528571       0.1652857
    Corrected Total             13     167.3292857

    Source                      DF     Type III SS     Mean Square    F Value    Pr > F
    year                         1       2.5709280       2.5709280      15.55    0.0028
    ddays                        1     149.5113969     149.5113969     904.56    <.0001
    ddays*year                   1       0.1443907       0.1443907       0.87    0.3720

                                                    Standard
         Parameter                 Estimate             Error    t Value    Pr > |t|
         Intercept              2.276215022 B      0.47936770       4.75      0.0008
         year       1956     -2.525364863 B      0.64031966      -3.94      0.0028
         year       1957      0.000000000 B       .                .         .
         ddays                 0.789175258 B      0.04127925      19.12      <.0001
         ddays*year 1956      0.050622829 B      0.05416197       0.93      0.3720
         ddays*year 1957      0.000000000 B       .                .         .
```

We immediately see that the interaction between year and day degress is not significant ($p = 0.37$), so there is no indication of different slopes.

The reported estimates are

$$\hat{\alpha}(1957) = 2.276; \qquad \hat{\beta}(1957) = \quad 0.789;$$
$$\hat{\alpha}(1956) - \hat{\alpha}(1957) = -2.525; \qquad \hat{\beta}(1956) - \hat{\beta}(1957) = 0.051$$

so we get

$$\hat{\alpha}(1957) = 2.276; \qquad \hat{\beta}(1957) = 0.789$$
$$\hat{\alpha}(1956) = 2.276 - 2.525 = -0.249; \qquad \hat{\beta}(1956) = 0.789 + 0.051 = 0.840$$

These estimates would have come out immediately (supplied with standard errors) had we specified the model without the main effect of day degrees and with no intercept:

```
proc glm data = cauli;
  class year;
  model leaves = year year*ddays / noint ss3 solution;
run;
```

with output

```
                                The GLM Procedure

Dependent Variable: leaves

                                        Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        4     1591.797143      397.949286    2407.64    <.0001
Error                       10        1.652857        0.165286
Uncorrected Total           14     1593.450000


Source                      DF     Type III SS     Mean Square    F Value    Pr > F
year                         2       3.7836322       1.8918161      11.45    0.0026
ddays*year                   2     155.2185715      77.6092857     469.55    <.0001

                                                Standard
    Parameter                Estimate           Error    t Value    Pr > |t|
    year       1956     -0.249149841      0.42451841      -0.59      0.5703
    year       1957      2.276215022      0.47936770       4.75      0.0008
    ddays*year 1956      0.839798087      0.03506484      23.95      <.0001
    ddays*year 1957      0.789175258      0.04127925      19.12      <.0001
```

It is important to realize that the two models are identital — only the parameterizations differ. Very often, just as here, one specification of the model is the most useful for hypothesis testing whereas another a most useful for reporting the estimates and confidence intervals (but not for testing).

Since the hypotheses of parallell lines was not rejected, we fit the model with a common slope. First a version suitable for testing:

```
proc glm data = cauli;
  class year;
  model leaves = year ddays / ss3 solution;
run;
```

with output

```
                                The GLM Procedure

Dependent Variable: leaves

                                        Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        2     165.5320379      82.7660190     506.57    <.0001
Error                       11       1.7972478       0.1633862
Corrected Total             13     167.3292857


Source                      DF     Type III SS     Mean Square    F Value    Pr > F
year                         1      13.4626351      13.4626351      82.40    <.0001
ddays                        1     155.0741808     155.0741808     949.13    <.0001

                                                Standard
    Parameter                Estimate           Error    t Value    Pr > |t|
    Intercept          1.952760141 B      0.32979633       5.92      0.0001
    year       1956   -1.962451499 B      0.21619295      -9.08      <.0001
```

```
year      1957      0.000000000 B        .              .         .
ddays               0.818580247        0.02657046      30.81      <.0001
```

We see that degree days and year are both significant. The α- and β-estimates are most easily obtained by fitting the model model with no intercept (`noint`):

```
proc glm data = cauli;
  class year;
  model leaves = year ddays / noint ss3 solution;
run;
```

with output (strongly edited, this time):

```
                                        Standard
Parameter                Estimate          Error     t Value    Pr > |t|

year      1956      -0.009691358      0.33654257       -0.03      0.9775
year      1957       1.952760141      0.32979633        5.92      0.0001
ddays                0.818580247      0.02657046       30.81      <.0001
```

Chapter 3

Non-linear regression

In Chapters 1 and 2 we have extended the simple linear regression model by incorporating more explanatory variables (of different kinds) in the model. In this chapter we will move in another direction, and consider non-linear regression models. Note that we then leave the nice world of linear models where all distribution results are exact (provided that the model is true). However, we still assume that the measurement errors are Gaussian and the results from the linear models hold approximately. For example, the test statistic (2.2) still makes sense and is approximately F-distributed.

3.1 General comments

Consider the situation with a response variable, y and a single covariate, x. Assume that the plot of y against x shows a non-linear relationship. We have already discussed two ways of handling non-linear relationships: In some cases a linear relationship can be obtained after transformation of the response and/or the covariate. In other cases a quadratic (or another polynomial) model is perhaps appropriate.

In yet other cases none of these solutions apply. Sometimes there is a theoretical model from the subject area, for example a physical law, suggesting a certain (non-linear) relationship. Or perhaps previous studies have shown a particular non-linear relationship. In both cases the knowledge can form the basis for a statistical model, in this case a non-linear regression model.

Assume that theory claims that $y \approx f(x, \theta_1, \ldots, \theta_p)$ for a known non-linear function f and unknown parameters $\theta_1, \ldots, \theta_p$. The corresponding non-linear regression model is then given by

$$y_i = f(x_i, \theta_1, \ldots, \theta_p) + e_i, \quad i = 1, \ldots, n$$

where as usual $e_1, \ldots, e_n \sim N(0, \sigma^2)$ are independent. In particular we assume variance homogeneity. We are interested in estimating the relationship between x and y, that is, we are interested in estimating the parameters $\theta_1, \ldots, \theta_p$.

We use the least squares estimates. In other words, $(\hat{\theta}_1, \ldots \hat{\theta}_p)$ is the vector of values for which

$$\sum_{i=1}^{n} \left(y_i - f(x_i, \theta_1, \ldots, \theta_p)\right)^2$$

is the least possible. For linear models (f linear in the θ's) the least squares estimator can be found explicit, but for non-linear models the least squares function must be mimimized by the use of some numerical algorithm.

In R parameters are estimated with the `nls`-function. Apart from the function f defining the model some starting values of $\theta_1, \ldots, \theta_p$ must be supplied to `nls`. It is the starting values for the numerical procedure. In SAS parameters are estimates with `proc nlin`. Again starting values, or at least a range of possible starting values, must be specified. See the computer sections for details on these issues.

For some models (as the one in the example below) the interpretation of the parameters is straightforward and good starting values can easily be guessed from the graph of y against x. For other models it might be much more difficult to come up with good starting values, and one may have to try different starting values before the algorithm converges.

Note that polynomial models are often chosen on the basis on empirical findings rather than on theoretical grounds: it turns out that a certain polynomial model fits the data well, but there is no theoretical justification for the model. On the other hand, a non-linear model is most often based on some previous knowledge and one has to check if the model fits the actual data reasonably well.

3.2 An example

Let us consider an example on chemical reactions.

Example 3.1 (Puromycin) The reaction time (or actually the reaction velocity) for a certain chemical process has been measured for six concentrations of one of the chemicals involved, namely the enzyme puromycin. For each of the six concentrations there are two independent measurements of the reaction time. The data is listed in Table 3.1, and the reaction time is plotted against concentration in Figure 3.1.

Concentration	Reaction time	
0.02	76	47
0.06	97	107
0.11	123	139
0.22	159	152
0.56	191	201
1.10	207	200

Table 3.1: The puromycin data.

Evidently, the reation time increases as the concentration of puromycin increases, and the relationship is clearly not linear. Rather, it looks like the reaction time reaches a plateau as the concentration increases. A possible model with this property is given by the so-called *Michaelis-Menten kinetics* which states that

$$y \approx \frac{\alpha \cdot x}{\beta + x}. \tag{3.1}$$

Here α is the value at the plateau and β is the concentration for which the reaction time is half the value at the plateau. (Check is yourself: what happens when x is very large and for $x = \beta$?)

First, note that if use the reciprocal function on both the right hand side and the left hand side, then (3.1) gives

$$\frac{1}{y} \approx \frac{\beta + x}{\alpha \cdot x} = \frac{1}{\alpha} + \frac{\beta}{\alpha} \cdot \frac{1}{x}$$

suggesting a simple linear regression of $1/y$ on $1/x$. The simple linear regression analysis is illustrated in Figure 3.2. The left part shows $1/y$ plotted against $1/x$, and the right part is the residual plot corresponding to the simple linear regression. Both plots indicate that the systematic part of the model is indeed appropriate both also that the variance increases as $1/x$ increases. This is confirmed by Bartlett's

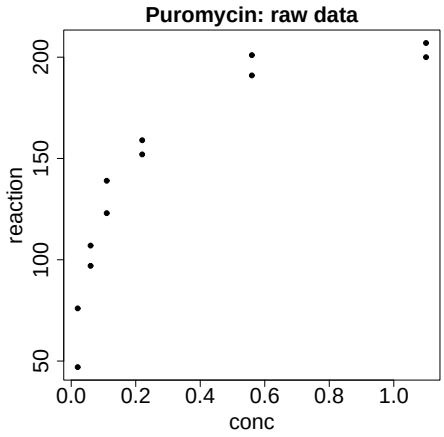


Figure 3.1: The puromycin data.

test for equal variances which compares the variances in the six concentration groups: the test statistic is 12.9 which should be compared to the $\chi^2(5)$-distribution, giving a p-value of 0.016. All together, we conclude that the simple linear regression of $1/y$ on $1/x$ does not describe the data well!

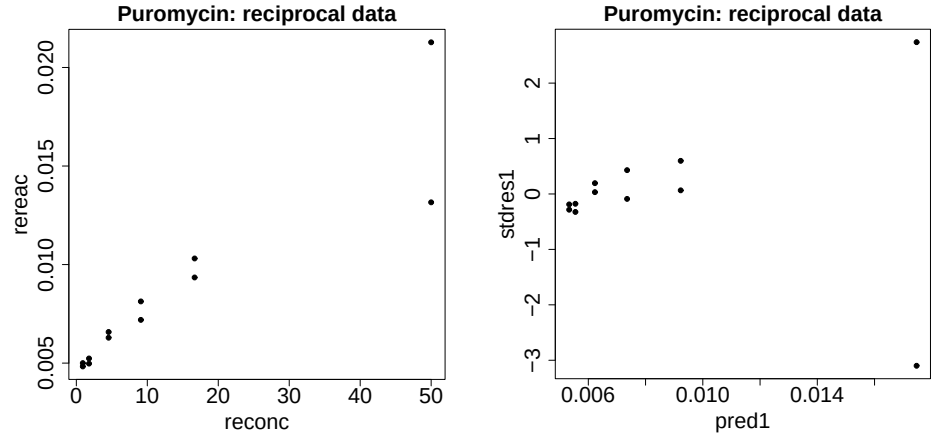


Figure 3.2: The reciprocal puromycin data (left), and the residual plot for the simple linear regression of the reciprocal data (right).

Looking at the plot of the raw data again (Figure 3.1) we see that there seems to variance homogeneity on the original scale. This is confirmed by Bartlett's test on the raw data: we now get a test statistic of 2.46 and a p-value of 0.78. This, together with the observation that the Michalis-Menten model seemed to catch the mean structure of the data, suggests a non-linear regression model based on (3.1). Hence, consider the model

$$y_i = \frac{\alpha \cdot x_i}{\beta + x_i} + e_i \tag{3.2}$$

with $e_i \sim N(0, \sigma^2)$ independent. We want to estimate the parameters α and β.

As mentioned already we use the least squares method for estimation. This requires numerical optimization and the procedures in `R` and `SAS` need starting values for their algorithms. In this case it is easy

to find good initial values, due to the interpretation of α and β. From Figure 3.1 200 and 0.1 seem to be reasonable estimates of the plateau value and the concentration with half the maximal reaction time. See the computer sections for details on how these values are used in `R` and `SAS`.

The least squares estimates turn out to be

$$\hat{\alpha} = 212.6837\ (6.9472); \qquad \hat{\beta} = 0.0641\ (0.0083)$$

and the variance estimate is $s^2 = 119.5449$. The 95% confidence intervals are $(197.3, 229.3)$ for α and $(0.0469, 0.0861)$ for β.

Model validation for non-linear regression models is carried out through investigation of residuals as for linear models, except that standardized residuals are no so easily accessible. Hence, we use the raw residuals (and remember that we cannot use the actual values to detect outliers since they are not standardized to have unit variance).

The residuals are plotted against the predicted values in the right part of Figure 3.3. The plot is not very informative because there are so few observations. More useful is perhaps the original plot of the reaction time against concentration with the estimated curve superimposed. This figure is shown in the left part of Figure 3.3. We see that the model describes the data reasonably well although it seems to sligtly underestimate reaction time for large concentrations. It does not make much sense to make a QQ-plot when there are only 12 observations.

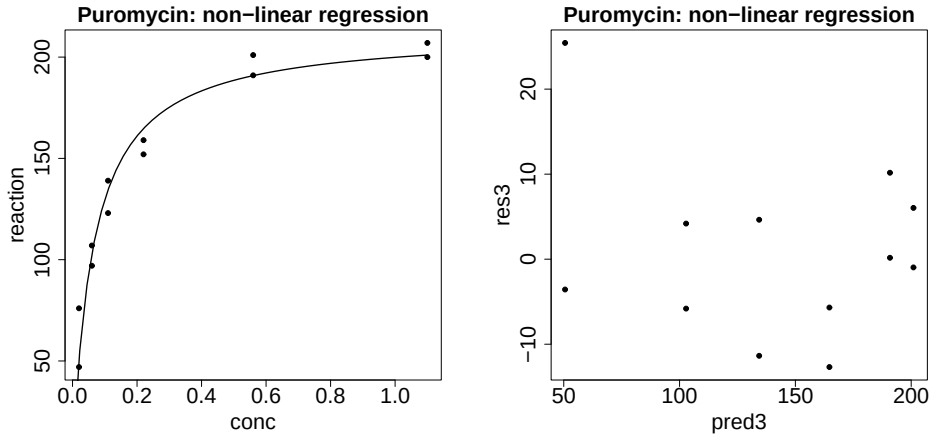


Figure 3.3: Non-linear regression for the puromycin data. Observed and predicted vales (left): residual plot (right).

For this particular dataset we can actually test the non-linear regression model against the oneway ANOVA defined by the six concentration groups:

$$y_i = \gamma(\texttt{conc}_i) + e_i$$

This model allows the six expected reaction times to take any six values whereas model (3.2) restricts the six expected values to satisfy the relation from the model. the ANOVA model with the F-test given by (2.2) with the oneway ANOVA as model A and the non-linear regression model as model B. We get

$$F = \frac{(1195.4 - 697.5)/(10-6)}{697.5/6} = 1.07.$$

This value should be compared to the the $F(4,6)$-distribution, giving a p-value of 0.45. Hence, the non-linear regression model is indeed a valid model. Note that such a test (and neither the Bartlett's

tests for equal variances) would not have been possible had we had different concentrations for each measurement.

Finally, let us illustrate what would have happened had Figure 3.1 inspired us to consider a quadratic relationship between concentration and reaction time. The corresponding model given by

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$$

is illustrated in Figure 3.4. The left part shows the data points with the estimated parabola superimposed and the right part shows the residual plot. Both clearly reveal that the mean structure is misspecified. □

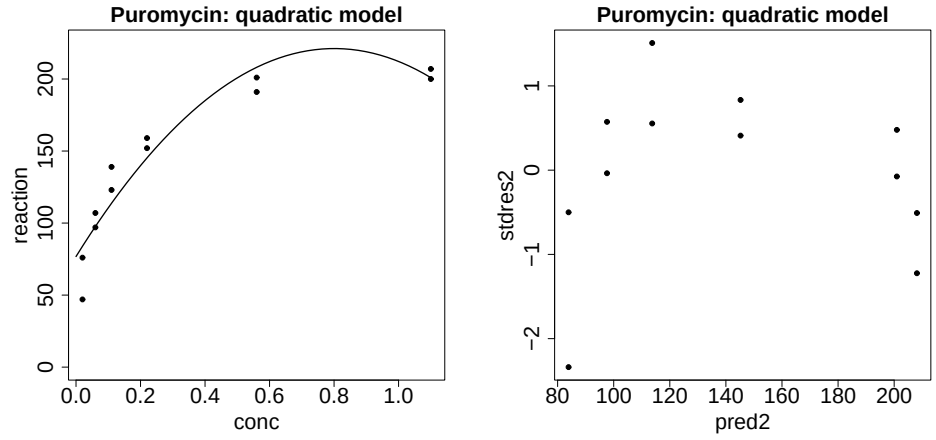


Figure 3.4: Quadratic regression for the puromycin data. Observed and predicted vales (left): residual plot (right).

3.2.1 R-programs and output

Example 3.1 *(continued)*

Reading the data into R *and construction of scatter plot*

Assume that the dataset is saved in the ascii-file `puromycin.txt` as follows:

```
conc reaction
0.02   76
0.02   47
0.06   97
  .     .
  .     .     [more datalines here]
  .     .
1.10  200
```

The dataset is read into R and attached, so we can use the variable names `conc` and `reaction`. Moreover the plot of reaction time against concentration (Figure 3.1) is constructed:

```
> puromycin = read.table("puromycin.txt",header=T)
> attach(puromycin)
> plot(conc,reaction)
```

The analysis of the reciproval data

The variables with the reciproval values are constructed and named `reconc` and `rereac`. The reciprocal variables are plotted against each other with `plot`, the simple linear regression of `rereac` on `reconc` is fitted with `lm`, and the residual plot for regression model is constructed as for Example 1.1 (see the figures in Figure 3.2). Moreover Bartlett's test is carried out with `bartlett.test`.

```
> reconc = 1/conc
> rereac = 1/reaction
> plot(reconc,rereac)

> model1 = lm(rereac ~ reconc)

> library(MASS)
> pred1 = predict(model1)
> stdres1 = stdres(model1)
> plot(pred1,stdres1)

> bartlett.test(rereac, factor(reconc))
        Bartlett test of homogeneity of variances
data:  rereac and factor(reconc)
Bartlett's K-squared = 13.8957, df = 5, p-value = 0.01629
```

Analysis with the non-linear model

First, Bartlett's test for equal variances:

```
> bartlett.test(reaction, factor(conc))
        Bartlett test of homogeneity of variances
data:  reaction and factor(conc)
Bartlett's K-squared = 2.4621, df = 5, p-value = 0.7822
```

The non-linear regression model is fitted with `nls`. `nls` requires a model formula, that is, the non-linear function to be fitted. It is specified similarly to way models are specified in `glm`, with the response on the left hand side of a "tilde" and the model expression on the right hand side. In this case we write `reaction ~ alpha*conc/(beta+conc)`. `R` also requires starting values for the algorithm that minimizes the least squares function. The starting values are specified via a `list` with a value for each parameter, in this case with values of `alpha` and `beta`. We use 200 and 0.1 which are specified as `start = list(alpha=200, beta=0.1)`.

In summary, the non-linear model is fitted as below, the model fit is summarized by `summary`, and confidence intervals are computed with `confint`.

```
> model3 = nls(reaction ~ alpha*conc / (beta+conc), start = list(alpha=200, beta=0.1))
> summary(model3)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
```

```
a 2.127e+02  6.947e+00  30.615 3.24e-11 ***
b 6.412e-02  8.281e-03   7.743 1.57e-05 ***

Residual standard error: 10.93 on 10 degrees of freedom

Correlation of Parameter Estimates:
       a
b 0.7651

> confint(model3)
Waiting for profiling to be done...
          2.5%        97.5%
a 197.30205011 229.29022954
b   0.04692625   0.08616203
```

As usual, `summary` gives the parameter estimates, their standard errors and test for the hypotheses $H_0 : \alpha = 0$ and $H_0 : \beta = 0$. We are not at all interested in these hypotheses in this example. Moreover the estimated correlation matrix of the parameter estimates, in this case just a single number, is given.

The predicted values and the raw residuals are computed and plotted as usual (right part of Figure 3.3). `R` cannot compute standardized residuals for non-linear models, so recall that the actual values cannot really be used for detecting outliers.

```
> pred3 = predict(model3)
> res3 = residuals(model3)
> plot(pred3,res3)
```

Test of non-linear regression against oneway ANOVA

The sum of squares from the non-linear regression is found from `model3`. Furthermore, the oneway ANOVA is fitted, and the F-test is computed:

```
> model3
Nonlinear regression model
  model:  reaction ~ a * conc/(b + conc)
   data:  parent.frame()
         a           b
212.6836297   0.0641211
residual sum-of-squares:  1195.449

> model4 = lm(reaction ~ factor(conc))
> anova(model4)
Response: reaction
             Df  Sum Sq Mean Sq F value    Pr(>F)
factor(conc)  5 30161.4  6032.3  51.891 7.386e-05 ***
Residuals     6   697.5   116.2

> (1195.4 - 697.5) / 4 / 116.2
[1] 1.071213
> 1- pf(1.07,4,6)
[1] 0.4471664
```

Plot with data and expected values

Now consider the left part of Figure 3.3 where the data points are plotted with the estimated function superimposed. First the data points are plotted with `plot` as usual. Next, the estimated function is superimposed with the `points`-command. `points` works similar to `plot`, except that it superimposes the points/curve onto the present plot, instead of making a new plot. In order to make a smooth curve we construct a vector `x` of length 50 with equidistant values from zero to 1.1. For each of the 50 `x`-values we compute the value of the estimated funtion

$$\frac{\hat{\alpha} \cdot x}{\hat{\beta} + x} = \frac{212.7 \cdot x}{0.06412 + x}$$

and store the results in the vector `y3`. Finally `y3` is plotted against `x` with `points`. The option `type="l"` has the effect that the 50 points are joined (by linear interpolation), so a curve rather than 50 points are plotted.

```
> plot(conc,reaction)

> x = seq(0,1.1,length=50)
> x
 [1] 0.00000000 0.02244898 0.04489796 0.06734694 0.08979592 0.11224490
 [7] 0.13469388 0.15714286 0.17959184 0.20204082 0.22448980 0.24693878
  .
  .
  .
[49] 1.07755102 1.10000000

> y3 = 212.7 * x / (0.06412 + x)
> points(x,y3,type="l")
```

Analysis with the quadratic model

We only show how to fit the model. The residual plot is made as usual (this is multiple linear regression model so we use the standardized residuals), and the plot with observed and expected points is constructed as above. Recall that this model was just for illustation; the mean structure is clearly misspecified.

```
> conc2 = conc*conc
> model2 = lm(reaction ~ conc + conc2)
```

3.2.2 SAS-programs and output

Example 3.1 *(continued)*

Reading the data into `SAS` *and construction of scatter plot*

The data is read into `SAS`, and the scatter plot of reaction time against concentration is created.

```
data puromycin;
  input conc @@;
  do rep=1,2;
```

```
    input reaction @@; output;
  end;
  cards;
  0.02  76  47
  0.06  97 107
  0.11 123 139
  0.22 159 152
  0.56 191 201
  1.10 207 200
  ;
run;

proc gplot data=puromycin;
  plot reaction*conc=1;
run;
```

Analysis of the reciprocal data

The reciprocal variables are defined as `rereac` and `reconc`. The simple linear regression of the reciprocal reaction time on the reciprocal concentration is fitted, and the corresponding residual plot is made. Finally, we test for identical variances with Bartlett's test. For this, the concentration should be thought of as a qualitative variable so it is included in a `class`-statement.

```
data puromycin2;
  set puromycin;
  rereac=1/reaction;
  reconc=1/conc;
run;

proc gplot data=puromycin2;
  plot rereac*reconc=1;
run;

proc glm data = puromycin2;
  model rereac = reconc;
  output out = reout predicted = p student = stdres;
run;

proc gplot data = reout;
  plot stdres * p;
run;

proc glm data=puromycin2;
  class reconc;
  model rereac=reconc / ss3;
  means reconc / hovtest=bartlett;
run;
```

The output from the final call to `proc glm` includes this (and all the usual stuff):

```
                          The GLM Procedure

        Bartlett's Test for Homogeneity of rereac Variance
```

```
Source        DF    Chi-Square    Pr > ChiSq
reconc         5       13.8957        0.0163
```

Analysis with the non-linear model

Before fitting the non-linear model we check for identical variances on the original scale:

```
proc glm data=puromycin;
  class conc;
  model reaction=conc / ss3;
  means conc / hovtest=bartlett;
run;
```

with output like this (and a lot other stuff):

```
                        The GLM Procedure

                                          Sum of
Source                DF         Squares     Mean Square    F Value    Pr > F
Model                  5     30161.41667      6032.28333      51.89    <.0001
Error                  6       697.50000       116.25000
Corrected Total       11     30858.91667

      Bartlett's Test for Homogeneity of reaction Variance

          Source        DF    Chi-Square    Pr > ChiSq
          Conc           5        2.4621        0.7822
```

Then, to the non-linear analysis with `proc nlin`. `proc nlin` of course need to know the particular non-linear model we have in mind. This is done in the `model`-statement where the expression for the response as a function of the explanatory variable(s) and the parameters is specified. Furthermore some initial values must be specified.

```
proc nlin data=puromycin;
  parms a=200 b=0.1;
  model reaction=a*conc/(b+conc);
  output out = nlinout predicted = p residual = res;
run;

proc gplot data=nlinout;
  plot res * p;
run;
```

The output looks as follows:

```
                        The NLIN Procedure
                    Dependent Variable reaction
                       Method: Gauss-Newton

                          Iterative Phase
                                                 Sum of
                  Iter           a           b     Squares
```

```
                    0       200.0      0.1000      7964.2
                    1       212.0      0.0543      1593.2
                    2       211.8      0.0623      1201.0
                    3       212.6      0.0639      1195.5
                    4       212.7      0.0641      1195.4
                    5       212.7      0.0641      1195.4
                    6       212.7      0.0641      1195.4

     NOTE: Convergence criterion met.

          NOTE: An intercept was not specified for this model.

                              The NLIN Procedure

                                       Sum of        Mean                Approx
  Source                     DF       Squares      Square    F Value     Pr > F
  Model                       2        270214      135107    1130.18     <.0001
  Error                      10        1195.4       119.5
  Uncorrected Total          12        271409

                                     Approx
        Parameter      Estimate     Std Error    Approximate 95% Confidence Limits
        a                 212.7        6.9471       197.2       228.2
        b                0.0641       0.00828      0.0457      0.0826

                     Approximate Correlation Matrix
                                       a               b
                    a       1.0000000       0.7650835
                    b       0.7650835       1.0000000
```

First, SAS summarizes the iterations it has been through in order to get a fit of the model. The message "Convergence criteria met" is very important. If SAS gives another message there has been problems in the optimization and SAS is not so sure that the results are correct. We cannot trust the results from such an analysis, but could try with other (better) starting values.

Further down in the output we find the variance estimate, $s^2 = 119.5$ and the parameter estimates from the non-linear structure. Also reported is the correlation matrix of the parameter estimates.

Note that all the ingredients for the F-test can be found in the above output from `proc glm` and `proc nlin`.

Instead of specifying the starting values, one may let SAS look for some numerically, but we have to tell SAS roughly where to look. Here is program where SAS looks for such starting values in a grid:

```
proc nlin data=puromycin method=dud;
  parms a=200 to 220 by 5 b=0.05 to 0.8 by 0.1;
  model reaction=(a*conc)/(b+conc);
  output out=par parms=a b;
run;
```

SAS finds the same parameter estimates as before (luckily enough), so we do not show the output.

In the above calls to `proc nlin` SAS makes numerical computations of the derivatives during the optimization. It is also possible to specify the derivatives directly, which makes the work easier for SAS, but it is usually not worthwhile with today's computers.

Plot of observations and predicted valued

Let us show how to make the plot of data points together with the graph for the fitted non-linear function (similar to Figure 3.3).

First pairs of fictious concentration values and the corresponding estimated reaction time values are constructed and saved in the dataset `fit`. In total 300 pairs are computed. Then the two datasets, `puromycin` and `fit` are merged because we want to plot variables from both in the same plot. Then the data points and the (`c`,`v`)-values are plotted in the same plot. This is obtained by `overlay`. Note that the data points are points whereas the (`x`,`y`)-values are joined to a curve; this is obtained with `symbol1` and `symbol2`.

```
data fit;
  do x=0 to 300;
    c = 0.02 + x*(1.10 - 0.02)/300;
    v = 212.684*c/(0.064121+c);
    output;
  end;
run;

data fit; merge puromycin fit ; run;

symbol1 i=none v=dot c=black;
symbol2 i=join v=none c=black l=1;

proc gplot data=fit;
  plot reaction*conc=1 v*c=2 / overlay;
run;
```

Analysis with the quadratic model

Finally, the quadratic model. Here we just show how the model is fitted. The residual plot is made as usual, and the plot with observed points and the fitted parabola is constructed as the similar plot for the non-linear model. Recall that this model was just for illustation; the mean structure is clearly misspecified.

```
data puromycin;
  set puromycin;
  concsq = conc*conc;
run;

proc glm data = puromycin;
  model reaction = conc concsq;
  output out=quadout predicted=p student=stdres;
run;
```

Chapter 4

Gaussian models with random effects

So far, the Gaussian (normal) models we have considered have included only fixed (or systematic) effects, apart from the residual. That is, all explanatory variables have been used for description of the mean structure ($\mathrm{E}\, y_i$, $i = 1, \ldots, n$). In this chapter we will discuss models with *both fixed and em random effects*, that is, models with multiple sources of variation. Some of the explanatory variables are used to describe the random variation. Gaussian models with random effects and a fixed part that is linear are called *mixed linear models.* We first give some general comments and recommendataions and then move on with the analysis of two datasets.

4.1 Some general considerations

4.1.1 Fixed effects and random effects

When we set up a mixed linear model we have to decide which explanatory variables to include in the model. Moreover, we must decide if the factors (qualititative explanatory variables) should be fixed or random.

The *fixed factors* are the factors which we believe describe the response variable in a systematic way. Typical examples are treatment, time, variety, dose, sex, breed (although factors like variety and breed may also occur as random effects depending on the design and the purpose of the experiment). Usually the experiment has been carried out in order to investigate the effects of some of these factors, whereas others are included in the analysis in order to adjust for their effects (this is often the case for sex, for example).

The fixed, or *systematic*, effects describe the mean structure of the model. That is, they are used to describe $\mathrm{E}\, y_i$, $i = 1, \ldots, n$, just as in the linear models with no random effects. Note that also covariates (such as time, dose or baseline measurements) may be used in the fixed part of the model.

A factor is used as a *random factor* if its levels can be thought of as being randomly selected from a population of possible levels. Typical examples of random factors are person, animal, block, litter, herd, and field. We are not interested in properties for the specific persons/animals/fields in the experiment but rather in the population properties.

The random effects describe the variance structure of the observations. Each random effect contributes to the variance with a term, so the variance of each y_i is a sum $\sigma^2 + \sigma_1^2 + \ldots + \sigma_m^2$ where σ^2 is the residual variance and σ_j^2 is the variance due to the j'th random factor. Note that all observations have the same

variance. In the linear normal models (with no random effects) all observations are assumed to be independent. This is not true in models with random effects. Observations on different levels of (all) random factor(s) are independent but *observations on same the level of a random factor are correlated*. Loosely speaking, observations with the same level of a random factor are "more alike" than observations with different levels.

Example 4.1 (A split-plot experiment) As an example, consider a field experiment with eight plots, called whole plots. Four of them (randomly selected among the eight) are treated with some fertilizer, four of them are not. Moreover, each plot is divided into two subplots; one of them is sown with one variety, the other subplot with another. At the end of the season the yield of each subplot is measured. This is a so-called split-plot experiment. The experimenter is interested in the effect of fertilizer and variety on yield. In this case, fertilizer and variety (and their interaction) are obviously fixed effects. There is no interest in the particular eight whole plots, so whole plot should be random. In other words: the yield on two subplots from the same whole plot are assumed to be correlated ("more alike"), which seems quite reasonable. □

In some cases the decision to use a certain factor as random rather than systematic is absolutely essential for the analysis, and only the random effects analysis is correct. This is true for the example above: since both subplots on a whole plot are either fertilized or not, the effect of fertilizer should be compared to the variation between whole plots. In other cases the tests are the same no matter whether a factor is used as systematic or random, and the difference is mainly a question of interpretation of the results: do the conclusions apply to the particular factor levels from the experiment or do they apply to a whole population of possible factor levels? The beech wood example below is an example of this situation.

4.1.2 Factor diagrams and mixed linear models

In order to decide on a statistical model we have to understand the structure of the data. For complex experimental designs with many factors a *factor diagram* is often a good help throughout the analysis. For mixed linear models in "nice" (balanced in a certain sense) designs the factor diagram moreover tells how to perform the tests for model reduction and tell how to compute degrees of freedom for each factor.

First, write up the relevant factors from the experiment, including relevant interactions. Second, identify the ordering of the factors; which factors are coarser/finer than others? Recall that a factor $\texttt{fac}_1$ is coarser than another $\texttt{fac}_2$ if knowing the level of $\texttt{fac}_2$ implies that you also know the level of $\texttt{fac}_1$. Third, make a diagram with all the factors and put an arrow from $\texttt{fac}_2$ to $\texttt{fac}_1$ if $\texttt{fac}_1$ is coarser than $\texttt{fac}_2$. Put the finer factors to the left and the coarser factors to the right. Draw also the identity factor, `I` (to the very left), and the trivial factor, `0` (to the very right) in the diagram. All factors are coarser than `I` and finer than `0`. Fourth, decide whether the factors should be fixed or random, and put brackets, [ ], around the random factors, also around `I`.

A factor diagram corresponds to a mixed linear model, namely the model with the factors without brackets as fixed effects and the factors in brackets as random effects. Note that covariates do not belong in a factor diagram, but that mixed linear models may very well include covariates.

In a model with random effects some of the F-tests from the usual ANOVA should be modified — in nice cases by using the mean square error (MSE) for another random factor than `I` in the denominator. The factor diagram shows which random factor the fixed effects should be tested against. A fixed effect should be tested against *the coarsest random factor that is finer than the systematic factor in question*. In the nice cases, the degreess of freedom may also be computed from the factor diagram. We will not go into details about that, however, since `SAS` or `R` will compute the degrees of freedom automatically.

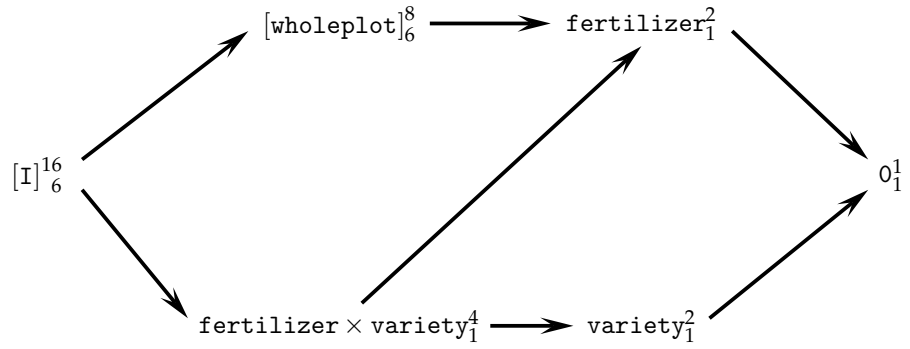


Figure 4.1: Factor diagram for the split-plot experiment.

Even in unbalanced cases where the above rules do not apply it is often useful to draw the factor diagram in order to get an overview of the experimental design. Let us consider three nice cases:

Example 4.1 *(continued)* The factor diagram for the split-plot experiment is drawn in Figure 4.1. The degrees of freedom are added. We see that the interaction should be tested against the residual. If the interaction is not significant, then the effect of fertilizer should be tested against whole plot, and the effect of variety against the residual. □

Example 4.2 (Humidity of beech wood) In order to investigate the effect on humidity of drying of beech wood the following experiment was carried out. Each of 20 planks was dried in a certain period of time. Then the humidity percentage was measured in five depths and three widths for each plank. The points of measurements are described in Figure 4.2. There are 15 meaurements for each plank and 300 observations in total, summarized in Table 4.1. The interest is in the variation of humidity across beech wood planks in general whereas one is not interested in the specific planks from the experiment.

There are three factors in this experiment: `planks` with 20 levels corresponding to the 20 planks. `width` with three levels and `depth` with five levels. It is natural to include the interaction between width and depth, `width` × `depth`. The interaction factor has 15 levels, corresponding to the grid in Figure 4.2. The factors `width`, `depth` and their interaction are obviously fixed. It is natual to consider plank as random as the interest lies in variation of humidity across planks in general. The 20 planks are randomly chosen from a large population of planks and the results from the analysis with plank as random will apply to the populatation of planks.

In the following let y_i be the humidity percentage for the i'th measurement, while $\texttt{plank}_i$ is the plank number, $\texttt{width}_i$ is the width and $\texttt{depth}_i$ is the depth. Then the above model is given by

$$y_i = \mu + \alpha(\texttt{width}_i) + \beta(\texttt{depth}_i) + \gamma(\texttt{width}_i, \texttt{depth}_i) + d(\texttt{plank}_i) + \epsilon_i,\ i = 1, \ldots, 300, \qquad (4.1)$$

where $d(j) \sim N(0, \sigma^2_{\texttt{plank}})$, $\epsilon_i \sim N(0, \sigma^2)$ and all $d(j)$'s and ϵ_i's are independent. Figure 4.3 shows the corresponding factor diagram. We see that all fixed effects should be tested against the residual. This is not surprising since all "treatments" (combinations of widths and depths) are present in all "blocks" (planks).

Note the variance structure in the above model: the variance of all y_i's is $\sigma^2_{\texttt{plank}} + \sigma^2$; observations from different planks are independent whereas two observations from the same plank are correlated with correlations $\sigma^2_{\texttt{plank}} / (\sigma^2_{\texttt{plank}} + \sigma^2)$. □

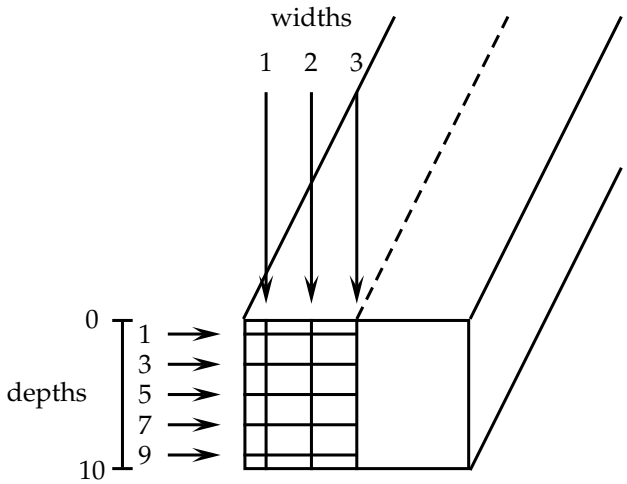


Figure 4.2: The beech wood experiment.

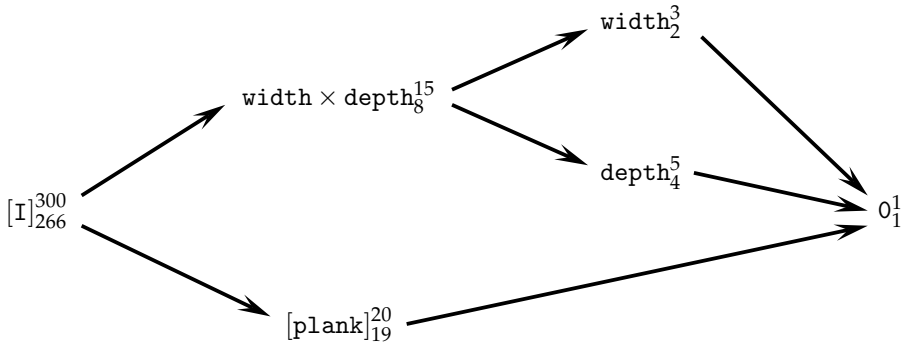


Figure 4.3: Factor diagram for the beech wood data.

Example 4.3 (Redness of pork meat) In order to compare the colour of meat from two breeds of pigs and for two light levels while storing, the following experiment was carried out: 20 pigs, 10 from an old breed not used commercially anymore and 10 from another breed, were slaughtered. Two pieces of meet from each pig were stored, one piece in dark and one piece in light. After six days, the "redness" of the meat was measured. The data is listed in Table 4.2. Large values of colour correspond to very red meat wheras low values correspond to less red meat.

There are three factors, `pigno` with 20 levels, `storage` and `breed` with two levels. It is natural to include the interaction factor `storage` $\times$ `breed` as well. Note that `breed` is coarser than `pigno`. We will use `storage`, and `breed` and `storage` $\times$ `breed` as fixed and `pigno` as random.

For the i'th observation, let y_i be the colour measurement and $\texttt{breed}_i$, $\texttt{storage}_i$ and $\texttt{pigno}_i$ be the levels of the explanatory factors. Then the model is the following:

$$y_i = \mu + \alpha(\texttt{storage}_i) + \beta(\texttt{breed}_i) + \gamma(\texttt{storage}_i, \texttt{breed}_i) + d(\texttt{pigno}_i) + \epsilon_i,\ i = 1, \ldots, 300, \qquad (4.2)$$

where $\epsilon_i \sim N(0, \sigma^2)$, $d(j) \sim N(0, \sigma^2_{\texttt{pigno}})$ and all ϵ_i's and $d(j)$'s are independent.

Width	1					2					3				
Plank / Depth	1	3	5	7	9	1	3	5	7	9	1	3	5	7	9
1	3.4	4.9	5.0	4.9	4.0	4.1	4.7	5.2	4.6	4.3	4.4	4.8	5.0	4.9	4.2
2	4.3	5.5	6.2	5.4	4.7	3.9	5.6	5.7	5.5	4.9	4.0	4.7	4.5	3.9	4.0
3	4.2	5.5	5.6	6.3	4.5	5.4	6.2	6.1	6.4	5.2	4.5	4.9	4.9	4.9	4.4
4	4.4	6.0	7.1	6.9	4.6	4.6	6.1	6.6	6.5	4.7	4.9	5.9	5.8	6.4	4.7
5	3.9	4.7	5.2	5.0	3.7	4.2	5.2	5.4	4.8	3.9	4.0	4.4	4.4	4.1	3.5
6	4.6	5.9	6.3	5.8	4.8	5.9	7.3	6.9	6.9	4.4	5.2	5.7	6.6	6.0	4.0
7	3.9	5.6	6.0	5.3	5.0	4.9	6.9	7.1	6.1	4.5	4.3	5.4	5.9	5.5	4.2
8	3.9	4.5	5.3	5.6	4.7	3.7	4.9	4.8	4.9	4.3	3.8	4.5	5.4	4.8	4.0
9	3.6	4.1	4.0	4.4	3.7	3.8	5.1	5.0	4.6	3.3	3.0	3.9	4.7	4.9	3.8
10	6.5	8.7	9.5	7.9	6.6	6.9	8.9	7.4	7.0	6.9	5.8	7.5	7.7	7.3	5.9
11	3.7	5.2	5.5	5.9	4.4	4.7	5.8	5.7	4.9	4.2	3.7	5.0	6.3	5.2	4.3
12	4.3	5.8	6.2	5.2	4.4	4.8	6.7	7.0	6.1	5.2	5.1	5.7	5.9	6.4	5.1
13	6.5	8.8	9.1	8.9	6.0	5.9	7.5	8.4	7.9	5.7	4.0	4.2	4.9	4.6	3.5
14	4.4	6.2	6.7	6.4	4.3	5.7	7.0	7.4	7.3	5.5	4.6	6.2	6.8	5.8	4.9
15	5.5	7.1	7.5	6.9	5.4	6.4	8.4	8.9	8.1	6.1	6.5	8.4	9.1	9.2	7.5
16	5.2	6.0	6.2	6.6	5.3	6.6	7.6	7.8	7.7	5.8	5.9	6.7	6.7	5.0	3.9
17	3.7	4.5	5.0	4.5	3.7	3.7	4.4	4.8	4.4	4.3	3.7	4.5	4.7	5.3	3.9
18	6.0	7.4	7.8	7.5	5.7	6.9	8.6	8.8	7.5	5.4	5.1	6.1	5.2	5.4	4.7
19	3.8	4.6	4.8	4.4	3.8	3.7	4.7	4.7	4.3	3.7	3.3	3.5	3.7	3.4	3.2
20	6.1	7.4	7.7	6.7	4.6	4.7	6.3	7.1	6.5	5.1	4.7	6.0	6.0	6.3	4.2

Table 4.1: The beech wood data.

pigno	storage	breed	colour	pigno	storage	breed	colour
13	light	old	5.5630	251	light	new	3.3360
13	dark	old	9.5280	251	dark	new	6.0190
41	light	old	5.5730	252	light	new	4.4060
41	dark	old	4.9450	252	dark	new	7.2740
55	light	old	3.4870	256	light	new	3.4580
55	dark	old	7.5490	256	dark	new	4.0540
66	light	old	4.0410	264	light	new	4.0320
66	dark	old	6.4500	264	dark	new	6.4510
74	light	old	4.3130	277	light	new	2.3770
74	dark	old	6.8880	277	dark	new	6.3430
84	light	old	5.2320	280	light	new	2.0220
84	dark	old	5.0420	280	dark	new	4.8940
129	light	old	5.1560	283	light	new	5.3420
129	dark	old	7.1570	283	dark	new	2.0790
138	light	old	2.9290	284	light	new	2.6660
138	dark	old	7.3460	284	dark	new	1.6060
181	light	old	4.1400	285	light	new	4.9530
181	dark	old	7.5740	285	dark	new	7.4610
190	light	old	5.3910	286	light	new	5.4800
190	dark	old	4.9710	286	dark	new	6.3960

Table 4.2: The pork meat data.

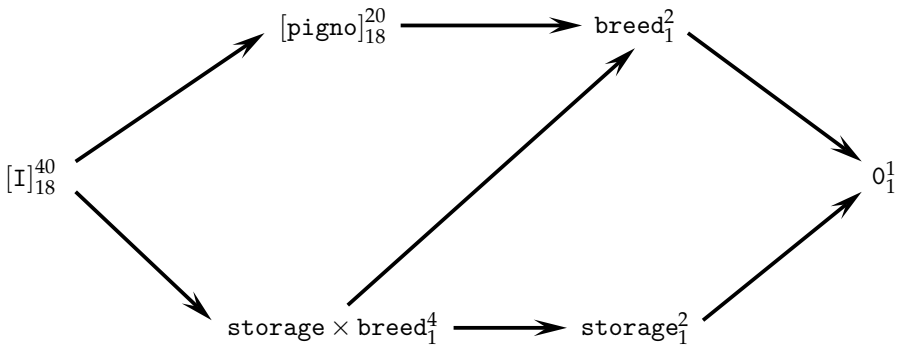


Figure 4.4: Factor diagram for the pork meat data.

The corresponding factor diagram is drawn in Figure 4.4. We see that the effect of `breed` should be tested against `pigno`, which is correct. Had we used `pigno` as a fixed effect we could not have tested for a `breed` effect if there was a significant difference between pigs (a significant effect of `pigno`).

Note that we can think of model (4.7) as a split-plot model (see Example 4.1) with `pigno` as the whole plot factor. There is a difference, though, since the levels of `breed` are not randomly allocated to the different pigs, but this does not matter for the analysis. Rather we like to think of the model as a model with *random intercepts*: d equal to zero corresponds to an average pig. This population average is displaced randomly for each pig, with $d(\texttt{pigno}_i)$ for pig number i. This random displacement follows the pig at all observations: the meat of the pig may be more red than average no matter the storage ($d(\texttt{pigno}_i) > 0$), or less red than average no matter the storage ($d(\texttt{pigno}_i) > 0$).

4.1.3 Model reduction: approximate methods and the likelihood ratio test

As already mentioned, hypothesis tests can be carried out by F-tests if the design is balanced in a certain sense. In these cases one may use the mean square errors (MSE's) from the corresponding linear model, that is, the model where the random factors are fixed instead. The factor diagram shows which MSE should appear in the denominator as well as the degrees of freedom for the F-test.

Many experiments are not balanced, though, due to the design itself or due to missing values. The factor diagram may still help to get an overview of the design, but approximate methods are needed for hypothesis testing and confidence intervals. Both `SAS` and `R` can make such approximations, so we can analyze data from unbalanced designs, too.

There is a lot of debate among experts about what approximations to use. In particular `SAS` and `R` do not support quite the same approximations. `proc mixed` in `SAS` makes approximate F-tests, where the degrees of freedom are computed by various methods which the user may choose. We generally recommend to use Satterthwaite's approximation, see the details in Section 4.2.2. Not all statisticians agree that these tests are valid, and they are not implemented in `R`. In `R` we will instead carry out likelihood ratio (LR) tests.

The rationale for the LR test is the following: For a given model, the maximum of the likelihood function measures (in a certain sense) how well the model fits the data. Hence, if we compare the maximum of the likelihood function under a given model with the maximum of the likelihood function under a null

model (assuming a hypothesis to be true), we have a measure of the discrepancy of the models. In other words we measure how much worse the null model fits the data compared to the full model.

To be precise, we use

$$LR = 2 \cdot \log L(\text{full model}) - 2 \cdot \log L(\text{null model}).$$

This statistic is approximately χ^2-distributed; the degrees of freedom is equal to the decrement in model dimensions (number of model parameters) from the full model to the null model. However, the experience is that these approximate p-values tend to be too small, thereby sometimes overestimating the importance of certain effects. Experts therefore recommend to compute a better approximation to the p-value by so-called parametric bootstrap if the approximate p-value is below the significance level, but not very small. This is quite easy to do in `R`, see two examples in Section 4.2.1 and 4.3.1. Likelihood ratio tests are also easily carried out in `SAS`, but bootstrap p-values are not so easily carried out.

The examples in this chapter are all of the nice kind, where we could use exact F-tests. We will use the general methods, anyway, in order to show what to do in the general case.

Finally, some comments on estimation method. In order to make likelihood ratio tests (for fixed effects) it is essestial that the models are fitted with ML (maximum likelihood). However, it is well-known that REML (Restricted Maximum Likehood) estimation generally produced better estimates. Hence, we recommend to always fit the final model with REML rather than ML; also if model reduction has been carried out by likelihood ratio test using ML. Note that both `R` and `SAS` use REML as default.

4.1.4 Model validation

It is not so obvious how to carry out model validation in mixed linear models. We recommend to do it in the corresponding linear model, that is, the model where the random factors are moved to the fixed part of the model. In other words, the procedure is as follows: first fit the linear model (with `lm` or `proc glm`) which as fixed effects include the fixed effects as well as the random effects from the mixed linear model we wish to validate. Make a residual plot and possibly a QQ-plot for this model in the usual way. Note that we are not interested in the linear model in itself, but only consider it for model validation purposes.

4.2 Analysis of the beech wood data

Example 4.2 *(continued)* Let us continue with the beech wood data. In order to get an idea about the dependence of width and depth on humidity, we make some *profile plots*, see Figure 4.5. In the top left plot the depth-humidity pattern is illustrated by plotting the average humidity (over width) for each plank against width. In the top right plot, the pattern for width is illustrated similarly. We see an extensive plank-to-plank variation. Moreover there is a clear relation between humidity and depth: the humidity is high in the center of the planks and low at the sides. The message about width is less clear. We have also plotted the 15 averages of humidity (over de 20 planks) against width and depth (the bottom plots). Graphically, there are no indications of an interaction between width and depth. We test the hypothesis formally below.

Recall model (4.1) with `width`, `depth` and the interaction `width` $\times$ `depth` as systematic factors and plank as random:

$$y_i = \mu + \alpha(\texttt{width}_i) + \beta(\texttt{depth}_i) + \gamma(\texttt{width}_i, \texttt{depth}_i) + d(\texttt{plank}_i) + \epsilon_i,\ i = 1, \ldots, 300, \qquad (4.3)$$

We first validate the model by investigating the standardized residuals from the corresponding linear model, that is, the model where plank is fixed rather than random. The standardized residuals are

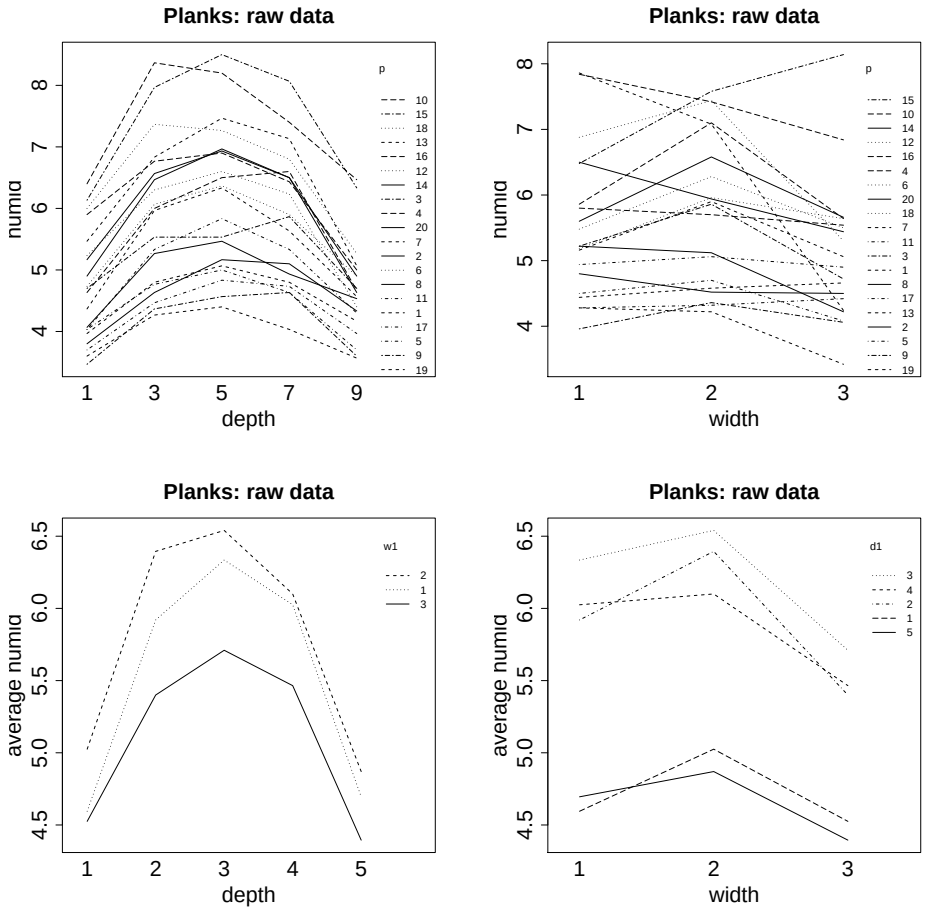


Figure 4.5: Humidity of beech wood: profile plots.

plotted against the predicted values in Figure 4.7. We see a clear pattern: variances tend to be larger for large predicted values than for small predicted values, so we do not believe in variance homogeneity.

Let us try if a logarithm transformation can fix the problem. We use $y = \log(\texttt{humid})$ as response variable instead of `humid`. Figure 4.7 shows the residual plot to the left and the QQ-plot for the standardized residuals. Neither makes us nervous about the appopriateness of the model.

Now to the actual analysis of model (4.3). First we fit the model and test if the interaction between width and depth is signficant, that is, if the model can be reduced to

$$y_i = \mu + \alpha(\texttt{width}_i) + \beta(\texttt{depth}_i) + d(\texttt{plank}_i) + \epsilon_i,\ i = 1, \ldots, 300, \tag{4.4}$$

It turns out that the interaction is not significant ($p \approx 0.45$, depending on the test method). This is completely in line with the bottom figures of Figure 4.5. Hence, we accept model (4.4). Both main effects are highly significant ($p < 0.001$), so there is a variation in humidity across the planks in both directions.

Take a look at the left plots of Figure 4.5 again. They indicate that the humidity is symmetric around the center (`depth` $= 5$). Let us make a formal test for this hypothesis. Let `dist` be a factor that partions the observations into three groups: observations in the center (`depth` $= 5$), observations from the side

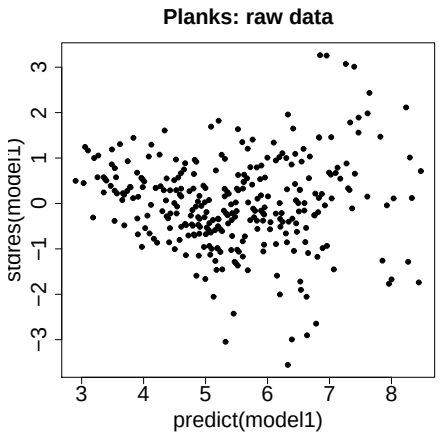


Figure 4.6: Humidity of beech wood: residual plot for the raw data.

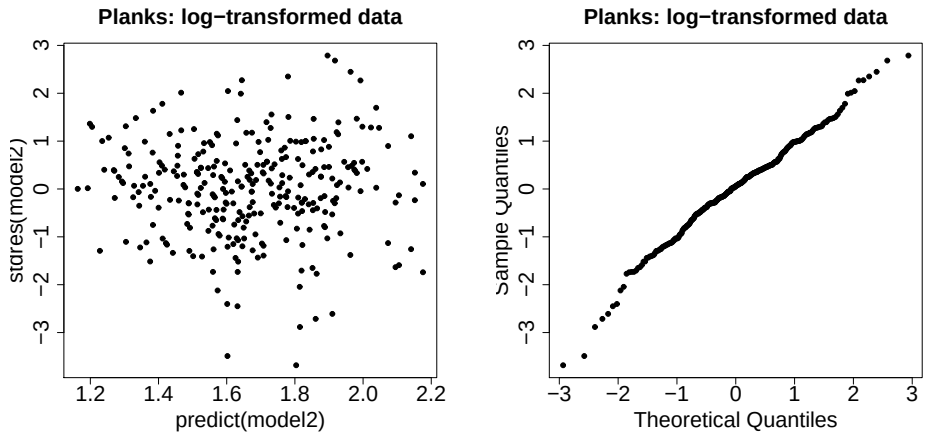


Figure 4.7: Humidity of beech wood: residual plot and QQ-plot for the log-transformed data.

(depth $= 1$ and depth $= 9$) and observations in between (depth $= 3$ and depth $= 7$). We can define it as

$$\texttt{dist}_i = |\texttt{depth}_i - 5|$$

with values 0, 2 and 4. The hypothesis of symmetry then corresponds to the model reduction from (4.5) to

$$y_i = \mu + \alpha(\texttt{width}_i) + \phi(\texttt{dist}_i) + d(\texttt{plank}_i) + \epsilon_i,\ i = 1, \ldots, 300, \tag{4.5}$$

The hypothesis is accepted with a p-value of 0.89. Hence, there is symmetry around the center.

The variance estimates (REML) in the final model are

$$\hat{\sigma}^2_{\texttt{plank}} = 0.0313; \quad \hat{\sigma}^2 = 0.011$$

We finish with the adjusted means, also called least squares (LS) means, for the main effects of width and distfac.

For width equal to 1, say, we get (with approximate 95% confidence limits in parenthesis)

$$\hat{\mu} + \hat{\alpha}(1) + \bar{\hat{\phi}} = 1.70\ (1.62, 1.78)$$

where we average over the three levels of `dist`. For `dist` equal to 0, say, we get

$$\hat{\mu} + \bar{\hat{\alpha}} + \hat{\phi}(0) = 1.80\ (1.72, 1.88)$$

where we average over the three levels of `width`. Similarly for the other levels of `width` and `depth`. We give them on the original scale (taking exponentials of both estimates and confidence limits):

Factor	Adj. mean	Adj. mean	Adj. mean
`width`	1: 5.48 (5.05, 5.94)	2: 5.75 (5.30, 6.24)	3: 5.07 (4.68, 5.50)
`dist`	0: 6.05 (5.57, 6.58)	2: 5.75 (5.30, 6.23)	4: 4.59 (4.24, 4.98)

In particular we find the expected pattern for `distfac`: the humidity decreases from the center and out. In the width-direction the pattern is less clear.

4.2.1 `R` programs and output

Mixed linear models can be analyzed with the `lme`-function from the `nlme`-package or with the `lmer`-function from the `lme4`-package. The packages must be loaded before we can use the functions, see below.

If there are non-nested random effects, then the specification of the random effects is much easier with `lmer` than with `lme`. However, `lmer` does not (yet) have all the "facilities" that `lme` does so will mainly use `lme`. In a few cases we also show how to use `lmer`. Note that `lmer` is more general than `lme` and can also be used for generalized linear models with random effects, for example logistic regression models with random effect, see *** Chapter **??**. ***

Some general comments on installation and loading of packages before we start the data analysis: There are a large number of additional R-packages available which are not automatically installed with the base package. `nlme` and `lmer` are two such packages. A package is a collection of R-functions. To access the functions from a package, the package should once and for all be installed on your local computer. On a computer with internet connection, clicking "Packages" in the R menu will give you the option "Install package(s) from CRAN", and first list a number of countries (choose "Denmark") and then list the possible packages. Click the wanted package and it is installed! This only needs to be carried out once on your computer. To actually use the functions from the add-on package (and for the help information to be visible) the package must also be loaded, either via the "Packages" menu or by the following `library`-command. You need to do this every time R is re-started. Actually, `nlme` is in the standard package so it should only be loaded (not installed). The `lme4`-package should be installed, too.

Example 4.2 *(continued)* We now show how to analyse the beech wood data with `R`.

Reading the data into `R`

Suppose that the data are available in the file `plank.txt` as follows:

```
plank width depth humid
1       1     1    3.4
1       1     3    4.9
.       .     .     .  [more datalines here]
.       .     .     .
20      3     9    4.2
```

Then, the dataset is read into R the explanatory variables are made factors and a variable with the logarithmic humidity values is constructed:

```
> plankdata = read.table("plank.txt",header=T)
> attach(plankdata)

> loghumid = log(humid)
> w = factor(width)
> d = factor(depth)
> p = factor(plank)
```

The profile plots

The profiles for each plank (the upper plots in Figure 4.5) are most easily constructed with `interaction.plot`. Note however, that the figures are only reasonable because the values of `depth` and `width` are equidistant. For the average plot in the bottom we need to construct the averages ourselves.

```
> interaction.plot(w,p,humid)
> interaction.plot(d,p,humid)

> meanhum = rep(0,15)
> w1 = rep(0,15)
> d1 = rep(0,15)

> for (i in 1:3) for (j in 1:5)
  {
   no = 5*(i-1)+j
   w1[no] = i
   dep = d[j]
   d1[no] = dep
   meanhum[no] = mean(humid[w==i & d==dep])
  }

> interaction.plot(w1,d1,meanhum)
> interaction.plot(d1,w1,meanhum)
```

Model validation

Model validation is carried out in the corresponding linear model with `lm`:

```
> library(MASS)

> model1 = lm(humid ~ w*d + p)
> plot(predict(model1),stdres(model1))

> model2 = lm(loghumid ~ w*d + p)
> plot(predict(model2),stdres(model2))
> qqnorm(stdres(model2))
```

Analysis of mixed linear model with `lme`

Since the design is balanced we could carry out exact F-tests (see below on how to so with `aov`). We will, however, use the (approximate) methods that are applicable for non-balanced designs as well.

First `nlme` is loaded so we can use `lme`. Second, we fit models (4.3) and (4.4) with `lme`. The fixed effect part of a model is written in the usual `lm`-way. The random part is specified by a one-sided expression followed by some grouping variables after the `|`. Here, we have only one grouping variable, `plank`. Since we want to use ML estimation (rather than REML which is default), we use the option `method="ML"`. `anova` gives us a test for the reduction from `model3` to `model4`.

```
> library(nlme)
> model3 = lme(loghumid ~ w+d+w:d, random = ~1|p, method="ML")
> model4 = lme(loghumid ~ w+d, random = ~1|p, method="ML")
> anova(model4,model3)
       Model df       AIC       BIC   logLik   Test  L.Ratio p-value
model4     1  9 -407.7940 -374.4599 212.8970
model3     2 17 -399.9446 -336.9803 216.9723 1 vs 2 8.150664  0.4189
```

We get $LR = 8.15$ and an approximate p-value computed from the $\chi^2(8)$-distribution of 0.42 ($8 = 17 - 9$). In this case there is no doubt that the correct p-value is above 5%, so there is no need to do bootstapping. We will do it anyway, however, just to show the method:

First, `simulate.lme` is used to to simulate 1000 datasets from the null model, using the estimates from the real dataset. In our case the null model corresponds to `model4`. For each simulated dataset, the null as well as the alternative model, here given by `model3` are fitted. `R` saved the maximum values of the log-likelihood function for each simulated dataset in a list. This takes a few minutes. We plug out the relevant values, compute the LR test statistic in `lrsim` and finally compute the frequency of simulated LR-values that are larger than our observed value, 8.151. In this case out bootstrap p-value is 0.436.

```
> sim = simulate.lme(model4, m2=model3, nsim=1000, method="ML")
> lrsim = 2*(sim$alt$ML - sim$null$ML)
> psim = sum(lrsim > 8.151)/1000
> psim
[1] 0.436
```

Next, we test for the main effects. We fit the models with no depth effect and no width effect, respectively, and use `anova`:

```
> model5 = lme(loghumid ~ w, random = ~1|p, method="ML")
> anova(model5,model4)
       Model df       AIC       BIC    logLik   Test  L.Ratio p-value
model5     1  5 -172.4461 -153.9272  91.22304
model4     2  9 -407.7940 -374.4599 212.89698 1 vs 2 243.3479  <.0001

> model6 = lme(loghumid ~ d, random = ~1|p, method="ML")
> anova(model6,model4)
       Model df       AIC       BIC   logLik   Test  L.Ratio p-value
model6     1  7 -347.5627 -321.6362 180.7813
model4     2  9 -407.7940 -374.4599 212.8970 1 vs 2 64.23127  <.0001
```

We see that both main effects are highly significant (no need for bootstrapping).

Now, to the test of symmetry around the center. The factor `distfac` is constructed, as a factor with level 0, 2, 4. We fit the model (4.5) and test it against model (4.4):

```
> dist = abs(depth-5)
> distfac = factor(dist)

> model7 = lme(loghumid ~ w+distfac, random = ~1|p, method="ML")
> anova(model7,model4)
       Model df       AIC       BIC   logLik   Test   L.Ratio p-value
model7     1  7 -411.5675 -385.6410 212.7838
model4     2  9 -407.7940 -374.4599 212.8970 1 vs 2 0.2264393   0.893
```

Estimation, adjusted means, etc.

We fit the final model with REML. The estimates are found with `summary` and `VarCorr`:

```
> model7a = lme(loghumid ~ w+distfac, random = ~1|p, method="REML")
> summary(model7a)
Linear mixed-effects model fit by REML

Fixed effects: loghumid ~ w + distfac
                 Value  Std.Error  DF   t-value p-value
(Intercept)  1.8093497 0.04272041 276  42.35328  0.0000
w2           0.0489567 0.01498030 276   3.26807  0.0012
w3          -0.0763691 0.01498030 276  -5.09797  0.0000
distfac2    -0.0517808 0.01674849 276  -3.09167  0.0022
distfac4    -0.2753288 0.01674849 276 -16.43903  0.0000

> VarCorr(model7a)
p = pdLogChol(1)
            Variance   StdDev
(Intercept) 0.03126445 0.1768176
Residual    0.01122047 0.1059267
```

The adjusted means for `width` and `distfac` are obtained by the function `estimable` from the `gmodels`-package. Recall that the package should be installed and loaded before we can use `estimable`. `estimable` enables us to estimate linear combinations of the parameters. We must specify the particular linear combinations we are interested in. For example, we write

$$\hat{\mu} + \hat{\alpha}(1) + \bar{\hat{\phi}} = 1 \cdot \hat{\mu} + 1 \cdot \hat{\alpha}(1) + 0 \cdot \hat{\alpha}(2) + 0 \cdot \hat{\alpha}(3) + \frac{1}{3} \cdot \hat{\phi}(0) + \frac{1}{3} \cdot \hat{\phi}(2) + \frac{1}{3} \cdot \hat{\phi}(4). \tag{4.6}$$

From the `summary`-output above we see that $\alpha(1)$ and and $\phi(0)$ have been set to zero by `R`. Hence, (4.6) is equal to

$$1 \cdot \hat{\mu} + 0 \cdot \hat{\alpha}(2) + 0 \cdot \hat{\alpha}(3) + \frac{1}{3} \cdot \hat{\phi}(2) + \frac{1}{3} \cdot \hat{\phi}(4).$$

so the relevant coefficients are (1,0,0,1/3,1/3). Similarly for the other linear combinations.

We get estimates, standard errors as follows. Recall that these are on the logarithmic scale, so they should be exponentiated in order to be on the original scale.

```
> library(gmodels)

> ls.w1 = c(1,0,0,1/3,1/3)
> ls.w2 = c(1,1,0,1/3,1/3)
> ls.w3 = c(1,0,1,1/3,1/3)
```

```
> ls.d0 = c(1,1/3,1/3,0,0)
> ls.d2 = c(1,1/3,1/3,1,0)
> ls.d4 = c(1,1/3,1/3,0,1)

> lsmeans = rbind(ls.w1,ls.w2,ls.w3,ls.d0,ls.d2,ls.d4)
> y = estimable(model7a, lsmeans, conf.int=0.95)

> y
      Estimate Std. Error  t value DF Pr(>|t|) Lower.CI Upper.CI
ls.w1 1.700313 0.04098271 41.48854 92        0 1.618918 1.781708
ls.w2 1.749270 0.04098271 42.68311 92        0 1.667875 1.830665
ls.w3 1.623944 0.04098271 39.62510 92        0 1.542549 1.705339
ls.d0 1.800212 0.04183575 43.03047 92        0 1.717123 1.883302
ls.d2 1.748431 0.04070290 42.95594 92        0 1.667592 1.829271
ls.d4 1.524883 0.04070290 37.46375 92        0 1.444044 1.605723

> exp(y)
      Estimate Std. Error      t value           DF Pr(>|t|) Lower.CI Upper.CI
ls.w1 5.475662   1.041834 1.042907e+18 9.017628e+39        1 5.047625 5.939995
ls.w2 5.750402   1.041834 3.443818e+18 9.017628e+39        1 5.300889 6.238034
ls.w3 5.073059   1.041834 1.617933e+17 9.017628e+39        1 4.676495 5.503252
ls.d0 6.050931   1.042723 4.874119e+18 9.017628e+39        1 5.568484 6.575178
ls.d2 5.745583   1.041543 4.524042e+18 9.017628e+39        1 5.299391 6.229343
ls.d4 4.594608   1.041543 1.863372e+16 9.017628e+39        1 4.237798 4.981459
```

Recall that only the estimates and the confidence limits make sense on the exponential scale (the `exp(y)`-object). In particular the standard errors do not! `R` also comes with a warning message which has to do with the computation of the degrees of freedom: the tests and confidence intervals are only approximate.

Analysis with `lmer`

We could have fitted the models with `lmer` instead of `lme`. For example:

```
> library(lme4)

> model3a = lmer(loghumid ~ w+d+w:d + (1|p), method="ML")
> model4a = lmer(loghumid ~ w+d + (1|p), method="ML")
> anova(model4a,model3a)
        Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
model4a  9 -407.79 -374.46  212.90
model3a 17 -399.94 -336.98  216.97 8.1507      8     0.4189
```

Analysis with `aov`

Finally, we show how the exact test for interaction can be carried out with `aov`. Recall that this test is only valid when the design is balanced.

```
> model8 = aov(loghumid ~ w*d + Error(p))
> summary(model8)
Error: p
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 19 9.1236  0.4802
```

```
Error: Within
           Df Sum Sq Mean Sq F value    Pr(>F)
w           2 0.7979  0.3989 35.3059 2.518e-14 ***
d           4 4.2849  1.0712 94.8067 < 2.2e-16 ***
w:d         8 0.0888  0.0111  0.9821    0.4503
Residuals 266 3.0056  0.0113
```

Since the interaction is not significant, the model without interaction should then be fitted, and tests for the main effects should be carried out.

4.2.2 SAS programs and output

Mixed linear models can be analyzed with `proc mixed` in sas. The call to `proc mixed` is quite similar to that of `proc glm`: The qualitative variables (factors) should appear in the `class` statement in order for SAS to know that they are indeed qualitative. The fixed part is specified as for `proc glm` whereas the random factor(s) are specified in the `random`-statement.

As default, SAS used the REML-method for estimation. We generally recommend to write `nobound`, allowing for negative variance components, and to include the option `ddfm=satterth` in `model` statement. This tells SAS to use a particular method for computation of degrees of freedom in non-balanced cases. If you prefer to make *likelihood ratio tests*, this is possible too. See below for a short explanation.

Example 4.2 *(continued)* We now show how to analyse the beech wood data with SAS.

Reading the data into SAS

Suppose that the data are available in the file `plank.txt` as follows:

```
plank width depth humid
1       1     1    3.4
1       1     3    4.9
.       .     .     .  [more datalines here]
.       .     .     .
20      3     9    4.2
```

Then, the dataset is read into SAS with the following program lines:

```
data planks;
infile 'c:\plank.txt' firstobs=2;
input plank width depth humid;
proc print;
run;
```

Model validation

Model validation is carried out in the corresponding linear model, that is, the model where plank is used as a fixed effect. Hence, we use `proc glm`. First, the variable `humid` is used as response, next the logarithmic humidity is computed and used as response.

```
proc glm data = planks;
  class plank width depth;
  model humid = width*depth  plank;
  output out=out1 predicted = pred1 student=sres1;
run;

symbol1 i=none v=dot c=black;
proc gplot data = out1;
  plot sres1*pred1;
run;

data planks;
  set planks;
  loghumid = log(humid);
run;

proc glm data = planks;
  class plank width depth;
  model loghumid = width*depth  plank;
  output out=out2 predicted = pred2 student=sres2;
run;

proc gplot data = out2;
  plot sres2*pred2;
run;

proc univariate data=out2;
  qqplot sres2;
run;
```

Analysis with the mixed linear model

Model (4.3) is fitted with `proc mixed` as follows. As mentioned we recommend to use Satterthwaite's approximation to the degrees of freedom (`ddfm=satterth`) and allow for negative variance components (`nobound`).

```
proc mixed data = planks nobound;
  class plank width depth;
  model loghumid = width depth width*depth / ddfm=satterth;
  random plank;
run;
```

The output goes like this (unedited, for this one time):

```
                        The Mixed Procedure

                        Model Information

      Data Set                     WORK.PLANKS
      Dependent Variable           loghumid
      Covariance Structure         Variance Components
      Estimation Method            REML
      Residual Variance Method     Profile
      Fixed Effects SE Method      Model-Based
```

Degrees of Freedom Method Satterthwaite

Class Level Information

Class	Levels	Values
plank	20	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
width	3	1 2 3
depth	5	1 3 5 7 9

Dimensions

Covariance Parameters	2
Columns in X	24
Columns in Z	20
Subjects	1
Max Obs Per Subject	300

Number of Observations

Number of Observations Read	300
Number of Observations Used	300
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	-45.97969832	
1	1	-352.69281063	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Estimate
plank	0.03126
Residual	0.01130

Fit Statistics

-2 Res Log Likelihood	-352.7
AIC (smaller is better)	-348.7
AICC (smaller is better)	-348.7
BIC (smaller is better)	-346.7

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq

```
                1        306.71          <.0001


                Type 3 Tests of Fixed Effects

                          Num     Den
Effect                     DF      DF    F Value    Pr > F
width                       2     266      35.31    <.0001
depth                       4     266      94.81    <.0001
width*depth                 8     266       0.98    0.4503
```

Note the message "Convergence criteria met". This means that the numerical procedure found a solution. If another message comes up, we cannot trust the results from the fit. We see that the interaction between width and depth is not significant. We fit model (4.4) without the interaction term (but leave out the output that shows that both main effects are highly significant).

```
proc mixed data = planks nobound;
  class plank width depth;
  model loghumid = width depth / ddfm=satterth;
  random plank;
run;
```

In order to test if model (4.4) can be reduced to model (4.5) we construct the factor with distances, and fit the model with both the original depth factor and the new distance factor.

```
data planks;
  set planks;
  dist = abs(depth-5);
proc print;
run;

proc mixed data = planks nobound;
  class plank width depth dist;
  model loghumid = width  dist depth / ddfm=satterth;;
  random plank;
run;
```

The output is like this (edited), where we see that depth is not significant as long as the distance factor is in the model.

```
                      The Mixed Procedure

                       Iteration History

    Iteration    Evaluations    -2 Res Log Like       Criterion
            0              1       -70.85875516
            1              1      -382.34966620      0.00000000
                  Convergence criteria met.


                    Covariance Parameter
                          Estimates

                    Cov Parm     Estimate
```

```
plank         0.03126
Residual      0.01129


Type 3 Tests of Fixed Effects

                Num     Den
Effect           DF      DF    F Value    Pr > F
width             2     274      35.32    <.0001
dist              0       .          .         .
depth             2     274       0.11    0.8951
```

Finally we fit model (4.5), ask for the parameter estimates (`solution`) and the adjusted means (`lsmeans`):

```
proc mixed data = planks nobound;
  class plank width depth dist;
  model loghumid = width  dist / ddfm=satterth solution;
  random plank;
  lsmeans width dist / cl;
run;
```

And the output:

```
                    The Mixed Procedure

                     Iteration History

    Iteration    Evaluations    -2 Res Log Like       Criterion
            0              1       -80.28630752
            1              1      -394.22748682      0.00000000
                  Convergence criteria met.


                   Covariance Parameter
                        Estimates

                   Cov Parm     Estimate
                   plank         0.03126
                   Residual      0.01122


                  Solution for Fixed Effects

                                                 Standard
Effect       width    dist    Estimate       Error      DF    t Value    Pr > |t|
Intercept                      1.4577     0.04161    22.2      35.03      <.0001
width        1                0.07637     0.01498     276       5.10      <.0001
width        2                 0.1253     0.01498     276       8.37      <.0001
width        3                      0           .       .          .          .
dist                 0         0.2753     0.01675     276      16.44      <.0001
dist                 2         0.2235     0.01368     276      16.35      <.0001
dist                 4              0           .       .          .          .

                    Type 3 Tests of Fixed Effects
```

Effect	Num DF	Den DF	F Value	Pr > F
width	2	276	35.55	<.0001
dist	2	276	190.83	<.0001

Least Squares Means

Effect	width	dist	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
width	1		1.7003	0.04098	20.9	41.49	<.0001	0.05	1.6151	1.7856
width	2		1.7493	0.04098	20.9	42.68	<.0001	0.05	1.6640	1.8345
width	3		1.6239	0.04098	20.9	39.63	<.0001	0.05	1.5387	1.7092
dist		0	1.8002	0.04184	22.7	43.03	<.0001	0.05	1.7136	1.8868
dist		2	1.7484	0.04070	20.4	42.96	<.0001	0.05	1.6636	1.8332
dist		4	1.5249	0.04070	20.4	37.46	<.0001	0.05	1.4401	1.6097

Finally, some useful comments if you prefer to do likelihood ratio tests. Model (4.3) if fitted with ML as follows:

```
proc mixed data = planks method=''ML'';
  class plank width depth;
  model loghumid = width depth width*depth;
  random plank;
run;
```

Then, SAS reports $-2 \cdot \log L(\text{null model})$. Fit the null model (the model under the hypothesis) similarly. Then compute the difference between the reported log-likehoods and compare the difference to the relevant χ^2-approximation. Recall that χ^2-approximation of the p-value tends to be too small, so be careful with the conclusisions.

4.3 Analysis of the pork meat data

Example 4.3 *(continued)* Let us continue with the pork meat data. Recall the random intercepts model (4.2) with `storage`, `breed` and `storage` × `breed` as systematic factors and `pigno` as a random factor:

$$y_i = \mu + \alpha(\texttt{storage}_i) + \beta(\texttt{breed}_i) + \gamma(\texttt{storage}_i, \texttt{breed}_i) + d(\texttt{pigno}_i) + \epsilon_i,\ i = 1, \ldots, 300, \tag{4.7}$$

First, we carry out model validation. We do so in the "corresponding linear model", that is, the model where `pigno` is fixed rather than random. The residual plot in Figure 4.8 does not indicate variance heterogeneity, and we are not worried about the QQ-plot either.

Second, we try to reduce the model. It turns out that the interaction `storage` × `breed` is not significant ($p \approx 0.45$ depending on the method). In other words, we accept the model

$$y_i = \mu + \alpha(\texttt{storage}_i) + \beta(\texttt{breed}_i) + d(\texttt{pigno}_i) + \epsilon_i,\ i = 1, \ldots, 300, \tag{4.8}$$

Both the effect of `storage` and the effect of `breed` turn out to be significant in this model ($p \approx 0.0003$ and $p \approx 0.02$, respectively) in this model, so (4.8) is the final model.

Third, we estimate interesting parameters (standard errors in parenthesis) in the final model and give

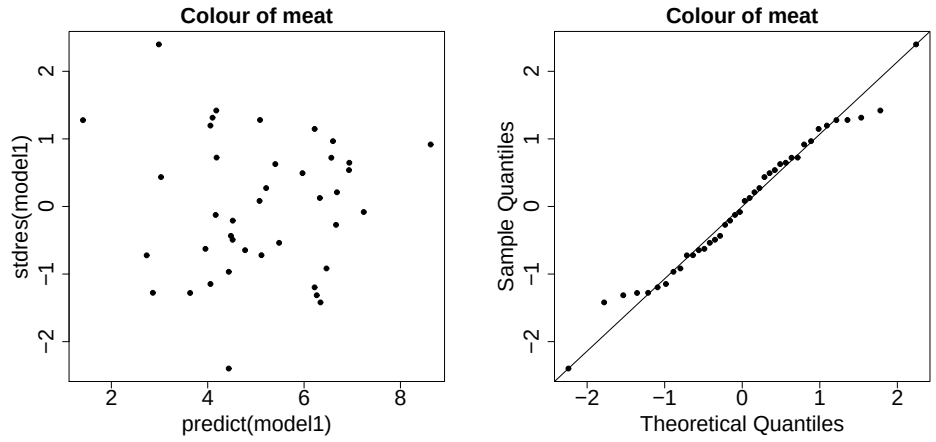


Figure 4.8: Residual plot and QQ plot for the pork meat data.

conclusions. We get

$$\hat{\alpha}(\texttt{dark}) - \hat{\alpha}(\texttt{light}) = 1.807\ (0.458)$$
$$\hat{\beta}(\texttt{old}) - \hat{\beta}(\texttt{new}) = 1.131\ (0.474)$$

so the redness is 1.8 larger for dark storage compared to light storage and the redness is 1.1 larger for the old breed than for the new breed. The variance parameters are estimated to

$$\sigma^2_{\texttt{pigno}} = 0.0716; \quad \sigma^2 = 2.100$$

Finally, we compute the *adjusted means*, also called least squares means (or LS means). For storage we find

$$\hat{\mu} + \hat{\alpha}(\texttt{dark}) + \bar{\hat{\beta}} = 6.00\ (0.33)$$
$$\hat{\mu} + \hat{\alpha}(\texttt{light}) + \bar{\hat{\beta}} = 4.19\ (0.33)$$

and for breed we get

$$\hat{\mu} + \bar{\hat{\alpha}} + \hat{\beta}(\texttt{old}) = 5.66\ (0.33)$$
$$\hat{\mu} + \bar{\hat{\alpha}} + \hat{\beta}(\texttt{new}) = 4.53\ (0.33)$$

4.3.1 R programs and output

Example 4.3 *(continued)* We now show how to analyse the pork meat data with R.

Reading the data into R

Suppose that the data are available in the file `redness.txt` as follows:

```
pigno  storage breed  colour
 13     light   old   5.5630
```

```
 13     dark    old   9.5280
 41     light   old   5.5730
  .       .       .     .       [more datalines here]
  .       .       .     .
286     dark    new   6.3960
```

The dataset is read into R and attached, so we can use the variable names. Moreover, the pigno-variable is made a factor.

```
> redness = read.table("redness.txt",header=T)
> attach(redness)
> pigno = factor(pigno)
```

Model validation

Model validation is carried out in the model where pigno is fixed rather than random. Hence, we can use lm in order to produce the residual plot and the QQ-plot in Figure 4.8.

```
> model1 = lm(colour ~ storage*breed + pigno)
> library(MASS)
> plot(predict(model1),stdres(model1))
> qqnorm(stdres(model1))
> qqline(stdres(model1))
```

Analysis of mixed model with lme

We use lme for the analysis, see Section 4.2.1 for details on the syntax. Here the grouping variable is pigno.

```
> library(nlme)
> model2 = lme(colour ~ storage + breed + storage:breed, random=~1|pigno,method="ML")
> model3 = lme(colour ~ storage+breed, random=~1|pigno, method="ML")
> anova(model3,model2)
       Model df      AIC      BIC    logLik   Test   L.Ratio p-value
model3     1  5 151.3832 159.8276 -70.69161
model2     2  6 152.7377 162.8710 -70.36886 1 vs 2 0.6455155  0.4217
```

The anova-output shows that the interaction between storage and breed is not significant. The p-value if large enough that there is no need to compute a more accurate approximation of the p-value by bootstrap methods.

We then test if there is a significant effect of breed. We fit the model without breed, and use anova to test the model against model (4.8). It turns out that that the approximate χ^2-value is 0.019. This in the area where we may doubt the approximation so we compute a more accurate approximation by bootstrap and get 0.027. We conclude that there is a (slightly) significant effect of breed.

```
> model4 = lme(colour ~ storage, random=~1|pigno, method="ML")
> anova(model4,model3)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model4     1  4 154.8893 161.6448 -73.44463
model3     2  5 151.3832 159.8276 -70.69161 1 vs 2 5.506023   0.019
```

```
> sim = simulate.lme(model4, m2=model3, nsim=1000, method="ML")
> lrsim = 2*(sim$alt$ML - sim$null$ML)
> psim = sum(lrsim > 5.506)/1000
> psim
[1] 0.027
```

Finally we test for the effect of storage which turns out to be significant:

```
> model5 = lme(colour ~ breed, random=~1|pigno, method="ML")
> anova(model5,model3)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model5     1  4 163.0274 169.7829 -77.51370
model3     2  5 151.3832 159.8276 -70.69161 1 vs 2 13.64417   2e-04
```

Hence, `model3` is the final one.

Estimates from final model

We estimate the final model with REML and get estimates and (approximate) confidence limits by `summary`, `intervals` and `VarCorr`.

```
> summary(model3a)
Linear mixed-effects model fit by REML

Fixed effects: colour ~ storage + breed
               Value Std.Error DF   t-value p-value
(Intercept)   5.4357 0.4058068 19 13.394796  0.0000
storagelight -1.8065 0.4582788 19 -3.941924  0.0009
breedold      1.1313 0.4736546 18  2.388449  0.0281

> intervals(model3a)
Approximate 95% confidence intervals

 Fixed effects:
                 lower    est.      upper
(Intercept)   4.5863365  5.4357  6.2850635
storagelight -2.7656884 -1.8065 -0.8473116
breedold      0.1361886  1.1313  2.1264114

> VarCorr(model3a)
pigno = pdLogChol(1)
            Variance   StdDev
(Intercept) 0.07164632 0.2676683
Residual    2.10019413 1.4492047
```

Adjusted means

The adjusted means for storage and breed are obtained by the function `estimable` from the `gmodels`-package. Recall that the package should be installed and loaded first. `estimable` enables us to estimate linear combinations of the parameters. We must specify the particular linear combinations we are interested in.

For example, we write

$$\hat{\mu} + \hat{\alpha}(\texttt{dark}) + \bar{\hat{\beta}} = 1 \cdot \hat{\mu} + 1 \cdot \hat{\alpha}(\texttt{dark}) + 0 \cdot \hat{\alpha}(\texttt{light}) + \frac{1}{2} \cdot \hat{\beta}(\texttt{new}) + \frac{1}{2} \cdot \hat{\beta}(\texttt{old}) \tag{4.9}$$

From the `summary`-output above we see that $\alpha(\texttt{dark})$ and $\beta(\texttt{new})$ have been set to zero by `R`. Hence, (4.9) is equal to

$$1 \cdot \hat{\mu} + 0 \cdot \hat{\alpha}(\texttt{light}) + \frac{1}{2} \cdot \hat{\beta}(\texttt{old}),$$

so the relevant coefficients are (1,0,1/2). Similarly, the coefficient for $\hat{\mu} + \hat{\alpha}(\texttt{light}) + \bar{\hat{\beta}}$ is (1,1,1/2). For breed we find the coefficients (1,1/2,1) for $\hat{\mu} + \bar{\hat{\alpha}} + \hat{\beta}(\texttt{old})$ and (1,1/2,0) for $\hat{\mu} + \bar{\hat{\alpha}} + \hat{\beta}(\texttt{new})$.

We get estimates, standard errors and (approximate) confidence limits as follows:

```
> library(gmodels)

> lsmean.dark = c(1,0,0.5)
> lsmean.light = c(1,1,0.5)
> lsmean.old = c(1,0.5,1)
> lsmean.new = c(1,0.5,0)

> lsmeans = rbind(lsmean.dark, lsmean.light, lsmean.old, lsmean.new)
> estimable(model3a,lsmeans,conf.int=0.95)
             Estimate Std. Error  t value  DF     Pr(>|t|) Lower.CI Upper.CI
lsmean.dark   6.00135  0.3295330 18.21168 9.0 2.070825e-08 5.255894 6.746806
lsmean.light  4.19485  0.3295330 12.72968 9.0 4.649902e-07 3.449394 4.940306
lsmean.old    5.66375  0.3349244 16.91053 9.5 2.073934e-08 4.912134 6.415366
lsmean.new    4.53245  0.3349244 13.53276 9.5 1.593678e-07 3.780834 5.284066
```

`R` also comes with a warning message which has to do with the computation of de degrees of freedom. The confidence intervals are only approximate.

Analysis with `lmer`

We could have fitted model (4.7) and (4.8) with `lmer` instead. The difference, compared to `lme`, is the specification of the random effect:

```
> library(lme4)
> model2a = lmer(colour ~ storage*breed + (1|pigno))
> model3a = lmer(colour ~ storage+breed + (1|pigno))
```

Analysis of mixed model with `aov`

Exact F-tests can be carried out since the design is balanced. Here we just show how to fit model (4.7) with `aov`. This gives the test for interaction.

```
> model6 = aov(colour ~ storage*breed + Error(pigno))
> summary(model6)
Error: pigno
           Df Sum Sq Mean Sq F value  Pr(>F)
breed      1 12.798  12.798  5.7047 0.02808 *
Residuals 18 40.383   2.243
```

```
Error: Within
              Df Sum Sq Mean Sq F value   Pr(>F)
storage        1 32.634  32.634 15.2038 0.001051 **
storage:breed  1  1.267   1.267  0.5904 0.452213
Residuals     18 38.636   2.146
```

Then the model without interaction is fitted and the relevant hypotheses are tested.

4.3.2 SAS programs and output

Example 4.3 *(continued)* We now show how to analyse the pork meat data with SAS.

Reading the data into SAS

Suppose that the data are available in the file `redness.txt` as follows:

```
pigno  storage breed  colour
 13     light   old   5.5630
 13     dark    old   9.5280
 41     light   old   5.5730
  .       .      .     .      [more datalines here]
  .       .      .     .
286     dark    new   6.3960
```

Then the dataset is read into SAS with the following program lines:

```
data redness;
  infile 'redness.txt' firstobs=2;
  input pigno storage $ breed $ colour;
run;
```

Model validation

A residual plot and a QQ-plot similar to those of Figure 4.8 are produced as follows. The output is not shown.

```
proc glm data=redness;
  class storage breed pigno;
  model colour = storage breed storage*breed pigno;
  output out=out1 predicted=pred student=sres;
run;

proc gplot data=out1;
  plot sres*pred;
proc univariate data=out1;
  qqplot sres;
run;
```

Analysis of mixed linear model

Model (4.7) is fitted with `proc mixed` (see Section 4.2.2 for details on the syntax):

```
proc mixed data=redness nobound;
  class storage breed pigno;
  model colour = storage breed storage*breed / ddfm=satterth;
  random pigno;
run;
```

The output goes like this (edited):

```
                        The Mixed Procedure

                         Iteration History

      Iteration    Evaluations    -2 Res Log Like       Criterion
              0              1       139.67605395
              1              1       139.66725932      0.00000000

                   Convergence criteria met.


                      Covariance Parameter
                           Estimates

                      Cov Parm     Estimate
                      pigno         0.04851
                      Residual       2.1465


                  Type 3 Tests of Fixed Effects

                         Num     Den
      Effect              DF      DF    F Value    Pr > F
      storage              1      18      15.20    0.0011
      breed                1      18       5.70    0.0281
      storage*breed        1      18       0.59    0.4522
```

Note the message "Convergence criteria met". This means that the numerical procedure found a solution. If another message comes up, we cannot trust the results from the fit. We see that the interaction between storage and breed is not significant. We fit model (4.8) without the interaction term. Moreover, we ask `SAS` for the adjusted means (LS means) with confidence limits.

```
proc mixed data=redness nobound;
  class storage breed pigno;
  model colour = storage breed / ddfm=satterth solution clparm;
  random pigno;
  lsmeans storage breed / cl;
run;
```

The output goes as follows, this time in an edited version:

The Mixed Procedure

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	141.95877041	
1	1	141.93862944	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Estimate
pigno	0.07165
Residual	2.1002

Solution for Fixed Effects

Effect	storage	breed	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept			4.7605	0.4058	32.1	11.73	<.0001	0.05	3.9340	5.5870
storage	dark		1.8065	0.4583	19	3.94	0.0009	0.05	0.8473	2.7657
storage	light		0	.	.	.	.	.	.	.
breed		new	-1.1313	0.4737	18	-2.39	0.0281	0.05	-2.1264	-0.1362
breed		old	0	.	.	.	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
storage	1	19	15.54	0.0009
breed	1	18	5.70	0.0281

Least Squares Means

Effect	storage	breed	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
storage	dark		6.0014	0.3295	36.9	18.21	<.0001	0.05	5.3336	6.6691
storage	light		4.1949	0.3295	36.9	12.73	<.0001	0.05	3.5271	4.8626
breed		new	4.5325	0.3349	18	13.53	<.0001	0.05	3.8288	5.2361
breed		old	5.6638	0.3349	18	16.91	<.0001	0.05	4.9601	6.3674

Chapter 5

Repeated measurements

In this chapter we will analyze data from experiments where several measurements of some physical quantitity is taken from each experimental unit or subject (person, animal, plant or similar) during the experimental period. This data type occurs quite often, and we refer to it as repeated measures.

Usually we think of repeated measures as observations on the same experimental unit at several time-points, but it might just as well be at different locations of the same unit, for example different locations in the blood stream. In the following we will think of the time situation. Often the experimental units are divided into groups each of which is given a particular treatment. The goal is then typically to compare the treatments. Moreover, one is often interested in the development over time (and the effect of treatment on the this time-response relation).

Example 5.1 (Growth of rats) An experiment was carried out in order to investigate the effect of *thyroxin* and *thiouracil* on the growth of rats. During a period of five weeks, 7 rats has thyroxin added to their drinking water, 10 rats has thiouracil added, and another 10 rats were used as controls. The rats were weighed each week. Table 5.1 shows the log-transformed weight. The goal is to compare the three treatment groups: is the growth similar or different in the three groups? □

Had all measurements been taken from different experimental units, a simple linear or non-linear regression analysis with treatment dependent coefficients (and independent observations) would have been natural. With repeated measures, however, it is not reasonable to assume independence between all observations. Rather, observations from same experimental units tend to be positively correlated, and observations taken close in time tend to more correlated than observations not close in time. This must be taken into account in the analysis. In principle we have less information about the treatment differences compared to the situation where all measurements come from different experimental units. On the other hand we have more information about the time development with repeated measures.

In the following, we will discuss three different analysis strategies: analysis of summary measures, analysis with the random intercepts model (corresponding to the split-plot model) and analysis with a model incorporating a more sophisticated correlation structure.

Finally, a warning: At first sight a natural idea might be to analyze the data for each point in time seperately. This is usually not a good idea, though (except for exploratory analysis)! First, since only few data are used in each analysis the analyses are not so strong as is an analysis of the full dataset. In other words, (treatment) differences should be large in order to come out significant. Second, the analyses (tests) are not independent as they are based on observations from the same subjects. Third, it is difficult to give an overall conclusion if the different analyses result in different conclusions. Fourth, the analyses would not tell much about the development over time. In summary, the analysis method

Rat	Week / Treatment	1	2	3	4	5
		log weight (g)				
1	Kontrol	4.0431	4.4543	4.7362	4.9345	5.1475
2	Kontrol	4.0943	4.5326	4.8122	4.9836	5.1761
3	Kontrol	3.9512	4.3438	4.7095	4.9698	5.2204
4	Kontrol	3.8918	4.2047	4.6052	4.8598	5.0999
5	Kontrol	4.0254	4.3944	4.6444	4.7958	5.0173
6	Kontrol	3.8286	4.2485	4.6250	4.8752	5.0304
7	Kontrol	3.9318	4.2627	4.5433	4.7005	4.9488
8	Kontrol	4.1431	4.5109	4.7185	4.8675	5.0370
9	Kontrol	3.8918	4.2047	4.4998	4.7185	4.9416
10	Kontrol	4.0431	4.4067	4.7005	4.9345	5.1299
11	Thyroxin	4.0775	4.4427	4.7958	4.9836	5.1985
12	Thyroxin	3.9890	4.2627	4.4998	4.7005	4.9273
13	Thyroxin	4.0254	4.3175	4.6821	5.0173	5.2417
14	Thyroxin	4.0775	4.4427	4.7536	4.9972	5.1761
15	Thyroxin	4.0431	4.2767	4.5747	4.7875	4.9698
16	Thyroxin	3.9512	4.2905	4.5747	4.7536	4.9416
17	Thyroxin	3.9512	4.2485	4.6540	4.9273	5.1417
18	Thiouracil	4.1109	4.4543	4.6913	4.7875	4.8598
19	Thiouracil	4.0775	4.3820	4.6151	4.7095	4.8040
20	Thiouracil	3.9703	4.3694	4.6052	4.6634	4.8903
21	Thiouracil	4.0775	4.4773	4.6052	4.7095	4.8040
22	Thiouracil	3.9318	4.3175	4.6151	4.8122	4.9416
23	Thiouracil	3.9318	4.3175	4.5218	4.6052	4.7791
24	Thiouracil	4.0254	4.3567	4.5539	4.6347	4.6821
25	Thiouracil	4.0604	4.2341	4.5326	4.7536	4.9416
26	Thiouracil	3.8286	4.1109	4.3567	4.4998	4.6728
27	Thiouracil	3.9703	4.2767	4.4886	4.6444	4.8040

Table 5.1: The growth of rats data.

is not plain wrong but not very informative, either.

5.1 Illustrative plots

As always it is a good idea to explore the data graphically before the actual analysis. Perhaps even more so for repeated measurements as it gives an idea about the development over time. We recommend two types of plots: subject profile plots and average profile plots.

The *subject profiles* show the observed curve, response against time, for each subject. Hence, we get an idea what the typical time-response relation is like. The curves could be marked after treatment so potential treatment differences may be spotted already here if the picture is not too blurred. Alternatively, and depending on the number of subjects, a graph for each treatment could be made. Note that strange-looking subjects (potential outliers) are often detected in the subject profile plot.

The *average profiles* show the average (over subjects) response against time for each treatment. Treatment differences are often depicted in these plots. These plots also illustrate the interaction between treatment and time. We can of course not see from the plots whether the interaction and potential different treatment effects are significant; a proper statistical analysis is needed for that.

Example 5.1 *(continued)* Figure 5.1 shows the profiles for each rat (to the left) and the average profiles (to the right).

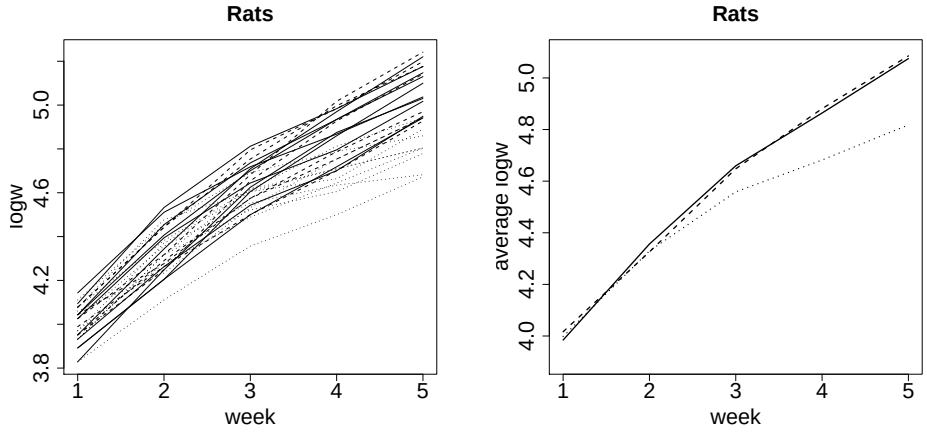


Figure 5.1: Profile plots for the rats data. Solid curves are control, dashed are thryroxin, dotted are thiouracil.

The rat profiles show that the logarithmic weight increases over time. The relationship does not seem to be linear; rather there seems to be a curvature suggesting a quadratic relationship. There is a single rat for which the curve is somewhat different compared to the others. From the plot it it difficult to judge if there is a treatment effect, but it may seem like the thiouracil rats (dotted curves) are slightly below the control rats and the thyroxin rats.

This impression is strengthened in the average profile plot. The thiouracil average is lower in the later weeks compared to the control and thyroxin averages. Moreover, there seems to be an interaction between treatment and time. This will of course be tested in the analysis. □

5.1.1 R programs

We here show how to read the data, construct appropriate variables, and profile plots. At first sight it might not be clear exactly what happens all the time. My best advice is: go home and try it line by line and see for yourself what happens!

Reading the data

Suppose that the data are available in the file `thyroxin.txt` as follows:

```
rat treat        logw1   logw2   logw3   logw4   logw5
 1  Control     4.0431   4.4543   4.7362   4.9345   5.1475
 2  Control     4.0943   4.5326   4.8122   4.9836   5.1761
 .     .         .        .        .        .        .    [more datalines here]
 .     .         .        .        .        .        .
27 Thiouracil   3.9703   4.2767   4.4886   4.6444   4.8040
```

We read the data into the dataset `rats`, and attach it as usual.

```
> rats = read.table("thyroxin.txt",header=T)
> attach(rats)
```

Note that the data file has a line for each subject (rat), rather than a line per measurement as usual. For some purposes this structure is nice, for other purposes we need the usual structure. The following commands create vectors of length 135, one per measurement. The new rat variable, `newrat` is made a factor. Moreover we create a week factor, `weekfac`, and a variable `week2` with the squared values of `week`. Note that `week` and `week2` are numerical variables. These are for later use.

```
> newrat = rep(rat,5)
> newrat = factor(newrat)

> logw = c(logw1,logw2,logw3,logw4,logw5)
> newtreat = rep(treat,5)

> week = c(rep(1,27),rep(2,27),rep(3,27),rep(4,27),rep(5,27))
> weekfac = factor(week)
> week2 = week*week
```

Subject profile plot

There are various ways to create subject profile plots similar to the left plot in Figure 5.1. Here we show one of them. First we make an "empty" plot (`type="n"`) with only labels and axes. Thereafter we add the profiles for the 27 rats: `points` makes `R` add to an existing plot instead of making a new one. We make curves rather than points (`type="l"`). For the control rats we use solid curves (`lty=1`), for the thyroxin rats we use dashed curves (`lty=2`), and for thiouracil rats we use dotted lines (`lty=3`).

```
> plot(week,logw,type="n")

> for (i in 1:27)
  {
   rati = c(logw1[i],logw2[i],logw3[i],logw4[i],logw5[i])

   if (i<=10) points(c(1:5),rati,type="l", lty=1)
   if (i>=11 && i<=17) points(c(1:5),rati,type="l", lty=2)
   if (i>=18) points(c(1:5),rati,type="l", lty=3)
  }
```

Average profile plot

There are also numerous ways to construct the average profile plots. We show one of them. First the 15 average values are computed, five for each of the three treatments, by plugging out the relavant values from the `logw`-variable. The values are stored treatment-wise in three vectors. Then an empty plot with appropriate axes are made, and finally the three treatment profiles are added.

```
> cont.mean = rep(0,5)
> for (j in 1:5) cont.mean[j] =  mean(logw[(27*(j-1)+1) : (27*(j-1)+10)])

> thy.mean = rep(0,5)
> for (j in 1:5) thy.mean[j] =  mean(logw[(27*(j-1)+11) : (27*(j-1)+17)])
```

```
> thiou.mean = rep(0,5)
> for (j in 1:5) thiou.mean[j] =  mean(logw[(27*(j-1)+18) : (27*(j-1)+27)])

> plot(c(1:5),type="n",ylim=c(3.9,5.1))
> points(c(1:5), cont.mean, type="l",lwd=2,lty=1)
> points(c(1:5), thy.mean, type="l",lwd=2,lty=2)
> points(c(1:5), thiou.mean, type="l",lwd=2,lty=3)
```

5.1.2 SAS programs

Reading the data

Suppose that the data are available in the file `thuroxin.txt` as follows:

```
rat treat        logw1   logw2   logw3   logw4   logw5
 1  Control     4.0431   4.4543   4.7362   4.9345   5.1475
 2  Control     4.0943   4.5326   4.8122   4.9836   5.1761
 .     .         .        .        .        .        .    [more datalines here]
 .     .         .        .        .        .        .
27 Thiouracil   3.9703   4.2767   4.4886   4.6444   4.8040
```

We read the data as usual:

```
data rats;
  infile 'c:\thyroxin.txt' firstobs=2;
  input rat treat $ v1 v2 v3 v4 v5;
proc print;
run;
```

Note that the data set has a line for each subject (rat), rather than a line per measurement as usual. For some purposes this structure is nice, for other purposes we need the usual structure. The program lines below create a new dataset, `rats1` with the usual structure: `1week` and `1week1` are identical, but we will need both in the analysis, one as a factor and one a as numerical variable. `week2` contains the squared `week`-values. The response is called `logw`.

```
data rats1;
  set rats;
  if v1 then do; week=1; week1=1; week2=1; logw=v1; output; end;
  if v2 then do; week=2; week1=2; week2=4; logw=v2; output; end;
  if v3 then do; week=3; week1=3; week2=9; logw=v3; output; end;
  if v4 then do; week=4; week1=4; week2=16; logw=v4; output; end;
  if v5 then do; week=5; week1=5; week2=25; logw=v5; output; end;
  drop v1 v2 v3 v4 v5;
proc print;
run;
```

Profile plots

The subject profile plots are constructed as follows:

```
symbol1 i=join v=none c=black repeat=27;
proc gplot data=rats1;
```

```
  plot logw * week = rat;
run;
```

The average profile plots are constructed as follows. First the 15 averages, one per combination of treatments and week, are computed. Then they are plotted against week.

```
proc means data=rats1 nway;
  var logw;
  class treat week;
  output out=out1 mean = meanlogw;
run;

proc gplot data=out1;
  plot meanlogw*week=treat;
run;
```

5.2 Analysis of summary measures

As already mentioned the dependence between observations from the same subject must be taken into account in a valid statistical model.

One way to overcome this "problem" is to analyze a so-called summary measure. The idea is to reduce the curve for each subject to a single value — a value that summarizes the curve. Some common choices are the average response (over time), the area under the curce, the slope of the curve, the increment (final measurement minus first measurement), the response at the final time-point, the size or position of a peak.

The analysis of such a summary measure is quite simple: since there is only one value per subject, we can assume independence and we are thus back to the linear models (unless there are other random effects to take into account, like family).

The hard thing is to choose reasonable summary measures: a good summary measures measures something important! That is, something that charachterizes the individual curves. Not all of the above examples would be reasonable in all cases. You may be inspired from the subject profile plot to see what charachterizes the individual profiles. You may not, however — and this is important, have an eye to the treatment effects when you choose a summary measure! This is called significance hunting: if you look hard enough for significant differences, you will end up finding some. In other words, *you are not allowed choose your choice of summary measure based on inspection of treatment differences.*

Analysis of a good summary measure (or perhaps two reflecting different aspects of the curves, but not too many) is often very useful, at least as a preliminary analysis. It is robust and easy to carry out. On the other hand it does not allow investigation of the development over time, and it it not quite satisfactory to use only a small part of the data (we have thrown good data away). In other words, we need methods that use all the data.

Example 5.1 *(continued)* For the growth of rats data we could for example use the increment of the logarithmic weight from week 1 to week 5 as a summary measure. That is, we have the observation 1.1044 for the first rat, 1.0818 for second rat etc. The analysis of this summary measure is left the reader as an exercise. □

5.3 Analysis with random intercepts (the split-plot model)

A quite common model for repeated measurements is the *random intercepts model*. In the random intercepts model subject is used as a random factor. We can think of this as a random displacement of the intercept for each subject: some subjects generally (at all times) have a high level of the response, others generally have a low level.

Measurements from different subjects are independent, but the random intercepts model introduces correlation between any two measurements from the same subject. Every two measurements from the same subject share the same correlation, namely $\sigma^2_{\texttt{sub}}/(\sigma^2 + \sigma^2_{\texttt{sub}})$ where σ^2 is the measurement error (residual) variance and $\sigma^2_{\texttt{sub}}$ is the between-subjects variance. If there are only a few measurements per subject, or if the measurements are far from each other in time, this may very well be an adequate description of the correlation structure in the model.

Generally we would, however, believe that measurements that are close in time tend to be "more alike" than observations far away in time from each other. In Section 5.4 we discuss a model with such a correlation structure. The random intercepts model is a submodel of this more complex model, so we can actually test whether the more advanced model is significantly better at describing our data.

Note that the random intercepts model is equivalent to the split-plot model with subjects as whole plots, treatments as whole plot factor, and time as subplot factor. Since time is not randomly allocated, the split-plot interpretaion does not make much sense, though, and it is better to think of random intercepts. But the analysis is just the same as for the split-plot model.

Example 5.1 *(continued)* The random intercepts analysis of the growth of rats data is left to the reader as an exercise. □

5.4 A model for repeated measures

As argued above it is often reasonable to believe that the correlation between pairs of measurements on the same subject decreases as the time-lag between the measurements increases. We want to take this into account in the analysis. In other words, we want models that describe the *seriel correlation structure* of our data. We will consider one such model here, which we will refer to as *the Diggle-model*.

Note, however, that there are numerous other models, and it is not straight-forward how to choose between them. `SAS` and `R` computes some information criteria (AIC and BIC) which can be used a guidelines. However, the quality of such model selection criteria is still a matter of debate, and we will not discuss them any further. A closely related issue is model validation: how do we check that the seriel correlation structure of our model is adequate for our data? This is a matter os debate, too. Graphical methods, comparing the empirical correlations and the estimated model correlations as a function of time-lag exist, but are most useful if there are more than just a few observations on each subject. An example is the semi-variogram which is quite easy to make in `R`.

Note, that the random intercepts model, assuming that the correlation between any two meaurements is constant, is a submodel of the Diggle-model. Hence, if we wish, we can test for the model reduction from the Diggle model to the random intercepts model.

Let us consider the growth of rats data again.

Example 5.1 *(continued)* The relevant explanatory variables are treatment, week and rat. We will include treatment and week as well as their interaction as fixed effects. Rat is obviously a random effect.

This suggests the model

$$y_i = \gamma(\texttt{treat}_i, \texttt{week}_i) + a(\texttt{rat}_i) + e_i,\ i = 1, \ldots, 135 \tag{5.1}$$

where $a(j) \sim N(0, \nu^2)$, $e_i \sim N(0, \sigma^2)$, and they are all independent. As usual, ν^2 measures the variation between rats, and σ^2 measures the variation between measurements from the same rat (the within-rat variation).

This model is the random intercepts model with correlation $\nu^2/(\nu^2 + \sigma^2)$ for any two pairs of measurements on the same subject. In order to get the property that the correlation decreases as the the time-lag increases we add another term to the model:

$$y_i = \gamma(\texttt{treat}_i, \texttt{week}_i) + a(\texttt{rat}_i) + b_i + e_i \tag{5.2}$$

where $b_i \sim N(0, \tau^2)$ and

$$\text{corr}(b_{i_1}, b_{i_2}) = \begin{cases} 0 & \texttt{rat}_{i_1} \neq \texttt{rat}_{i_2} \\ \exp\left(-(\texttt{week}_{i_1} - \texttt{week}_{i_2})^2/\phi^2\right), & \texttt{rat}_{i_1} = \texttt{rat}_{i_2} \end{cases}$$

Then the variance of each y_i is a sum of three terms,

$$\text{Var} y_i = \nu^2 + \tau^2 + \sigma^2.$$

Measurements from different rats are independent (uncorrelated) whereas measurements, y_{i_1} and y_{i_2}, from the same rat are correlated with correlation

$$\frac{\nu^2 + \tau^2 \exp\left(-(\texttt{week}_{i_1} - \texttt{week}_{i_2})^2/\phi^2\right)}{\nu^2 + \tau^2 + \sigma^2} \quad (i_1 \neq i_2)$$

Recall that $\texttt{week}_{i_1} - \texttt{week}_{i_2}$ measures the time-lag between measurement i_1 and i_2, so the correlation structure has the property that we were looking for. The parameter ϕ determines how fast the correlation decreases.

In the left part of Figure 5.2 we have shown the residual plot for the corresponding linear model. It looks quite okay. In the right part we have plotted the semi-variogram, illustrating the appropriateness of the correlation structure of the models. The x-axis is week-distance and the y-axis is a function of correlation. The dots are empirical quantities (that is, computed from the data only), whereas the curve illustrates the model estimates of the same quantities. If the curve fits the points nicely, then the model catches the features in the data. It looks all right, but other models might fit just as well, since there are only five measurements for each rat. Anyway, the plots do not make us worried.

That was the definition of the Diggle model. Now, let us turn to model reduction. First, we might try to reduce model (5.2) to the random intercepts model (5.1).

Generally, we recommend to carry out model reduction in the random part of the model by comparing the restricted log-likelihood functions. This is similar to likelihood ratio tests, *except that we use REML estimation rather than ML estimation*. In other words: fit both the Diggle model and the random intercepts model with REML, compute the difference in maximum restricted log-likelihoods, and multiply it by two. We then just need to compare it to its distribution on order to get a p-value. Unfortunately this is not straight-forward as the asymptotic distrubution (for this particular hypothesis) is not the $\chi^2(2)$-distribution as one might believe. An approximate p-value may rather be obtained by simulation.

For the rats data, we are "lucky", though: The maxmum restricted log-likelihoods turn out to be 162.6 for the Diggle model and 127.0 for the random intercepts model, so our test statistic is as large as

$$\textit{REML-LR} = 2 \cdot (162.6 - 127.0) = 71.2.$$

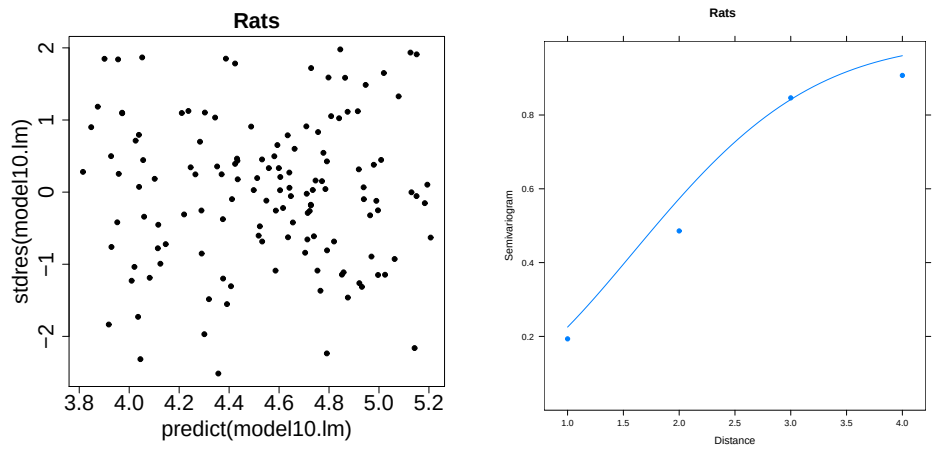


Figure 5.2: Residual plot and semi-variogram for the rats data.

There is no doubt that this value is highly significant, so we cannot accept the random intercepts model. Comparing with the semi-variogram form Figure 5.2 this is not surprising: the random intercepts model corresponds to constant variogram, which clearly does not fit the data.

Then, we try to reduce the systematic part of the model. We first test if the interaction between treatment and time is significant, that is, test if model (5.2) can be reduced to

$$y_i = \alpha(\texttt{treat}_i) + \beta(\texttt{week}_i) + a(\texttt{rat}_i) + b_i + e_i. \tag{5.3}$$

This model reduction turns out not to be possible, $p < 0.0001$. In other words, the treatment-effect is not the same at all five time-points.

Inspired by the subject profile plot in Figure 5.1 (and the random intercepts analysis from an exercise), we then test if the relationship between week (time) and logarithmic weight can be described with second-order polynomials, one per treatment. This corresponds to

$$y_i = \alpha(\texttt{treat}_i) + \delta(\texttt{treat}_i) \cdot \texttt{week}_i + \psi(\texttt{treat}_i) \cdot \texttt{week}_i^2 + a(\texttt{rat}_i) + b_i + e_i. \tag{5.4}$$

This is not possible either ($p < 0.0001$).

Hence, model (5.2) has not been reduced. We report the differences in expected log-weights between control and thyroxin and between control and thiouracil for each week in Table 5.2 together with p-values for each difference being zero. We see that there is no effect of Thyroxin at any time, whereas there is an effect of Thiouracil from week 3 and onwards.

It would perhaps have been better to test the effect of thyroxin in *one test* rather than one test per week. It is left to the reader to consider how to do this!

It is often a good idea to illustrate the results graphically by plotting the expected log-weights (or weigths) in a figure similar to the right plot of Figure 5.1. However, since the interaction between treatment and week is still in the model, the relevant figure would be almost indistinguishable from that plot.

Finally, we give the estimates of the variance parameters in model (5.2):

$$\hat{\sigma}^2 = 0.000334, \qquad \hat{v}^2 = 0.00325, \quad \hat{\tau}^2 = 0.00575, \quad \hat{\phi} = 2.2433$$

Week	Control – Thiouracil	Control – Thyroxin
1	-0.0140 (0.0432) 75%	-0.0320 (0.0476) 50%
2	0.0267 (0.0432) 54%	0.0304 (0.0476) 52%
3	0.1009 (0.0432) 2%	0.0116 (0.0476) 81%
4	0.1820 (0.0432) 0%	-0.0170 (0.0476) 72%
5	0.2570 (0.0432) 0%	-0.0104 (0.0476) 83%

Table 5.2: Estimated expected differences (s.e) for the growth of rate data and p-values for the hypothesis of no difference.

5.4.1 R programs and output

Fit of the Diggle model and model validation

First, the linear model corresponding to (5.2) is fitted and the residual plot is constructed:

```
> model10.lm = lm(logw ~ newtreat*weekfac+newrat)
> plot(predict(model10.lm),stdres(model10.lm))
```

Next, model (5.2) is fitted and the semi-variogram is constructed. The model is fitted quite similarly to the random intercepts model, except that the seriel correlation structure should be specified. This is done by the `corr`-statement. `corGaus` specifies that we use the Diggle model, `form =~ week | newrat` that our subjects are given by `newrat` and that `week` is our time variable. `nugget = T` specifies that there is also measurement errors (the e_i's) in the model.

```
> model10 = lme(logw ~ newtreat*weekfac, random =~ 1|newrat,
                corr = corGaus(form = ~ week | newrat, nugget=T),
                method="REML")

> plot(Variogram(model10, form =~ week), ylim=c(0,1))
```

Test of random intercepts model

The random intercepts model is fitted in the usual way, and `anova` is used for the test. Note that we have fitted both models with REML. This is what we recommend for model reduction in the random part of the model.

```
> model11 = lme(logw ~ newtreat*weekfac, random =~ 1|newrat, method="REML")
> anova(model11,model10)
        Model df       AIC       BIC   logLik   Test  L.Ratio p-value
model11     1 17 -219.9473 -172.5599 126.9736
model10     2 19 -287.1166 -234.1542 162.5583 1 vs 2 71.16929  <.0001
```

For this particular hypothesis it is known that the test statistics is not $\chi^2(2)$-distributed as one should think (not even approximately). Hence, we would have liked to compute an approximate p-value by

bootstrap. Unfortunately `simulate.lme` does not apply to models with seriel correlation structure. This is not a problem for this particular dataset, though, as the REML-LR test statistic is very large so there is no doubt that the random intercepts model is not appropriate for the data.

Reduction of the systematic part of the model

We cannot use REML when testing hypotheses about the systematic part of the model. Hence, we fit all models with ML.

First we fit the full model (5.2) and model (5.3) without interaction, and test for model reduction:

```
> model10a = lme(logw ~ newtreat*weekfac, random =~ 1|newrat,
                 corr = corGaus(form = ~ week | newrat, nugget=T),
                 method="ML")

> model12 = lme(logw ~ newtreat+weekfac, random =~ 1|newrat,
                corr = corGaus(form = ~ week | newrat, nugget=T),
                method="ML")

> anova(model12,model10a)
         Model df       AIC       BIC   logLik   Test  L.Ratio p-value
model12      1 11 -356.7327 -324.7746 189.3663
model10a     2 19 -380.5067 -325.3065 209.2533 1 vs 2 39.77401  <.0001
```

Next, we fit the quadratic model (5.4) and test for model reduction:

```
> model13 = lme(logw ~ newtreat + week + week2 + newtreat:week + newtreat:week2,
                random =~ 1|newrat,
                corr = corGaus(form = ~ week | newrat, nugget=T),
                method="ML")
> anova(model13,model10a)
         Model df       AIC       BIC   logLik   Test  L.Ratio p-value
model13      1 13 -358.3250 -320.5564 192.1625
model10a     2 19 -380.5067 -325.3065 209.2533 1 vs 2 34.18168  <.0001
```

Estimation in the final model

Since model (5.2) could not be reduced we fit it with REML. We choose the parameterization without intercept. `VarCorr` only gives us estimates for ν^2 and $\sigma^2 + \tau^2$. You can have the remaining information from `model10b`. Alternatively, all that information (and a whole lot more) comes out with `summary`.

```
> model10b = lme(logw ~ newtreat:weekfac-1, random =~ 1|newrat,
                 corr = corGaus(form = ~ week | newrat, nugget=T),
                 method="REML")

> VarCorr(model10b)
newrat = pdLogChol(1)
            Variance    StdDev
(Intercept) 0.003251906 0.05702549
Residual    0.006080664 0.07797861

> model10b
```

```
Correlation Structure: Gaussian spatial correlation
 Formula: ~week | newrat
 Parameter estimate(s):
    range    nugget
2.2432381 0.0548862
```

The output requires some explanation: The "intercept variance" is the estimate of ν^2. The "residual variance" is the estimate of $\sigma^2 + \tau^2$. The "range" parameter estimate ϕ , and the "nugget" parameter is $\sigma^2/(\sigma^2 + \tau^2)$. Hence, we have the following equations:

$$\hat{\nu}^2 = 0.00325, \quad \sigma^2 + \tau^2 = 0.00608, \quad \hat{\phi} = 2.24, \quad \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\tau}^2} = 0.0549$$

Solving for the original parameters we find

$$\hat{\nu}^2 = 0.00325, \quad \hat{\sigma}^2 = 0.0549 \cdot 0.00608 = 0.000334, \quad \hat{\tau}^2 = 0.00608 - 0.000334 = 0.00575, \quad \hat{\phi} = 2.24,$$

Finally, we estimate the expected differences from Table 5.2 with `estimable`. We show only the difference between Control and Thiouracil here, the difference Control-Thyroxin differences are computed similarly.

```
> library(gmodels)

> thy1 = c(1,0,-1,0,0,0,0,0,0,0,0,0,0,0,0)
> thy2 = c(0,0,0,1,0,-1,0,0,0,0,0,0,0,0,0)
> thy3 = c(0,0,0,0,0,0,1,0,-1,0,0,0,0,0,0)
> thy4 = c(0,0,0,0,0,0,0,0,0,1,0,-1,0,0,0)
> thy5 = c(0,0,0,0,0,0,0,0,0,0,0,1,0,-1)

> thydifs = rbind(thy1,thy2,thy3,thy4,thy5)
> thyest = estimable(model10b, thydifs)

> thyest
        Estimate Std. Error    t value DF  Pr(>|t|)
thy1 -0.03199429 0.04760758 -0.6720419 94 0.5032055
thy2  0.03043000 0.04760758  0.6391840 94 0.5242578
thy3  0.01164571 0.04760758  0.2446189 94 0.8072847
thy4 -0.01703000 0.04760758 -0.3577162 94 0.7213577
thy5 -0.01035286 0.04760758 -0.2174624 94 0.8283194
```

5.4.2 SAS programs and output

We use `proc mixed` for analysis of the Diggle model. The seriel correlation structure is given in a `repeated`-statement. Note that we do not use `nobound` and `ddfm = sattherth` in case of repeated measurements.

Model validation

A residual plot similar to that of Figure 5.1 is produced as follows, fitting the linear model corresponding to (5.1):

```
proc glm data=rats1;
  class rat week treat;
```

```
  model logw = treat*week rat;
  output out=out2 predict = pred student = stdres;
run;

proc gplot data = out2;
  plot stdres*pred;
run;
```

Fit of the Diggle model and test of random intercepts model

First, let us fit the random intercepts model:

```
proc mixed data=rats1;
  class rat week treat;
  model logw = treat*week;
  random rat;
run;
```

In this analysis we are only going to use the maximum restricted log-likelihood, so we only show a little part of the output here. We see that $-2 \log \text{Res}L = 253.9$.

```
                    Fit Statistics

          -2 Res Log Likelihood          -253.9
          AIC (smaller is better)        -249.9
          AICC (smaller is better)       -249.8
          BIC (smaller is better)        -247.4
```

The Diggle model is fitted with `proc mixed`. The syntax is the usual, except that a `repeated`-statement should be included. `week` tells `SAS` which variable describes time, `subject=rat` tells `SAS` which factor is the subject factor. This is general for all models with a seriel correlation structure. The `type` tells `SAS` which model to use for the seriel structure. `SAS` has a lot of possibilities (around 30!); `sp(gau)` is the Diggle model; again `SAS` needs to know what is the time variable. `local` makes `SAS` include measurement errors as well (the e_i's), and `R` makes `SAS` write the estimated correlation matrix for measurements on the same rat. The latter is sometimes useful. Note that we do not use `nobound` and `ddfm = sattherth` in case of repeated measurements.

```
proc mixed data=rats1;
  class rat week treat;
  model logw = treat*week;
  random rat;
  repeated week / subject=rat type=sp(gau)(week1) local R;
run;
```

The output goes like this:

```
                        The Mixed Procedure

                        Iteration History

    Iteration    Evaluations    -2 Res Log Like       Criterion
            0              1      -177.80311592
```

```
          1              4      -191.72670851        .
          .              .           .
          .              .           .
         13              1      -325.11657152      0.00000000
                  Convergence criteria met.


       Estimated R Matrix for rat 1

Row        Col1        Col2        Col3        Col4        Col5

  1    0.006081    0.004711    0.002596    0.000961    0.000239
  2    0.004711    0.006081    0.004711    0.002596    0.000961
  3    0.002596    0.004711    0.006081    0.004711    0.002596
  4    0.000961    0.002596    0.004711    0.006081    0.004711
  5    0.000239    0.000961    0.002596    0.004711    0.006081


        Covariance Parameter Estimates

       Cov Parm     Subject    Estimate
       rat                     0.003252
       Variance     rat        0.005747
       SP(GAU)      rat          2.2433
       Residual                0.000334


                Fit Statistics

     -2 Res Log Likelihood          -325.1
     AIC (smaller is better)        -317.1
     AICC (smaller is better)       -316.8
     BIC (smaller is better)        -311.9


          Type 3 Tests of Fixed Effects

                    Num     Den
     Effect          DF      DF    F Value    Pr > F
     week*treat      14      96     198.33    <.0001
```

We see that $-2 \log \text{Res}L = 325.1$. Compared to the random intercepts model we find the test statistic

$$REML\text{-}LR = 325.1 - 253.9 = 71.2.$$

For this particular hypothesis it is known that the test statistics is not $\chi^2(2)$-distributed as one should think (not even approximately). This is not a problem for this particular dataset, though, as the test statistic is very large so it is clear that the random intercepts model is not appropriate for the data.

Reduction of systematic part of the model

We then move on to the systematic part of the model. First, the test for interaction between week and treatment. It has already been carried out by the above fit of the Diggle model. The above output shows a F-value of 360.5, and a p-value less than 0.0001.

Next, the test for model (5.4) against (5.2). One way to carry out this test is to fit (5.4) again, but this time with the "original" interaction term `treat` × `week` (with `week` as factor) as well as the interaction terms `treat` × `week1` (`week1` a covariate) and `treat` × `week2` (`week2` the squared values, also a covariate) in the `model` statement.

It turns out that `SAS` has numerical problems: we get a warning that `SAS` "stopped beacuse of infinite likelihood". We fix the problem by giving `SAS` some starting values, namely those obtained from the above output.

```
proc mixed data=rats1;
  class rat week treat;
  model logw = treat week1 week2 treat*week1 treat*week2 treat*week;
  random rat;
  repeated week / subject=rat type=sp(gau)(week1) local R;
  parms (0.003) (0.005) (2.24) (0.0003) / noprofile;
run;
```

From the output we see that the interaction term `treat` × `week` is still significant:

```
                    The Mixed Procedure

                    Iteration History

   Iteration    Evaluations    -2 Res Log Like       Criterion
           1              2      -320.95766697      0.00000008
           2              1      -320.95768845      0.00000000
                  Convergence criteria met.


                Estimated R Matrix for rat 1

Row        Col1        Col2        Col3        Col4        Col5
  1    0.006081    0.004711    0.002595    0.000961    0.000239
  2    0.004711    0.006081    0.004711    0.002595    0.000961
  3    0.002595    0.004711    0.006081    0.004711    0.002595
  4    0.000961    0.002595    0.004711    0.006081    0.004711
  5    0.000239    0.000961    0.002595    0.004711    0.006081


            Covariance Parameter Estimates

            Cov Parm     Subject    Estimate

            rat                     0.003252
            Variance     rat        0.005747
            SP(GAU)      rat          2.2432
            Residual                0.000334


                      Fit Statistics

           -2 Res Log Likelihood          -321.0
           AIC (smaller is better)        -313.0
           AICC (smaller is better)       -312.6
           BIC (smaller is better)        -307.8
```

```
                Type 3 Tests of Fixed Effects

                          Num     Den
          Effect           DF      DF    F Value    Pr > F
          treat             2      96       0.58    0.5633
          week1             0       .        .       .
          week2             0       .        .       .
          week1*treat       0       .        .       .
          week2*treat       0       .        .       .
          week*treat        6      96       6.48    <.0001
```

Estimation in the final model

Hence, model (5.2) is the final model. We fit it again, this time asking for comparison of the adjusted means, since this also gives us the estimated expected differences from Table 5.2.

```
proc mixed data=rats1;
  class rat week treat;
  model logw = treat*week / solution;
  random rat;
  repeated week / subject=rat type=sp(gau)(week1) local R;
  lsmeans treat*week / pdiff;
run;
```

The output is enormous beacuse of all the comparisons. We only show little part of it. From the output we also take out the variance estimates.

```
                      Iteration History

   Iteration    Evaluations    -2 Res Log Like       Criterion

           0              1      -177.80311592
           1              4      -191.72670851       .
           .              .          .               .
           .              .          .               .
          13              1      -325.11657152      0.00000000
                  Convergence criteria met.


                 Estimated R Matrix for rat 1

 Row        Col1        Col2        Col3        Col4        Col5
   1    0.006081    0.004711    0.002596    0.000961    0.000239
   2    0.004711    0.006081    0.004711    0.002596    0.000961
   3    0.002596    0.004711    0.006081    0.004711    0.002596
   4    0.000961    0.002596    0.004711    0.006081    0.004711
   5    0.000239    0.000961    0.002596    0.004711    0.006081


               Covariance Parameter Estimates

              Cov Parm     Subject    Estimate
              rat                     0.003252
```

Variance	rat	0.005747
SP(GAU)	rat	2.2433
Residual		0.000334

Fit Statistics

-2 Res Log Likelihood	-325.1
AIC (smaller is better)	-317.1
AICC (smaller is better)	-316.8
BIC (smaller is better)	-311.9

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
week*treat	14	96	198.33	<.0001

Least Squares Means

Effect	treat	week	Estimate	Standard Error	DF	t Value	Pr > |t|
week*treat	Control	1	3.9844	0.03055	96	130.43	<.0001
week*treat	Thiourac	1	3.9985	0.03055	96	130.89	<.0001
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
week*treat	Thyroxin	5	5.0852	0.03651	96	139.27	<.0001

Differences of Least Squares Means

Effect	treat	week	_treat	_week	Estimate	Standard Error	DF	t Value	Pr > |t|
week*treat	Control	1	Thiourac	1	-0.01403	0.04320	96	-0.32	0.7461
week*treat	Control	1	Thyroxin	1	-0.03199	0.04761	96	-0.67	0.5032
week*treat	Control	1	Control	2	-0.3719	0.01655	96	-22.47	<.0001
week*treat	Control	1	Thiourac	2	-0.3452	0.04320	96	-7.99	<.0001
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
week*treat	Control	2	Thiourac	2	0.02669	0.04320	96	0.62	0.5382
week*treat	Control	2	Thyroxin	2	0.03043	0.04761	96	0.64	0.5242
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
week*treat	Control	5	Thiourac	5	0.2570	0.04320	96	5.95	<.0001
week*treat	Control	5	Thyroxin	5	-0.01035	0.04761	96	-0.22	0.8283
week*treat	Thiourac	5	Thyroxin	5	-0.2673	0.04761	96	-5.61	<.0001

Recall that is always problematic to make such an awful lot of pairwise comparisons. One ought to adjust the p-values, taking into account that we are making numerous tests. One such adjustment is the Tukey adjustment. Tukey adjusted p-values are obtained in SAS by the option `adjust = Tukey` in the `lsmeans`-statement. However, we use only a few of the p-values so the problem is not quite as big as it may seem at first sight, and in fact it would not change our conclusions in the present case.

Chapter 6

Models for binary response data

The main purpose of the present chapter is to introduce the class of logistic regression models. We leave the framework of the previous chapters where a common feature of all models has been the assumption that the response variable is continuous and normally distributed. The logistic regression models have been developped in order to model data from experiments where the output is binary (dead/alive, no/yes, 0/1, etc.).

The distribution of a binary random variable, Y, is entirely described by the probability of success

$$P(Y = 1) = p.$$

This distribution is often referred to as the *binomial* distribution with parameters $(1, p)$ or the *Bernouilli* distribution with parameter p. The basic assumption throughout this chapter is that our dataset consists of independent binary random variables $Y_1, Y_2, \ldots, Y_N$. If we denote by p_i the probability $P(Y_i = 1)$ of success for the i'th variable then finding a suitable statistical model for the data amounts to describing the probabilities p_i. As each response Y_i is typically associated with a number of explanatory variables (sex, treatment group, age, dose, etc.) the goal of the statistical analysis is to understand how those affect the probability p_i of success.

6.1 Tables of counts

Example 6.1 Exploring the effect of watering and light conditions on germination.

Let us consider a dataset from a growth experiment with 72 pots. 12 pots are associated with each combination of the factors `water` with levels `little, moderate, much` and `light` with levels `8 hours` and `12 hours`. After a fixed amount of time we observe the response, `germination`, that takes one of the values `yes` or `no`. The full dataset contains 72 datalines (one for each put) summarizing the joint configuration of the three variables `water, light,` and `germination`. However, it is usually more convenient to report the results of such an experiment in a tabular as given below.

	little		moderate		much	
	8 hours	12 hours	8 hours	12 hours	8 hours	12 hours
yes	3	5	6	8	4	0
no	9	7	6	4	8	12

Here we have counted the number of observations for each combination of `water`, `light`, and `germination`. A table of this form is referred to as a *contingency table*.

If we want to test if `germination` depends on the level of watering one may consider the tabular obtained by aggregating the full tabular above over the levels of the factor `light`.

	little	moderate	much	total
yes	$Y_{11} = 8$	$Y_{12} = 14$	$Y_{13} = 4$	$Y_{1.} = 26$
no	$Y_{21} = 16$	$Y_{22} = 10$	$Y_{23} = 20$	$Y_{2.} = 46$
total	$Y_{.1} = 24$	$Y_{.2} = 24$	$Y_{.3} = 24$	$n = 72$

The probability of observing germination in a pot that received little watering is estimated to

$$\frac{8}{24} = 0.33,$$

whereas the estimates for watering levels `moderate` and `much` become 0.58 and 0.17. To determine whether the estimates are significantly different we need a statistical model.

Let us denote by Y_{11}, Y_{12}, and Y_{13} the number of pots with germination for the three different watering levels. Since it is reasonable to believe that the outcome in different pots do not influence each other we assume that the variables are independent and binomially distributed

$$Y_{1j} \sim b(n_j, p_j), \quad j = 1, 2, 3.$$

Here we have that $n_1, n_2, n_3 = 24$ as there are 24 pots receiving either dose of water. The hypothesis of homogeneity of the occurence of germination in the groups given by the watering level may be formulated as $H_0 : p_1 = p_2 = p_3$.

A test of H_0 can be based on the Pearson χ^2-statistic informally defined by

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}},$$

measurering the difference between the *expected* counts in the cells under H_0 and the *observed* counts summed over all cells in the tabular above. The Pearson test statistic for H_0 turns out to be 9.1505 which is far beyond the 95%-quantile of a $\chi^2(2)-$distribution. We may also use the likelihood ratio test statistic given by

$$G^2 = 2 \sum_{i=1}^{2} \sum_{j=1}^{3} Y_{ij} \log \left(\frac{Y_{ij} n}{Y_{.j} Y_{i.}} \right),$$

where $n = 72$ is the total number of pots. One finds that

$$G^2 = 9.4030 \sim \chi^2(2)$$

and since the 95%-quantile of a $\chi^2(2)-$distribution is 5.991 we reject the hypothesis H_0. We conclude that the watering level does influence the germination in the pots.

But how do we quantify the effect of using `moderate` amount of water instead of `little` water? As an answer to this question it is customary to report the *odds ratio* defined by

$$OR = \frac{p_2}{1 - p_2} / \frac{p_1}{1 - p_1}.$$

If we plug in the estimates $\hat{p}_1$ and $\hat{p}_2$ we obtain the following estimate

$$\widehat{OR} = \frac{0.5833}{1-0.5833} / \frac{0.3333}{1-0.3333} = 2.8,$$

hence the odds for germination increases by a factor 2.8 if we increase the watering level from `little` to `moderate`. □

We may repeat the analysis of the growth experiment to examine the effect of `light` but clearly it would be more desirable to consider a model that takes both of the variables `water` and `light` into account. One possible way to do so would be to consider the combination of `water` and `light` as a factor on six levels and test for equal germination probability between the groups. This results in a Pearson test statistic of $G^2 = 13.485$ yielding af p-value of $p = 0.019$ in a $\chi^2(5)$-distribution. However, the statistical analysis should reflect the type of questions we would like to answer. In this experiment we might like to answer questions of the form: "Is the effect of changing the watering level from `little` to `moderate` the same across all levels of `light`?" If the experiment contains more than two explanatory factors the complexity of the questions we might like to answer increases even further and clearly a more flexible class of statistical models is needed.

Section 6.2 and 7.3 present classes of models that allow us to analyze tables of counts grouped according to several factors. One important feature of the models is that it is easy to quantify the effect of individual variables in the model in terms of odds ratios.

6.1.1 R-programs and output

Example 6.1 *(continued)*

Read in data

When reading data into R remember to use `attach` so that we may refer to variables of `data` using the variable names `light, water, germination,` and `count`.

```
> data=read.table(file="Growth.txt",header=T)
> attach(data)
> data
   light water germination count
1      8   low         yes     3
2      8   low          no     9
3      8   mod         yes     6
4      8   mod          no     6
5      8  high         yes     4
6      8  high          no     8
7     12   low         yes     5
8     12   low          no     7
9     12   mod         yes     8
10    12   mod          no     4
11    12  high         yes     0
12    12  high          no    12
```

To produce a two-by-two table of the data cross-classified after `germination` and `water` write

```
> observed=xtabs(count~water+germination)
```

```
> observed
      germination
water  no yes
  high 20   4
  low  16   8
  mod  10  14
```

Test for effect of watering level

The program line

```
> chisq.test(observed)

        Pearson's Chi-squared test

data:  data.water
X-squared = 9.1505, df = 2, p-value = 0.01030
```

calculates the Pearson χ^2 test statistic and the corresponding p-value for a test of no effect of `water`. The likelihood ratio test may be constructed using the code

```
> expected=outer(rowSums(observed),colSums(observed))/sum(observed)
>
> lr=-2*sum(observed*log(expected/observed))
> lr
[1] 9.402998
>
> pchisq(lr,2,lower=FALSE)
[1] 0.009081654
```

Here `rowSums` and `colSums` are the vector of row and column counts, and the matrix `expected` contains the expected cell frequencies of the table of count under the hypothesis, H_0, of equal germination frequency for each watering level. The following program line computes minus two times the log likelihood ratio for H_0 against the full model. Note that the likelihood ratio, `lr`, is obtained automatically if the test is performed using `glm` as described in section 6.2.1. Finally, the p-value for the likelihood ratio test is calculated in a χ^2-distribution with 2 degrees of freedom.

6.1.2 SAS-programs and output

Example 6.1 *(continued)*

Read in data

The SAS program below reads and prints the data for the growth experiment.

```
data growth;
input light $ water $ germination $ count;
cards;
8       low     yes             3
```

```
8       low     no              9
8       mod     yes             6
8       mod     no              6
8       high    yes             4
8       high    no              8
12      low     yes             5
12      low     no              7
12      mod     yes             8
12      mod     no              4
12      high    yes             0
12      high    no              12
;
run;

proc print;
run;
```

Test for effect of watering level

`proc freq` produces tables of counts and tests for the effect of watering.

```
proc freq;
weight count;
tables water*germination/chisq;
run;
```

Note that the `chisq` option in the `tables` statement implies that the Pearson χ^2 test as well as the likelihood ratio test statistic are printed as part of the output.

```
  water     germination

Frequency
Percent
Row Pct
Col Pct  no      yes       Total

high          20       4      24
           27.78    5.56   33.33
           83.33   16.67
           43.48   15.38

low           16       8      24
           22.22   11.11   33.33
           66.67   33.33
           34.78   30.77

mod           10      14      24
           13.89   19.44   33.33
```

	41.67	58.33	
	21.74	53.85	
Total	46	26	72
	63.89	36.11	100.00

Statistics for Table of water by germination

Statistic	DF	Value	Prob
Chi-Square	2	9.1505	0.0103
Likelihood Ratio Chi-Square	2	9.4030	0.0091
Mantel-Haenszel Chi-Square	1	8.9047	0.0028
Phi Coefficient		0.3565	
Contingency Coefficient		0.3358	
Cramer's V		0.3565	

6.2 Logistic regression models

In the present section we introduce a class of statistical models suitable for datasets with a binary response variable (no/yes, 0/1, etc.). As an example let us for a while return to the growth experiment discussed in section 6.1 and let us recode the response so that 1 means germination and 0 means no germination. We want to build a joint statistical model that allows us to simulatneously quantify the effect of `water` and `light`. Before proceeding we reorganize the data into a tabular as given below.

`water`	`light`	Group size	No. of positives
little	8 hours	12	$Y_1 = 3$
little	12 hours	12	$Y_2 = 5$
moderate	8 hours	12	$Y_3 = 6$
moderate	12 hours	12	$Y_4 = 8$
much	8 hours	12	$Y_5 = 4$
much	12 hours	12	$Y_6 = 0$

The entries of last column are labelled $Y_1, \ldots, Y_6$ and we replicate the notation for factors from the previous chapters, so that $\texttt{water}_3$ denotes the level of the watering factor for Y_3 which here takes the value `moderate`. The number of observations in each group is denoted $n_1, \ldots, n_6$, which are all $= 12$ in this case. We assume that each variable Y_i follows a binomial distribution

$$Y_i \sim b(n_i, p_i), \quad i = 1, \ldots, 6,$$

and that they are mutually independent. Under this model the estimate for p_i is simply the frequency of positive responses in group i so that for instance

$$\hat{p}_3 = \frac{6}{12} = 0.50.$$

We can think of this as a model with interaction between the factors `water` and `light`. In the logistic regression model for no interaction (but possibly main effects) of `water` and `light` we assume that

$$\log\left(\frac{p_i}{1-p_i}\right) \quad = \quad \alpha(\texttt{water}_i) + \beta(\texttt{light}_i), \quad i = 1, \ldots, 6. \tag{6.1}$$

It may seem ackward that we model the logarithm of the odds $\eta_i = \log(p_i/(1-p_i))$ on the left hand side instead of just p_i. The reason is that the p_i's have to be numbers between zero and one. Rearranging the terms of (6.1) we find that

$$p_i = \frac{\exp\left((\alpha(\texttt{water}_i) + \beta(\texttt{light}_i)\right)}{1 + \exp\left((\alpha(\texttt{water}_i) + \beta(\texttt{light}_i)\right)},$$

which belongs to the interval $(0, 1)$ no matter the values of $\alpha(\texttt{water}_i)$ and $\beta(\texttt{light}_i)$. This would not have been the case had we written $p_i = \alpha(\texttt{water}_i) + \beta(\texttt{light}_i)$ which might at first sight seem more natural. In terms of log-odds the model with interaction we also be written as

$$\log\left(\frac{p_i}{1-p_i}\right) \quad = \quad \gamma(\texttt{water}_i, \texttt{light}_i), \quad i = 1, \ldots, 6 \tag{6.2}$$

Note the difference between the two models: in (6.2) the 6 p_i's are allowed to take any values between 0 and 1 whereas in (6.1) the p_i's are forced to satisfy a certain relationship.

Example 6.1 *(continued)* Let us try to test the hypothesis (6.1) of no interaction between `water` and `light` in growth experiment. The R- and SAS-programs in section 6.2.1 and 6.2.2 explain in details how to read in data, fit a logistic model and perform the test for the data of the growth experiment introduced in example 6.1.

A test of the model with no interaction against the full model introduced in section 6.1 yields a test statistic of $G^2 = 7.7960 \sim \chi^2(1)$ correpsonding to a p-value of approximately 2%. Though this makes us reject the model (6.1) with additive effect of `water` and `light` lets us discuss for a while how to interpret the parameter estimates of this model. From the output of section 6.2.1 and 6.2.2 we find

Parameter	Estimate	95%-conf. int.
$\alpha(\texttt{little})$	-0.6931	[-1.7416,0.2727]
$\alpha(\texttt{moderate})$	0.3365	[-0.6188,1.3306]
$\alpha(\texttt{much})$	-1.609	[-2.9489,-0.5158]
$\beta(\texttt{12hours}) - \beta(\texttt{8hours})$	0.000	[-1.0379,1.0379].

Note that the model (6.1) is overparameterized implying that one parameter, $\beta(\texttt{8hours})$, is used as reference. From the tabular we may read off the estimate for the group given by `water=moderate` and `light=12 hours` which turns out to be

$$\log\left(\frac{\hat{p}_4}{1-\hat{p}_4}\right) \quad = \quad 0.3365 + 0 = 0.3365.$$

This implies that $\hat{p}_4 = 0.583$ which is exactly the same as for the model of section 6.1 where we decided to ignore `light`. We stress that as we reject the hypothesis of no interaction we do not trust the estimate given above, but the example shows that if we insist on using a wrong model the estimate $\beta(\texttt{12hours}) - \beta(\texttt{8hours})$ for the effect of `light` is zero.

It is instructive to consider how there can be an effect of the combination of `water` and `light` when there is no *marginal* effect of `light` under the additive model. Inspecting the table of counts reveals what is going on. For `little` and `moderate` watering level there appears to be a positive effect of increasing the amout of `light` from `8 hours` to `12 hours`. In contrast, for pots exposed to `much` water there seems to be a negative effect of increasing `light`. The fact that changing the light conditions does not have the same effect for different levels of `water` is exactly what must be described by an interaction term `water`×`light`. □

So far we have considered a logistic regression model with two factors but it could of course be extended to include also continuous variables. The example below discusses how to do a logistic regression analysis wiht both a factor and a continuous explanatory variable. However, we remind that the use of logistic regression models still requires a binary response.

Example 6.2 Effect of insecticide on moth.

Different groups of moths consisting of 20 males and 20 females have been exposed to the insecticide *trans-cypermethrin*. After three days the number of moths that have died or collapsed has been observed. The data is summarized in the tabular below.

	Dose (μg)					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

If we introduce the variables

$$Y_i = \begin{cases} 1 & \text{if } i\text{-th moth has died or collapsed} \\ 0 & \text{otherwise,} \end{cases}$$

then as before we assume that

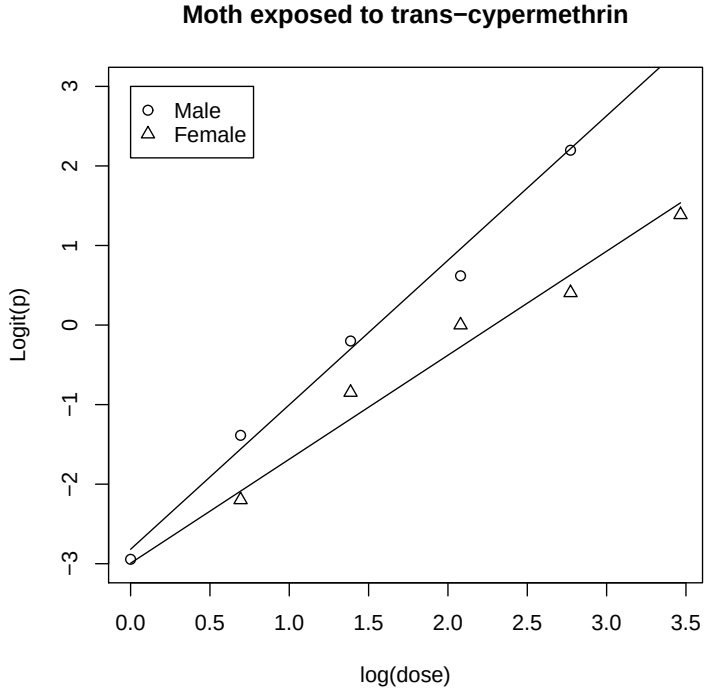
$$Y_i \sim b(1, p_i),$$

and that they are independent. Here p_i denotes the probability that the $i-$th moth collapses or dies within three days. We want to use a statistical model that allows the probability, p_i, to depend on the factor `sex` *and* the covariate `dose`. One such model is given by

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha(\texttt{sex}_i) + \beta(\texttt{sex}_i)\log(\texttt{dose}_i), \tag{6.3}$$

where $\texttt{sex}_i$ and $\texttt{dose}_i$ are the sex and the dose corresponding to the $i-$th moth. The model expresses that for each `sex` the logit of p_i depends linearly on the logarithm of the dose of insecticide.

To understand the model it is a good idea to make the following plot.



For each combination of `sex` and `dose` we count the number of moths and calculates the estimated probability of dying or collapsing within that group by dividing by the group size of 20. For `dose`=2 and `sex`=Female we get the frequency $\hat{p} = 2/20 = 0.1$. We then plot

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$$

against the logarithm of the dose, i.e. $\log(2)$, using a different plotting symbol for each level of `sex`. The logistic model (6.3) expresses that the points corresponding to same sex must fall around a straight line. The lines on the plot are based on (6.3). From the output of section 6.2.1 and 6.2.2 the slope of the lines corresponding to male and female moths are found to be

$$\begin{aligned} \hat{\beta}(\texttt{Male}) &= 1.8163 \\ \hat{\beta}(\texttt{Female}) &= 1.8163 - 0.5091 = 1.3072 \end{aligned}$$

and the intersections with the y-axis are

$$\begin{aligned}\hat{\alpha}(\texttt{Male}) &= -2.8186\\ \hat{\alpha}(\texttt{Male}) &= -2.8186 - 0.1750 = -2.9936.\end{aligned}$$

A test for reduction of the full model with no restrictions on the 12 probabilities to the logistic model (6.3)is accepted as we get that

$$G^2 = 4.9937 \sim \chi^2(8)$$

corresponding to a p-value of 76%.

We continue to test the hypothesis $H_0 : \alpha(\text{Male}) = \alpha(\text{Female})$ about the intersection of the lines with the y-axis being equal for both levels of `sex`. The model may formally be written as

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta(\texttt{sex}_i)\log(\texttt{dose}_i) \tag{6.4}$$

and the test statistic for the hypothesis is

$$G^2 = 0.0505 \sim \chi^2(1)$$

yielding a p-value of 82%. We therefore accept the model given by (6.4) and further analysis shows that it can not be reduced anymore. In particular the slope is not the same for the two sexes ($G^2 = 11.940$, $p = 0.0005$). Parameter estimates as well as confidence intervals for the model (6.4) become

$$\begin{array}{rcll}\hat{\alpha} & = & \text{-2.9073} & \text{[-3.7247,-2.1917],}\\ \hat{\beta}(\texttt{Male}) & = & 1.8601 & \text{[1.4220,2.3708],}\\ \hat{\beta}(\texttt{Female}) - \hat{\beta}(\texttt{Male}) & = & \text{-0.5872} & \text{[-0.9572,-02474].}\end{array}$$

There may be minor differences in the 95%-confidence intervals obtained by R and SAS as they use slightly different approximations. Those reported above have been taken from the R-output. But how do we use the estimates above to quantify the effect of the insecticide? Suppose for instance that we want to estimate the effect of doubling the dosis from `dose`$= 1$ to `dose`$= 2$. The corresponding increase of the log-odds for a male moth is estimated to

$$(\hat{\alpha} + \hat{\beta}(\texttt{Male})\log(2)) - (\hat{\alpha} + \hat{\beta}(\texttt{Male})\log(1)) = \hat{\beta}(\texttt{Male})\log(2/1) = 1.29$$

with a confidence interval given by $\log(2) * [1.4220, 2.3708] = [0.99, 1.64]$. (What is the increase of log-odds for `Male` moths when `dose` increases from 2 to 4?)

It may also be of interest to predict the outcome for a particular value of the covariate `dose`. In the moth example the log-odds that a `female` moth receiving a `dose`$= 10$ dies or collapses within three days is estimated to

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\alpha} + \hat{\beta}(\texttt{Female})\log(10) = -2.9073 + (1.8601 + (-0.5872))\log(10) = 0.0236$$

implying that

$$\hat{p} = \frac{\exp(0.0236)}{1+\exp(0.0236)} = 0.506.$$

A 95 %-confidence interval for the estimated probability becomes $[0.394, 0.618]$ which may be calculated using either SAS or R. □

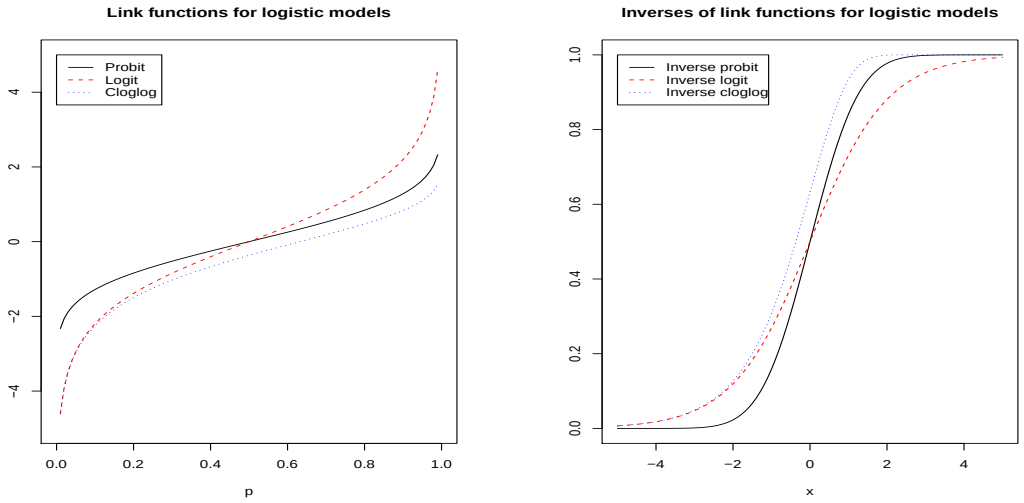
As discussed in the introduction of the chapter a statistical model for binary response data is equivalent to a description of how the probability of success depends on explanatory variables. For a logistic regression model we have so far used the *logit* link function

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$

and the idea has been to model $logit(p_i)$ as a linear function of factors and covariates associated to the $i-$th observation. It is straight forward to include more than two factors/covariates and as in the example 6.2 it may be preferable to transform some of the covariates. For an experiment with factors `A` and `B` and a covariate `C` another example could be

$$\text{logit}(p_i) = \alpha(\mathtt{A}_i, \mathtt{B}_i) + \beta(\mathtt{A}_i)\sqrt{\mathtt{C}_i}.$$

Yet another modification that is often encountered in the literature is to use other link functions than 'logit' on the left hand side of the model equation.



Two such examples are the `probit` (Ψ^{-1}) function (inverse cumulative density function for a $N(0,1)$-distribution and the `cloglog` function as displayed in the figure above. Both R and SAS allow for different choices of 'links', c.f. section 6.2.1 and 6.2.2.

6.2.1 R-programs and output

Example 6.1 *(continued)*

For a description of the dataset we refer to the program of section 6.1.1.

Fit logistic regression model

Before analysing the data we change the status of `light` to a factor. Logistic regression models are fitted using the `glm` function. Note that when the response variable (`germination`) is not coded as 0/

1 R automatically uses the value in the first dataline as response group '1'. Thus, for this example `germination = yes` is interpreted as the outcome '1', c.f. section 6.1.1.

When using `glm` one must remember to specify the family option (here `binomial`) to indicate which distribution is to be used for the response. When using the `binomial` family R uses the `logit` link function as the default. To use other link functions write `binomial(link=''cloglog'')` or `binomial(link=''probit'')` instead of just `binomial`.

```
> light=factor(light)
> glm0=glm(germination~water*light,weights=count,binomial)
> glm.add=glm(germination~water+light-1,weights=count,binomial)
```

`glm0` is the model allowing for individual probabilities of positive response for each combination of `water` and `light`. The `glm.add` model is the logistic regression model with main effects of `water` and `light`. The '`water+light-1`' in the call of `glm` implies that the first level of the factor `light` will be used as reference for the parameter estimates.

To obtain estimates for the parameters and 95%-confidence intervals under the model `glm.add` write

```
> summary(glm.add)

Call:
glm(formula = germination ~ water + light - 1, family = binomial,
    weights = count)

Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
waterhigh -1.609e+00  6.074e-01   -2.650  0.00806 **
waterlow  -6.931e-01  5.064e-01   -1.369  0.17109
watermod   3.365e-01  4.903e-01    0.686  0.49255
light12    2.967e-14  5.252e-01 5.65e-14  1.00000

> confint(glm.add,level=0.95)

               2.5 %     97.5 %
waterhigh -2.9488934 -0.5158302
waterlow  -1.7415677  0.2727403
watermod  -0.6188004  1.3306372
light12   -1.0378636  1.0378636
```

Note that the estimate for the difference between the groups given by `light` is zero.

Testing the logistic regression model.

To test the logistic regression model against the full model use the program line

```
> anova(glm.add,glm0,test="Chisq")
```

that generates the output

```
Analysis of Deviance Table
```

```
Model 1: germination ~ water + light - 1
Model 2: germination ~ water * light
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         7     84.781
2         5     76.985  2    7.796     0.020
```

From the tabular we read off the test statistic (7.7960) and the p-value (0.0203) for the test.

Example 6.2 *(continued)*

Read in data

```
> data=matrix(nrow=12,ncol=2)
> data[,1]=c(1,4,9,13,18,20,0,2,6,10,12,16)
> data[,2]=20-data[,1]

> sex=gl(2,6,labels=c("Male","Female"))
> dose=c(1,2,4,8,16,32,1,2,4,8,16,32)
> logdose=log(dose)
```

Data are read in and stored in object `data`. The first column contains the number of moths that collapsed or died within 3 days and the second column is the number of moths still alive. When using `glm` to fit a logistic regression model one is allowed to enter the response data as a two column matrix with the columns denoting positive/negative responses.

Variables `sex` and `dose` are created. The variable `logdose` is the covariate obtained by taking the logarithm of the dose. Note that R considers `dose` as a covariate and that we have to specify explicitly when it has to be regarded as a factor (grouping variable).

Fit logistic regression models.

```
> glm0=glm(data~sex*factor(dose),family=binomial)
> glm.linear=glm(data~sex*logdose,family=binomial)
> glm.parallel=glm(data~sex+logdose,family=binomial)
> glm.intersect=glm(data~sex*logdose-sex,family=binomial)
> glm.dose=glm(data~logdose,family=binomial)
```

The full model (`glm0`) assigns individual frequencies of positive response to all 12 combinations of `sex` and `dose`. `glm.linear` is the logistic model where $logit(p_i)$ is modelled as a linear function of `logdose` and different lines are fitted for `Male` and `Female` moths. The plot in example 6.2 may be generated from the code

```
> freq=fitted(glm0)
> logitfreq=log(freq/(1-freq))
> plot(logdose,logitfreq,pch=as.numeric(sex),ylim=c(-3,3),
   main="Moth exposed to trans-cypermethrin",ylab="Logit(p)",xlab="log(dose)")
> ylab="Logit(p)",xlab="log(dose)")
> legend(0,3,c("Male","Female"),pch=1:2)
> dev.print(device=pdf,file="ExMoth.ps")
```

The vector `logitfreq` contains the values of

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$$

for the model `glm0`. The last program line tells R to save the currect plot in a pdf-file. Lines corresponding to the fit of the `glm.linear` object may be added to the plot using the following code.

```
freql=fitted(glm.linear)
lines(logdose[1:6],log(freql/(1-freql))[1:6])
lines(logdose[7:12],log(freql/(1-freql))[7:12])
```

Testing and obtaining relevant estimates.

To test for the logistic regression model and obtain estimates for the parameters use the program below.

```
> summary(glm.linear)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.8186     0.5480  -5.143 2.70e-07 ***
sexFemale          -0.1750     0.7783  -0.225    0.822
logdose             1.8163     0.3059   5.937 2.91e-09 ***
sexFemale:logdose  -0.5091     0.3895  -1.307    0.191

> anova(glm.linear,glm0,test="Chisq")

Analysis of Deviance Table

Model 1: data ~ sex * logdose
Model 2: data ~ sex * factor(dose)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         8     4.9937
2         0  5.239e-10  8   4.9937    0.7582
```

Succesive tests for same intersection of the regression line for `Male` and `Female` and for no effect of `sex` at all is given by

```
> anova(glm.intersect,glm.linear,test="Chisq")

Analysis of Deviance Table

Model 1: data ~ sex * logdose - sex
Model 2: data ~ sex * logdose
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         9     5.0443
2         8     4.9937  1   0.0505    0.8221

> anova(glm.dose,glm.intersect,test="Chisq")

Analysis of Deviance Table
```

```
Model 1: data ~ logdose
Model 2: data ~ sex * logdose - sex
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        10    16.9840
2         9     5.0443  1  11.9398    0.0005
```

Our final model is `glm.intersect` and parameter estimates are given by

```
> summary(glm.intersect)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.9073     0.3893  -7.468 8.12e-14 ***
logdose            1.8601     0.2408   7.723 1.13e-14 ***
sexFemale:logdose -0.5872     0.1799  -3.264   0.0011 **
```

A statement of the form

```
> confint(glm.intersect,level=0.95)
```

generates confidence intervals for the parameters of the model given by the object `glm.intersect`.

To predict the probability that a `female` moth receiving `dose`$= 10$ survives more than three days one has to understand the parameterization used by R. The final model `glm.intersect` contains three parameters α, $\beta($`male`$)$, and $\beta($`female`$)$ but the estimates from the output of R

```
> coef(glm.intersect)
      (Intercept)           logdose sexFemale:logdose
       -2.9072994         1.8601238        -0.5871858
```

corresponds to α, $\beta($`male`$)$, and $\beta($`female`$) - \beta($`male`$)$. The estimate we are looking for may therefore be found by evaluating

$$1 \cdot (-2.9073) + \log(10) \cdot 1.8601 + \log(10) \cdot (-0.5872)$$

and then transforming this quantity from the log-odds scale to a probability. A convenient way to obtain the estimate and a confidence interval is to run the following program lines

```
> dose10=c(1,log(10),log(10))
> param=rbind(dose10)
> library(gmodels)
> est=estimable(glm.intersect,param,conf.int=0.95)
> est
         Estimate Std. Error  X^2 value DF Pr(>|X^2|)   Lower.CI  Upper.CI
dose10 0.02374847  0.2327583 0.01041024  1  0.9187324 -0.4380374 0.4855344
```

Here `dose10` is the vector of coefficents used to obtain the relevant estimate from the original parameter estimates in the model object `glm.intersect`. The function `estimable` from the package `gmodels` calculates the estimate and 95%-confidence interval corresponding to the vector `dose10`. All results are given on the log-odds scale. To translate the estimates on the log-odds scale to probabilities finally use the program below that defines the inverse logit function, `invlogit`, and evaluates the desired probability and limits for a 95 % confidence interval.

```
> invlogit=function(u){exp(u)/(1+exp(u))}
> pred=invlogit(est$Estimate)
> predlow=invlogit(est$Lower.CI)
> predup=invlogit(est$Upper.CI)
> print(c(predlow,pred,predup))
[1] 0.3922087 0.5059368 0.6190539
```

6.2.2 SAS-programs and output

Logistic regression models may be fitted using the `GENMOD` procedure in SAS.

Example 6.1 *(continued)*

For a description of the dataset we refer to the program of section 6.1.2.

Fit logistic regression model

To fit a logistic regression model with interaction between `light` and `water` use the program

```
proc genmod data=growth;
class light water;
weight count;
model germination=light water light*water/dist=binomial link=logit type3;
run;
```

The `model` statement expresses that the variable `germination` should be modelled as a logistic regression model (`dist=binomial`) with `logit` as link function and that `count` describes the number of pots for each combination of the variables `light, water,` and `germination`.

```
       LR Statistics For Type 3 Analysis

                          Chi-
Source              DF    Square    Pr > ChiSq

light                1      1.53        0.2157
water                2     12.78        0.0017
light*water          2      7.80        0.0203
```

The output shows that the interaction can not be removed (p-value: 2%).

Estimates and confidence intervals

Below we fit the logistic regression model *without* the interaction term.

```
proc genmod data=growth;
class light water;
weight count;
model germination=water light/noint dist=binomial link=logit type3 lrci;
```

```
run;

[The following displays a part of the output]
```

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95 Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		0	0.0000	0.0000	0.0000	0.0000	.	.
water	high	1	1.6094	0.6074	0.5156	2.9488	7.02	0.0081
water	low	1	0.6931	0.5064	-0.2727	1.7413	1.87	0.1711
water	mod	1	-0.3365	0.4903	-1.3304	0.6188	0.47	0.4926
light	12	1	-0.0000	0.5252	-1.0378	1.0378	0.00	1.0000
light	8	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

The `noint` option of the `model` statement ensures that one parameter is fitted for each level of `water` and one parameter for the difference $\beta(\texttt{light8}) - \beta(\texttt{light12})$. Note that the estimates for the effects of `water` is given for the reference group given by `light` $= 12$. Therefore the estimates given in Example 6.1 in section 6.2 have opposite signs. The `lrci` option of the `model` statement forces SAS to print the likelihood ratio based confidence limits for the parameters.

Example 6.2 *(continued)*

Read in data

The following program reads in the dataset of example 6.2. The third column (`n`) is the total number of moths in each `dose` group and the last column is the number of dead/collapsed moths with three days. A new variable `logdose` is constructed.

```
data moth;
input dose sex $ n a;
logdose=log(dose);
cards;
1 m 20 1
1 f 20 0
2 m 20 4
2 f 20 2
4 m 20 9
4 f 20 6
8 m 20 13
8 f 20 10
16 m 20 18
16 f 20 12
32 m 20 20
32 f 20 16
;
run;

proc print data=moth;
run;
```

Fit logistic regression models and test for reduction.

To fit the full logistic regression model where `sex` is used as a factor and `logdose` as a covariate use the following program.

```
proc genmod data=moth;
class sex;
model a/n=sex logdose sex*logdose/noint dist=binomial link=logit lrci type3;
run;
```

It is important to specify the family option `dist=binomial` indicating that the response is binary and follows a binomial distribution. In exercise 9.4 you are supposed to try to use other `link` functions than the logit function. Part of the output is displayed below.

The GENMOD Procedure
Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	8	4.9937	0.6242
Scaled Deviance	8	4.9937	0.6242
Pearson Chi-Square	8	3.5047	0.4381
Scaled Pearson X2	8	3.5047	0.4381
Log Likelihood		-105.7388	

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		0	0.0000	0.0000	0.0000	0.0000	.	.
sex	f	1	-2.9935	0.5527	-4.1865	-2.0017	29.34	<.0001
sex	m	1	-2.8186	0.5480	-4.0054	-1.8360	26.46	<.0001
logdose		1	1.8163	0.3059	1.2740	2.4860	35.24	<.0001
logdose*sex	f	1	-0.5091	0.3895	-1.3077	0.2399	1.71	0.1912
logdose*sex	m	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
sex	1	0.05	0.8221
logdose	1	112.73	<.0001
logdose*sex	1	1.76	0.1842

The results of the test of the logistic model against the full model is given by the line beginning with `Deviance`. The test statistic is 4.9937 and the test has 8 degrees of freedom. Remember that the test of the logistic model only makes sense when multiple individuals are exposed to the same treatment (here same combination of `dose` and `sex`).

When fitting the parameters of the logistic regression model SAS uses the group given by `sex=Male` as reference for the slope parameter. Thus in the notation specified by (6.3) in example 6.2 the parameters

from the SAS output are estimates of

$$\alpha(\texttt{Male}) \quad \alpha(\texttt{Female}) \quad \beta(\texttt{Male}) \quad \beta(\texttt{Female}) - \beta(\texttt{Male}).$$

Finally, the output shows that the p-value for removing the effect of `sex` is 82 %. The model obtained by removing `sex` from the `model` statement i SAS is given by (6.4) and expresses that the lines for `Males` and `Females` intersect the y-axis at the same point while the slopes are allowed to be different.

Parameter estimates under the model with same intersection for `Male` *and* `Female`

To fit the model specified by (6.4) of example 6.2 and obtain relevant estimates one may write

```
proc genmod data=moth;
class sex;
model a/n=logdose sex*logdose/dist=binomial link=logit lrci;
estimate 'femaledose10' int 1 logdose  2.302585 sex*logdose  2.302585/exp;
run;
```

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-2.9073	0.3893	-3.7245	-2.1916	55.78	<.0001
logdose		1	1.8601	0.2408	1.4219	2.3707	59.65	<.0001
logdose*sex	f	1	-0.5872	0.1799	-0.9571	-0.2474	10.65	0.0011
logdose*sex	m	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
femaledose10	0.0237	0.2328	0.05	-0.4324	0.4799	0.01	0.9187
Exp(femaledose10)	1.0240	0.2384	0.05	0.6489	1.6160		

Three parameters are estimated: common intercept, slope of `Males`, and difference between slope of `Females` and `Males`. The `lrci` option of the `model` statement tells SAS to print 95%−confidence intervals for the parameters.

The probability that a `female` moth receiving `dose` $= 10$ survives more than three days may be expressed as

$$\delta_{f,10} = 1 \cdot \alpha + \log(10) \cdot \beta(\texttt{male}) + \log(10) \cdot (\beta(\texttt{female}) - \beta(\texttt{male}).$$

The `estimate` statement tells us that we want an estimate of 1 times the intercept, $\log(10) = 2.302585$ times `logdose` ($\beta(\texttt{Male})$), and $\log(10) = 2.302585$ times `logdose*sex` ($\beta(\texttt{Female}) - \beta(\texttt{Male})$). The `exp` option of the `estimate` statement indicates that we want estimates for both log-odds and odds. An estimate and a 95 %-confidence interval for $\delta_{f,10}$ is then given as $\hat{\delta}_{f,10} = 0.0237\ [-0.4324, 0.4799]$. Taking the inverse logit function

$$\text{invlogit} : x \to \frac{\exp(x)}{1 + \exp(x)}$$

we may tranform the confidence interval from the log odds-scale to the scale of the response (probability on $(0, 1)$) and we get $\hat{p} = 0.506$ with 95%-confidence interval $[0.392, 0.619]$.

6.3 Overdispersion in logistic regression models

Scientific experiments are carried out to determine which conditions (covariate) affect the outcome. In the present framework our starting point for testing whether a given varible is important for the binary response (0/1) has been to assume a logistic regression model for the frequency of getting outcome 1. In the case where multiple individuals are subjected to exactly the same treatment we described how to test the logistic regression model where the log-odds depends linearly on a covariate. We stressed that this test was to be reported before examining the effect of other treatment factors.

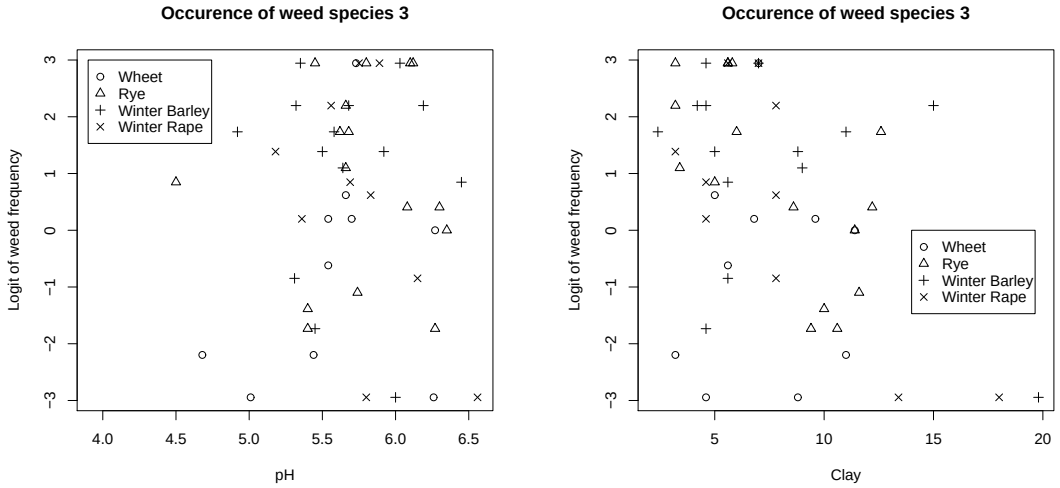
It may happen that the test rejects the logistic regression model or that we are in a case where the logistic regression model may not be tested. Does this mean that our experiment is of no value for exploring the effect of the treament factor? In principle it implies that it is highly questionable and not recommendable to trust the results of the logistic regression analysis concerning the effect of treament factors. However, we present below two modifications of the model that may be used when the model assumptions of the classical model are clearly not satisfied. The drawback of the two approaches is that it is not obvious how to check the validity of the assumptions.

Before continuing with the mathematical details we present an example to illustrate what may be the consequences of neglicting the fact that the logistic regression model does not describe the data. The examples further discuss what may be the cause of deficiencies from the logistic regression model and motivates the methods introduced in section 6.3.1 and 6.3.2.

Example 6.3 Occurence of weed.

The occurence of weed has been examined on several fields using the method proposed by Raunkiær. The idea is to repeatedly throw a ring into the field at random and to count how many times the different weed species occur within the area surrounded by the ring. The data for this example has kindly been disposed by Christian Andreasen and is a small part of a huge experiment.

The data set discussed here stems from 68 fields and consists of the total number of rings that contains *Ager-stedmoderblomst*. On each field the ring has been thrown a total number of 20 times. Related to each observation we consider the factor `crop` with four levels and the two covariates `pH` and `clay`.



Considering the plots above displaying our data it seems difficult to spot an effect of either of the ex-

planatory variables. Let us try to fit the following logistic model to the data

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha \cdot \texttt{pH}_i + \beta_0(\texttt{crop}_i) + \beta_1(\texttt{crop}_i) \cdot \texttt{clay}_i, \tag{6.5}$$

where $Y_i \sim b(20, p_i)$ denotes the count on the $i-$th field and all 68 observations are assumed to be independent. The p-values for removing either `pH` or the interaction `crop`$\times$`clay` are indistinguisable from zero.

The conclusion may appear surprising as the plots above are messy and do not clearly reveal a marginal effect of `clay, pH,` or `crop`. The explanation for this is that the data are not well described by the model (6.5). Note that we cannot test if the model is okay as in the previous examples because `pH` and `clay` take different values (not set by the experimenter). We shall mention two possible reasons that (6.5) may be inappropriate for the data.

The model (6.5) requires the counts Y_i to be binomially distributed. This is reasonable only if there is independence between individual counts on the same field. As we will see later in section 6.3.1 strong (positive) correlation is likely to cause too much variation in the data compared to what can be explained by the binomial distribution.

Another possible explanation for the strange results is that there could be huge differences between the 68 fields due to conditions that we have no chance to include in our model. In section 6.3.2 we discuss how to include a random effect to take this possibility into account. □

6.3.1 Including an overdispersion parameter

Suppose that the experiment is carried out by collecting the experimental units (individuals) in k different groups. As our dataset we consider the variables Y_i counting the number of times the outcome 1 turned up in each group. When we decide to describe the data using a logistic regression model we assume that

$$Y_i \sim b(n_i, p_i), \quad i = 1, \ldots, k,$$

where n_i denotes the number of individuals (or experimental units) in group i. This assumption is reasonable in the case where the response of individuals in *same* group does not influence each other - in statistical terms when outcomes associated with different individuals are independent.

One may often argue that the response for individuals in same group are more related than individuals from different groups. This is likely to conflict with the assumption of the Y_i's following a binomial distribution. If we decide to parameterize the variance of Y_i by

$$Var(Y_i) = \sigma^2 n_i p_i (1 - p_i)$$

then estimates of σ^2 far from 1 indicates deviations from the binomial assumption since we know that

$$Var(Y_i) = n_i p_i (1 - p_i),$$

for $Y_i \sim b(n_i, p_i)$. The parameter σ^2 is denoted the *overdispersion parameter*.

The following example explains how to obtain an estimate of σ^2 for a logistic regression model. It is further illustrated how the estimate may be included in the analysis, so that is possible to test for the effect of the explanatory variables when the binomial assumption does not hold. More formally:

- all parameter estimates are unchanged (compared to usual logistic regression)
- all s.e.'s are multiplied by $\hat{\sigma}$
- all Wald test statistics are divided by $\hat{\sigma}^2$

Example 6.3 *(continued)* For the weed example we now redo the statistical analysis where we allow for overdispersion. Using (6.5) as our initial model for p_i the overdispersion parameter is estimated to $\hat{\sigma}^2 = 10.5431$. This is far from 1 indicating that the binomial assumption is not met. The corrected Wald test for removing the interaction `crop`$\times$`clay` becomes

$$55.12/10.5431 = 5.2281 \sim \chi^2(3)$$

corresponding to a p-value of 16%. For succesive reduction of (6.5) it may be convenient to construct table of the form (0 meaning the model with constant p for all observations).

Model effects	Deviance	Df	$\hat{\sigma}^2$	G^2	p-value
`pH` + `crop` $\times$ `clay`	-	59	10.5431	-	-
`pH` + `crop` + `clay`	55.12	3	11.7081	5.2281	0.1558
`crop` + `clay`	25.04	1	11.5956	2.1387	0.1436
`clay`	41.99	3	11.1160	3.6212	0.3054
0	152.21	1	11.7740	13.6929	2.153e-4

We end up with the model

$$\log\left(\frac{p_i}{1-p_i}\right) \quad = \quad \beta_0 + \beta_1 \cdot \texttt{clay}_i,$$

expressing that the occurence of weed only depends on the amount of `clay` in the soil. The parameter estimates for the effect of clay together with 95 %-confidence intervals are given by

$$\begin{array}{ll} \hat{\beta}_0 = 1.7137 & \text{[0.8713,2.5561]} \\ \hat{\beta}_1 = -0.1709 & \text{[-0.2694,-0.0724].} \end{array}$$

The confidence intervals are constructed simply by multiplying the lengths of the intervals obtained for the model *without* overdispersion by a factor $\hat{\sigma} = \sqrt{11.116} = 3.334$. □

It may be illuminating to examine the effect of the correction for overdispersion in a simple teoretical example.

Example 6.4 Effect of correlation within groups

Suppose that we want to examine the effect of a pesticide by conducting a field experiment on 50 fields of which 25 are treated with the pesticide while the remaining 25 are used as controls. By looking at previous experiments we have selected two different weeds, A and B, that are both present with probability 50% on a randomly chosen untreated field. As our data we count for each of the 50 fields how many of the species A and B that are observed. Thus, our data consist of 50 observations, $Y_i, i = 1, \ldots, 50$, each taking one of the values 0, 1, or 2.

To analyze the data we plan to use a logistic regression model where $Y_i \sim b(2, p_i)$ are independent and

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta \cdot x_i,$$

where x_i is 1 or 0 depending on whether the $i-$th field has been treated with the pesticide. If a test of the hypothesis $H_0 : \beta = 0$ is rejected we conclude that the pesticide influences the occurence of weed. We decide to use a 5 % significance level for the test meaning that if there is no effect of the pesticide the test will only yield the wrong result and detect an effect with probability 5%. We will say that the test has level 0.05 (or 5%).

If the assumptions for the logistic model are not satisfied the level of the test may be different from 5% even when there is no effect of the pesticide. For this particular example we assume that whether weed A is observed on a field is independent of whether weed B is observed on the same field. This is probably very unrealistic!

Below we present a simulation study that examines the level of the test when we simulate from a model that allows for dependence between weed A and B but where there is no effect of the pesticide. Since weed A and B both occur with probability 50 % we claim that the *joint* occurence of (A,B) must be described as

Weed A	Weed B	probability
+	+	p
+	-	$1/2 - p$
-	+	$1/2 - p$
-	-	p

where $p \in (0, 1/2)$. In particular Y_i is not binomial and one can show (try!) that under this assumption

$$Var(Y_i) = 2 \cdot p.$$

Under the assumption of the logistic regression model - and no effect of the pesticide - we have that $Y_i \in b(2, 1/2)$ so

$$Var(Y_i) = 2 \cdot 1/2 \cdot (1 - 1/2) = 1/2$$

hence the overdispersion parameter becomes

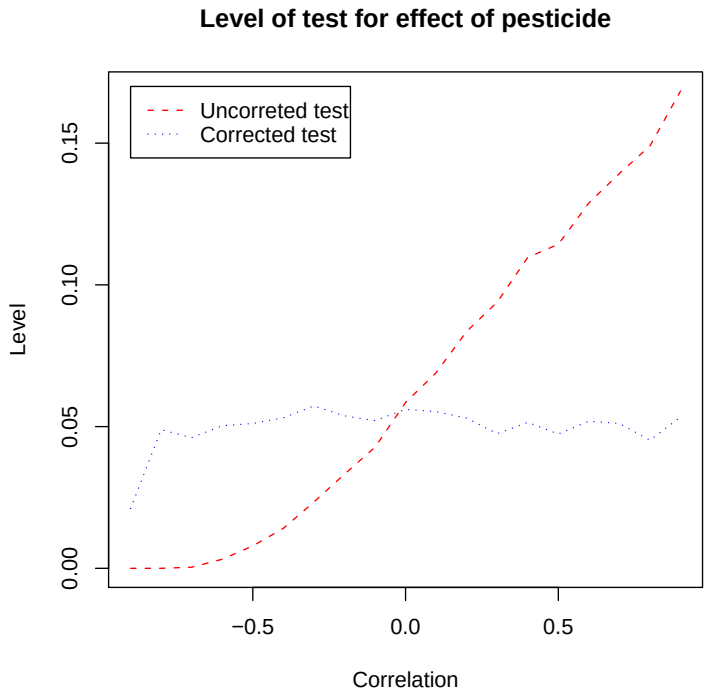
$$\sigma^2 = \frac{2 \cdot p}{1/2} = 4p.$$

If $Y_{i,A}$ (respectively $Y_{i,B}$) is 1 if weed A (respectively weed B) is observed on field i one finds that their correlation is given by

$$\rho = Corr(Y_{i,A}, Y_{i,B}) = 4p - 1.$$

The correlation is often used to measure dependence between random variables and a correlation far from zero (and close to -1 or 1) conflicts an assumption of independence.

For different values of the correlation 10.000 datasets were simulated from the distribution given by the table above. For each of these the hypothesis of no pesticide effect was tested, both with and without correction for overdispersion. The rejection frequency was then computed. The figure below displays the level of the test for effect of the pesticide both with and without correction for overdispersion.



The case $\rho = 0$ corresponds to independence between occurence of weed A and B. Here the assumptions for the logistic regression model hold and the level of the corrected as well as the uncorrected test are close to 5 % as desired. However, for $\rho = 0.6$ the power of the uncorrected test is 12.9 % implying that the test would claim an effect of the pesticide in more that one out of eight experiments even when this is not the case. We further observe, that if the test is corrected for overdispersion we obtain a test with power close to 5 % almost no matter what is the correlation between the two types of weed. □

6.3.2 Random effects in logistic regression models

In the following section we present a more specific model that may be used when the logistic regression model seems inappropriate. Think of the situation where there is a source of variation which is not important for the actual research question (the effect of treatments on the response), but which hinders the analysis of these matters. The idea in to replicate the terminology from gaussian models with random effects. Typical examples are experiments where individuals (persons/animals) are exposed to different treatments and large differences between patients make it diffucult to detect the effect of the treatment. The following example explains how to include a random effect in a logistic regression model.

Example 6.3 *(continued)* Let us try to fit the model given by (6.5) where we further allow for a random `field` effect to account for differences between the fields. We introduce the variables $X_i, i = 1, \ldots, 20 \cdot 68$, indicating the (binary) response (1/0) of *individual* throws with the ring. The model is formally specified by assuming that $X_i \sim b(1, p_i)$ with

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha \cdot \texttt{pH}_i + \beta_0(\texttt{crop}_i) + \beta_1(\texttt{crop}_i) \cdot \texttt{clay}_i + \gamma(\texttt{field}_i) \tag{6.6}$$

where $\gamma(1), \ldots, \gamma(68) \sim N(0, \sigma_F^2)$ are independent random variables. A table for successive reduction of the systematic part of the model is given below. Note that for R we report the p-value based on the likelihood ratio test and for SAS we give the p-value for the F-test.

Model effects	p-value (R)	p-value (SAS)
`crop` $\times$ `clay`	0.0911	0.1131
`crop` $+$ `clay`	0.0306	0.0601
`clay`	0.4228	0.4448
0	0.0001	< 0.0001

If we maintain a significance level of 5% we may not remove the interaction `crop` $\times$ `clay` since the p-value of the F-test is only 3%. Let us for a while ignore this fact and consider the model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \texttt{clay}_i + \gamma(\texttt{field}_i) \tag{6.7}$$

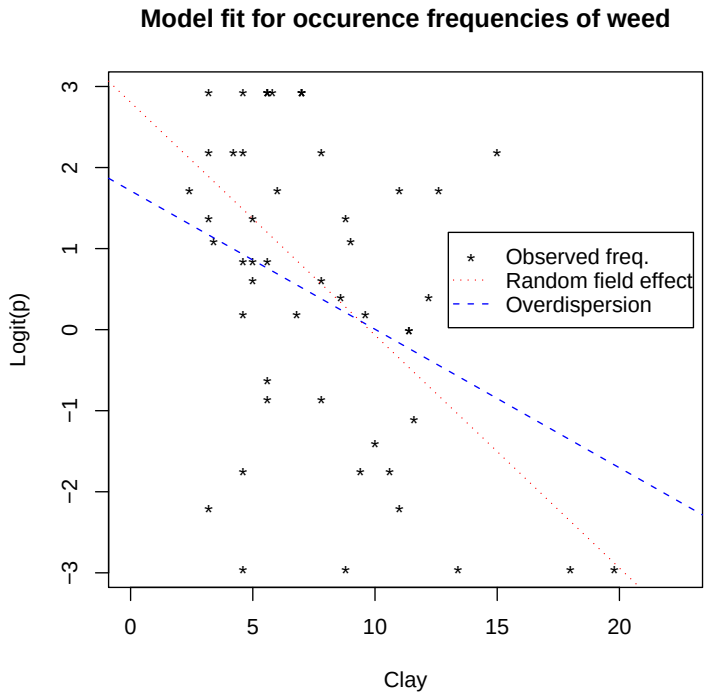
with common slope (wrt. `clay`) for all `crop`s. The parameter estimates become

$$\begin{aligned} \hat{\beta}_0 &= 2.8054 \\ \hat{\beta}_1 &= -0.2873 \\ \hat{\sigma}_F^2 &= 7.287. \end{aligned}$$

Comparing the estimates above with those obtained for the model with overdispersion in section 6.3.1 we note that there are clear differences. If we quantify the effect of `clay` by calculating the change of the odds by increasing the level of `clay` by 1 we get

Model	$\hat{\beta}_1$	$\exp(\hat{\beta}_1)$
Overdispersion	-0.1709	0.843
Random field effect	-0.2873	0.750

Thus, depending on which model we decide to trust increasing `clay` by 1 reduces odds of observing weed in the ring by either 16% or 25 %. The plots below help us visualizing the differences between the two approches.



The model in section 6.3.1 predicts the observations to fall on the dotted line, and uses the overdispersion parameter to describe the variation around the line. The model including a random `field` effect predicts the observations to be located on the dashed line, that is chosen so that the deviations of the actual observations are normally distributed. It seems impossible to argue which model that must be preferred to the other. However, remember that one difference between the models is that the interaction `crop`×`clay` is actually significant in the random effect model. □

6.3.3 R-programs and output

Example 6.3 *(continued)*

Description of dataset

The data are read into R and stored as a data frame called `data`. The data frame includes the number of rings, `c3`, containing the weed species 3,*Ager-stedmoderblomst*, the factor `crop`, and the covariates `pH` and `ler` (=clay) measuring chemical properties of the soil. The total number of datalines are 68. Initially we make a matrix, `d3`, with two columns collecting the number of positive and negative responses using the following code

```
> d3=cbind(c3,20-c3)
```

Fit logistic regression models with overdispersion parameter

Logisitic regression models with overdispersion parameters are fitted using `glm()` with `quasibinomial` as the family option.

```
> glm1.over=glm(d3~pH+crop*ler,family=quasibinomial)
> glm2.over=glm(d3~ler+crop+pH,family=quasibinomial)
> glm3.over=glm(d3~ler+crop,family=quasibinomial)
> glm4.over=glm(d3~ler,family=quasibinomial)
> glm5.over=glm(d3~1,family=quasibinomial)
```

Parameter estimates for the model `glm1.over` are found by

```
> summary(glm1.over)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.76590    2.77286  -0.637   0.5267
pH                  0.85540    0.49537   1.727   0.0894 .
cropRug            -0.74618    1.31326  -0.568   0.5721
cropVinterbyg       0.42443    1.62203   0.262   0.7945
cropVinterraps     -3.11336    1.45706  -2.137   0.0368 *
ler                -0.28066    0.10876  -2.581   0.0124 *
cropRug:ler         0.03065    0.16483   0.186   0.8531
cropVinterbyg:ler  -0.15660    0.19524  -0.802   0.4257
cropVinterraps:ler  0.22940    0.14190   1.617   0.1113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 10.54309)
```

in particular we may read off the dispersion parameter (10.54309).

A test for the effect of `crop`×`ler` is produced by

```
> anova(glm2.over,glm1.over,test="Chisq")
Analysis of Deviance Table

Model 1: d3 ~ ler + crop + pH
Model 2: d3 ~ pH + crop * ler
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        62     749.89
2        59     694.77  3    55.12      0.16
```

and the tests for further reductions of the model are obtained by writing

```
> anova(glm3.over,glm2.over,test="Chisq")
> anova(glm4.over,glm3.over,test="Chisq")
> anova(glm5.over,glm4.over,test="Chisq")
```

Including a random field effect

Logistic regression models with random effects may be fitted using the `lmer()` function included in the `lme4` package. The first challenge is to get the data on a suitable form.

```
> block=rep(1:68,rep(20,68))
```

```
> c3rand=c()
> for(i in 1:68){for(j in 1:2){if(d3[i,j]>0){c3rand=c(c3rand,rep(2-j,d3[i,j]))}}
> lerrand=rep(ler,rep(20,68))
> pHrand=rep(pH,rep(20,68))
> croprand=rep(crop,rep(20,68))
```

The R-script above produces a dataset with $68 \cdot 20$ data lines - one for *each* of the moths in the experiment. The objects `lerrand, pHrand,`and `croprand` are vectors of length $68 \cdot 20$ containing the value of `ler, pH,` and `crop` for individual rings in the experiment.

To fit the models and perform the tests reported in section 6.3.2 the following R-script may be used.

```
> library(lme4)
> glm1.b=lmer(c3rand~pHrand+croprand*lerrand+(1|block),family=binomial)
> glm2.b=lmer(c3rand~croprand*lerrand+(1|block),family=binomial)
> glm3.b=lmer(c3rand~croprand+lerrand+(1|block),family=binomial)
> glm4.b=lmer(c3rand~lerrand+(1|block),family=binomial)
> glm5.b=lmer(c3rand~1+(1|block),family=binomial)
> anova(glm5.b,glm4.b,glm3.b,glm2.b,glm1.b,test="Chisq")
Data:
Models:
glm5.b: c3rand ~ 1 + (1 | block)
glm4.b: c3rand ~ lerrand + (1 | block)
glm3.b: c3rand ~ croprand + lerrand + (1 | block)
glm2.b: c3rand ~ croprand * lerrand + (1 | block)
glm1.b: c3rand ~ pHrand + croprand * lerrand + (1 | block)
       Df     AIC     BIC  logLik   Chisq Chi Df Pr(>Chisq)
glm5.b  2 1111.14 1121.57 -553.57
glm4.b  3 1098.12 1113.77 -546.06 15.0138      1  0.0001067 ***
glm3.b  6 1101.32 1132.61 -544.66  2.8039      3  0.4228618
glm2.b  9 1098.41 1145.35 -540.21  8.9027      3  0.0306130 *
glm1.b 10 1097.56 1149.71 -538.78  2.8535      1  0.0911725 .
```

Parameter estimates under the model given by (6.7) in section 6.3.2 can be found using

```
> summary(glm4.b)
Generalized linear mixed model fit using PQL
Formula: c3rand ~ lerrand + (1 | block)
 Family: binomial(logit link)

Random effects:
 Groups Name        Variance Std.Dev.
 block  (Intercept) 7.287    2.6995
number of obs: 1360, groups: block, 68

Fixed effects:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)  2.80540    0.72713  3.8582 0.0001142 ***
lerrand     -0.28728    0.08498 -3.3805 0.0007234 ***
```

6.3.4 SAS-programs and output

Example 6.3 *(continued)*

Description of dataset

The data are read into SAS from the file 'WeedExChristianAndreasen.txt' and stored as a dataset called `weed`. The dataset contains a number of variables of which we shall only consider the number of rings, `c3`, containing the weed species 3 (*Ager-stedmoderblomst*), the factor `crop`, and the covariates `pH` and `ler` (=clay) measuring chemical properties of the soil. The total number of datalines are 68. When reading in the data we add a variable `total` with the value 20 corresponding to the number of rings thrown at each field.

```
data weed;
infile 'C:\WeedExChristianAndreasen.txt' firstobs=15;
input crop $ c3 c7 c18 pH K Mg N C P ler silt grovsand finsand orgstof;
field=_N_;
total=20;
run;
```

Note that the entire path to the file should be specified, hence on your computer the line `'C:/WeedExChristianAndreasen.txt'` must correspond to the subdirectory, where you have stored the file with the dataset. The statement `firstobs=15` tells SAS to ignore the first 14 lines of the file, since these are only comments and not part of the dataset. The line `field=_N_` creates a new variable `field` taking the values $1, \ldots, 68$ corresponding to the dataline number.

Fit logistic regression models with overdispersion parameter

To include an overdispersion parameter when fitting a logistic regression model use the `pscale` option in the `model` statement of the `PROC GENMOD`.

```
proc genmod;
class crop;
model c3/total=pH crop ler crop*ler/dist=binomial link=logit type3 pscale;
run;
```

The program above fits the logistic model (6.5) and estimates an overdispersion parameter ($\hat{\sigma}^2 = 3.247^2 = 10.543$). Part of the output is given below.

```
 Parameter              DF   Estimate      Error          Limits          Square   Pr > ChiSq

  ler*crop    Vinterra    0     0.0000     0.0000     0.0000     0.0000       .          .
  Scale                   0     3.2470     0.0000     3.2470     3.2470

  Source       Num DF    Den DF    F Value    Pr > F     Square    Pr > ChiSq

  pH                1        59       3.05    0.0858       3.05        0.0806
  crop              3        59       2.42    0.0750       7.26        0.0640
  ler               1        59      16.14    0.0002      16.14        <.0001
  ler*crop          3        59       1.74    0.1681       5.23        0.1558
```

The p-value for removing the interaction `crop` $\times$ `ler` is 0.1558.

Successively reducing the fixed effects of the model we end up with the model containing only the covariate `ler`. Parameter estimates with 95 %-confidence intervals are:

```
proc genmod;
class crop;
model c3/total=ler/dist=binomial link=logit type3 pscale;
run;

[Part of the output]
                        Analysis Of Parameter Estimates

                                 Standard     Wald 95% Confidence       Chi-
   Parameter    DF    Estimate      Error           Limits             Square    Pr > ChiSq

   Intercept     1      1.7137     0.4243      0.8820      2.5453      16.31        <.0001
   ler           1     -0.1709     0.0495     -0.2678     -0.0739      11.93        0.0006
   Scale         0      3.3341     0.0000      3.3341      3.3341

NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

                                                                    Chi-
          Source       Num DF    Den DF    F Value    Pr > F      Square    Pr > ChiSq

          ler               1        66      13.69    0.0004      13.69        0.0002
```

We find that the test statistic for removing the effect of `ler` is $G^2 = 13.69 \sim \chi^2(1)$ (p-value: 0.02 %). The dispersion parameter is estimated to $\hat{\sigma} = 3.3341$ and estimates for the two other parameters become

$$\begin{aligned} \hat{\beta}_0 &= 1.7137 \quad [0.8820, 2.5453] \\ \hat{\beta}_1 &= -0.1709 \quad [-0.2678, -0.0739]. \end{aligned}$$

Including a random field effect

Logistic regression models with random effects are fitted using the `glimmix` macro. The first challenge is to get the data on a suitable form.

```
data weed2;
set weed;
v=1; do i=1 to c3; output; end;
v=0; do i=1 to total-c3; output; end;
keep field pH ler crop v;
proc print;
run;
```

The SAS-program above produces a dataset, `weed2`, with $68 \cdot 20$ datalines - one for *each* of the rings in the experiment.

To fit the models and perform the tests reported in section 6.3.2 the following SAS-program may be used.

```
%include "C:\glimmix.sas";

%glimmix(data=weed2,
```

```
stmts=%str(class crop field;
model v=crop ler crop*ler pH/solution;
random field;),
error=binomial,
link=logit
)
```

Note that when calling the `glimmix` macro "C:\glimmix.sas" must refer to a subdirectory on your own computer, where you have saved the program `glimmix.sas` (which may be downloaded from the course homepage).

Chapter 7

Models for polytomous response data

A common feature of all the examples discussed in Chapter 7 is that the response data can be thought of as being binary. However, all of the models have extensions to the case where the response takes values in a set with J levels $1, \ldots, J$. In this case the outcome of the $i-$th experiment can be summarized into a vector

$$Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})$$

where exactly one of the coordinates is one and the others are zero. Y_i will be assumed to follow a multinomial distribution

$$Y_i \sim m(1, (p_{i1}, \ldots, p_{iJ}))$$

where p_{ij} is the probability that the i-th outcome is j. We further suppose that the Y_i's are all independent. Thus the models discussed below only differ in the way we decide to model the dependence of the probability parameters

$$p_{i1}, p_{i2}, \ldots, p_{iJ}$$

on factors and covariates associated with the $i-$th experiment.

In certain cases there is a natural ordering on the set of the categories for the response. An important example of this kind occurs if the response is the result of a discretization procedure where a latent continuous response variable is assigned to one of a number of different categories according to different thresholds.

7.1 Tables of counts

If multiple individuals are associated with the same configuration of the explanatory variables the data are usually summarized into a table of counts. The following table displays the result of an experiment concerning the effect of spraying against eyespot (knækkefodssyge). Groups of approximately 50 plants have been exposed to one of ten different treatments and after some time each plant has been judged on a scale ranging from 1 to 4. The experiment has been repeated four times on different blocks. Note that response category 1 corresponds to the plant being completely healthy, 4 to severe lesions on the whole stem.

	Block															
	1				2				3				4			
	Judgment				Judgment				Judgment				Judgment			
Treatment	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	4	16	17	9	11	9	20	9	10	8	18	13	4	13	19	9
2	28	7	11	2	23	11	14	4	22	11	12	4	18	16	14	0
3	3	12	21	9	9	8	19	10	3	12	23	12	7	13	27	10
4	3	15	22	7	12	9	17	11	6	13	19	8	8	16	18	11
5	10	9	16	11	7	5	16	11	5	13	16	10	4	8	24	12
6	30	4	10	2	24	9	10	4	26	10	11	3	26	8	11	1
7	9	9	19	11	6	6	20	17	1	13	20	14	8	8	24	11
8	5	8	23	12	8	12	15	11	8	6	21	13	6	7	20	14
9	3	15	19	12	8	15	15	12	14	12	13	8	4	9	16	19
10	9	13	16	8	9	9	20	12	5	7	25	14	6	9	15	18

Example 7.1 Eyespot (Knækkefodssyge).

The variables of the experiment are denoted `treatment`, `block`, and `judge`. Suppose that we want to examine the effect of `treatment` on the judgement of the plant and let us ignore the `block` factor. The resulting data are summarized into the following table.

	Judgement				
Treatment	1	2	3	4	Total
1	29	46	74	40	189
2	91	45	51	10	197
3	22	45	90	41	198
4	29	53	76	37	195
5	26	35	72	44	177
6	106	31	42	10	189
7	24	36	83	53	196
8	27	33	79	50	189
9	29	51	63	51	194
10	29	38	76	52	195

We denote by

$$Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})$$

the number of plants in each response category for the $i-$th `treatment` group and assume that $Y_1, \ldots, Y_{10}$ are independent and follow a multinomial distribution

$$Y_i \sim m(n_i, (p_{i1}, \ldots, p_{i4}))$$

where n_i is the total number of plants receiving `treatment` $= i$. The hypothesis of homogeneity between treatment groups may be expressed as

$$H_0 : \quad p_{1j} = p_{2j} = \ldots = p_{10j},$$

for all response categories $j = 1, 2, 3, 4$. The (likelihood ratio) test statistic of H_0 against the unrestricted model with individual probabilities for each treatment is $G^2 = 280.27 \sim \chi^2(27)$ which is highly significant (p-value$= 0\%$). The parameter estimates under the full model is simply the tabular of observed frequencies corresponding to the table of counts presented above. □

The conclusion of our analysis above is that there are differences between the judgement of plants in different `treatment` groups. There are, however, at least two problems with our model.

Firstly, we ignore the `block` factor which may potentially affect the judgement of the plant and consequently make it questionable to interpret the parameter estimates for `treatment` groups. As for the case with binary response data it is not clear how to build a model where the response `judge` is allowed to depend on both `block` and `treatment`. In section 7.2 we explain how the logistic regression model may be extended to the case with multiple response categories.

Another problem with the analysis in example 7.1 is that it may be difficult to interpret differences between treatment groups. This is due to the fact that the model allows the effect of the treatment to vary over different response categories. The *proportional odds-model* in section 7.3 explains how to build a model where the effect of explanatory variables is described in a unified way over all response categories. This allows for a simple interpretation of the effect of factors and covariates in the experimental design.

7.1.1 R-programs and output

Example 7.1 *(continued)*

Read in data

The data are read into R from an ASCII file organised as follows.

```
Treat   Block   A       B       C       D
1       1       4       16      17      9
2       1       28      7       11      2
3       1       3       12      21      9
.       .
.       .
[more datalines here]
.       .
.       .
6       4       26      8       11      1
7       4       8       8       24      11
8       4       6       7       20      14
9       4       4       9       16      19
10      4       6       9       15      18
```

The columns with names `A` to `D` are the number of plants in each of the four judgement categories. The data set is stored as a six column matrix labelled `data`.

Testing for effect of `treatment`

The program lines below construct a two-way table of counts (`table1`) by classifying the data according to `treatment` and response group.

```
> table1=xtabs(data[,3:6]~Treat)
> table1

Treat   A   B   C   D
   1   29  46  74  40
   2   91  45  51  10
```

```
   3   22  45  90  41
   4   29  53  76  37
   5   26  35  72  44
   6  106  31  42  10
   7   24  36  83  53
   8   27  33  79  50
   9   29  51  63  51
   10  29  38  76  52
> chisq.test(table1)

        Pearson's Chi-squared test

data:  table1
X-squared = 301.586, df = 27, p-value < 2.2e-16
```

A Pearson χ^2-test for homogeneity of the response distribution for different `treatment` groups is obtained using the `chisq.test()` method exactly as described in section 7.1 where the response was binary.

A likelihood ratio test for the same hypothesis may be constructed using the `multinom()` method of the `nnet` library.

```
> library(nnet)
> Tmod1=multinom(data[,3:6]~factor(Treat))
> Tmod2=multinom(data[,3:6]~1)
> anova(Tmod2,Tmod1)
Likelihood ratio tests of Multinomial Models

          Model Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
1             1       117   5188.982
2 factor(Treat)        90   4908.716 1 vs 2    27 280.2651       0
```

The `multinom` method is used to fit multinomial distributions with at least three different response groups (for binary response use `glm` with `family = binomial`). The object `Tmod1` contains a fit of a multinomial model with individual parameters for each level of `treatment` whereas `Tmod2` has fitted one common set of parameters for all `treatment` levels.

7.1.2 SAS-programs and output

Example 7.1 *(continued)*

Read in data

The data are stored as a dataset `knaekke` in SAS so that each dataline contains the three variables `Treat` (`treatment`), `Block` (`block`) and `Judge` (`judgement` group) and the corresponding number of `counts`. The first six lines of `knaekke` looks like:

```
 Obs    Block    Treat    Judge    count

 1      1         1       A         4
 2      1         1       B        16
```

```
3      1         1       C        17
4      1         1       D         9
5      1         2       A        28
6      1         2       B         7
```

Testing for effect of `treatment`

Use `proc freq` to obtain test statistics for the hypothesis of the distribution on judgement groups being the same for all levels of `treatment`.

```
proc freq;
weight count;
tables Judge*Treat/chisq;
run;
```

Statistics for Table of Judge by Treat

Statistic	DF	Value	Prob
Chi-Square	27	301.5860	<.0001
Likelihood Ratio Chi-Square	27	280.2651	<.0001
Mantel-Haenszel Chi-Square	1	24.0571	<.0001
Phi Coefficient		0.3964	
Contingency Coefficient		0.3685	
Cramer's V		0.2289	

Note that the output contains the likelihood ratio test statistic as well as the Pearson χ^2-test statistic.

7.2 Multinomial logistic regression modela

We consider in this section a data set where the response is categorized into one of J different groups and where each observation further contains recordings of a number of explanatory variables. The model below is described using the terminology of the eyespot (knækkefodssyge) experiment introduced in section 7.1 where we have four different response groups and two factors `treatment` and `block`. However, we stress that the model may include an arbitrary number of explanatory variables some of which may be continuous covariates.

The full data set consists of 40 variables

$$Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}),$$

counting the number of plants in each judgement group for each of the 40 combinations of the factors `treatment` and `block`. We assume that the Y_i's are independent and that Y_i follows a multinomial distribution with probability parameter

$$p_i = (p_{i1}, p_{i2}, p_{i3}, p_{i4}),$$

where the parameters must obey the constraint $\sum_j p_{ij} = 1$. In the following the model will be referred to as the full (polytomous) logistic regression model and we will use instead a parameterization given by the log-odds

$$\eta_{ij} \quad = \quad \log\left(\frac{p_{ij}}{p_{i1}}\right) \quad j = 2, \ldots, J-1, \tag{7.1}$$

wrt. some reference (baseline) category, here response group 1.

Example 7.1 *(continued)* The (polytomous) logistic regression model with no interaction between `block` and `treatment` can be expressed as

$$\log\left(\frac{p_{ij}}{p_{i1}}\right) \quad = \quad \alpha(\texttt{treatment}_i, j) + \beta(\texttt{block}_i, j), \quad j = 2,3,4, \quad i = 1, \ldots, 40. \tag{7.2}$$

The model expresses that for each response group the effect of the `treatment` and `block` factors enters additively on the log-odds scale. Tedious computations show that

$$p_{ij} = \frac{\exp\left(\alpha(\texttt{treament}_i, j) + \beta(\texttt{block}_i, j)\right)}{\sum_{j=1}^{4} \exp\left(\alpha(\texttt{treament}_i, j) + \beta(\texttt{block}_i, j)\right)} \quad j = 1, \ldots 4,$$

where one should put $\alpha(\texttt{treat}_i, 1) = \beta(\texttt{block}_i, 1) = 0$. Unfortunately, it is difficult to use this mathematical structure of the model to draw any interesting conclusions concerning the efffect of the two factors `treat` and `block`.

For fixed levels of `treat` and `block` the parameter estimates corresponding to (7.2) may be interpreted in the following way: the ratio between the frequency of a plant being placed into judgement group `judge`$= 4$ or `judge`$= 2$, respectively, is given by

$$\frac{p_{i4}}{p_{i2}} = \exp\left(\alpha(\texttt{treat}_i, 4) - \alpha(\texttt{treat}_i, 2) + \beta(\texttt{block}_i, 4) - \beta(\texttt{block}_i, 2)\right).$$

As submodels of (7.2) we may consider

$$\log\left(\frac{p_{ij}}{p_{i1}}\right) \quad = \quad \alpha(\texttt{treament}_i, j) \tag{7.3}$$

$$\log\left(\frac{p_{ij}}{p_{i1}}\right) \quad = \quad \beta(\texttt{block}_i, j) \tag{7.4}$$

with only a main effect of one of the two factors.

The likelihood ratio test of the model (7.2) with no interaction against the full model (7.1) yields a test statistic of $LR = 93.43 \sim \chi^2(81)$ corresponding to a p-value of 16%. We further find that the `block` factor may be removed ($LR = 8.61 \sim \chi^2(9)$, p-value $= 47\%$) but that `treatment` has a significant effect ($LR = 280.27 \sim \chi^2(27)$, p-value $= 0\%$). Our final model is therefore given by (7.3). The parameter estimates become

	Response category		
Effect	2	3	4
`treatment` $= 1$	0.4613462	0.9367278	0.3215106
`treatment` $= 2$	-0.7042333	-0.5790669	-2.2086934
`treatment` $= 3$	0.7157091	1.4088488	0.6225709
`treatment` $= 4$	0.6028764	0.9634392	0.2435296
`treatment` $= 5$	0.2968240	1.0181676	0.5256048
`treatment` $= 6$	-1.2292959	-0.9258565	-2.3604394
`treatment` $= 7$	0.4051431	1.2404833	0.7919260
`treatment` $= 8$	0.2010237	1.0738824	0.6165557
`treatment` $= 9$	0.5648464	0.7761404	0.5648464
`treatment` $= 10$	0.2701929	0.9632942	0.5839398

We have one set of parameters for each response category (expect for the reference group) and ten different levels of `treatment` yielding a total of $10 \cdot (4-1) = 30$ parameters. For a plant recieving treatment `treat`$= 6$ the ratio between the probabilities of being placed into judgement groups `judge`$= 3$ or `judge`$= 1$, respectively, is

$$\frac{p_{i3}}{p_{i1}} = \exp(-0.9259) = 0.396.$$

The probability of `judge`$= 3$ may be computed as

$$p_{i3} = \frac{\exp(-0.9259)}{1 + \exp(-1.2293) + \exp(-0.9259) + exp(-2.3604)} = 0.222.$$

The statistical analysis takes into account both of the explanatory factors and we conclude that there is only an effect of `treatment`. But does the analysis allows us to compare the effect of different `treatments`? As the estimate for `treatment` $= 6$ is the lowest for all response categories we may probably say that this treatment is the most efficient. However, it may be more difficult to compare `treatment` 4,5 and 9 since the ordering of their effects seems to depend strongly on which response category we decide to consider. The fact that treatment effects must be evaluated separately for each response group is a major drawback related to the use and in particular the interpretation of the logistic regression model for polytomous response data. □

7.2.1 R-programs and output

Example 7.1 *(continued)*

Description of dataset

The data are read into R and stored as a data frame called `data`. The data frame contains two factors `T` (treatment) and `Bl` (block) and four response variables `A, B, C,` and `D` containing the number of plants in each of four response categories.

```
> data
   Treat Block  A  B  C  D
1      1     1  4 16 17  9
2      2     1 28  7 11  2
3      3     1  3 12 21  9
4      4     1  3 15 22  7

[more datalines here]

37     7     4  8  8 24 11
38     8     4  6  7 20 14
39     9     4  4  9 16 19
40    10     4  6  9 15 18
```

Logistic regression with polytomous response

Logistic regression models are fitted using `multinom()` in the `nnet` package. The response is arranged as a matrix y with four columns containing the number of plants in each response category.

```
> library(nnet)
> y=cbind(A,B,C,D)
> mod.full<-multinom(y~Bl*T)
> mod.add<-multinom(y~Bl+T)
> mod.treat<-multinom(y~T)
> mod.block<-multinom(y~Bl)
> mod.1<-multinom(y~1)
```

The models may be tested against each other using the `anova()` method. A test for the effect of interaction factor may look like

```
> anova(mod.add,mod.full,test="Chisq")
Likelihood ratio tests of Multinomial Models

Response: y
   Model Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1 Bl + T        81   4900.106
2 Bl * T         0   4806.675 1 vs 2    81 93.43107 0.1629195
```

Parameter estimates

Parameter estimates for the final model (7.3) may be obtained by writing

```
> coef(mod.treat)
         T1         T2        T3        T4
B 0.4613462 -0.7042333 0.7157091 0.6028764
C 0.9367278 -0.5790669 1.4088488 0.9634392
D 0.3215106 -2.2086934 0.6225709 0.2435296
         T5         T6        T7        T8
B 0.2968240 -1.2292959 0.4051431 0.2010237
C 1.0181676 -0.9258565 1.2404833 1.0738824
```

```
D 0.5256048 -2.3604394 0.7919260 0.6165557
         T9       T10
B 0.5648464 0.2701929
C 0.7761404 0.9632942
D 0.5648464 0.5839398
```

7.2.2 SAS-programs and output

Example 7.1 *(continued)*

Description of dataset

The dataset is read into SAS as 160 datalines containing the variables `Block, Treat, Judge` (response category) and `count`.

```
   Obs    Block    Treat    Judge    count

    1       1        1        A        4
    2       1        1        B       16
    3       1        1        C       17
    4       1        1        D        9
    5       1        2        A       28

  [more datalines here]

  154       4        9        B        9
  155       4        9        C       16
  156       4        9        D       19
  157       4       10        A        6
  158       4       10        B        9
  159       4       10        C       15
  160       4       10        D       18
```

Logistic regression with polytomous response

Logistic regression models for data with polytomous response data may be fitted by `proc logistic` using the `link=glogit` option.

```
proc logistic;
freq count;
class Treat Block;
model Judge = Treat Block Treat*Block/link=glogit;
run;

[part of the output displayed below]

                    The LOGISTIC Procedure


                                Model Fit Statistics
```

Criterion	Intercept Only	Intercept and Covariates
AIC	5194.982	5046.674
SC	5211.660	5713.822
-2 Log L	5188.982	4806.674

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Treat	27	219.9731	<.0001
Block	9	11.7118	0.2300
Treat*Block	81	78.5987	0.5549

Note that SAS prints the Wald test statistic for the interaction `Treat` $\times$ `Block`. To calculate the likelihood ratio test statistic (as reported in example 7.1) we need to calculate the difference between $-2\log L$ for the model with no interaction (output given below) and for the model with interaction (output above). The likelihood ratio test statistic is found to be

$$\text{LR} = 4900.106 - 4806.674 = 93.432 \sim \chi^2(81); \quad \text{p-value} = 0.1629.$$

```
proc logistic data=knaekke;
freq count;
class Treat Block;
model Judge = Treat Block/link=glogit;
run;

[part of the output]
```

The LOGISTIC Procedure

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	5194.982	4978.106
SC	5211.660	5194.929
-2 Log L	5188.982	4900.106

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Treat	27	250.5445	<.0001
Block	9	8.4995	0.4847

The output shows that the Wald test for removing the `Block` factor is 8.50 corresponding to a p-value of 48%. The likelihood ratio test presented in the notes yields a p-value of 47% and may be obtained as described above.

The 30 parameter estimates for the final model (7.3) are given as part of the output below but note that all effects are given relative to the group `Treat` $= 10$ which is used as reference (Intercept). The output further shows that the effect of `Treat` is significant (Wald test$= 250.80 \sim \chi^2(27)$, p-value$= 0\%$). Note that SAS uses the fourth response category as reference. Thus, the output above displays estimates for $\log(p_{ij}/p_{i4}), j = 1, 2, 3, i = 1, \ldots, 40$.

```
proc logistic;
freq count;
class Treat;
model Judge = Treat/link=glogit;
run;

[Part of the output given below]

          The LOGISTIC Procedure

                          Model Fit Statistics

                                              Intercept
                                 Intercept            and
                  Criterion           Only     Covariates

                  AIC              5194.982       4968.716
                  SC               5211.660       5135.503
                  -2 Log L         5188.982       4908.716

                        Type 3 Analysis of Effects

                                        Wald
                    Effect      DF    Chi-Square    Pr > ChiSq

                    Treat       27      250.8018        <.0001


                                                 Standard          Wald
Parameter       Judge    DF    Estimate            Error    Chi-Square    Pr > ChiSq

Intercept       A         1      0.0298           0.0835        0.1276        0.7209
Intercept       B         1      0.1883           0.0792        5.6569        0.0174
Intercept       C         1      0.7175           0.0725       97.9269        <.0001
Treat     1     A         1     -0.3514           0.2336        2.2633        0.1325
Treat     1     B         1     -0.0485           0.2089        0.0539        0.8163
Treat     1     C         1     -0.1023           0.1899        0.2901        0.5901
Treat     2     A         1      2.1784           0.3095       49.5521        <.0001
Treat     2     B         1      1.3158           0.3226       16.6398        <.0001
Treat     2     C         1      0.9118           0.3177        8.2354        0.0041
```

Treat	3	A	1	-0.6524	0.2507	6.7710	0.0093
Treat	3	B	1	-0.0952	0.2087	0.2081	0.6483
Treat	3	C	1	0.0688	0.1835	0.1405	0.7078
Treat	4	A	1	-0.2735	0.2370	1.3310	0.2486
Treat	4	B	1	0.1711	0.2073	0.6809	0.4093
Treat	4	C	1	0.00233	0.1934	0.0001	0.9904
Treat	5	A	1	-0.5559	0.2365	5.5259	0.0187
Treat	5	B	1	-0.4171	0.2175	3.6782	0.0551
Treat	5	C	1	-0.2250	0.1859	1.4653	0.2261
Treat	6	A	1	2.3310	0.3074	57.4832	<.0001
Treat	6	B	1	0.9431	0.3348	7.9364	0.0048
Treat	6	C	1	0.7176	0.3230	4.9370	0.0263
Treat	7	A	1	-0.8221	0.2354	12.1975	0.0005
Treat	7	B	1	-0.5751	0.2088	7.5876	0.0059
Treat	7	C	1	-0.2689	0.1732	2.4117	0.1204
Treat	8	A	1	-0.6460	0.2294	7.9332	0.0049
Treat	8	B	1	-0.6038	0.2157	7.8389	0.0051
Treat	8	C	1	-0.2601	0.1772	2.1549	0.1421
Treat	9	A	1	-0.5944	0.2242	7.0302	0.0080
Treat	9	B	1	-0.1883	0.1940	0.9419	0.3318
Treat	9	C	1	-0.5062	0.1834	7.6159	0.0058

7.3 Proportional odds models

The proportional odds model is an example of a model for polytomous response data where the explanatory variables have a unified effect on all response categories.

We present the model through an application to the data of example 7.1. Thus,

$$Y_i = (Y_{i1}, \ldots, Y_{i4}) \sim m(n_i, (p_{i1}, \ldots, p_{i4}))$$

are the number of plants belonging to the individual judgement groups for a particular combination of `treatment` and `block`. In the proportional odds model we consider the cumulative probabilities

$$\gamma_{ij} \quad = \quad P(Y_i \leq j), \quad j = 1, 2, 3,$$

and model the logit transform

$$\eta_{ij} = \log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right) \quad = \quad \theta_j - \delta(\texttt{treatment}_i, \texttt{block}_i) \tag{7.5}$$

as a difference between a parameter related to the response group, j, and a parameter for the joint configuration of `treatment` and `block`. The only restriction on the parameters is that we must have $\theta_1 \leq \theta_2 \leq \theta_3$ (which follows since clearly $\gamma_{i1} \leq \gamma_{i2} \leq \gamma_{i3}$).

The big advantage of the proportional odds model lies in the interpretation of the effect of explanatory variables. Suppose for instance that we want to compare treatment groups `treatment`$= 3$ and `treatment`$= 7$ in `block` $= 1$. The change of the (cumulative) odds for response category 1, 2, or 3, when comparing the two treatment groups are

$$\begin{aligned}
\exp\left((\theta_1 - \delta(3,1)) - (\theta_1 - \delta(7,1))\right) &= \exp\left(\delta(3,1) - \delta(7,1)\right) \\
\exp\left((\theta_2 - \delta(3,1)) - (\theta_2 - \delta(7,1))\right) &= \exp\left(\delta(3,1) - \delta(7,1)\right) \\
\exp\left((\theta_3 - \delta(3,1)) - (\theta_3 - \delta(7,1))\right) &= \exp\left(\delta(3,1) - \delta(7,1)\right),
\end{aligned}$$

in particular the treatment effect is the same no matter which response category we consider.

Example 7.1 *(continued)* For the example with eyespot (knækkefodssyge) the proportional odds model described by (7.5) may be tested against the full multinomial logistic regression model given by (7.1). The model (7.5) contains $10 \cdot 4 = 40$ parameters for the levels of `treatment` $\times$ `block` and 3 parameters for the cumulative log-odds. However, as the two sets of parameters enter additively into (7.5) only $40 + 3 - 1 = 42$ parameters may be estimated and the number of degrees of freedom associated with the test is $120 - 42 = 78$. The test statistic turns out to be

$$G^2 = 94.07 \sim \chi^2(78)$$

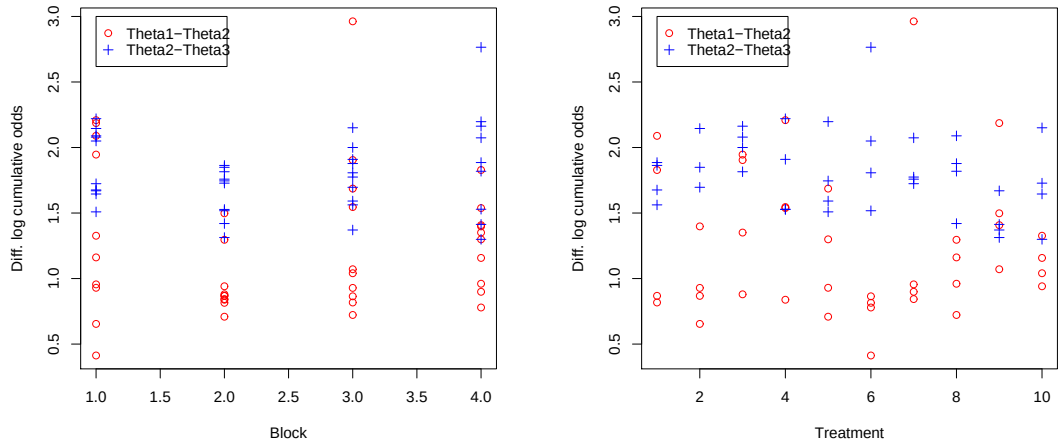
implying that the proportional odds model is accepted with a p-value of 10%. As a graphical check of the model one may calculate estimates, $\hat{\eta}_{ij}$, for the log-odds of cumulative probabilities

$$\eta_{ij} = \log\left(\frac{\gamma_{ij}}{1-\gamma ij}\right), \quad j = 1, 2, 3, \quad i = 1, \ldots, 40.$$

Under the proportional odds model the differences

$$\eta_{i2} - \eta_{i1} = \theta_2 - \theta_1 \quad \text{and} \quad \eta_{i3} - \eta_{i2} = \theta_3 - \theta_2$$

should be independent of i (this is the explanatory variables).



The figure above displays the estimated differences for the 40 levels of `treatment` × `block` against each of the individual factors. The proportional odds model requires that the level of the points does not depend in a systematic way on explanatory variables. This is clearly confirmed from the plot.

We may continue by considering the hypothesis H_0 :

$$\log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right) \quad = \quad \theta_j - (\alpha(\texttt{treatment}_i) + \beta(\texttt{block}_i)) \tag{7.6}$$

about the effect of `treatment` and `block` entering linearly on the scale of log-odds to cumulative probabilities. The model contains $3+10+4=17$ parameters of which only 15 may be estimated. The test of H_0 against the full proportional odds model (7.5) therefore follows a χ^2-distribution with $42-15=27$ degrees of freedom. The test statistic is

$$G^2 = 31.25 \sim \chi^2(27)$$

yielding a p-value of 26 %.

Continuing the analysis we find that the `block` factor may be removed

$$G^2 = 4.69 \sim \chi^2(3) \qquad p = 20\%$$

but that the `treatment` effect is highly significant with

$$G^2 = 252.3 \sim \chi^2(9) \qquad p = 0\%.$$

Our final model is given by

$$\log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right) \quad = \quad \theta_j - \alpha(\texttt{treatment}_i). \tag{7.7}$$

Choosing `treatment` $=1$ as reference group the estimated effect of `treatment` with approximate 95 %-confidence intervals become

Parameter	Estimate	95 %-conf. int
$\hat{\alpha}(2) - \hat{\alpha}(1)$	-1.387	[-1.758,-1.019]
$\hat{\alpha}(3) - \hat{\alpha}(1)$	0.156	[-0.201,0.513]
$\hat{\alpha}(4) - \hat{\alpha}(1)$	-0.084	[-0.443,0.275]
$\hat{\alpha}(5) - \hat{\alpha}(1)$	0.191	[-0.181,0.563]
$\hat{\alpha}(6) - \hat{\alpha}(1)$	-1.713	[-2.101,-1.331]
$\hat{\alpha}(7) - \hat{\alpha}(1)$	0.343	[-0.019,0.706]
$\hat{\alpha}(8) - \hat{\alpha}(1)$	0.286	[-0.081,0.653]
$\hat{\alpha}(9) - \hat{\alpha}(1)$	0.079	[-0.286,0.444]
$\hat{\alpha}(10) - \hat{\alpha}(1)$	0.231	[-0.134,0.596]

We observe that the only `treatment` groups with a clear effect compared to the reference group is `treatment` $= 2$ and `treatment` $= 6$. From model (7.7) we deduce that for any response category, j, the change of the cumulative odds by replacing `treatment`$= 1$ with either `treatment`$= 2$ or `treatment`$= 6$ is given by

$$\begin{aligned} \hat{OR}_{21} &= \frac{\exp(\theta_j - \alpha(2))}{\exp(\theta_j - \alpha(1))} = \exp(\alpha(1) - \alpha(2)) = \exp(1.387) = 4.00 \quad [2.77, 5.80] \\ \hat{OR}_{61} &= \frac{\exp(\theta_j - \alpha(6))}{\exp(\theta_j - \alpha(1))} = \exp(\alpha(1) - \alpha(6)) = \exp(1.713) = 5.55 \quad [3.78, 8.17]. \end{aligned}$$

In particular, the odds that a plant is judged better than a given category will with a probability of 95 % be 4 to 8 times larger if it is exposed to `treatment` number 6 rather than to `treatment` number 1. □

Formulating a statistical model for polytomous data in terms of log-odds of cumulative probabilities may not seem to be the most obvious thing to do.

However, one motivation for studying the proportional odds model is that it arises naturally as a so-called threshold model. To see this assume that the true outcome of the experiment is given by independent continuous random variables $Z_1, \ldots, Z_n$ who are not directly observable. Instead (or as a direct consequence of the way the data is collected) the individual responses are grouped according to a number of thresholds

$$\theta_1 < \theta_2 < \ldots < \theta_{J-1}.$$

and the analysis must be based solely on records of the discretized data. Assume for instance that $Z_i \sim N(\alpha_i, 1)$ and let γ_{ij} be the probability that Z_i is smaller than θ_j. Then

$$\gamma_{ij} = P(Z_i < \theta_j) = P(Z_i - \alpha_i < \theta_j - \alpha_i) = \Phi(\theta_j - \alpha_i),$$

where Φ is the cumulative distribution function for a standard normal distribution. We deduce that

$$\Phi^{-1}(\gamma_{ij}) = \theta_j - \alpha_i,$$

and since Φ^{-1} and the logit function tt logit: $p \to \log(p/1-p)$ are almost identical (see figure of section 7.2) this is almost equivalent to

$$\log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right) = \theta_j - \alpha_i.$$

We conclude that the assumption of the latent variables Z_i being normally distributed automatically leads to a proportional odds model for the cumulative probabilities of the related threshold model. To be very precise we get a proportional odds model with the `probit` ($= \Phi^{-1}$) link function and not the `logit` link function used in the definition (7.5).

7.3.1 R-programs and output

Example 7.1 *(continued)*

For a description of the data set we refer to the R-program in section 7.2.1.

Fit proportional odds models

Proportional odds models are fitted using `polr()` in the `MASS` package. Initially we bring the data on a suitable form by replicating the data so that each dataline corresponds to one plant and contains the variables `x` (response with four levels: 1-4), `Tvec` (treatment), and `Bvec` (block). The first 10 datalines look like

```
      x Tvec Bvec
 [1,] 1    1    1
 [2,] 1    1    1
 [3,] 1    1    1
 [4,] 1    1    1
 [5,] 2    1    1
 [6,] 2    1    1
 [7,] 2    1    1
 [8,] 2    1    1
 [9,] 2    1    1
[10,] 2    1    1
```

The full proportional odds model with interaction between `Tvec` and `Bvec` is fitted by

```
> library(MASS)
> pomod.full=polr(x~Tvec*Bvec)
```

where of course the first line is only to be run once for each R session.

A test for effect of the interaction `Tvec` $\times$ `Bvec` can be produced by

```
> pomod.add=polr(x~Tvec+Bvec)
> anova(pomod.add,pomod.full,test="Chisq")
Likelihood ratio tests of ordinal regression models

Response: x
        Model Resid. df Resid. Dev   Test    Df   LR stat.   Pr(Chi)
1 Tvec + Bvec      1904   4931.992
2 Tvec * Bvec      1877   4900.743 1 vs 2    27   31.24874 0.2610487
```

Obtaining parameter estimates

The final model turns out to be the one with only a main effect of `Tvec`. Parameter estimates and confidence intervals may be obtained as follows

```
> pomod.treat=polr(x~Tvec)
> summary(pomod.treat)
```

```
Re-fitting to get Hessian

Call:
polr(formula = x ~ Tvec)

Coefficients:
             Value Std. Error    t value
Tvec2  -1.38717731  0.1883990 -7.3629766
Tvec3   0.15585034  0.1818799  0.8568860
Tvec4  -0.08362947  0.1831287 -0.4566705
Tvec5   0.19082717  0.1898748  1.0050155
Tvec6  -1.71329713  0.1963044 -8.7277560
Tvec7   0.34349571  0.1848669  1.8580705
Tvec8   0.28596909  0.1870551  1.5287960
Tvec9   0.07889693  0.1863414  0.4234000
Tvec10  0.23099318  0.1860247  1.2417337

Intercepts:
    Value    Std. Error t value
A|B  -1.5940   0.1388   -11.4844
B|C  -0.4562   0.1331    -3.4278
C|D   1.3109   0.1371     9.5587

Residual Deviance: 4936.683
AIC: 4960.683

> confint(pomod.treat)

             2.5 %     97.5 %
Tvec2  -1.75808654 -1.0192749
Tvec3  -0.20069012  0.5125035
Tvec4  -0.44273121  0.2753637
Tvec5  -0.18130204  0.5632525
Tvec6  -2.10051545 -1.3306232
Tvec7  -0.01875220  0.7061552
Tvec8  -0.08060793  0.6528828
Tvec9  -0.28632210  0.4443697
Tvec10 -0.13358844  0.5958591
```

In the output `Tvec6` refers to the effect of treatment group number 6 compared to treatment group 1 which is used as reference. The `Intercepts` are estimates for the parameters θ_j in the proportional odds model.

7.3.2 SAS-programs and output

Example 7.1 *(continued)*

For a description of the data set we refer to the SAS-program in section 7.2.2.

Fit proportional odds model

The full proportional odds model with interaction between `Treat` and `Block` is fitted by

```
proc genmod data=knaekke;
weight count;
class Treat Block;
model Judge = Treat Block Treat*Block/dist=mult link=cumlogit type3;
run;
```

```
[part of the output]
```

```
                              The GENMOD Procedure

                    Criteria For Assessing Goodness Of Fit

Criterion                 DF           Value        Value/DF

Log Likelihood                     -2450.3716

                 LR Statistics For Type 3 Analysis

                                        Chi-
Source              DF     Square    Pr > ChiSq

Treat                9     256.91        <.0001
Block                3       5.70        0.1273
Treat*Block         27      31.25        0.2610
```

When fitting the proportional odds model it is important to specify the `link` = `cumlogit` option. Note that since our response is polytomous we need to select the multinomial distribution for the response (`dist` = `mult`). We read off that

$$-2 \cdot \log L = -2 \cdot (-2450.3716) = 4900.7432.$$

For the full model we have from section 7.2.2 that $-2 \log L = 4806.674$ hence the likelihood ratio test statistic form testing the proportional odds model become

$$LR = 4900.7432 - 4806.674 = 94.0692 \sim \chi^2(120 - 42); \quad \text{p-value} = 10\%.$$

The output further shows that the interaction `Treat` $*$ `Block` may be removed.

Running `proc genmod` with the interaction removed from the `model` line yields (among others) the output

```
  LR Statistics For Type 3 Analysis

                                        Chi-
Source              DF     Square    Pr > ChiSq

Treat                9     253.30        <.0001
Block                3       4.69        0.1958
```

showing that `Block` has no significant effect of the judgement of the plant.

Obtaining parameter estimates

Finally running the program lines

```
proc genmod data=knaekke;
weight count;
class Treat;
model Judge = Treat /dist=mult link=cumlogit type3;
estimate 'LogOR21' Treat -1 1 /exp;
estimate 'LogOR61' Treat -1 0 0 0 0 1 /exp;
run;
```

[part of the output]

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept1		1	-1.8251	0.1407	-2.1007	-1.5494	168.36	<.0001
Intercept2		1	-0.6872	0.1344	-0.9506	-0.4239	26.16	<.0001
Intercept3		1	1.0799	0.1361	0.8131	1.3467	62.93	<.0001
Treat	1	1	0.2310	0.1860	-0.1336	0.5956	1.54	0.2143
Treat	2	1	1.6182	0.1895	1.2468	1.9895	72.94	<.0001
Treat	3	1	0.0751	0.1822	-0.2820	0.4323	0.17	0.6801
Treat	4	1	0.3146	0.1837	-0.0454	0.6746	2.93	0.0867
Treat	5	1	0.0402	0.1902	-0.3326	0.4129	0.04	0.8327
Treat	6	1	1.9443	0.1974	1.5574	2.3312	97.00	<.0001
Treat	7	1	-0.1125	0.1851	-0.4753	0.2502	0.37	0.5433
Treat	8	1	-0.0550	0.1873	-0.4221	0.3121	0.09	0.7691
Treat	9	1	0.1521	0.1868	-0.2140	0.5182	0.66	0.4154
Treat	10	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Treat	9	252.30	<.0001

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
LogOR21	1.3872	0.1884	0.05	1.0179	1.7564	54.21	<.0001
Exp(LogOR21)	4.0035	0.7543	0.05	2.7674	5.7917		
LogOR61	1.7133	0.1963	0.05	1.3285	2.0980	76.17	<.0001
Exp(LogOR61)	5.5471	1.0889	0.05	3.7755	8.1501		

shows that `Treat` has a significant effect and display parameter estimates. If the use the parameteriza-

tion

$$\log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right) \quad = \quad \theta_j - \alpha(\texttt{treatment}_i).$$

of model (7.7) the output contains estimates of $\theta_1, \theta_2, \theta_3$ and the contrasts

$$\alpha(10) - \alpha(1), \alpha(10) - \alpha(2), \ldots, \alpha(10) - \alpha(9).$$

The lines

```
estimate 'LogOR21' Treat -1 1 /exp;
estimate 'LogOR61' Treat -1 0 0 0 0 1 /exp;
```

make SAS give estimates and confidence intervals for the log-odds ratio and the odds ratio of the effect of `treament` 2 or 6 in relation to `treatment` 1.

Chapter 8

Exercises

8.1 Juiciness of peas

In a sensoric experiment a number of assessors were asked to evaluate the juicyness of 15 batches of peas on a continuous scale from zero to 15. Small values indicate dryness and large values juiciness. The average assessments are listed in the last column in Table 8.1. Also measured for each batch were the starch concent (percentage of dry matter) and the content of sucrose (g per 100 g fresh weight). In the table the starch and sucrose measurements after subtraction of the average values (18.10 for starch and 5.62 for sucrose) are listed.

Batch	Starch	Sucrose	Juiciness
1	−2.80	0.04	8.00
2	−2.98	−0.27	8.57
3	4.04	−0.33	4.37
4	2.59	−0.17	6.57
5	−0.91	0.84	7.34
6	−2.98	0.56	8.60
7	−2.75	0.01	9.59
8	−2.28	−0.96	3.81
9	−0.75	0.99	7.55
10	4.80	−1.36	3.73
11	−0.45	0.59	8.77
12	2.87	−0.29	4.85
13	2.71	−0.11	6.72
14	−0.45	0.48	7.62
15	−0.66	−0.02	9.60

Table 8.1: The pea data: starch content, sucrose content and average assessment of juiciness. The average values 18.10 and 5.62 have been subtracted from starch and sucrose, respectively.

1. The interest is on prediction of the assessment of juiciness from concent of starch and sucrose. What would be a reasonable model for this purpose? Write up the model formally. What is the interpretation of the parameters?
2. A model is fitted with `R` and `SAS` below, and some output is given, too. Read off the parameter

estimates. Do both starch and sucrose contribute significantly to the assessment of juicyness?

3. What is the expected difference in juiciness for two batches with the same sucrose content but with a difference in starch content of 2%? Give also an 95% confidence interval for the expected difference.

4. Consider the plots in Figures 8.1 and the residual plots and the Cook's plot in 8.2. Is the model is appropriate for the data? Are there highly influential observations?

5. The same model as before is fitted again but without observation no. 8. Does the conclusion change? How would you proceed the data analysis?

6. What is the expected juiciness of a batch of peas with 20% starch in dry matter and a sucrose content of 6 g per 100 g fresh weight? Usually we would supply such an expectation/prediction with a confidence interval or a prediction interval. Explain the difference between these two intervals. Which is the widest?

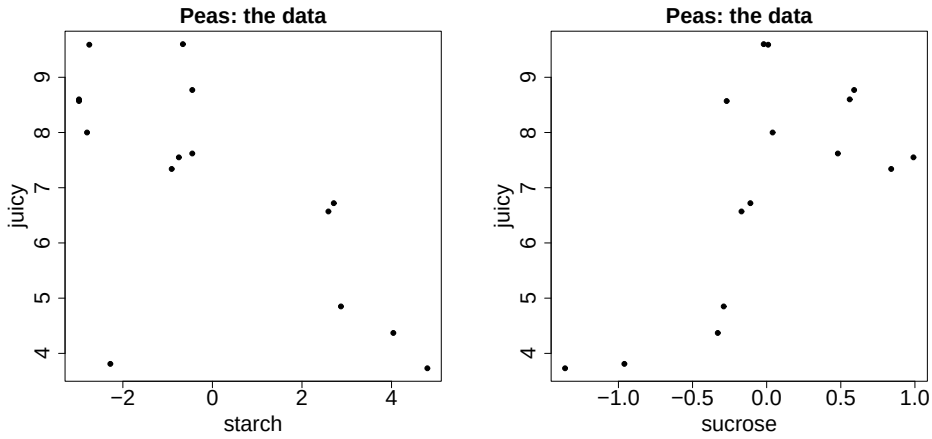


Figure 8.1: Plot of juiciness against starch and sucrose for the peas data.

R output

```
> model1 = lm(juicy ~ starch + sucrose)
> summary(model1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0460     0.3477  20.262 1.20e-10 ***
starch       -0.3428     0.1471  -2.330   0.0381 *
sucrose       1.4596     0.6255   2.333   0.0378 *

Residual standard error: 1.347 on 12 degrees of freedom

> confint(model1)
                 2.5 %      97.5 %
(Intercept)  6.28833783  7.80366217
starch      -0.66344217 -0.02224553
```

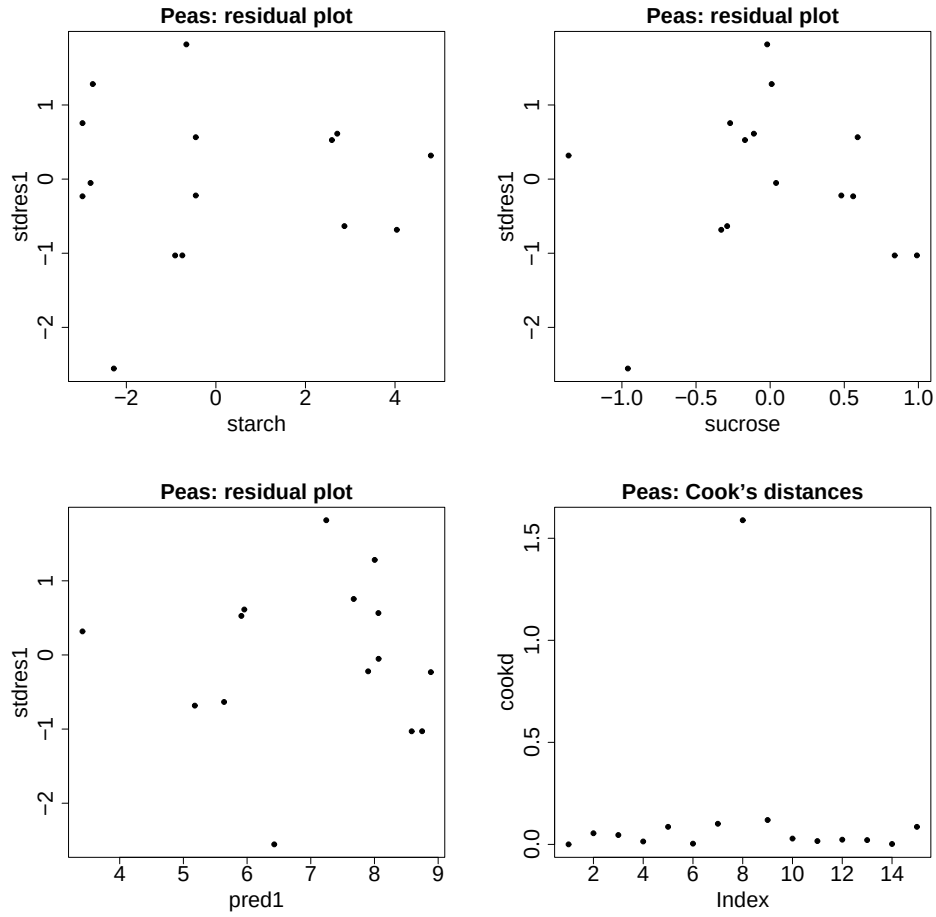


Figure 8.2: Peas: residual plots and plot of Cook's distances for the multiple regression model of juiciness on starch and sucrose.

```
sucrose      0.09673893  2.82243753

> starch2 = starch[-8]
> sucrose2 = sucrose[-8]
> juicy2 = juicy[-8]

> model2 = lm(juicy2 ~ starch2 + sucrose2)
> summary(model2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.3477     0.2590  28.370 1.22e-11 ***
starch2      -0.5597     0.1198  -4.672  0.00068 ***
sucrose2      0.2998     0.5450   0.550  0.59326

Residual standard error: 0.9498 on 11 degrees of freedom
```

SAS programs and output

```
proc glm data = peas;
  model juicy = starch sucrose / clparm ;
  output out = glmout predicted = p student = stdres cookd = cook;
run;

data peas2;
  set peas;
  if juicy = 3.81 then delete;
run;

proc glm data = peas2;
  model juicy = starch sucrose;
run;
```

with output:

```
                                The GLM Procedure

                          Number of Observations Read          15
                          Number of Observations Used          15

Dependent Variable: juicy

                                        Sum of
        Source                     DF       Squares     Mean Square    F Value    Pr > F
        Model                       2    34.30209735    17.15104868       9.46    0.0034
        Error                      12    21.76626265     1.81385522
        Corrected Total            14    56.06836000


        Source                     DF    Type III SS     Mean Square    F Value    Pr > F
        starch                      1     9.84719873     9.84719873       5.43    0.0381
        sucrose                     1     9.87660077     9.87660077       5.45    0.0378

                                   Standard
   Parameter         Estimate          Error    t Value    Pr > |t|      95% Confidence Limits
   Intercept      7.046000000     0.34774083      20.26      <.0001     6.288337829  7.803662171
   starch        -0.342843847     0.14714358      -2.33      0.0381    -0.663442166 -0.022245527
   sucrose        1.459588229     0.62550086       2.33      0.0378     0.096738926  2.822437532


                                The GLM Procedure

                          Number of Observations Read          14
                          Number of Observations Used          14

Dependent Variable: juicy

                                        Sum of
        Source                     DF       Squares     Mean Square    F Value    Pr > F
        Model                       2    34.92484849    17.46242425      19.36    0.0002
        Error                      11     9.92383722     0.90216702
        Corrected Total            13    44.84868571


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

```
starch                     1      19.69334472      19.69334472       21.83     0.0007
sucrose                    1       0.27298093       0.27298093        0.30     0.5933


                                                 Standard
Parameter                 Estimate                  Error     t Value    Pr > |t|
Intercept              7.347739455            0.25899902       28.37      <.0001
starch                -0.559725002            0.11980045       -4.67      0.0007
sucrose                0.299813165            0.54503955        0.55      0.5933
```

8.2 Outdoor Running World Records

The world records on ten distances (outdoor running) are listed for men and women in Table 8.2. The records were taken from the webiste of the International Association of Athletics Federation (IAAF), `http://www.iaaf.org` on October 18, 2006. You are not going use the dates for the records in the following, only the distances and the times.

	Men		Women	
Distance (m)	Time (sec)	Date	Time (sec)	Date
100	9.77	14/06/05	10.49	16/07/88
200	19.32	01/08/96	21.34	29/09/88
400	43.18	26/08/99	47.60	06/10/85
800	101.11	24/08/97	113.28	26/07/83
1500	206.00	14/07/98	230.46	11/09/93
3000	440.67	01/09/96	486.11	13/09/93
5000	757.35	31/05/04	864.53	03/06/06
10000	1577.53	26/08/05	1771.78	08/09/93
21097.5	3535.00	15/01/06	4004.00	15/01/99
42195	7495.00	28/09/03	8125.00	13/04/03

Table 8.2: World records on October 18, 2006.

We want to examine the dependence on the record (the time) of the distance, and to examine the difference between men and women. In particular we want answers to the following questions:

1. What is the expected increment in the world record when the distance is doubled? And is the answer the same for all distances? And the same for men and women?
2. How much faster are men compared to women? And does the difference depend on the distance?

First of all, you have to consider how to analyse the data in order to be able to answer these questions. You may consider the following questions:

- What kind of model is appropriate for these data?
- Plot time against distance and log-time against log-distance, perhaps for men and women seperately. Do you think it is a good idea to transform the data?

- Identify the parameter(s) that have to do with the dependence of distance. What is the interpretation of the parameter(s)? What does it tell you about the first "research question"?
- Identify the parameter(s) that have to do with the difference between the sexes. What is the interpretation of the parameter(s)? What does it tell you about the second "research question"?

Finally, an extra question:

3. On the IAAF-webpage it also appears that the world records for 100 km is 22413 seconds (6 hours, 13 minutes, 33 seconds) for men and 23591 second (6 hours, 33 minutes and 11 seconds) for women. How does this compare to the data material and the model you have used?

8.3 Growth of turkeys

The following experiment was carried out in order to investigate the effect of vitamin A on growth of turkey. 48 turkeys were randomly allocated to six groups of eight turkeys, and each group was given a certain amount of vitamin A in their feed. The amount of vitamin A and the average weekly increment in logarithmic weight for each group are listed in Table 8.3.

Vitamin A	Increment of log-weight
1.5	0.159
3.0	0.226
6.0	0.299
12.0	0.316
24.0	0.330
48.0	0.298

Table 8.3: The turkey data: amount of vitamin A in feed (IU/g) and average weekly increment in logarithmic weight (gram).

1. Plot the average increment in logarithmic weight against the logarithm to the amount of vitamin A.
2. Write up a model for the data. Compute estimates and confidence intervals for the parameters.
3. Estimate the optimal amount of vitamin A in the sense that the expected increment in logarithmic weight is the largest possible.
4. Do you have any objections against this analysis?

8.4 Phosphor in plants

In a field experiment the concentration of phosphor available for plant was measured for each of 18 plants. Moreover, the concentration of unorganic phosphor was chemically determined and the concentration of an organic phosphor component was measured for each plant. The primary interest is to describe the concentration of phosphor available for the plants as a function of the concentration of unorganic and organic phosphor. The results are listed in Table 8.4.

Analyze the data!

Unorganic	Organic	Plant
0.4	53	64
0.4	23	60
3.1	19	71
0.6	34	61
4.7	24	54
1.7	65	77
9.4	44	81
10.1	31	93
11.6	29	93
12.6	58	51
10.9	37	76
23.1	46	96
23.1	50	77
21.6	44	93
23.1	56	95
1.9	36	54
26.8	58	168
29.9	51	99

Table 8.4: Concentration of phosphor in plants.

This includes proper specification of the model, model validation, check for strongly influential observations, model reduction (if possible), parameter estimation and interpretation, conclusion of the analysis.

8.5 Accumulation of drug in liver

An experiment with rats was carried out in order to investigate the accumulation of a certain drug in the liver. Each rat was given a dose of the drug, approximately proportional to their bodyweight. After a period the rats were slaughtered, their livers weighed and the drug dose in the liver was measured. The result for the 19 rats are given in Table 8.5.

Analyze the data.

This includes proper specification of the model, model validation, check for strongly influential observations, model reduction (if possible), parameter estimation and interpretation, conclusion of the analysis.

8.6 Yield of barley

The following experiment was carried out in a greenhouse. 15 pots were sown with barley seeds: 3, 7, 15, 34, 77 barley seeds per pot, respectively, with three pots for each number of barley seeds. After harvest, the total fresh weight yields (in grams) was measured for each pot. The results are listed in Table 8.6.

1. Plot the yield against the number of barley seeds and test for identical variances in the five groups.

2. Consider instead the logarithmic yield as response, and repeat question 1 with this response.

Bodyweight	Liver weight	Dose	Dose in liver
176	6.5	0.88	0.42
176	9.5	0.88	0.25
190	9.0	1.00	0.56
176	8.9	0.88	0.23
200	7.2	1.00	0.23
167	8.9	0.83	0.32
188	8.0	0.94	0.37
195	10.0	0.98	0.41
176	8.0	0.88	0.33
165	7.9	0.84	0.38
158	6.9	0.80	0.27
148	7.3	0.74	0.36
149	5.2	0.75	0.21
163	8.4	0.81	0.28
170	7.2	0.85	0.34
186	6.8	0.94	0.28
146	7.3	0.73	0.30
181	9.0	0.90	0.37
149	6.4	0.75	0.46

Table 8.5: Accumulation of drug in liver.

No. of seeds	Yield		
3	7.5	9.8	9.0
7	18.8	27.7	27.1
15	64.7	30.2	37.0
34	84.3	110.0	71.2
77	125.8	85.7	91.9

Table 8.6: The barley data.

We want to use the following non-linear model for the relationship between the number of barley seeds (x) and the logarithmic yield (y):

$$y \approx a - b \cdot e^{-cx} \tag{8.1}$$

3. What is the interpretation of the parameters a and b? (*Hint: what happens for* $x = 0$ *and* x *very large?*) Suggest starting values for a and b.

4. We also need a starting value for c. From the graph, choose a pair (x, y) which you believe satisfy equation (8.1), that is, a pair nearby the graph. Use this point, the starting values from question 3, and equation (8.1) to compute a starting value for c.

5. Fit the non-linear model to data, and report the estimates.

6. Model validation: Make a test of the non-linear regression model against the oneway ANOVA with five groups. Make also a plot of observed and fitted values as well as a residual plot. What is the conclusion regarding the appropriateness of the model?

8.7 Production of milk powder

In an experiment about production of milk powder two factors were varied: water activity on three levels (1, 2, 3) and temperature while drying (100, 110, 120, 140 Celcius degrees). Only 9 of the 12 combinations were tested in the experiment.

There were three replications in the experiment, in the sense that milk powder was prepared in three rounds. This gives 27 samples of milk powder in total. Each of these was stored and measurements were taken after 4, 6 and 8 weeks. Each time the concentration of maillard reaction products as well as a sensoric taste score (high values means good taste) were measured. The data are listed in Table 8.7.

First consider only measurements from week4, that is, the first 27 observations.

1. Set up a model for the concentration of maillard reactions products: which factors are relevant; should they be fixed or random? Draw also the corresponding factor diagram.
2. Analyze the data in order to investigate the effect of water activity and temperature on the concentration of maillard reaction products.

From now on, consider the full dataset with 81 observations.

3. Set up a model for the concentration of maillard reaction products. Draw also the corresponding factor diagram.

   Hint: Which (three-factor) interaction corresponding to the grouping of the 81 observations into the 27 different samples?

   Another hint: make sure to keep track of the ordering of the factors.

4. Write a few lines of `R` code or `SAS` code that would fit the model. Consider how you would try to reduce the model.

   Hint for `R`*-users:* if there are two random factors `fac1` and `fac2` where `fac1` is coarser than `fac2`, then the model may be fitted with `lme` as follows:

   ```
   lme(response ~ linear part, random =~ 1|fac1/fac2)
   ```

5. A particular analysis ended up with a final model which is fitted with `R` and `SAS` below. Write up the corresponding model. Use the output to estimate the variance parameters. Also, estimate the expected concentration of maillard reaction products for water activity level 2, after 4 weeks of storage at 140 Celcius degress.
6. Make an analysis of the taste score variable in order to examine how temperature, water activity and storage time affects the taste. What is the conclusion?

`R` programs and output for question 5.

```
> sample = rnd:temp:water
> model5a = lme(maillard ~ water+week+temp, random=~ 1|rnd/sample)

> summary(model5a)
Linear mixed-effects model fit by REML

Fixed effects: maillard ~ water + week + temp
                Value  Std.Error DF   t-value p-value
```

Round	Week	Maillard	Taste	Water	Temp	Round	Week	Maillard	Taste	Water	Temp
1	4	2.90	10.1	1	100	2	6	2.11	11.2	3	100
1	4	2.13	11.0	1	110	2	6	1.98	11.8	3	110
1	4	2.00	11.1	1	120	2	6	2.20	11.0	3	140
1	4	2.13	11.1	2	100	3	6	2.20	7.0	1	100
1	4	2.38	11.9	2	120	3	6	2.34	10.7	1	110
1	4	2.56	10.7	2	140	3	6	2.49	10.3	1	120
1	4	2.60	10.8	3	100	3	6	2.63	9.7	2	100
1	4	1.91	11.0	3	110	3	6	3.06	9.0	2	120
1	4	2.27	10.8	3	140	3	6	3.28	9.6	2	140
2	4	2.19	11.0	1	100	3	6	2.34	10.2	3	100
2	4	2.32	11.0	1	110	3	6	2.51	9.2	3	110
2	4	2.41	11.6	1	120	3	6	2.77	10.2	3	140
2	4	2.49	11.1	2	100	1	8	2.39	9.6	1	100
2	4	2.61	11.7	2	120	1	8	2.41	9.8	1	110
2	4	2.63	10.8	2	140	1	8	2.71	11.4	1	120
2	4	2.06	11.0	3	100	1	8	2.49	11.2	2	100
2	4	1.98	10.0	3	110	1	8	2.06	11.2	2	120
2	4	2.27	11.2	3	140	1	8	3.10	9.8	2	140
3	4	2.13	10.1	1	100	1	8	2.32	10.8	3	100
3	4	2.13	9.4	1	110	1	8	2.29	9.4	3	110
3	4	2.22	10.7	1	120	1	8	2.72	12.0	3	140
3	4	2.80	8.3	2	100	2	8	2.27	11.0	1	100
3	4	2.77	10.9	2	120	2	8	2.25	11.2	1	110
3	4	2.99	9.2	2	140	2	8	2.46	9.6	1	120
3	4	1.98	10.3	3	100	2	8	2.53	9.2	2	100
3	4	1.98	9.3	3	110	2	8	2.70	11.0	2	120
3	4	2.20	10.5	3	140	2	8	2.81	11.6	2	140
1	6	2.13	10.0	1	100	2	8	2.20	11.8	3	100
1	6	2.34	10.5	1	110	2	8	2.15	10.6	3	110
1	6	2.49	11.2	1	120	2	8	2.41	11.4	3	140
1	6	2.41	10.8	2	100	3	8	2.41	9.6	1	100
1	6	2.85	11.2	2	120	3	8	2.42	9.0	1	110
1	6	2.84	11.2	2	140	3	8	2.73	10.2	1	120
1	6	2.24	8.4	3	100	3	8	3.33	7.8	2	100
1	6	2.06	11.4	3	110	3	8	3.25	9.4	2	120
1	6	2.42	11.6	3	140	3	8	3.75	9.6	2	140
2	6	2.20	9.3	1	100	3	8	2.80	10.6	3	100
2	6	2.27	11.3	1	110	3	8	2.81	10.2	3	110
2	6	2.49	11.7	1	120	3	8	3.06	10.0	3	140
2	6	2.34	11.2	2	100						
2	6	2.70	10.8	2	120						
2	6	2.61	11.0	2	140						

Table 8.7: The milk powder data.

```
(Intercept)  2.1929136 0.11340945 52 19.336250  0.0000
water2       0.2828889 0.08142711 19  3.474136  0.0025
water3      -0.0908889 0.08142711 19 -1.116199  0.2783
week6        0.1207407 0.05777774 52  2.089745  0.0415
week8        0.2885185 0.05777774 52  4.993593  0.0000
temp110     -0.0461111 0.08636649 19 -0.533900  0.5996
temp120      0.1058889 0.08636649 19  1.226041  0.2352
temp140      0.2907778 0.08636649 19  3.366789  0.0032

> VarCorr(model5a)
            Variance     StdDev
rnd =       pdLogChol(1)
(Intercept) 0.020328503  0.14257806
sample =    pdLogChol(1)
(Intercept) 0.009841698  0.09920533
Residual    0.045066615  0.21228899
```

SAS **programs and output for question 5.**

```
proc mixed data=milk nobound;
  class water week temp rnd;
  model maillard = water week temp / ddfm=satterth solution;
  random rnd temp*water*rnd;
run;
```

with output

The Mixed Procedure

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
1	1	18.30982410	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Estimate
rnd	0.02033
water*temp*rnd	0.009842
Residual	0.04507

Fit Statistics

-2 Res Log Likelihood	18.3
AIC (smaller is better)	24.3
AICC (smaller is better)	24.7
BIC (smaller is better)	21.6

Solution for Fixed Effects

Effect	water	week	temp	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept				2.6813	0.1170	6.07	22.92	<.0001
water	1			0.09089	0.08143	19	1.12	0.2783
water	2			0.3738	0.08143	19	4.59	0.0002
water	3			0	.	.	.	.
week		4		-0.2885	0.05778	52	-4.99	<.0001
week		6		-0.1678	0.05778	52	-2.90	0.0054
week		8		0	.	.	.	.
temp			100	-0.2908	0.08637	19	-3.37	0.0032
temp			110	-0.3369	0.09973	19	-3.38	0.0032
temp			120	-0.1849	0.09973	19	-1.85	0.0793
temp			140	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
water	2	19	11.46	0.0005
week	2	52	12.58	<.0001
temp	3	19	4.91	0.0108

8.8 Disease in cucumbers

A greenhouse experiment was carried out to investigate how the spread of a disease ("agurkesyge") in cucumbers depends on climate and on the amount of fertilizer for two varieties. The following data (kindly supplied by Eigil de Neergård, Department of Plant Pathology, KVL) are an extract from the experiment. Two climates were used: 1 (change to day temperature 3 hours before sunrise) and 2 (normal change to day temperature). Three amounts of fertilizer were applied: normal (2.0 units), high (3.5 units), and extra high (4.0 units). The two varieties were Aminex and Dalibor.

At a certain time the plants were "standardized" to have equally many leaves, and then (on day 0, say) the plants were contaminated with the disease. On 8 particular subsequent days the amount of infection (in percent) was registered. From the resulting curve of infection two summary measures were calculated (in a way not specified here), namely the rate of spread of the disease, and the level of infection at the end of the period.

There were 3 blocks each consisting of 2 sections, a section being a part of the greenhouse. Each section consisted of 3 plots, which were each divided into 2 subplots, each of which had 6-8 plants. Thus there were a total of 36 subplots. Results were recorded for each subplot.

The experimental factors were randomly allocated to the different units as follows: the 2 climates were allocated to the 2 sections within each block, the 3 amounts of fertilizer were allocated to the 3 plots within each section, and finally the 2 varieties were allocated to the 2 subplots within each plot. Thus, in summary, there were

- 3 blocks
- 2 sections per block (given 2 different climates)
- 3 plots per section (given 3 different amounts of fertilizer)
- 2 subplots per plot (with 2 different varieties)

The results are given in Table 8.8.

1. Write up a statistical model and draw the corresponding factor diagram.
2. Analyze the end level of disease in order to investigate the effect of the different factors in the experiment.
3. Analyze the rate of spread in order to investigate the effect of the different factors in the experiment.

8.9 Tenderness of pork

In an experiment concerning chilling of pork two chilling methods (tunnel-chilling and fast-chilling) were compared. 24 porks were sampled from two pH-groups (high and low pH, 12 porks from each). After slaughtering the 24 porks were divided into two sides. One side was tunnel-chilled, the other fast-chilled. After some time the tenderness of the 48 meat pieces was measured. The data is listed in Table 8.9. The experiment was made by Anders Juel Møller, KVL.

Analyze the data in order to investigate the effect of pH and chilling method on tenderness of pork meat.

8.10 Summary measure analysis of the growth of rats data

Consider the growth of rats data from Example 5.1.

Use the increment in log-weight from week 1 to week 5 as response and carry out an analysis to see if there is an effect of treatment on the increment.

In particular:

- create the variable with the increments.
- what is a reasonable model?
- is there a significant effect of treatment on the increments?
- be careful with the conclusion: quantify the treatment differences (if any). For each treatment group, estimate the factor with which the weight of a rat increases from week 1 to week 5.

8.11 Random intercepts analysis of the growth of rats data

Consider again the growth of rats data from Example 5.1. Analyze the data with a random intercepts model with rat as a random effect. This includes considering the following issues:

- What is a reasonable model for the fixed part of the model? You may get inspiration from the subject profile plot (left plot of Figure 5.1).
- Reduce the model as much as possible.
- What is the conclusion about treatment differences? Report relevant estimates from the final model. Could you think of a way to graphically report the results of the analysis?

Block	Section	Climate	Fertilizer	Variety	End level	Rate
1	1	2	2.0	aminex	48.8981	0.06915
1	1	2	2.0	dalibor	42.2463	0.06595
1	1	2	3.5	aminex	48.2108	0.04679
1	1	2	3.5	dalibor	41.6767	0.04881
1	1	2	4.0	aminex	55.4369	0.04025
1	1	2	4.0	dalibor	40.9562	0.04859
1	2	1	2.0	aminex	51.5573	0.09353
1	2	1	2.0	dalibor	36.7739	0.10353
1	2	1	3.5	aminex	47.9937	0.05327
1	2	1	3.5	dalibor	47.8723	0.04397
1	2	1	4.0	aminex	57.9171	0.05225
1	2	1	4.0	dalibor	37.7185	0.09324
2	3	2	2.0	aminex	60.1747	0.04182
2	3	2	2.0	dalibor	45.6937	0.06983
2	3	2	3.5	aminex	51.0017	0.08863
2	3	2	3.5	dalibor	52.2796	0.03622
2	3	2	4.0	aminex	51.1251	0.05875
2	3	2	4.0	dalibor	48.7217	0.08169
2	4	1	2.0	aminex	51.6001	0.07001
2	4	1	2.0	dalibor	50.4463	0.09907
2	4	1	3.5	aminex	48.3387	0.05788
2	4	1	3.5	dalibor	38.6538	0.06834
2	4	1	4.0	aminex	51.3147	0.05695
2	4	1	4.0	dalibor	38.2488	0.07908
3	5	1	2.0	aminex	49.6958	0.07218
3	5	1	2.0	dalibor	29.6786	0.11351
3	5	1	3.5	aminex	46.6692	0.08825
3	5	1	3.5	dalibor	36.5892	0.09107
3	5	1	4.0	aminex	56.0320	0.04532
3	5	1	4.0	dalibor	36.0955	0.08712
3	6	2	2.0	aminex	45.9790	0.08882
3	6	2	2.0	dalibor	37.2489	0.12796
3	6	2	3.5	aminex	40.7277	0.06418
3	6	2	3.5	dalibor	38.4831	0.08540
3	6	2	4.0	aminex	44.5242	0.06215
3	6	2	4.0	dalibor	34.3907	0.09651

Table 8.8: The cucumber data.

Pork	pH-group	Tunnel	Fast
1	low	7.22	5.56
2	low	3.11	3.33
3	low	7.44	7.00
4	low	4.33	4.89
5	low	6.78	6.56
6	low	5.56	5.67
7	low	7.33	6.33
8	low	4.22	5.67
9	low	3.89	4.00
10	low	5.78	5.56
11	low	6.44	5.67
12	low	8.00	5.33
13	high	8.44	8.44
14	high	7.11	6.00
15	high	6.00	5.78
16	high	7.56	7.67
17	high	5.11	4.56
18	high	8.67	8.00
19	high	5.78	7.67
20	high	6.11	5.67
21	high	7.44	7.56
22	high	7.67	6.11
23	high	8.00	8.22
24	high	8.78	8.44

Table 8.9: The tenderness of pork data

8.12 A test for the thyroxin effect across weeks

Consider the repeated measures analysis of the growth of rats data from Example 5.1. As mentioned in the text, a test of the thyroxin effect across all weeks could have been appropriate. What is the hypothesis? What is the corresponding reduced model? How would you carry out the test in practice? Do it.

8.13 Growth of guinea pigs

In order to investigate the effect of vitamin E on growth of guinea pigs 15 animals were followed during a 7 week experimental period. In week 1 all animals were given a growth inhibiting substance and in week 5 their feed was supplemented by varying doses (0, low, high) of vitamin E. There were five animals in each treatment group. The weight of each animal was recorded at the end of weeks 1, 3, 4, 5, 6 and 7. The data is given in Table 8.10.

		Week					
Dose	Animal	1	3	4	5	6	7
0	1	455	460	510	504	436	466
	2	467	565	610	596	542	587
	3	445	530	580	597	582	619
	4	485	542	594	583	611	612
	5	480	500	550	528	562	576
low	6	514	560	565	524	552	597
	7	440	480	536	484	567	569
	8	495	570	569	585	576	677
	9	520	590	610	637	671	702
	10	503	555	591	605	649	675
high	11	496	560	622	622	632	670
	12	498	540	589	557	568	609
	13	478	510	568	555	576	605
	14	545	565	580	601	633	649
	15	472	498	540	524	532	583

Table 8.10: The growth of guinea pigs data

1. Make some illustrative plots of the data.
2. Use the weight increment as a summary statistic and analyze it in order to see if there is an effect of vitamin E.
3. Write up a model for all the data (repeated measurements) and test whether it can be reduced to a model with random intercepts.
4. Reduce the systematic part of the model (if possible).
5. Formulate conclusions in the final model.

8.14 Activity of rats

The data comes from an investigation of the effect of a certain type of exposure to the activity of rats. The experimental unit was a cage with two rats. During the entire experimental period the rats were daily exposed to the matter under investigation, in one of three concentrations (treatment 1, 2 and 3, respectively). Once per month month during 10 months the activity of the rats was measured by placing the rats from one cage in a chamber in which each intersection of a light beam was counted. The total count through a period of 57 hours was used as the result for that cage. The data is listed in Table 8.11.

In the following, use the logarithmic counts for the analysis.

1. Make some illustrative plots of the data.
2. Choose a summary statistic and analyze it in order to see if there is an effect of treatment.
3. Write up a model for all the data (repeated measurements). Fit the model.
4. Can the model be reduced to a model with random intercepts?
5. Reduce the systematic part of the model.

   Hint: There are a number of possibilities. Is the interaction between month and treatment significant? Are all treatments significantly different? Can the relation between month and rat activity be simplified?
6. What is the conclusion regarding the effect of the treatments? Report the relevant estimates from the final model. What would be a good way to report the results?

8.15 Photosynthesis in pines

An experiment with 40 shoots of pines was carried out in order to investigate the effect of a salt treatment on photosynthesis. The 40 shoots came from 20 families, 10 of which were placed in a green house and 10 were placed outside. From each family two shoots were selected at random. One shoot was treated with salt, the other was not. Phosynthesis was measured before the treatment and again 1, 4, 7, 15 and 29 days after the treatment. The experiment was carried out by Anders Rebild, Arboretet, KVL.

The results from the experiment are listed in Table 8.12. In the table, treatment 2 is the salt treatment and treatment 1 is control. Note that observations are missing from four shoots from the green house.

1. Make some illustrative plots of the data.
2. Choose a summary statistic and analyze it in order to see if there is an effect of treatment. Remember to include other relevant variables in the analysis.
3. Write up a model for all the data (repeated measurements). How can the photosynthesis measurement before treatment be included in the model?
4. Fit the model.
5. Can the model be reduced to a model with random intercepts?
6. Is it possible to reduce the fixed effects part of the model?
7. What is the conclusion regarding the effect of salt treatment on photosynthesis? Report relevant estimates from the final model. What would be a good way to report the results?

8.16 Slagteriernes Svinesundhedstjeneste

When inspection of meat at the slaughteries reveals too many problems with a particular pig breeder the breeder is offered a visit by *Slagteriernes Svinesundstjeneste (SST)*. At present the expenses of the visits are covered by a mandatory fee payed to the slaugtheries by all breeders. To examine whether the breeders are willing to pay a fee directly to SST for the visit an investigation has been made. The data has been grouped according to the size of the herd of pigs and whether the breeder has previously been offered a visit by SST (Data from S. Andersen: Analyse af tælledata, 1998).

Size of herd	Offered visit	Willing to pay fee	
		yes	no
< 500	yes	8	10
< 500	no	58	131
500-1000	yes	21	30
500-1000	no	15	33
> 1000	yes	12	17
> 1000	no	6	14

1. Examine how the willingness to pay a fee to SST depends on the two explanatory variables.
2. Formulate a conclusion of the statistical analysis where you report relevant odds ratios accompagnied by 95%−confidence intervals.

8.17 Mortality of beetles exposed to CS_2

Eight groups each consisting of approximately 60 beetles have been exposed to different doses of the gas CS_2 for a period of 5 hours. At the end of the experiment the number of living beetles has been observed. The data comes from ***Bliss C. I. (1935): *The calculation of the dosage-mortality curve* and is shown below.

Dose (mg/l)	No. of beetles	No. of deaths
49.05689	59	6
52.99074	60	13
56.91150	62	18
60.84151	56	28
64.75898	63	52
68.69103	59	53
72.61060	62	61
76.54203	60	60

1. Calculate log odds for each of the eight groups and plot these against the dose.
2. Does the plot confirm that the data can be described by a logistic regression model with dose as covariate:
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta \cdot \texttt{dose}_i.$$
3. Perform a statistical test of the model indicated under question 2. Explain how to derive the number of degrees of freedom for the test.
4. Answer questions 1.-3. where you consider $\log(\texttt{dose})$ instead of `dose`.

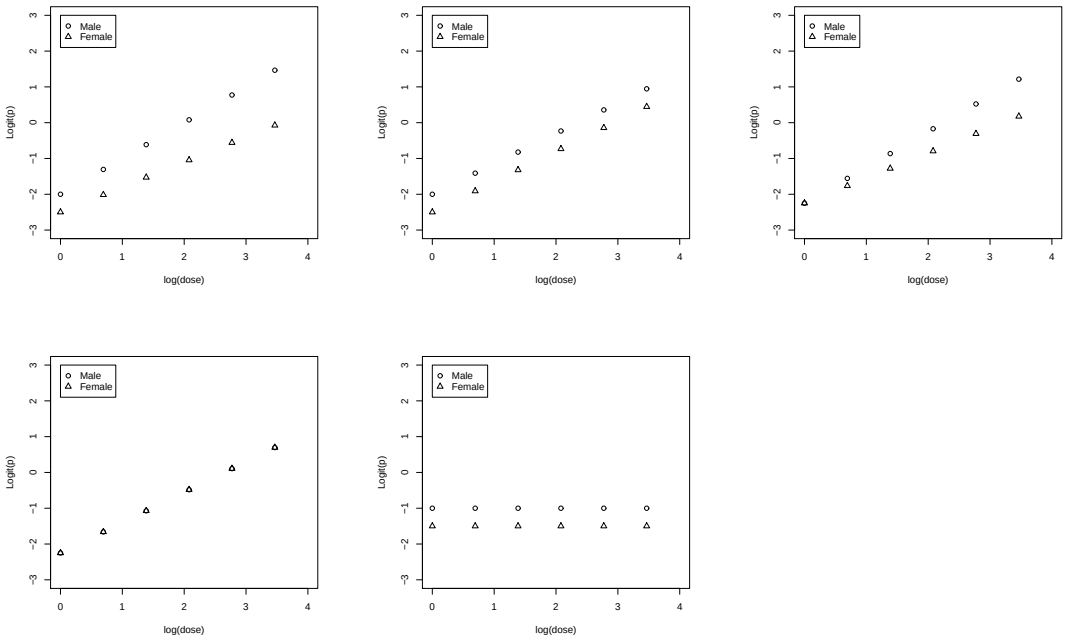
5. Formulate a conclusion of the statistical analysis where you include parameter estimates and confidence intervals for relevant parameters.
6. Give an estimate and a 95 %-confidence interval for the probability that a beetle receiving a dose of 60 mg/l will die in the experiment.
7. LD_{50} is the dose that will kill 50% of the beetles. Find an estimate for LD_{50}.

8.18 Effect of insecticides on moths: submodels of the full logsitic model

Consider the insecticide data of example 7.2. The logistic regression model

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha(\texttt{sex}_i) + \beta(\texttt{sex}_i)\cdot\log(\texttt{dose}_i)$$

expresses that the log-odds depends linearly on the logarithm of the `dose`, and that a line is fitted for each level of `sex` (both intercept and slope depend on `sex`).



Do the following for each of the models 1-5 below: find the plot corresponding to the model, find the number of (free) parameters and write some R or SAS cose that fits the model.

1. $$\log\left(\frac{p_i}{1-p_i}\right) = \alpha(\texttt{sex}_i) + \beta(\texttt{sex}_i)\cdot\log(\texttt{dose}_i)$$
2. $$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta(\texttt{sex}_i)\cdot\log(\texttt{dose}_i)$$

3.
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha(\texttt{sex}_i) + \beta \cdot \log(\texttt{dose}_i)$$

4.
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta \cdot \log(\texttt{dose}_i)$$

5.
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha(\texttt{sex}_i)$$

8.19 Different link functions

Both R and SAS allow to fit regression models for binary response data using the link functions (see for instance the plots in the end of section 7.2.)

$$\begin{aligned} \text{logit} &: p \to \log\left(\frac{p}{1-p}\right) \\ \text{probit} &: p \to \Psi^{-1}(p) \\ \text{cloglog} &: p \to \log(-\log(1-p)). \end{aligned}$$

Here the `logit`-link function corresponds to the the usual logistic regression model. Consider the data set from exercise 8.17 and denote by $\hat{p}_i, i = 1, \ldots, 8$, the fraction of dead beetles for the eight different doses of CS_2. Above answer the following questions for each of the above link functions.

1. Make a plot of $f(\hat{p}_i)$ against $\log(\texttt{dose}_i)$.
2. Calculate the p-value for the test of the logistic regression model with link function f, where $\log(\texttt{dose})$ is used as a covariate. Which link function gives the best description of the data?
3. What is the effect of increasing $log(\texttt{dose})$ by 0.1?

8.20 Experiment with two different diets

20 persons have participated in an experiment where two different diets are to be compared. By randomization 10 persons have been assigned to each diet and every week a weight gain or weight loss has been observed. The observations are the number of weeks where the diet resulted in a weight loss for each of the 20 persons in the experiment. The table below displays the results for a period of eight weeks showing the number of person for each combination of diet and weeks with weight loss.

Weeks with weight loss	0	1	2	3	4	5	6	7	8
diet 1	1	0	2	0	1	1	2	0	3
diet 2	2	1	0	1	2	1	2	1	0

1. Use a logistic regression model to test whether there is a difference between the two diets.
2. Discuss why the assumptions of the logistic regression model might not be satisfied and come up with an alternative way to analyse the data.

8.21 Moth experiment with three different response groups

Consider the experiment concerning the effect of insecticide on moths discussed in example 7.2. The explanatory variables are `sex` and `logdose` and we stress that the latter may (and will) play the role as a factor *or* as a covariate at different places of the present exercise.

Suppose that the experiment was carried out by classifying each of the moths into one of the three groups: `unaffected` by insecticide, `collapsed` but alive, or `dead`. Denote by

$$Y_i = (Y_{i1}, Y_{i2}, Y_{i3}), \quad i = 1, \ldots, 12,$$

the responses for each of the 12 combinations of `sex` and `logdose` and assume that the Y_i's are independent and multinomially distributed

$$Y_i \sim m(20, (p_{i1}, p_{i2}, p_{i3})).$$

Logistic regression for polytomous response

Consider the log-odds

$$\log\left(\frac{p_{ij}}{p_{i1}}\right) = \eta_{ij}, \quad j = 2, 3, \quad i = 1, \ldots, 12,$$

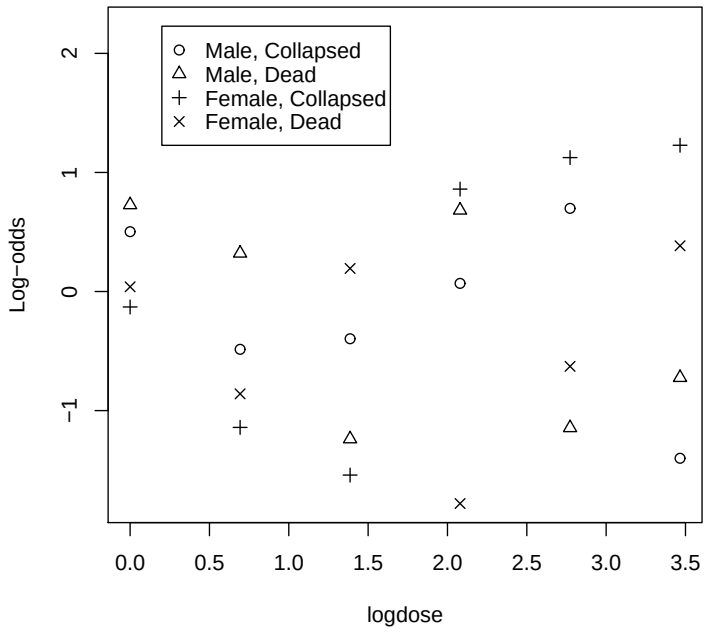
wrt. response category number 1 and discuss the following two questions for each of the models described below.

The questions below must be answered for each of the following models

1. $$\eta_{ij} = \alpha(\texttt{sex}_i, \texttt{logdose}_i, j)$$
2. $$\eta_{ij} = \alpha(\texttt{sex}_i, j) + \beta(\texttt{logdose}_i, j)$$
3. $$\eta_{ij} = \alpha(\texttt{sex}_i, j)$$
4. $$\eta_{ij} = \beta(\texttt{logdose}_i, j)$$
5. $$\eta_{ij} = \alpha(j)$$
6. $$\eta_{ij} = \alpha(\texttt{sex}_i, j) + \beta(\texttt{sex}_i, j) \cdot \texttt{logdose}_i$$
7. $$\eta_{ij} = \alpha(j) + \beta(\texttt{sex}_i, j) \cdot \texttt{logdose}_i$$
8. $$\eta_{ij} = \alpha(\texttt{sex}_i, j) + \beta(j) \cdot \texttt{logdose}_i$$
9. $$\eta_{ij} = \alpha(j) + \beta(j) \cdot \texttt{logdose}_i$$

where $j = 2, 3$ and $i = 1, \ldots, 12$.

Question 1
Make a plot with `logdose` on the x-axis and each of the 24 log-odds (η_{ij}) on the y-axis that indicates the restriction put forth by the particular model. Use different plotting symbols for each of the four combinations of `sex` and the response group $(j = 2, 3)$. The plot for a completely unrestricted model (model 1 below) may look like



Question 2
Discuss how many parameters that may be estimated for each model and give examples of models that are submodels of each other.

8.22 Effect of different substitutes on the taste of cheese

In order to examine the effect of 4 different substitutes on the taste of cheese 52 referees where asked to classify the taste into one of 9 response groups. The data is taken from McCullagh and Nelder: "Generalized Linear Models".

	Taste category									
Cheese	1	2	3	4	5	6	7	8	9	Total
A	0	0	1	7	8	8	19	8	1	52
B	6	9	12	11	7	6	1	0	0	52
C	1	1	6	8	23	7	5	1	0	52
D	0	0	0	1	3	7	14	16	11	52
Total	7	10	19	27	41	28	39	25	12	208

Table 8.13: Number of referees placing the different cheeses in each of the 9 taste categories. Bad taste corresponds to category 1 whereas good taste corresponds to group 9.

1. Fit a logistic regression model to the data.
2. Show by a test that the data may be described by a proportional odds model.
3. Verify the choice of the proportional odds model by calculating the cumulative odds ratios and plotting the logarithm of these against estimated thresholds.
4. Is there an effect of the treatment factor?
5. Give the conclusion of the analysis by presenting relevant odds ratios under your final model. How many parameters does the model contain?

8.23 Difference between fertilizers

In the following exercise we consider a thought field experiment where we wish to compare the effect of two different fertilizers A and B. The experiment is implemented by using each of the products on 50 plants and after a fixed period of time the condition of each plant is evaluated on a scale ranging from 1 to 5 with 5 being best.

Evaluation	1	2	3	4	5
Fertilizer A	10	4	16	15	5
Fertilizer B	20	15	5	2	8

The table above displays the number of plants receiving each grade at the final evaluation. Use a polynomial distribution

$$Y_i = (Y_{i1}, \ldots, Y_{i5}) \sim m(50, (p_{i1}, \ldots, p_{i5})), \quad i = 1, 2,$$

to describe the distribution of the evaluation score for the plants corresponding to the two different fertilizers.

1. How many free parameters does the model contain?
2. Formulate in terms of the parameters the hypothesis that the fertilizer does not affect the evaluation of the plants.
3. Show that the test for no effect of the fertilizer is rejected.

So far we have only concluded that there are differences between the groups of plants receiving fertilizer A and B. But which fertilizer is the best and how do we quantify a possible difference?

4. By looking at the tabular above, which fertilizer would you prefer?

Below we try to fit a proportional odds model to the data. Denote by

$$\gamma_{ij} = p_{i1} + \ldots + p_{ij}, \quad j = 1,2,3,4, \quad i = 1,2,$$

the probability that a plant receiving fertilizer i gets an evaluation score less than or equal to j.

5. Make a table of estimated cumulative probabilities γ_{ij} based on the data above.
6. Make a table of estimated log-cumulative odds $\eta_{ij} = \log(\gamma_{ij}/(1-\gamma_{ij}))$ based on the data above.
7. The proportional odds model says that the differences

   $$\eta_{2j} - \eta_{1j}$$

   between the log-cumulative odds are all equal. Calculate these 4 differences from your table of log-cumulative odds.
8. Formally the proportional odds model is formulated by

   $$\eta_{ij} = \theta_j - \alpha_i.$$

   Estimate the parameters of the model. How many free parameters can be estimated? Quantify the difference between the two fertilizers.
9. Show by a test that the data can not be described by a proportional odds model.

As the proportional odds model was rejected we have not yet been able to answer the main question about which fertilizer is to be preferred. Below we try to transform the data by collapsing some of the response categories before redoing the analysis above. As the process of collapsing groups throws away information there is a risk that the transformation will make it impossible to detect differences between the fertilizers. Further, there is no canonical way to collapse groups and different groupings need not lead to the same conclusion. However, when the response categories are ordered (as is the case here!) grouping adjacent groups may be reasonable.

10. Collapse response categories $2+3$ and $4+5$ and make a table that summaries the outcome of the coarser response variable for the two fertilizers.
11. Show that the new data may be described by a proportional odds model.
12. Show that the model may not be reduced any further and conclude that there is a difference between the two fertilizers.
13. Give an estimate and a 95 %-confidence interval for the parameter that quantifies the difference between the two fertilizers.
14. By how many times does the odds of receiving grade 3 or lower increase if we switch from fertilizer A to fertilizer B. Remember to give both an estimate and a corresponding 95 %-confidence interval.

8.24 Pneumoconiosis among coalminers

The following data explores the degree of pneumconiosis in coalface workers as a function of exposure time. Severity of disease is rated into 3 categories. The data are taken from ***Ashford (1959)***.

Period (yr)	CatI (normal)	CatII	CatIII (severe)
5.8	98	0	0
15	51	2	1
21.5	34	6	3
27.5	35	5	8
33.5	32	10	9
39.5	23	7	8
46	12	6	10
51.5	4	2	5

Denote by

$$Y_i = (Y_{i1}, Y_{i2}, Y_{i3}), \quad i = 1, \ldots, 8,$$

the number of workers in each response category for the 8 different values of `Period`. We assume that the Y_i's are independent and follow a multinomial distribution

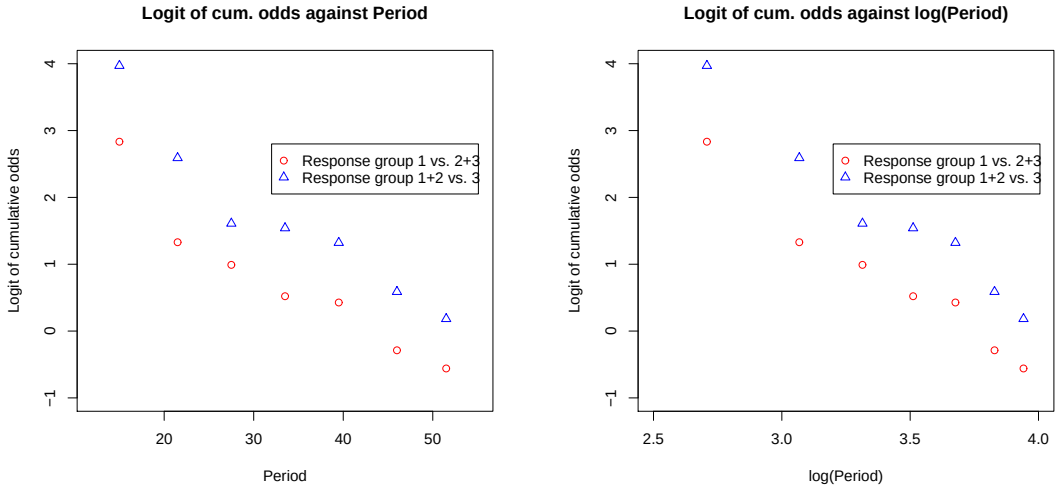
$$Y_i \sim m(n_i, (p_{i1}, p_{i2}, p_{i3})),$$

where n_i is the number of observations for each value of `Period`. We want to describe the data by one of the two proportional odds models given below.

$$\log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right) = \theta_j - \alpha \cdot \texttt{Period}_i, \quad j = 1, 2, i = 1, \ldots 8, \tag{8.2}$$

$$\log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right) = \theta_j - \alpha \cdot \log(\texttt{Period}_i), \quad j = 1, 2, i = 1, \ldots 8, \tag{8.3}$$

The only difference between the models is that for the latter the covariate `Period` has been transformed by the logarithm.



Discuss by considering the figure above

- if `Period` should be log-transformed
- if you think a proportional odds model would describe the data well

Analyse the data by answering the questions below.

1. Examine by a test whether the data can be described by the proportional odds model (8.3).
2. Give estimates and 95 %-confidence intervals for the (three) parameters of the model.
3. Does the exposure time (`Period`) affect the prevalence of pneumoconiosis?
4. Estimate the odds of having severe pneumoconiosis after 20 years of exposure.
5. Estimate the odds of having pneumoconiosis (CatII+CatIII) after 5, 10, and 20 years of exposure.

Note: The model (8.3) expresses that doubling the exposure time has the effect of increasing *either of the odds*

$$\frac{\gamma_{i1}}{1-\gamma_{i1}} \quad \textit{or} \quad \frac{\gamma_{i2}}{1-\gamma_{i2}}$$

by a factor $2^{-\alpha}$.

		Month									
Treat.	Cage	1	2	3	4	5	6	7	8	9	10
1	1	20584	15439	17376	14785	11189	10366	8725	9974	9576	6849
1	3	23265	16956	16200	12934	13763	11893	9949	10490	8674	7153
1	5	17065	12429	14757	10524	11783	8828	9016	9635	8028	8099
1	7	19265	19316	20598	16619	16092	13422	10532	10614	9466	9494
1	9	21062	14095	13267	12543	12734	12268	12219	11791	10379	8463
1	11	23456	10939	13270	14089	12986	13723	11878	13338	12442	10094
1	13	13383	11899	12531	15081	14295	13650	9988	11518	11915	7844
1	15	22717	22434	23151	13163	10029	10408	9119	10188	9549	11153
1	17	17437	13950	15535	14199	11540	9568	8481	9143	8117	5765
1	19	18546	12520	15394	10137	9218	7343	6702	7173	7257	5708
2	37	18536	16827	19185	12445	13227	10412	9855	9169	9639	6853
2	39	18831	14043	16493	12562	10397	8568	8599	8818	6011	5062
2	41	15016	13765	16648	14537	13929	10778	9897	9225	9491	5523
2	43	22276	15497	22024	15616	12440	11454	10290	9456	9567	7003
2	45	18943	14834	18403	16232	13085	12679	10489	9495	10896	8836
2	47	13598	10233	13392	10457	9236	8847	9445	9501	8509	5656
2	49	20498	22136	22094	19825	18157	11452	14809	14564	14503	10643
2	51	19586	12710	12745	7294	15757	15296	14097	14308	13933	10210
2	53	11474	8108	17714	16795	17364	16766	15016	13475	14349	8698
2	55	10284	10760	15628	10692	8420	5842	6138	10271	8435	4486
3	73	18459	15805	19924	18337	24197	18790	19333	22234	18291	11595
3	75	16186	11750	16470	18637	14862	14695	14458	14228	12909	9079
3	77	9614	8319	11375	9446	13157	11153	10540	11476	8976	6123
3	79	15688	15016	20929	12706	17351	15089	14605	15952	14795	10434
3	81	15864	13169	20991	20655	19763	19180	19003	18172	15025	11790
3	83	17721	14489	19085	21333	17011	16148	15280	14762	15745	10477
3	85	17606	7558	15646	15194	13036	10316	8172	8977	8378	3962
3	87	34907	29247	35831	15093	9754	10061	9042	11732	8716	4922
3	89	15189	14046	14909	14713	14999	14201	13184	13073	14639	10330
3	91	16388	14538	17548	19416	22034	17761	14488	16068	14773	10595

Table 8.11: The activity of rats data.

			Day					
Lokality	Family	Treatment	0	1	4	7	15	29
greenhouse	1	1	8.4123	8.8670	7.9576	8.6396	8.1849	7.2755
greenhouse	1	2	10.0257	4.8843	3.4704	3.3419	3.3419	3.4704
greenhouse	2	1	8.4225	9.8930	8.1551	8.4225	7.8877	5.8824
greenhouse	2	2	9.1849	4.7714	2.5050	3.9364	3.8171	3.6978
greenhouse	3	1	7.9542	7.9542	8.7398	7.5614	7.3650	5.7938
greenhouse	3	2	7.1589	4.9977	2.2963	3.2418	1.6209	1.2157
greenhouse	4	2	5.8320	5.1322	4.0824	4.1991	2.7994	2.6827
greenhouse	5	2	5.1792	3.8111	5.2769	3.3225	2.0521	1.7590
greenhouse	6	2	6.7961	6.1783	6.6196	3.5305	4.4131	3.5305
greenhouse	7	2	8.3650	6.7174	7.7313	6.2104	7.3511	6.0837
greenhouse	8	1	10.2015	11.5239	11.3350	10.9572	10.3904	9.4458
greenhouse	8	2	7.4460	6.0048	7.0857	5.8847	6.6053	7.6861
greenhouse	9	1	7.8084	8.1208	8.1208	7.4961	7.4961	7.4961
greenhouse	9	2	7.3911	2.6397	3.5636	4.4875	3.8275	4.3555
greenhouse	10	1	7.9004	9.8485	9.1991	8.2251	7.9004	8.4416
greenhouse	10	2	5.6162	6.0842	2.8081	4.5632	1.5211	2.1061
outside	11	1	10.3995	11.0928	11.0928	10.8947	10.0033	6.9330
outside	11	2	8.1933	8.6660	1.7332	1.0242	1.3393	2.1271
outside	12	1	8.4173	8.7442	8.5808	8.4173	6.7829	5.5571
outside	12	2	11.0723	10.3730	5.1282	3.8462	3.0303	1.9814
outside	13	1	6.2163	5.9677	7.5839	6.3407	2.9838	2.2379
outside	13	2	8.3109	8.5417	5.3097	4.6172	3.6937	1.3851
outside	14	1	10.0572	10.3025	11.5290	9.8119	8.3401	5.8872
outside	14	2	7.7540	7.6203	5.8824	5.2139	4.1444	2.8075
outside	15	1	9.8028	9.4648	9.9718	9.1268	6.7606	4.2254
outside	15	2	9.7605	8.1338	0.4067	0.0000	0.1356	-0.4067
outside	16	1	10.9008	11.8092	11.1279	10.5602	8.8569	7.0401
outside	16	2	8.4703	3.0341	0.5057	0.0000	-0.2528	-0.3793
outside	17	1	10.4496	11.0571	9.7205	8.0194	6.8044	6.4399
outside	17	2	9.6115	8.6504	4.8058	1.9223	0.9612	0.9612
outside	18	1	7.9083	8.3381	7.9943	7.9943	6.7049	4.4699
outside	18	2	10.1396	7.6047	2.3145	1.1021	0.5511	0.1102
outside	19	1	10.1992	11.6415	11.0234	10.8173	8.7569	7.7266
outside	19	2	9.0366	8.8029	5.1415	3.0382	1.8696	0.8569
outside	20	1	9.6980	10.0808	9.8256	9.3152	8.6772	7.0183
outside	20	2	10.2407	10.8972	3.6761	2.1007	0.7877	0.2626

Table 8.12: The photosynthesis data.