Forest biometrics with examples in R

Lauri Mehtätalo
Univ. of Joensuu, Faculty of Forest Sciences

May 16, 2008

II

Contents

Preface

These are lecture notes for course Biometrics: R-statistics in Forest Sciences. In writing these notes, I have used several sources. In writing chapter 1, the main sources have been the second edition of the classical book "Statistical Inference" by George Casella and Roger L Berger(Casella and Berger 2002) and "Methods for Forest Biometrics" by Juha Lappi (Lappi 2006a). In addition, I have used lecture notes of courses on mathematical statistics (by Eero Korpelainen) and statistical inference (by Jukka Nyblom). For chapters 2 and 3, the main sources have been the lecture notes of Jukka Nyblom on Regression analysis and Linear models, the book of Jose Pinheiro and Douglas Bates on Linear models with R (Pinheiro and Bates 2000), and the book "Generalized, Linear and Mixed Models" by CharlesMcCulloch and Shayle R Searle (McCulloch and Searle 2001). For the GLMs and GLMMs, I have used McCulloch and Searle (2001), too. The last chapter on models systems is based mainly on book "Econometric analysis" by William H Greene (Greene 1997). In addition, good sources of information have been the lecture notes of Annika Kangas on Forest Biometrics, the books of of Julian Faraway (Faraway 2004, 2006), Keith E Muller and Paul W Stewart (Muller and Stewart 2006), and William N. Venables and Brian D. Ripley (Venables and Ripley 2002).

Chapter 1

Random variables

1.1 Introduction to random variables

Random variables are real-valued variables with value specified by a random process. The value the random variable gets is called a realization of the random variable. Thus, there is a distinction between the random variable and the realized value, the former being usually denoted by a capital letter and the latter by a lowercase letter.

Example 1.1 *As an example, let us think of trees of a large forest stand s. The diameter of a tree in stand s is a random variable, which can be denoted by X. If we go to the stand s and observe that the diameter of a randomly selected tree is x, we have obtained observation $X = x$.*

There are two types of random variables: discrete and continuous. For a discrete random variable, all possible values can be enumerated. A continuous random variable can get any values between given minimum and maximum values. The minimum and maximum can also be infinite.

Example 1.2 *In the large forest stand, the tree diameter is a continuous random variable, getting any values between $[0, \infty]$. This interpretation assumes that the number of trees in the stand is infinite. In reality, only those values can be obtained that occur in the finite set of trees in the stand.*

Example 1.3 *Let us define random variable Y as tree species. This random variable is discrete, as it can get only values $y \in \{1, 2, 3, 4\}$, where the numbers correspond to "Pine", "Spruce", "Birch" and "Other". Another example of a discrete random variable is Z, the number of trees on a sample plot, which can get only integer values. In this case, all possible values can be numerated, even though the number of possible values is infinite.*

We may also have two or more random variables that are somehow related. For example, we may have observed both diameter and height for a tree, thus resulting our observations to be vectors of length two. The multiple variables may also have been produced by the same process. For example, Y_1 and Y_2 may be random variables "diameter of tree 1" and "diameter of tree 2" from a given stand. The n multivariate random variables are often combined into a n-dimensional random vector.

Example 1.4 *In example 1.3 the observations on tree species and diameter provide observations of a bivariate random vector.*

1.2 Distribution of a random variable

1.2.1 Univariate distribution

The probability of a random variable for getting a specific value is described by the distribution of the random variable. The distribution of a random variable, also called the cumulative distribution function (c.d.f) expresses the probability that the random variable X gets a value that is smaller than or equal to x:

$$F(x) = P(X \leq x) \quad \text{for all } x$$

Apparently, the distribution function of a discrete variable has jumps at the possible values of the random variable, and it is constant elsewhere. In contrast, the distribution function of a continuous random variable is a continuous function.

Example 1.5 *In the large forest stand, we are interested in the distribution of species, and diameter of individual trees. In addition, we are interested about how many trees we would get into a sample plot of size 0.01 ha. The left panel of figure 1.1 shows examples on the distribution functions for tree species, tree diameter and the number of tree in a large forest stand.*

The R-code for Figure 1.1

```
> windows(width=6, height=7)
> par(mfcol=c(3,2),
      mai=c(0.6,0.5,0.1,0.1),
      mgp=c(2,0.7,0))
> y<-c(1,2,3,4)
> Fy<-c(0,0.5,0.8,0.95,1.0)
> plot(stepfun(y,Fy),
+      xlab="y",
+      ylab="F(y)",
+      verticals=FALSE,
+      pch=16,
+      main=NA)
>
> # Diameter distribution, c.d.f.
> x<-seq(0,25,0.1)
> Fx<-pweibull(x,5,15)
```

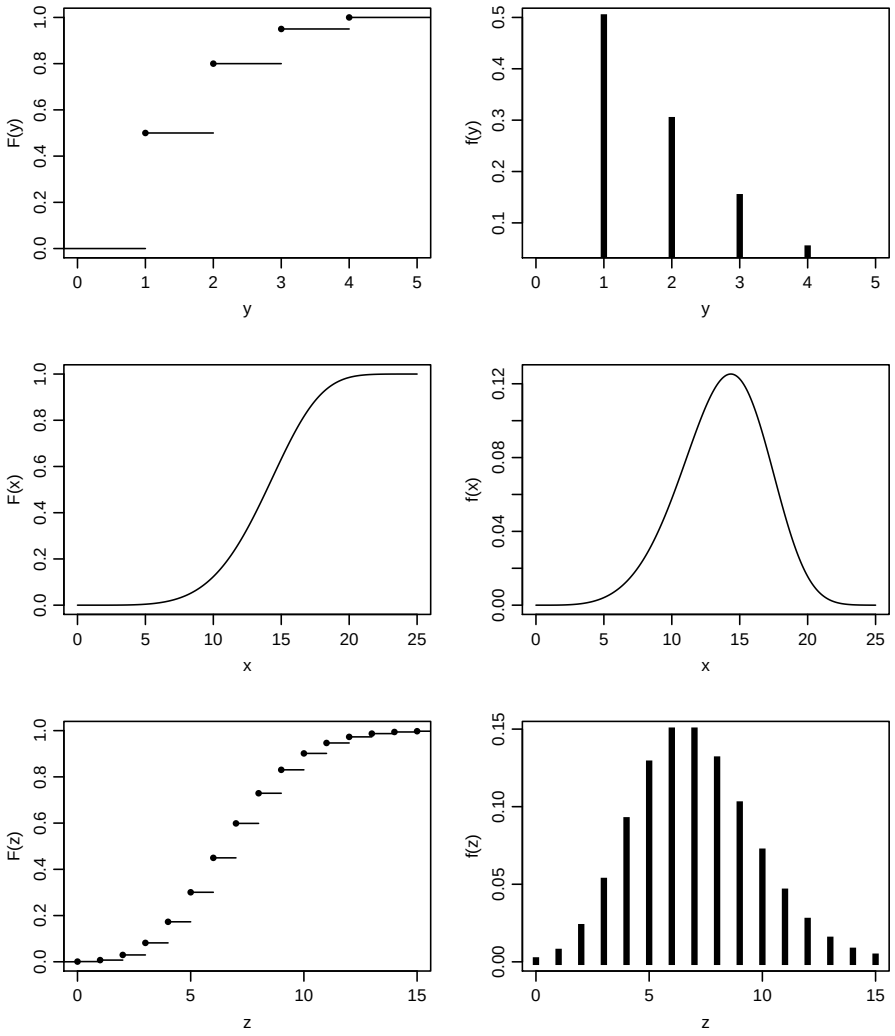


Figure 1.1: The cumulative distribution function of tree species (Y), tree diameter (X), and the number of trees on a sample plot (Z) (left), and the corresponding probability mass functions for Y and Z and probability density function for X.

```
> plot(x, Fx, type="l",
+      xlab="x", ylab="F(x)")
>
> # Number of trees, c.d.f.
> z<-seq(0,20,1)
> Fz<-ppois(z,7)
> plot(stepfun(z,c(0,Fz)),
+      xlab="z",
+      ylab="F(z)",
+      verticals=FALSE,
+      pch=16,
+      main=NA,
+      xlim=c(0,15))
>
> # Tree species, density.
> fy<-c(0.5,0.3,0.15,0.05)
> plot(y, fy, type="n",
+      xlim=c(0,5), xlab="y", ylab="f(y)")
> sapply(1:4,
+        function(i) lines(y[c(i,i)],c(0,fy[i]),lwd=4,lend="square"))
>
> # Diameter distribution, density.
> fx<-dweibull(x,5,15)
> plot(x,fx,type="l",xlab="x",ylab="f(x)")
>
> # Number of trees, density.
> fz<-dpois(z,7)
> plot(z,fz,xlim=c(0,15),type="n",xlab="z",ylab="f(z)")
> sapply(1:20,
+        function(i) lines(z[c(i,i)],c(0,fz[i]),lwd=4,lend="square"))
```

For a function $F(x)$ to be a cdf, the following three conditions need to hold (Casella and Berger 2002).

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

2. $F(x)$ is a nondecreasing function of x.

3. $F(x)$ is right continuous, i.e., for any x_0, $\lim_{x\to x_0} F(x) = F(x_0)$

Vice versa, any function fulfilling the above conditions is a proper distribution function. Thus, the function may have jumps, as the distribution functions of Y and Z in figure 1.1 have. The function may also be a mixture of continuous pieces and jumps. For a continuous random variable, the cdf does not need to be differentiable, and it may be defined by parts.

Example 1.6 *The percentile-based diameter distribution is defined by stating that the 0th, 25th 50th 75th and 100th diameter percentiles are 5, 10, 13, 17, and 24 cm, respectively, and the intermediate values are obtained through interpolating between these percentiles. These assumptions imply that it is assumed that the cdf satisfies* $F(5) = 0$*,* $F(10) = 0.25$*,* $F(13) = 0.5$*,* $F(17) = 0.75$*,* $F(24) = 1$*, and is linear between these.*

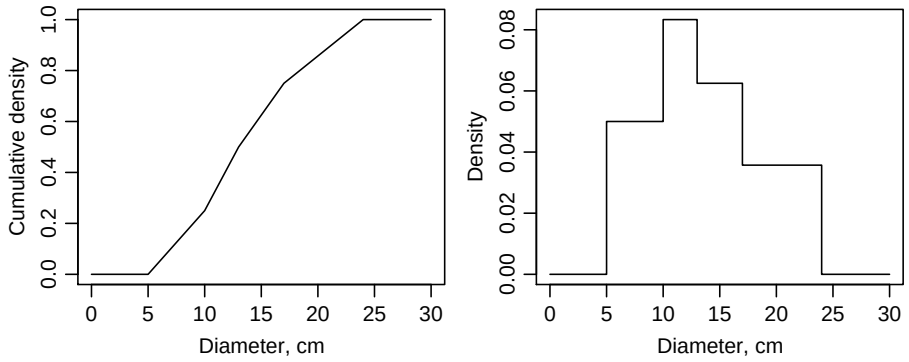


Figure 1.2: Illustration of the percentile-based distribution of Example 1.6.

Thus, the cdf is defined by parts:

$$F(x) = \begin{cases} 0 & x < 5 \\ -0.25 + 0.050x & 5 \le x < 10 \\ -0.58 + 0.083x & 10 \le x < 13 \\ -0.31 + 0.063x & 13 \le x < 17 \\ 0.14 + 0.036x & 17 \le x < 24 \\ 1 & x \ge 24 \end{cases}$$

The left plot of Figure 1.2 shows that the function is increasing, right continuous, and has a limit of 1 as x goes to infinity. Thus, the function is a proper distribution function. The matemathical proof is omitted.

The R-code for Figure 1.2

```
> windows(width=6,height=2.5)
> par(mfcol=c(1,2),
+     mai=c(0.6,0.5,0.1,0.1),
+     mgp=c(2,0.7,0),
+     cex=0.8)
> d<-c(5,10,13,17,24)
> p<-c(0,0.25,0.5,0.75,1)
> a<-(p[-1]-p[-5])/(d[-1]-d[-5])
> b<-p[-5]-a*d[-5]
> plot(c(0,d,30),
+      c(0,p,1),
+      type="l",
+      ylab="Cumulative density",
+      xlab="Diameter, cm")
> plot(c(0,rep(d,each=2),30),
+      c(0,0,rep(a,each=2),0,0),
+      type="l",
+      ylab="Density",
+      xlab="Diameter, cm")
```

The probability for a random variable to get a value between specified limits $[a, b]$ can be computed by subtraction

$$P(a \le\ x \le b) = F(b) - F(a). \tag{1.1}$$

Example 1.7 *In the previous example, the proportion of trees between diameters 20 and 10 cm is*

$$\begin{aligned} P(10 \leq x \leq 20) &= F(20) - F(10) \\ &= 0.14 + 0.036 \times 20 - (-0.25 + 0.050 \times 10) \\ &= 0.61 \end{aligned}$$

It was computed using

```
> F1020<-b[4]+a[4]*20-(b[1]+a[1]*10)
> F1020
[1] 0.6071429
```

The cumulative distribution function of X expresses the probability that the random variable gets the value of x or less. However, we are often interested in the probability that X would get exactly a certain value x. For discrete random variables, this probability is given by the probability mass function. It is defined only for a discrete random variable as

$$f(x) = P(X = x) \text{ for all } x$$

Thus, the pmf gives the point probabilities for the possible values of the discrete random variable. The relationship between pmf and cdf is simply $P(X \leq x) = \sum_{k<x} f(k) = F(x)$.

Example 1.8 *The pdf and pmf of example 1.5 are shown in the right panel of figure 1.1.*

For continuous distributions, a similar relationship holds for continuous distributions, except for replacement of sum operator by integral. For continuous random variables

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)dt\,. \tag{1.2}$$

Stating it in different way, the probability density function (pdf) of continuous random variable x is defined as the first derivative of $F(x)$

$$f(x) = \frac{d}{dx}F(x).$$

An important distinction appears between pdf and pmf. The value of pmf for x gives the probability for random variable X to get value x. With continuous distribution, probability of any distinct value is always 0. For example, the probability of getting a tree with diameter of exactly 20 cm is always zero. However we have a positive probability of getting a tree with diameter between 19.95 and 20.05cm, which would be classified to 20 cm using one millimeter classification in calipering. To compute such probability, we would need to integrate the density over the one-millimeter diameter class. This leads to the use of equation (1.1).

Example 1.9 *The density of percentile-based diameter distribution of example 1.6 is*

$$f(x) = \begin{cases} 0 & x < 5 \\ -0.25 + 0.050 & 5 \leq x < 10 \\ -0.58 + 0.083 & 10 \leq x < 13 \\ -0.31 + 0.063 & 13 \leq x < e17 \\ 0.14 + 0.036 & 17 \leq x < 24 \\ 0 & x \geq 24, \end{cases}$$

which is shown in the right plot of Figure 1.2.

The following conditions for density (or pmf) $f(x)$ can be deduced from the conditions of a cdf.

1. $f(x) \geq 0$ for all x
2. $\sum_x f(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f(x)dx = 1$ (pdf)

In many cases, we are doing transformations to random variables. For example, tree volumes may be transformed to log scale before fitting a volume model. It is important to understand the effect of transformation into the distribution of a random variable. The usually used transformations are one-to-one or one-to-many transformations, which mean that a single value of the original variable X corresponds only one value of Y. This means that the transformation function is always monotonic, either increasing or decreasing. Let X have cdf $F_X(x)$ and define the random variable Y as $g(X)$. The cdf of Y is

$$F_Y(y) = \begin{cases} F_X(g^{-1}(y)) & \text{if} g(x) \text{is increasing} \\ 1 - F_X(g^{-1}(y)) & \text{if} g(x) \text{is decreasing} \end{cases}$$

Example 1.10 *Let X be tree diameter, having cdf according to the Weibull distribution function*

$$F(x|\alpha, \beta) = 1 - \exp\left\{ -\left(\frac{x}{\beta}\right)^{\alpha} \right\},$$

with values $\alpha = 5$ and $\beta = 15$ for the shape and scale parameters, respectively. Assume that tree height Y depends on tree diameter according to the power function

$$Y = g(X) = aX^b,$$

where the parameters are is $a = 8$ and $b = 0.3$. With these values for parameters, the assumed H-D curve is an increasing function of tree diameter. For the distribution of tree heights, we need the inverse of the H-D curve, $g^{-1}(y) = \frac{1}{a} y^{1/b}$. The distribution of tree heights is obtained directly by writing the inverse transformation into the cdf of

diameter. We get

$$F(y|\alpha,\beta,a,b) = 1-\exp\left\{-\left(\frac{\frac{1}{a}y^{1/b}}{\beta}\right)^{\alpha}\right\} \quad (1.3)$$

$$= 1-\exp\left\{-\left(\frac{y}{(a\beta)^b}\right)^{\alpha/b}\right\}$$

$$= F(y|\alpha/b, a\beta^b), \quad (1.4)$$

Which shows that, assuming a Waibull distribution for tree height and a power equation as the H-D curve, the tree height is also distributed according to Weibull distribution with shape and scale parameters α/b and $a\beta^b$. Figure 1.3 illustrates the applied transformations and distributions of diameter and height. Note that this is not a general results, but it holds only for these specific functions. In general, there is no rule that the functional form of the distributions of tree height and diameter would be the same.

The R-code for Figure 1.3

```
> windows(width=6,height=7)
> par(mfcol=c(3,2),mai=c(0.6,0.5,0.5,0.1),mgp=c(2,0.7,0))
> x<-seq(0,25,0.10)
> a<-8
> b<-0.2
> alpha<-5
> beta<-15
> Fx<-pweibull(x,alpha,beta)
> fx<-dweibull(x,alpha,beta)
> # plot the cdf and density of diameter distribution
> plot(x,Fx,xlab="x",ylab=expression(F[X](x)),type="l",main="cdf of diameter")
> plot(x,fx,xlab="x",ylab=expression(f[X](x)),type="l",main="pdf of diameter")
> # define the power function
> powerf<-function(x,a,b) {
+         a*x^b
+         }
> # the inverse of power function
> power.inv<-function(y,a,b) {
+         (y/a)^(1/b)
+         }
> # plot the HD-curve
> plot(x, powerf(x,a,b), type="l",
+      xlab="x", ylab="g(x)" main="HD-curve")
> y<-seq(8,17,0.1)
> Fy<-pweibull(y,alpha/b,a*beta^b)
> fy<-dweibull(y,alpha/b,a*beta^b)
> # plot the cdf and density of height distribution
> plot(y, Fy, type="l",
+      xlab="y, m", ylab=expression(F[Y](y)), main="cdf of height")
> plot(y, fy, type="l",
+      xlab="y, m", ylab=expression(f[Y](y)), main="pdf of height")
> # compute the cdf in another way
> Fy2<-pweibull(power.inv(y,a,b),alpha,beta)
> # Check graphically that the two ways gave the same result
> plot(Fy, Fy2,
+      main=expression("Match between "*F[X]*({g^-1}(y))*" and "*F[Y](y)))
> abline(0,1)
```

Another way of deriving the distribution of a transformation random variables with a continuous density can be stated through pdf. If X has pdf $f_X(x)$ and $g(X)$ is a monotone transformation, with $g^{-1}(y)$ having a continuous derivative, then the pdf of

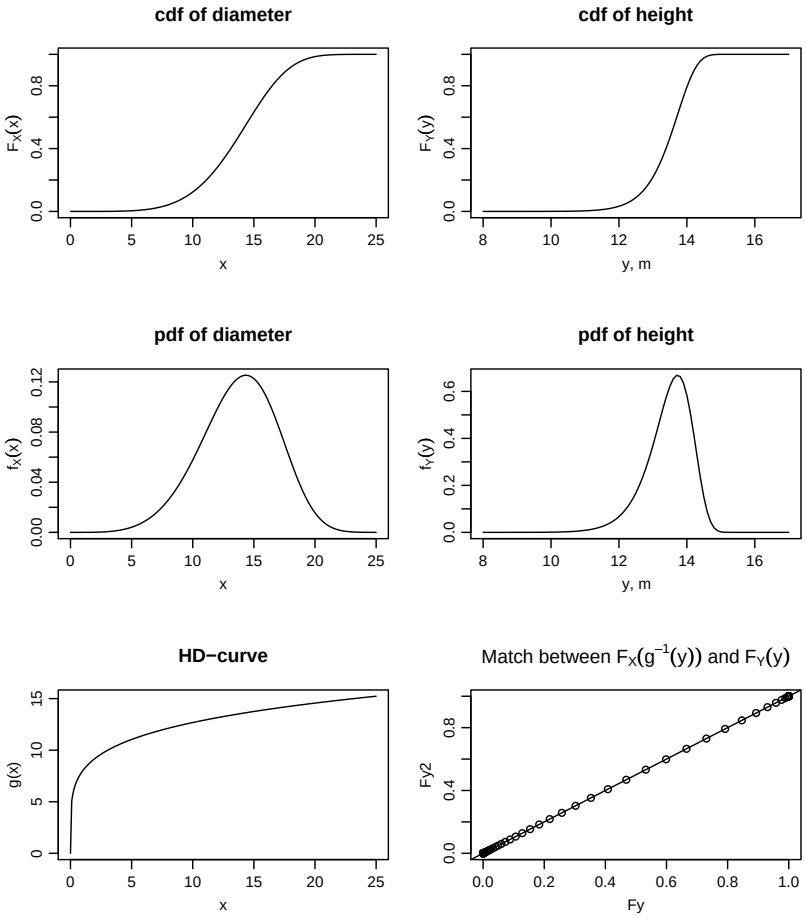


Figure 1.3: Illustration of the example on transformations.

y in the support of Y is

$$f_Y(y) = f_X(g^{-1}(y)) \left\| \frac{d}{dy} g^{-1}(y) \right\| , \tag{1.5}$$

and 0 elsewhere.

Example 1.11 *Forestry student has taken observations on the thickness of Norway spruce needles. The observations are skewed to the right, and are naturally all positive. Based on these observations, she assumes that the needle thickness is distributed according to the lognormal distribution, which has the pdf*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{e^{-(\log(x)-\mu)^2/(2\sigma^2)}}{x}.$$

For modeling purposes, she makes the logarithm transformation $Y = g(X) = \ln(X)$ to the observations. The inverse function of the applied transformation $g^{-1}(y) = e^y$ is continuous and increasing, and has derivative $\frac{d}{dy} g^{-1}(y) = e^y$. The distribution of logarithmic observations is

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{e^y} e^y \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} , \end{aligned}$$

which is the density of normal distribution with mean μ and variance σ^2.

Example 1.12 *The radius of a tree crown between the height of maximal crown radius and total height h can be expressed by an ellipsoid centered at $(x_0 h, y_0 h)$, $x_0 < 1, y_0 < 0$, and having half-axes ah and bh. Having an airborne laser scanning (ALS) in mind, we can assume that the observable crown radius remains constant after that, even though the actual radius decreases. The function based on these assumptions is defined by parts as*

$$Y(z) = \begin{cases} h\,(y_0 + b) & z \le x_0 h \\ h\left\{ y_0 + b\sqrt{1 - \frac{\left(\frac{z}{h} - x_0\right)^2}{a^2}} \right\} & x_0 h < z \le h \\ 0 & z > h \end{cases} , \tag{1.6}$$

where $a = \sqrt{\frac{b^2(1-x_0)^2}{b^2 - y_0^2}}$ is defined so that $Y(z, h)$ passes through point $(h, 0)$, the tree top. The assumed function is illustrated in figure 1.4 using parameter values $h = 20$, $x_0 = 0.2$, $y_0 = -0.1$, and $b = 0.25$.

Airborne laser scanner could be used to produce observations on the height of canopy surface at given points. Let random variable Z describe the height of tree

crown at random location within a single tree crown. The distribution of Z would be the proportion of crown area below z of the total area:

$$P(Z < z) = \frac{A_{max} - A(z)}{A_{max}} = 1 - \frac{A(z)}{A_{max}}.$$

With the ellipsoid tree crown, the maximum area is obtained at height x_0, $A_{max} = \pi h^2(y_0 + b)^2$. The crown area at any height between x_0 and h is $A(z) = \pi Y(z)^2$. The distribution of Z becomes

$$F_Z(z) = 1 - \frac{y_0^2 + 2y_0 b\sqrt{1 - \frac{\left(\frac{z}{h} - x_0\right)^2}{a^2}} + b^2\left(1 - \frac{\left(\frac{z}{h} - x_0\right)^2}{a^2}\right)}{(y_0 + b)^2}$$

The density is obtained by differentiating with respect to z

$$f_Z(z) = \frac{1}{(y_0 + b)^2}\left[\frac{2b^2}{a^2 h}\left(\frac{z}{h} - x_0\right) + y_0 b\left(1 - \frac{(\frac{z}{h} - x_0)^2}{a^2}\right)^{-1/2}\frac{2}{a^2 h}(z/h - x_0)\right] \tag{1.7}$$

An interesting result is obtained when $y_0 = 0$, i.e., when the center of ellipsoid is at the x-axis:

$$f_Z(z) = \frac{2}{a^2}\left(\frac{z}{h} - x_0\right),$$

which is the density of the triangular distribution illustrated in the lower right plot of Figure 1.4.

The R-code for Example 1.12

```
> # Ellipsoidial tree crown
> radius.ellipse<-function(x,b,x0=0,y0=0,h) {
+                 x<-x
+                 a<-sqrt(b^2*(1-x0)^2/(b^2-y0^2))
+                 r<-rep(h*(y0+b),length(x))
+                 r[x>x0*h&x<=h]<-h*(y0+b*sqrt(1-(x[x>x0*h&x<=h]/h-x0)^2/a^2))
+                 r[x>h]<-0
+                 r
+                 }
>
> # cdf of laser observations
> cdf.laser<-function(x,b,x0,y0,h) {
+     a<-sqrt(b^2*(1-x0)^2/(b^2-y0^2))
+     value<-rep(0,length(x))
+     value[x>=x0*h&x<h]<-1-
+          (y0^2+2*y0*b*sqrt(1-(x[x>=x0*h&x<h]/h-x0)^2/a^2)+
+          b^2*(1-(x[x>=x0*h&x<h]/h-x0)^2/a^2))/(y0+b)^2
+     value[x>=h]<-1
+     value
+     }
>
> # corresponding density
> pdf.laser<-function(x,b,x0,y0,h) {
+     a<-sqrt(b^2*(1-x0)^2/(b^2-y0^2))
+     value<-rep(0,length(x))
+     value[x>x0*h&x<h]<--1/(y0+b)^2*
+          (y0*b*(1-(x[x>x0*h&x<h]/h-x0)^2/a^2)^(-1/2)*
+          (-2/(a^2*h)*(x[x>x0*h&x<h]/h-x0))+
+          b^2*(-2/(a^2*h)*(x[x>x0*h&x<h]/h-x0)))
+     value
+     }
> x0<-0.2
```

```
> y0<--0.1
> b<-0.25
> h<-20
> x<-seq(0,22,0.01)
> r<-radius.ellipse(x,b,x0,y0,h)
> windows(width=6, height=5)
> par(mfcol=c(2,2), mai=c(0.6,0.5,0.1,0.1), mgp=c(2,0.7,0), cex=0.8)
> plot(x, r, type="l",
+      xlab="Height, m", ylab="Crown radius, m", ylim=h*c(-0.1,0.15))
> points(h*x0,h*y0)
> text(h*x0, h*y0,
+      expression((hx[0]*paste(",")*hy[0])), pos=4)
> lines(h*rep(x0,2), h*c(y0,y0+b), lty="dashed")
> points(h*rep(x0,2), h*c(y0,y0+b), pch="--")
> text(h*x0, h*(y0+b/2), "hb", pos=4)
> lines(c(0,h), c(0,0), lty="dashed")
> points(c(0,h), c(0,0), pch="I")
> text(h/2, 0, "h", pos=3)
>
> plot(x, cdf.laser(x,b,x0,y0,h), type="l",
+      xlab="Height, m", ylab="Cumulative density")
>
> plot(x, pdf.laser(x,b,x0,y0,h), type="l",
+      xlab="Height, m", ylab="Density")
>
> plot(x, pdf.laser(x,b,x0,0,h), type="l",
+      xlab="Height, m", ylab="Density")
```

In addition to cdf and pdf (or pmf), one function describing th distribution is the quantile function, which is defined as the inverse of the cdf

$$q_X(u) = F_X^{-1}(y)$$

This function is useful, for example in random number generation. However, it may not be possible to solve it analytically. Assume that X has continuous cdf $F_X(x)$ and let random variable Y be $Y = F_X(x)$. Random variable Y is uniformly distributed between 0 and 1. Assume that we want to generate random variables that have cdf $F_(x)$, and assume that we have observations of random variable U, which has uniform distribution between 0 and 1. A random sample from the desired distribution is obtained through transformation $q_X(U)$.

Example 1.13 *Random number generation from percentile-based diameter distribution*

1.2.2 Multivariate distribution

The distribution of a random vector is expressed with a multivariate distribution. The multivariate distribution is the joint distribution of the component variables of the random vector. It completely defines the probabilities of all possible value combinations of the component variables of a random vector. For a bivariate case, let us denote these components by X_1 and X_2, the random vector being then $\boldsymbol{X} = (X_1, X_2)'$. In this case, the joint distribution gives the probabilities for all possible realizations of pair X_1 and X_2. The joint distribution can be used to construct *conditional* and *marginal*

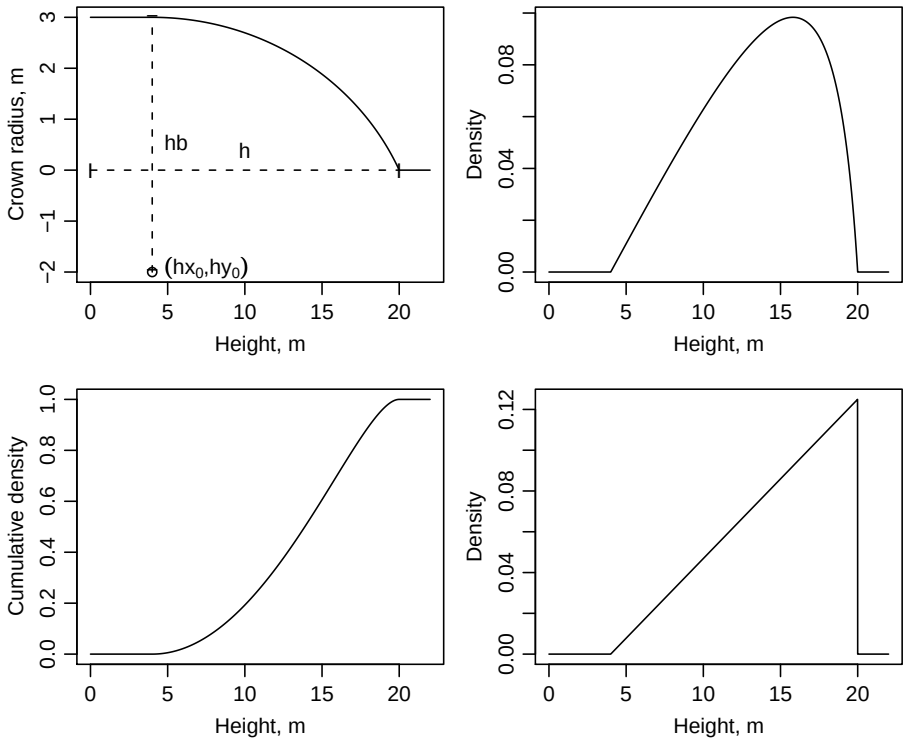


Figure 1.4: Illustration for example 1.12 using parameter values $h = 20$, $x_0 = 0.2$, $y_0 = -0.1$, and $b = 0.25$. The upper left figure shows the crown radius as a function of height. The cdf of laser observations is shown in the lower left figure, and the corresponding density in the upper right figure. The lower right figure illustrates the triangular density, obtained using $y_0 = 0$.

probability distributions of any single component variable of the random vector joint distribution

The conditional distribution expresses probability of one component variable for given values of the other variables. For example, the distribution of of $X_1|X_2 = x_2$ defines the distribution of X_1 given that X_2 has got the value of x_2. Thus, the conditional distribution is a univariate distribution in this case. For a random vectors of length 3 or more, a univariate conditional distribution is obtained by conditioning on all other component variables. For example, for random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$,we may be interested in conditional distribution of $X_1|X_2 = x_2, \ldots, X_n = x_n$. The conditional distribution may also be multivariate. For example, the distribution of $(X1, X_2)|X_3 = x_3$ is a bivariate distribution.

The marginal distribution of a random variable is the distribution in the whole population, assuming that no information on other random variables is available or used. The marginal probability is obtained as the sum of probabilities of $Y = y$ over the joint distribution. Thus, in a cross table of a joint distribution, the marginal distribution is in the lower or right marginals of the table, which gives the name "marginal". For a discrete random variable, the probabilities for any distinct value of marginal distribution is obtained as weighted average of conditional probabilities. For a continuous variable, the sums are replaced with integration.

The marginal distribution can be derived from the joint distribution. However, the joint distribution cannot be derived from marginal distribution without additional assumptions on the relationship between the components.

Example 1.14 *Assume that trees have been observed for species and health. The species is coded in the classes of example 1.3, and health in classes 1, 2 and 3, indicating dead, weakened, and healthy, respectively. The joint pmf of the two discrete variables can be expressed in tabular form*

		Y				
		1	*2*	*3*	*4*	$\sum$
	1	*0.1*	*0*	*0.1*	*0.01*	*0.21*
X	*2*	*0.1*	*0*	*0.05*	*0.02*	*0.17*
	3	*0.3*	*0.3*	*0*	*0.02*	*0.62*
	$\sum$	*0.5*	*0.3*	*0.15*	*0.05*	

The conditional distribution of $X|Y = 1$, i.e., the distribution of Scots pine trees into three classes can obtained by dividing the values of the first column by the marginal probability of Scots pine. The conditional probabilities are 0.2, 0.2 and 0.6 for dead, weakened, and healthy, respectively.

The conditional distribution $Y|X = 1$ is the distribution of dead trees by species. The conditional probabilities, obtained by dividing the values of the first row with the row sum of 0.21, are 0.476, 0,000, 0.476, 0.048 for tree species 1, 2, 3 and 4, respectively.

The marginal distributions of tree species are given in last column and row of the table.

Let $\boldsymbol{X}$ be a random vector

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \ddots \\ X_n \end{pmatrix}$$

In the continuous bivariate case, the joint probability density function is a function that expresses that the value of random vector (X, Y) falls within $A \in \Re$:

$$P((X, Y) \in A) = \int_A f(x, y)dxdy\,.$$

For example, the probability that X gets a value between a and b, and Y bets a value between c and d is $P(a < X \leq b, c < Y \leq d)) = \int_a^b \int_c^d f(x, y)dxdy$.

The marginal pdf is obtained by integrating out the other component variables. For a bivariate case, this implies that

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx\,.$$

This generalizes also for the multivariate random vector of length n. The marginal density of the k the component variable is

$$\begin{aligned} f(x_k) \quad = \quad & \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \\ & f(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_n)dx_n \ldots dx_{k+1}, dx_{k-1} \ldots dx_1\,. \end{aligned} \tag{1.8}$$

This is a function of only x_k, i.e., it is a density of one-dimensional random variable. For such a distribution, all results of section 1.2.1 apply.

From the conditions set for a univariate pdf, it can be generalized that any function for which $f(x, y) \geq 0$ for all $(x, y) \in \Re^2$, and which satisfies

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1$$

is a bivariate joint density function for some random vector (X, Y).

The definition of joint distribution function is analogous to the univariate case

$$\begin{aligned} F(\boldsymbol{X}) &= P(\boldsymbol{X} \leq \boldsymbol{x}) \\ &= P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \\ &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_n} f(u_1, u_2, \ldots, u_n) du_n \ldots du_2 du_1 \\ &= \int_{-\infty}^{\boldsymbol{x}} f(\boldsymbol{u}) d\boldsymbol{u} \end{aligned}$$

The marginal pdf of the kth component is obtained using (1.8) in (1.2). Thus, we just integrate the marginal density from $-\infty$ to x_k. But this integral is same as the conditional distribution function at $\boldsymbol{x} = (\infty, \ldots, \infty, x_k, \infty, \ldots, \infty)$. Thus, the marginal cdf of the kth component variable is

$$F_{X_k}(x_k) = F\infty, \ldots, \infty, x_k, \infty, \ldots, \infty) \,.$$

Example 1.15 *Let the joint distribution of tree diameter x and height y be expressed by a bivariate density $f(x, y)$. The marginal density of diameter is*

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

. The marginal density of height is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

Two random variables are *independent* if

$$p(x, y) = p(x)p(y) \qquad \text{for discrete } X \tag{1.9}$$

$$f_{XY}(x, y) = f_X(x f_Y(y) \quad \text{for continuous } X \,. \tag{1.10}$$

1.3 Expectation, variance and covariance

1.3.1 Expectation

In many cases, the distribution of a random variable is unknown, and we don't want to make strong assumptions on it. However, we may want to make inference about the behavior of the random variable. In such case, we can summarize the most interesting and important characteristics to some summary figures, that describe the most important properties of the random variable we are interested in.

The most commonly used characteris for describing the properties of a random variable is probably its *expected value*. The expected value is merely the mean of the random variable, which describes the average of the distribution, or the center of

gravity. The expected value is a number that used in the hope that it would as well as possible summarize the typical or expected value of an observation drawn from the underlying distribution.

The expected value of random variable X is defined as

$$\mathrm{E}X = \begin{cases} \sum_j p_j x_j & \text{for discrete } X \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{for continuous } X \end{cases} \tag{1.11}$$

It is also possible that the expected value is Infinity, if the tails of the distribution are thick or heavy enough.

Example 1.16 *The mean diameter in the stand of example 1.9 is*

$$\begin{aligned} \mathrm{E}(x) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_5^{10} 0.05 x dx + \ldots + \int_{17}^{24} 0.036 x dx \\ &= \Big/_5^{10} \frac{0.05}{2} x^2 + \ldots + \Big/_{17}^{24} \frac{0.036}{2} x^2 \\ &= \frac{1}{2}\left[0.05(10^2 - 5^2) + \ldots + 0.036(24^2 - 17^2)\right] \\ &= 13.625 \end{aligned}$$

It was computed using

```
> # The mean diameter
> mu<-sum(a/2*(d[-1]^2-d[-5]^2))
> mu
[1] 13.625
```

The expected value for function $g(X)$ is defined analogously

$$\mathrm{E}[g(X)] = \begin{cases} \sum_j p_j g(x_j) & \text{for discrete } X \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{for continuous } X \end{cases} \tag{1.12}$$

Example 1.17 *In the large forest stand of Example 1.3, the random variable X, tree species, gets values 1, 2, 3 and 4 with probabilities 0.5, 0.3, 0.15, 0.05, respectively. The mean height by species is expressed as*

$$g(x) = \begin{cases} 18 & x = 1 \\ 15 & x = 2 \\ 12 & x = 3 \\ 22 & x = 4 \end{cases}$$

The mean height is

$$\mathrm{E}H = \sum_j p_j g(X_j) = 0.5 \times 18 + 0.3 \times 15 + 0.15 \times 12 + 0.05 \times 22 = 16.4$$

Expected value has the following properties

$$\begin{aligned} \mathrm{E}(c) &= c && (1.13)\\ \mathrm{E}(cX) &= c\mathrm{E}(X) && (1.14)\\ \mathrm{E}(X+Y) &= \mathrm{E}(X)+\mathrm{E}(Y)\,, && (1.15) \end{aligned}$$

where c is a constant and Y and Y random variables. These properties hold also when X and Y are correlated.

The expected value of a bivariate random variable is just a vector of the expectations of the component variables, where expectations are defined through the marginal distribution. Thus, computing the expectation of a joint density requires only knowledge of the marginal distributions. However, if a transformation depends on several components, the joint density of the related components is needed. For example, the expectation of a bivariate transformation $g(X,Y)$ which depends on a bivariate random vector (X,Y) is analogously to (1.12)

$$\mathrm{E}[g(X,Y)] = \begin{cases} \sum_i \sum_j p_{ij} g(x_i, y_j) & \text{for discrete } (X,Y) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dyx & \text{for continuous } (X,Y) \end{cases} \qquad (1.16)$$

Example 1.18 *Consider again example 1.15. The expected value of diameter is just the expected value over the marginal distribution of diameters*

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx\,.$$

Correspondingly, the mean height is the expected value over the marginal distribution of heights

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy\,.$$

As an example of the expected value of a bivariate transformation, assume that the volume as a function of diameter and height is given by $v(x,y)$. The mean volume is the expected value of $v(x,y)$

$$\mathrm{E}(v(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} v(x,y) f(x,y) dxdy\,.$$

1.3.2 Variance and standard deviation

In addition to the mean of a random variable, we may be interested in how much the values of a random variable vary around the mean. This variation is described by the variance, which is defined as the expected value of the squared difference between the random variable and its mean:

$$\mathrm{var}X = \mathrm{E}\left(X - \mathrm{E}X\right)^2.$$

For the shake of simplicity, denote constant $\mathrm{E}X$ by μ. A computationally better form is obtained through simple algebra as

$$\begin{aligned} \mathrm{var}X &= \mathrm{E}\left(X-\mu\right)^2 \\ &= \mathrm{E}\left(X^2 - 2\mu\mathrm{E}X + \mu^2\right) \\ &= \mathrm{E}\left(X^2 - 2\mu^2 + \mu^2\right) \\ &= \mathrm{E}\left(X^2\right) - \mu^2 \end{aligned} \tag{1.17}$$

Variance is not easy to interpret. A more easy-to-interpret measure for the variation around mean is standard deviation, which is the square root of *standard deviation*

$$\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)} \tag{1.18}$$

It has the same unit as X. In practice, variance is usually used in computations, and the results are transformed to standard deviations for interpretation.

Example 1.19 *For the variance of tree diameter in Example 1.9, we need to compute* $\mu = E(X)$ *and* $E(X^2)$. *From example 1.16,* $\mu = 13.625$. *Equation* (1.12) *gives*

$$\begin{aligned} \mathrm{E}(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)dx \\ &= \int_5^{10} 0.05x^2dx + \ldots + \int_{17}^{24} 0.036x^2dx \\ &= \Big/_5^{10} \frac{0.05}{3}x^3 + \ldots + \Big/_{17}^{24} \frac{0.036}{3}x^3 \\ &= \frac{1}{3}\left[0.05(10^3-5^3) + \ldots + 0.036(24^3-17^3)\right] \\ &= 210.5 \end{aligned}$$

The variance is

$$\begin{aligned} \mathrm{varX} &= \mathrm{E}\left(X^2\right) - \mu^2 \\ &= 210.5 - 13.625^2 \\ &= 24.86\,, \end{aligned}$$

which gives standard error $sd(X) = \sqrt{24.86} = 4.99$.

These results were computed using

```
> # The second moment
> mu2<-sum(a/3*(d[-1]^3-d[-5]^3))
> mu2
[1] 210.5
>
> # Variance
> sigma2<-mu2-mu^2
> sigma2
[1] 24.85938
```

These results mean that taking randomly trees from a stand with the assumed percentile-based distribution, the mean diameter would in long run be 13.6 cm. The sampled tree diameters would differ from 13.6 cm, on average, by 5 cm in long run.

1.3.3 Covariance and correlation

The average linear joint variation of random variables X and Y is called covariance. Consider the difference of X from the expected value $\mathrm{E}(X)$, and correspondingly, the difference of Y from its expected value $\mathrm{E}(Y)$. The covariance is the expected value of the product of these differences:

$$\mathrm{cov}(X,Y) = \mathrm{E}\left\{[\mathrm{E}(X - \mathrm{E}(X))]\,[\mathrm{E}(Y - \mathrm{E}(Y))]\right\} \tag{1.19}$$

Computing the product by components and simplifying gives an alternative form of

$$\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \tag{1.20}$$

Covariance is a special case of variance. It can be seen by computing the covariance of X with itself,

$$\begin{aligned} \mathrm{cov}(X,X) &= \mathrm{E}(XX) - \mathrm{E}(X)\mathrm{E}(X) \\ &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \\ &= \mathrm{var}(X)\,. \end{aligned} \tag{1.21}$$

Example 1.20 *To compute the covariance between tree diameter and height in example 1.15, we need to compute the expected value of the product of diameter and height. It is obtained using equation* (1.16) *as (in this case, $g(X,Y) = XY$)*

$$\mathrm{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(xy)dydx$$

The expected values of X and Y were given in example 1.18. The covariance can be computed as $\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$.

As was variance, also covariance is hard to interpret, because it depends on the variation of the original variables. However, the covariance can be scaled to a much more easy-to-interpret figure, called correlation. The correlation is defined as

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\mathrm{sd}(X)\mathrm{sd}(Y)}\,. \tag{1.22}$$

The correlation varies between -1 and 1. The higher the absolute value of correlation is, the stronger is the linear relationship between X and Y. If the absolute value of correlation is 1, the relationship between X and Y can be exactly expressed by $Y =$

$aX + b$, where a and b are constants. If correlation is negative, a is negative, and if correlation is positive, a is positive. If correlation (and covariance) is 0, the random variables are uncorrelated. Correlation is related to independence so, that independent random variables are uncorrelated. However, uncorrelated variables are not necessarily independent.

Example 1.21 *Correlations for example 1.15*

The properties related to the mean of a random variable are sometimes called the *first order properties*. The properties related to variance and covariance are called the *second order properties*. These are related to the moments of the distribution: expected value is the first moment and variance is the second moment of the distribution. The moments for a distribution function can be computed using the moment generating function. The third and fourth moment, which are seldom used are the skewness and kurtosis. The analysis of a statistical model is mostly based on the first and second order properties of the random variable, and more detailed assumptions on the distribution are usually needed only for testing purposes.

1.3.4 Algebraic rules for variance, and covariance

The following rules are useful in calculating of variance

$$\begin{aligned} \mathrm{var}(aX) &= a^2\mathrm{var}(X) && (1.23)\\ \mathrm{var}(X+Y) &= \mathrm{var}(X)+\mathrm{var}(Y)+2\mathrm{cov}(X,Y) && (1.24)\\ \mathrm{var}(X-Y) &= \mathrm{var}(X)+\mathrm{var}(Y)-2\mathrm{cov}(X,Y) && (1.25) \end{aligned}$$

Combining rules (1.24) and (1.25) with rule (1.23) gives

$$\begin{aligned} \mathrm{var}(aX+bY) &= a^2\mathrm{var}(X)+b^2\mathrm{var}(Y)+2ab\mathrm{cov}(X,Y) && (1.26)\\ \mathrm{var}(aX-bY) &= a^2\mathrm{var}(X)+b^2\mathrm{var}(Y)-2ab\mathrm{cov}(X,Y) && (1.27) \end{aligned}$$

The following rules are useful in computations of covariance

$$\begin{aligned} \mathrm{cov}(X,Y) &= \mathrm{cov}(Y,X) && (1.28)\\ \mathrm{cov}(X,Y+Z) &= \mathrm{cov}(X,Y)+\mathrm{cov}(X,Z) && (1.29)\\ \mathrm{cov}(aX,bY) &= ab\mathrm{cov}(X,Y) && (1.30) \end{aligned}$$

Combining rules (1.29) and (1.30) gives

$$\mathrm{cov}(aX+bY,cZ+dW) = ac\mathrm{cov}(X,Z)+ad\mathrm{cov}(X,W)+bc\mathrm{cov}(Y,Z)+bd\mathrm{cov}(Y,W)\,, \quad (1.31)$$

where X, Y, Z, and W are random variables and a, b, c, and d are constants. It can be shown that all rules (1.23) – (1.30) are special cases of rule (1.31).

Adding a constant to a random variable has no effect on variance and covariance

$$\begin{aligned} \mathrm{var}(X+a) &= \mathrm{var}(X) && (1.32)\\ \mathrm{cov}(X+a, Y+b) &= \mathrm{cov}(X,Y)) && (1.33) \end{aligned}$$

1.3.5 Covariance and correlation matrices

The second order properties of a random vector can be described with a matrix called *variance-covariance matrix*, also called *variance matrix* or *covariance matrix*. The matrix summarizes all variances and covariances into a single matrix. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ be a random vector. The covariance matrix of is defined as

$$\begin{aligned} \mathrm{var}(\boldsymbol{X}) &= \mathrm{E}\{[\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})][\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})]'\} \\ &= \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2) & \cdots & \mathrm{cov}(X_1,X_k) \\ \mathrm{cov}(X_2,X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2,X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_k,X_1) & \mathrm{cov}(X_k,X_2) & \cdots & \mathrm{var}(X_k) \end{pmatrix} \end{aligned}$$

Applying (1.28) shows that the variance-covariance matrix is symmetric. It also positive semidefinite (proof is omitted).

The covariance matrix of random vectors $\boldsymbol{X}_k$ and $\boldsymbol{Y}_p$ is a $k \times p$ matrix, defined as

$$\begin{aligned} \mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y'}) &= \mathrm{E}\{[\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})][\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})]'\} \\ &= \begin{pmatrix} \mathrm{cov}(X_1,Y_1) & \mathrm{cov}(X_1,Y_2) & \cdots & \mathrm{cov}(X_1,Y_p) \\ \mathrm{cov}(X_2,Y_1) & \mathrm{cov}(X_2,Y_2) & \cdots & \mathrm{cov}(X_2,Y_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_k,Y_1) & \mathrm{cov}(X_k,Y_2) & \cdots & \mathrm{cov}(X_k,Y_p) \end{pmatrix}. \end{aligned}$$

Remembering from (1.21) that variance is a special case of covariance, it can be seen that the variance-covariance matrix of $\boldsymbol{X}$ is just a special case of the covariance matrix: $\mathrm{var}(\boldsymbol{X}) = \mathrm{cov}(\boldsymbol{X}, \boldsymbol{X})$.

In the univariate case, correlation was obtained from covariance by dividing the covariance by the product of standard deviations. Correspondingly, the correlation matrix is obtained by multiplying the covariance matrix from by a diagonal matrix including the inverted standard deviations

$$\mathrm{cor}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{diag}[\mathrm{var}(\boldsymbol{X})]^{-1/2} \mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}) \mathrm{diag}[\mathrm{var}(\boldsymbol{Y})]^{-1/2},$$

where

$$\mathrm{diag}[\mathrm{var}(\boldsymbol{X})]^{-1/2} = \begin{pmatrix} \frac{1}{sd(X_1)} & 0 & \cdots & 0 \\ 0 & \frac{1}{sd(X_2)} & \cdots & 0 \\ & & \ddots & \\ \cdots & \cdots & & \cdots \\ 0 & 0 & \cdots & \frac{1}{sd(X_k)} \end{pmatrix},$$

and correspondingly for y.

The algebraic rules of the previous section generalize also for matrices. Equations (1.14), (1.23), and (1.30) generalize to

$$\begin{aligned} \mathrm{E}(\boldsymbol{a'X}) &= \boldsymbol{a}'\mathrm{E}(\boldsymbol{X}) && (1.34)\\ \mathrm{var}(\boldsymbol{a'X}) &= \boldsymbol{a}'\mathrm{var}(\boldsymbol{X})\boldsymbol{a} && (1.35)\\ \mathrm{cov}(\boldsymbol{a'X}, \boldsymbol{b'Y}) &= \boldsymbol{a}'\mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y'})\boldsymbol{b} && (1.36) \end{aligned}$$

1.3.6 Sample estimators

The previous sections have shown several important rules for computations on random variables. However, the distributions of random variables are seldom known, nor are the expectations, variances and covariances. In most cases, we have collected observations on the realizations of the random variables. Using these realizations, we can *estimate* interesting *statistics* such as *sample means* and *sample variances*.

Let us define random variables $X_1, X_2, \ldots, X_n$ to describe a certain characteristic of n different sampling units. The sample mean of is the random variable

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The randomness of $\bar{X}$ is due to the randomness of X_i. The sample mean $\bar{X}$ is just a sum of random variables, $X_1, \ldots, X_n$, which is multiplied by constant $1/n$. If the expectation and variance of $X_1, \ldots, X_n$ are known, the expected value and variance of $\bar{X}$ can be derived from rules (1.14), (1.15), (1.23), and (1.24).

The sample variance is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2 \,.$$

The sample standard deviation is defined analogously to (1.18) as $S = \sqrt{S^2}$.

Example 1.22 *Let $X_1, X_2, \ldots, X_n$ be describe the diameters of trees $1, 2, \ldots, n$ from a forest stand, all of which have the same expected value μ and variance σ^2. We assume that the diameters are independent. The mean diameter of n trees, $\bar{X}$, is a*

random variable that has expectation

$$\begin{aligned} \mathrm{E}(\bar{X}) &= \mathrm{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n}\mathrm{E}(\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) \\ &= \frac{1}{n} n\mathrm{E}(X_i) \\ &= \mu\,. \end{aligned}$$

The second line resulted from application of rule (1.14) *and the third row from application of rule* (1.15)*.*

The variance of mean diameter $\bar{X}$ *would be*

$$\begin{aligned} \mathrm{var}(\bar{X}) &= \mathrm{var}(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \left(\frac{1}{n}\right)^2 \mathrm{var}(\sum_{i=1}^{n} X_i)\,, \end{aligned}$$

which was obtained through application of rule (1.23)*. Independence of* X_i *implies that* $\mathrm{cov}(X_i, X_j) = 0$ *for all* $i \neq j$*. Using rule* (1.24) *we get*

$$\begin{aligned} \mathrm{var}(\bar{X}) &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n}\mathrm{var}(X_i) \\ &= \frac{1}{n^2} n\sigma^2 \\ &= \frac{1}{n}\sigma^2 \end{aligned}$$

The standard deviation of mean diameter of the sample would be $\mathrm{sd}(\bar{X}) = \sigma/\sqrt{n}$*.*

The above definitions used uppercase letters to show that these quantities are random variables, not realizations. The realizations of sample statistics are obtained by replacing random variable X with its realization x. Those realized values of $\bar{X}$, S^2, and S would be denoted by $\bar{x}$, s^2, and s.

Example 1.23 *Assume that we have independently taken observations of diameter for 4 trees in stand,* $x_1 = 15$*,* $x_2 = 4$*,* $x_3 = 12$*, and* $x_4 = 11$*cm. The sample mean is*

$$\bar{x} = \frac{1}{4}(15 + 4 + 12 + 11) = 10.5\mathrm{cm}\,.$$

The sample variance is

$$s^2 = \frac{1}{4}\left[(15-10.5)^2 + (4-10.5)^2 + (12-10.5)^2 + (11-10.5)^2\right] = 21.67\text{cm}^2\,.$$

The sample standard deviation is $s = \sqrt{21.67} = 4.65$*cm.*

Consider the sum of squares $\sum_i = 1^n(x_i - a)^2$ as a function of a. The value of a that minimizes is the sample mean is $\bar{x}$. This result is related to the least squares estimation, which will be introduced in the next chapter. The sample variance can also be expressed in a form corresponding to (1.17) as $(n-1)s^2 = \sum_i = 1^n(x_i - \bar{x})^2 = \sum_i = 1^n x_i^2 - n\bar{x}^2$. This is useful in computations, as it expresses

Let random variables $Y_1, Y_2, \ldots, Y_n$ describe some other property of the same sampling units as $X_1, X_2, \ldots, X_n$. The sample covariance is defined as

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\,.$$

The sample correlation can be defined using sample covariance and sample standard deviation in equation (1.22). The expression reduces to

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}\,.$$

As with other sample statistics, C and R are random variables. The realizations c and r are obtained by utilizing the observed sample values in the equations.

A form corresponding to (1.20) is obtained as $(n-1)s^2 == \sum_i = 1^n x_i y_i - n\bar{x}\bar{y}$.

Note that the nominator includes $n-1$ instead of n, which would intuitively sound better justified. The reason for having $n-1$ is due to one degree of freedom that was used for estimating $\bar{x}$. As noted in equation a few paragraphs earlier, $\bar{x}$ yields the minimum value for the sum of squares applied in the estimation of variance. Thus, the sum of squares having the true unknown expectation, $\sum_i (X_i - E(X))^2$, is necessarily greater than or equal to $\sum_i (X_i - \bar{X})^2$. Thus, the utilized sum of squared is a downward biased estimator for the true sum of squares. This underestimation is accounted for by using $n-1$ instead of n in the denominator. The presented formulas yield unbiased estimators for the population variance and covariance, whereas the alternative estimators having n in the nominator would yield downward biased estimates.

1.4 Common distribution functions

This section shortly presents distribution functions that are important in linear modeling applications. The distributions are classified in three classes: Discrete distributions, continuous distributions, and distributions of important functions of Standard normal

variable. The last two classes differ in that the third class is derived for inference on normally distributed populations.

1.4.1 Discrete distributions

Bernoulli and Binomial distribution

A bernoulli trial has two possible discrete outcomes: success and failure. A random variable X has *Bernoulli(p)* distribution if it takes the value of 1 ("success") wit probability p, and value 0 ("failure") with probability $1-p$. The expected value and variance of X are obtained using equations (1.11) and (1.17) as

$$\begin{aligned} \mathrm{E}(X) &= 1p + 0(1-p) = p \\ \mathrm{E}(X^2) &= 1^2 p + 0^2(1-p) = p \\ \mathrm{var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = p - p^2 = p(1-p) \end{aligned}$$

Assume that we make n independent bernoulli trials, and let Y be the number of successes. The variable Y is said to have a *Binomial(n, p)* distribution with parameters n and p, which has pmf

$$P(Y = y|n, p) = \binom{n}{x} p^y (1-p)^{n-y}$$

In forestry, binomial distribution may be used, for example, to present plant health or survival, tree species, etc. The expectation and variance of binomial distribution can be deduced using rules (1.15) and (1.24) as

$$\begin{aligned} \mathrm{E}(Y) &= np \\ \mathrm{var}(Y) &= np(1-p) \end{aligned}$$

A generalization of binomial distribution is the *Multinomial($n, p_1, \ldots, p_k$)* distribution, where the number of classes is k instead of 2, the number of classes in the binomial distribution. Let random variable Z be "the number of bernoulli trials needed to get r successes". Then Z would have the *Negative binomial(r, p)* distribution. Such a distribution could be used, for example, in assessing how many trees should be bored in a field experiment to get at least r trees with a butt rot infection.

In R, functions `dbinom, pbinom, qbinom` can be used for computing pmf, cdf and inverse cdf of the binomial distribution. Function `rbinom` can be used for generating a sample from a specified binomial distribution. The corresponding functions for multinomial and negative binomial are (guess what they do) `dmultinom, rmultinom, rnegbin`.

Poisson distribution

Poisson distribution is a discrete distribution, that can be used for modeling of counts. It takes only nonnegative integer values. It can be used, for example, in modeling the number of individuals in a line, if new individuals enter and leave the line at random. If trees in a forest stand are located completely randomly, the number of trees within a fixed area is distributed according to the *Poisson(λ)* distribution. Random variable X has the Poisson distribution if the pmf is

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where λ is the parameter, which is often called intensity. In the random forest example, it is the stand density (trees per area unit). Both mean and variance of Poisson random variable can be shown to be λ. The lower plots in figure 1.1 show an example of the Poisson distribution.

R has functions `dpois, ppois` and `rpois` for computing pmf, cdf and inverse cdf of the Poisson distribution. There is also function `rpois` for generating Poisson distributed random numbers.

1.4.2 Continuous distributions

Continuous uniform distribution

The *Uniform(a, b)* distribution is defined by spreading the probability mass uniformly over interval $[a, b]$. The cdf of uniform distribution is

$$F(x|a,b) = \begin{cases} 0 & x < a \\ \frac{a}{a-b} + \frac{1}{b-a}x & a \le x < b \\ 1 & x \ge b \end{cases},$$

and the density is

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & a \le x < b \\ 0 & \text{otherwise} \end{cases}.$$

The expectation and variance are $\mathrm{E}(X) = \frac{b-a}{2}$ and $\mathrm{var}(X) = \frac{(b-a)^2}{12}$.

R has functions `dunif, punif, qunif` and `runif` for density, distribution function, quantile function and random number generation with uniform distribution. Uniform distribution is seldom a good assumption for a distribution of a random variable. However, uniform random numbers are very often needed in simulations. They are also needed in generating random numbers from any specific distribution using the probability integral transformation approach.

Example 1.24 *One wants to simulate 10 tosses of a coin. One alternative to proceed is to simulate 10 random numbers from the Uniform($0, 1$) distribution. Realizations*

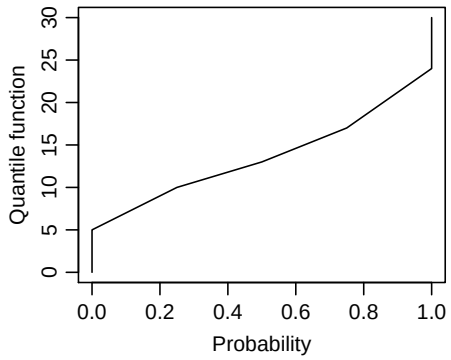


Figure 1.5: The left plot shows the quantile function of the percentile-based diameter distribution. The Right plot shows the simulated sample (1 cm classes), and the true underlying density (thick line)

smaller than 0.5 are interpreted as heads (H) and realizations equal to or above 0.5 as tails (T). The random numbers obtained using `runif(10)` *are 1 1 0 1 1 0 0 1 1 0, results simulated tosses H H T H H T T H H T.*

Example 1.25 *Generate 1000 random numbers from the percentile-based distribution of example 1.6. The quantile function of the percentile-based diameter distribution is*

$$q(p) = F^{-1}(p) = \begin{cases} 5 + 20p & 0 \leq p < 1/4 \\ 7 + 12p & 1/4 \leq p < 1/2 \\ 5 + 16p & 1/2 \leq p < 3/4 \\ -4 + 28p & 3/4 \leq p < 1 \end{cases}$$

First, 1000 random numbers, denoted by $u_1, \ldots, u_{1000}$*, were generated from Uniform*$(0, 1)$ *distribution. These values were transformed to percentile-based random numbers using* $y = q(u)$*. Figure 1.5 shows the utilized quantile function and the histogram of the simulated random sample.*

Normal distribution

The normal (Gaussian) distribution is probably the most well-known statistical distribution. (Casella and Berger 2002) mention three main for the special role of Normal distribution among the body of statistics. First, Normal distribution is very tractable analytically. Second, the shape of Normal distribution is the symmetric Bell-shape, which makes it an appealing alternative for many population models. Third, the central limit theorem shows that the normal distribution is a good approximation for several distributions under mild conditions and with large sample sizes.

The density of normal distribution *N*(μ, σ^2) is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where $\mathrm{E}(X) = \mu$ and $\mathrm{var}(X) = \sigma^2$. The cdf of normal distribution cannot be expressed in a closed form. The standard Normal distribution is the normal distribution with mean of 0 and variance of 1, *N(0,1)*.

It is interesting and important to note that the sum of independent normal random variables $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ is also normal with $X_1 + X_2 \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$. In several statistical tests, including the tests related to regression analysis and the linear model, the random variables to be analyzed are assumed to follow the normal or multinormal distribution. However, it should be pointed out that normal distribution needs not to be assumed in the estimation of parameters using the methods based on least squares. The central limit theorem states that the distribution of the means from samples of same size approaches the normal distribution as sample size gets large, regardless of the distribution of the random variable in the population.

Example 1.26 *Assume that the lifetime of a butterfly is highly skewed to the right, having distribution shown in the upper left plot of Figure 1.6. 1000 samples of size 5, 10 and 20 were taken from the population. The other plots of 1.6 show the distributions of sample means with the three applied sample sizes. The sample mean approaches the normal distribution quite fast, being fairly close to normal even with sample size of 10.*

Multivariate normal distribution

Random vector $\boldsymbol{Y}_{n\times 1} = (Y_1, Y_2; \ldots, Y_n)'$ follows the n-dimensional multivariate normal distribution with expectation $\boldsymbol{\mu}_{n\times 1}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{n\times n}$, $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if Y has (joint) density

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n}} |\Sigma|^{-1/2} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}(\boldsymbol{y}-\boldsymbol{\mu})}.$$

An alternative definition for normal distribution is obtained by defining that $Y \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if $\boldsymbol{t'Y} \sim N(\boldsymbol{t'\mu}, \boldsymbol{t'\Sigma t})$ with any $\boldsymbol{t} \in \Re$. If Σ is a diagonal matrix, then the result that the sum of independent normal variates are also normally distributed can be proved through appropriate selection of $\boldsymbol{t}$. The marginal distributions of a multinormal distribution are univariate normal distributions. Furthermore, the correlation between component variables is always linear. This reults in that in prediction, the Best Linear Prediction is always also the Best Predictor.

Lognormal distribution

If $\ln X$ is normally distributed with mean μ and variance σ^2, then X is said to be lognormally distributed. The pdf of lognormal distribution is obtained straightforwardly

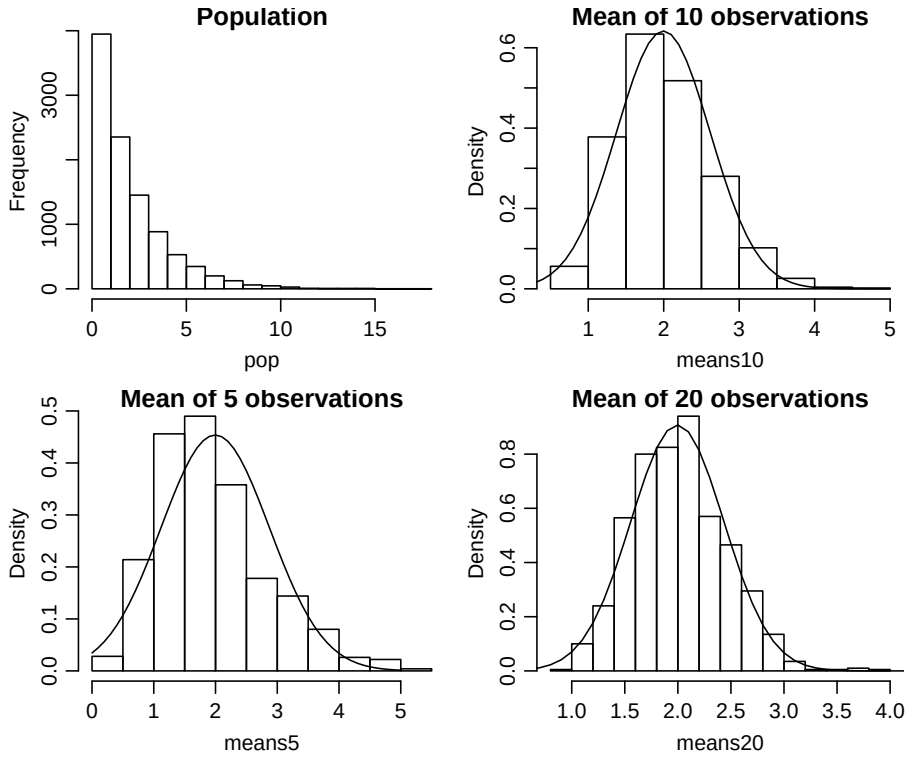


Figure 1.6: Illustration of Example 1.26.

using (1.5) with normal distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{e^{-(\log(x)-\mu)^2/(2\sigma^2)}}{x} \quad 0 \leq x < \infty$$

The expectation and variance are $\mathrm{E}X = e^{\mu+\sigma^2/2}$ and $\mathrm{var}X = e^{2(\mu+\sigma^2}$.

In many forestry applications, the response to be modeled does not express a normal distribution. A very common strategy for those cases is to take a logarithmic transformation of the response, and assume the residuals to be normally distributed. This implies that the original response is assumed to be lognormally distributed. The predictions from such a model are unbiased in the logarithmic scale. However, in the back-transformed scale, the predictions are downward biased, due to that for lognormal X, $\mathrm{E}X = e^{\mu+\sigma^2/2}$, not $\mathrm{E}X = e^{\mu}$. If the normality of residuals holds, a suitable strategy for correction of bias is to add half of the residual variance to the logarithmic predictions before applying the exponential transformation to get unbiased predictions in the original scale. However, this may not be a good strategy if normality of residuals does not hold. An alternative has been presented in (Lappi et al. 2006, p. 103 – 105).

Example 1.11 demonstrated the relationship of normal and lognormal distribution. For making computations with lognormal distribution in R, the standard R includes functions `dlnorm, plnorm, qlnorm` and `rlnorm`.

Weibull distribution

The (two-parameter) *Weibull(α, β)* distribution has distribution function

$$F(x|\alpha,\beta) = 1 - \exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\},$$

and density

$$f(x|\alpha,\beta) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}},$$

where α and β are the shape and scale parameters, respectively. The expectation and variance are $\mathrm{E}X = \beta\Gamma(1+\frac{1}{\alpha})$ and $\mathrm{var}X = \beta^2\left[\Gamma(1+\frac{2}{\alpha}) - \Gamma^2(1+frac1\alpha)\right]$.

The Weibull has been used for modeling failure time data. A special case of Weibull distribution is obtained with $\alpha = 1$, called the *Exponential(β)* distribution. The exponential distribution is used for modeling lifetimes. The population distribution of Example 1.26 was generated using the exponential distribution with $\beta = 2$. In forestry, the Weibull distribution has been used for modeling distributions of tree DBH since the paper of Bailey and Dell (1973). Also a three-parameter version with an additional parameter for location has been used. The reason for the popularity of Weibull distribution its flexibility with only two parameters. In addition, it has a closed form solution for the cdf, which eases the computation of diameter class frequencies. In particular, there are no theoretical reasons why Weibull should be favored over other alternatives.

R has function `dweibull, pweibull, qweibull` and `rweibull` for computations with Weibull distribution

Other distributions used in forestry

In addition to the Weibull distribution, many other distributions have been used for modeling tree size within a plot or stand. These distributions include include the beta, Jonson's SB and the logit-logistic distribution. The *beta(*p, q*)* distribution has density

$$f(x|\alpha,\beta) = \frac{1}{B(p,q)} x^{\alpha-1}(1-x)^{beta-1}$$

for $a \leq x \leq b$ and $p, q > 0$. The statistical literature usually expresses the beta-distribution in a two-parameter form, which is obtained from the above expression by setting $a = 0$ and $b = 1$.

$$f(x|p,q) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}$$

The expectation and variance of this constrained form are $\mathrm{E}X = \frac{p}{p+q}$ and $\mathrm{var}X = \frac{pq}{(p+q)^2(p+q+1)}$. The beta function can be defined through gamma function $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$. In forestry, the beta distribution has been used in a form where it has been bounded between minimum and maximum diameter instead of constants 0 and 1. In such applications, also the minimum and maximum diameters are treated as parameters rather than interpreting them as fixed constants, resulting in that the distribution has a total of four parameters.

The density of Jonsons *SB(*$\delta, \gamma, \lambda, \xi$*)* distribution using notations of Siipilehto (1999) is

$$f(x|\delta,\gamma,\lambda,\xi) = \frac{\delta}{\sqrt{2\pi}} \frac{\lambda}{(\xi+\lambda-x)(x-\xi)} e^{-\frac{1}{2}\left[\gamma+\delta\ln\left(\frac{x-\xi}{\xi+\lambda-x}\right)\right]^2}.$$

where ξ is the location parameter, λ the scale parameter, and δ and γ are the shape parameters affecting in kurtosis and asymmetry of the distribution, respectively.

The logit-logistic distribution (Tadikamalla et al. 1982, Wang and Rennolls 2005) can represent larger variation of skewness-kurtosis combinations than several other distributions including Weibull, beta and Johnsons SB distribution. In contrast to beta and SB, it also has a closed-form expression for cdf. The p.d.f. and density of the *logit-logistic(*$\psi, \lambda, \phi, \sigma$*)* distribution between the minimum and maximum diameters ψ and λ are

$$\begin{aligned} F(x|\psi,\lambda,\phi,\sigma) &= \frac{1}{1+\exp\left(\frac{\phi}{\sigma}\right)\left(\frac{x-\psi}{\lambda-x}\right)^{-\frac{1}{\sigma}}} \\ f(x|\psi,\lambda,\phi,\sigma) &= \frac{\lambda-\psi}{\sigma(x-\psi)(\lambda-x)} \frac{1}{\exp\left(-\frac{\phi}{\sigma}\right)\left(\frac{x-\psi}{\lambda-x}\right)^{\frac{1}{\sigma}} + \exp\left(\frac{\phi}{\sigma}\right)\left(\frac{x-\psi}{\lambda-x}\right)^{-\frac{1}{\sigma}} + 2}. \end{aligned}$$

The SB distribution is formulated by assuming that logit-transformation of x that is scaled to range $[0,1]$, $z = \text{logit}(\frac{x-\xi}{\lambda}) = \text{logit}(y) = \ln\left(\frac{y}{1-y}\right)$ follows the normal distribution with $\mu = -\frac{\gamma}{\delta}$ and $\sigma^2 = \frac{1-\gamma}{\delta^2}$. Thus, it could be called logit-normal distribution, by an analogy to the lognormal distribution. Correspondingly, the logit-logistic distribution is obtained by a similar approach, assuming that the logit follows the logistic distribution instead of the normal distribution.

Exponential family

One important group of distributions is the exponential family of distributions. It includes those distributions whose density or pmf can be expressed as

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{theta})\exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right), \tag{1.37}$$

where $h(x) \geq 0$, $c(\boldsymbol{\theta}) \geq 0$, $t_1, \ldots, t_k$ are real-valued functions of x that do not depend on $\boldsymbol{\theta}$, and $w_1, \ldots, w_k$ are real-valued functions of $\boldsymbol{\theta}$ that do not depend on x. The exponential family is important in that the expectations, variances and covariances of the random variables, as well as the maximum likelihood estimators for the parameter vector $\boldsymbol{\theta}$ can be expressed in a form that eases the computations. Furthermore, the ML estimators have certain desired properties. For further details, see Casella and Berger (2002). For these reasons, applications of generalized linear models and generalized linear mixed models, which are strongly based on ML theory, have been developed for the distributions belonging to the exponential family. Of the distributions presented in this paper, the binomial, multinomial, negative binomial, Poisson, normal, lognormal, and beta (with fixed a and b) distributions belong to the exponential family.

Example 1.27 *The lognormal distribution can be expressed as*

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi\sigma^2}}\frac{e^{-(\ln(x)-\mu)^2/(2\sigma^2)}}{x} \\ &= \frac{1}{x}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\ln(x)^2+\frac{\mu}{\sigma^2}\ln(x)-\frac{\mu^2}{2\sigma^2}} \end{aligned}$$

By defining $h(x) = 1/x$, $c(\boldsymbol{\theta}) = 1/\sqrt{2\pi\sigma^2}$, $w_1(\boldsymbol{\theta}) = -1/(2\sigma^2)$, $w_2(\boldsymbol{\theta}) = \mu/\sigma^2$, $w_3(\boldsymbol{\theta}) = -\mu^2/(2\sigma^2)$, $t_1(x) = \ln(x)^2$, $t_2(x) = ln(x)$, *and* $t_3(x) = 1$ *we see that the lognormal density is of form 1.37.*

R library `MASS` has function `mvrnorm` for generating multivariate normal random numbers.

1.4.3 Distributions of important transformations of a standard Normal variate

χ^2 distribution

If Z follows the standard normal distribution *N(0, 1)*, then the sum of squares $Y = \sum_{i=1}^{p} Z_i^2$ follows the $\chi^2(p)$ distribution with p degrees of freedom. It has density

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{p/2-1} e^{-x/2} \quad 0 \leq x < \infty, p = 1, 2, \ldots ,$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ is the gamma function that needs to be evaluated numerically. The expected value and variance of $\chi^2(p)$ distribution are p and $2p$, respectively. The expectation results directly from that

$$\begin{aligned} \mathrm{E}(Z) &= \mathrm{E}\left(\sum_{i=1}^{p} Z_i^2\right) \\ &= \sum_{i=1}^{p} \mathrm{E}\left(Z_i^2\right) \\ &= \sum_{i=1}^{p} (\mathrm{var}(Z) + \mathrm{E}(Z_i)^2) \\ &= p(1+0) = p \end{aligned}$$

Chi square distribution is used in construction of tests for the residual sum of squares for a regression model with normally distributed residuals. R has functions `dchisc, pchisc, rchisc` and `qchisc` for different computations with the χ^2-distribution. There is also function `gamma` for evaluating the gamma function.

Students t-distribution

If $Z \sim N(0, 1)$ and $u \sim \chi^2(n)$, then the ratio $v = \frac{z}{\sqrt{u/n}}$ follows the Students t-distribution, *t(n)*, with n 4 degrees of freedom. The density of students t-distribution is

$$f(x|n) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{\sqrt{n\pi}} \frac{1}{(1 + \frac{x^2}{n})^{\frac{1}{2}(n+1)}} \quad -\infty < x < \infty, \quad n = 1, 2, \ldots$$

The mean and variance are $\mathrm{E}(X) = 0$ for $(n > 1)$ and $\mathrm{var}(X) = \frac{n}{n-2}$ for $(n > 2)$.

The t-distribution was originally developed for testing if the mean of a sample of size n from a normally distributed population significantly differs from a fixed mean. The Students t-distribution is utilized instead of the normal because the variance is estimated from a sample, thus being a random variable with χ^2 distribution. The density is bell-shaped, and symmetric as normal distribution is too, but it is wider than the standard normal distribution. However, as n gets large, the t-distribution approaches the

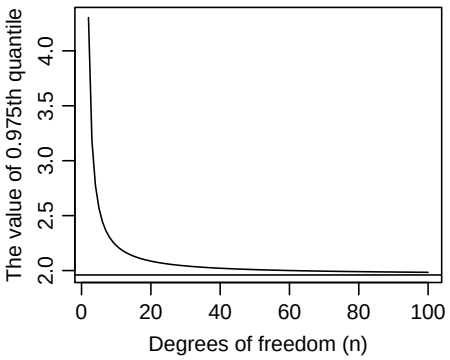


Figure 1.7: The 0.975th quantile is needed for constructing two-sided 95% confidence intervals for means. The horizontal line at 1.96 level shows the 0.975th quantile of normal distribution, and the decreasing line that of t-distribution as the degrees of freedom change from 2 to 100.

standard normal distribution. Thus, normal distribution leads to very similar inference as t distribution with large sample sizes tests on a large sample can be based on normal distribution (Figure 1.7).

R has functions `dt, pt, qt` and `rt` for computations with Student's t-distribution.

F distribution

Let X and Y be independent random variables that follow the χ^2 distribution: $X \sim \chi^2(m)$ and $Y \sim \chi^2(n)$. As an example, X and Y could be two different sums of squares. The distribution of ratio $F = \frac{X/m}{Y/n}$ follows the *F(m, n)* distribution with m and n degrees of freedom. The density is

$$f(z|m,n) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{m}{n}\right)^{m/2} \frac{x^{\frac{1}{2}m-2}}{(1+\frac{m}{n}x)^{\frac{1}{2}(m+n)}} \quad 0 \leq z < \infty, \quad m,n = 1,2,\ldots .$$

The mean and variance are $\mathrm{E}(Z) = \frac{n}{n-2}$ for $(n > 2)$ and $\mathrm{var}(Z) = 2\frac{n}{n-2}^2 \frac{m+n-2}{m(n-4)}$ for $(n > 4)$.

The F-distribution is used in testing two regression models against each other. R has function `pf, df, qf` and `rf` for computations with F distribution.

1.5 Fitting distribution functions to data

1.5.1 Maximum likelihood

In developing the distribution theory, the probability of getting a value of random variable between specified values is expressed with a distribution function, that is specified through parameters. Those parameters are thought of as fixed constants that specify

the distribution of the random variable in the population we are interested in. In reality, we observe realizations of random variable, and the interesting question is often "What values do the parameters of the underlying distribution have". This leads to change in the role of parameters and variables.

Treating the distribution function as a function of its parameters for the sample we have in hands leads to the definition of likelihood. Assume that observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ have been observed, and the joint distribution function of those observations is $f(\boldsymbol{x}|\boldsymbol{\theta})$, where vector $\boldsymbol{\theta}$ includes the parameters that specify the joint distribution function. If the observations are independent, the joint pdf is just the product of univariate densities: $f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$. The likelihood function is the joint pdf for the fixed $\boldsymbol{x}$ as a function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}) \tag{1.38}$$

If the components of $\boldsymbol{x}$ are independent, the likelihood becomes

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\theta_1, \theta_2, \ldots, \theta_k)$$

The value of likelihood cannot be interpreted as a probability, even though it is somehow related to it. However, it can be used to compare two parameter values against each other. If we observe that $L(\boldsymbol{\theta}_1|\boldsymbol{x}) > L(\boldsymbol{\theta}_2|\boldsymbol{x})$, we could say that θ_2 is more likely the value of the parameter of the underlying population than θ_2. This deduction leads to the principle of maximum likelihood in estimating the parameters of an underlying distribution of an assumed functional form.

As the likelihood itself has no clear meaning except for being a tool for comparing two estimates, it does not have any effect if we make an increasing transformation to the likelihood. In many situations, the computations get simpler if we use logarithmic likelihood instead of the likelihood. The log-likelihood is defined as

$$l(\boldsymbol{\theta}|\boldsymbol{x}) = \ln(L(\boldsymbol{\theta}|\boldsymbol{x})) = \ln(f(\boldsymbol{x}|\boldsymbol{\theta}))$$

Especially, with independent observations, minimizing the product of densities is equivalent to minimizing the sum of logarithmic densities

$$l(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln(f(x_i|\theta_1, \theta_2, \ldots, \theta_k))$$

The maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ is the value of $\boldsymbol{\theta}$ that maximizes the likelihood 1.38. If the likelihood function is differentiable with respect to the components of $\boldsymbol{\theta}$, then the k likelihood equations are formulated as

$$\frac{\partial}{\partial \theta_i} L(\boldsymbol{\theta}|\boldsymbol{x}) = 0 \quad k = 1, \ldots, k$$

which yields systems of k equations. The solutions for that system of equations gives candidates of likelihood estimates. Those candidates may be global minima or maxima, local minima or maxima, or inflection points. Other candidates are at the boundary of the parameter space. Of these candidates, the *maximum likelihood estimate* (ML estimate) is the one that gives the global maximum value of likelihood within the parameter space. The *ML estimator* is the expression that yields the ML estimate when evaluated using the data, $\boldsymbol{x}$.

One important feature of the ML estimator is the invariance property. Let $\widehat{\theta}$ be the MLE of θ, and $\tau(\theta)$ any function. The invariance means that the MLE of $\tau(\theta)$ is $\widehat{\tau(\theta)} = \tau\left(\widehat{\theta}\right)$.

There are no guarantee that the ML estimates would be unbiased or best in any sense. However, with certain distributions, there are results that show they are unbiased minimum variance estimators. Especially, these hold with distributions of the exponential family. In addition, if the first and second derivatives of the log-likelihood exist, and the Fisher information matrix is not zero, the ML estimators are asymptotically normal, minimum-variance unbiased estimators. The term asymptotically means that these results are true with large samples, whereas they may be badly violated with small samples. The asymptotic variance of ML estimate is that defined by the Crao-Ramer lower bound:

$$\mathrm{var}(\widehat{\boldsymbol{\theta}}) = \left\{ -\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta}^2} \right] \right\}^{-1}$$

where $-\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta}}\right]$ is the Fisher information matrix. In applications, the variance is estimated by replacing $\boldsymbol{\theta}$ with ML estimates $\widehat{\boldsymbol{\theta}}$.

Example 1.28 *Assume that $\boldsymbol{x} = x_1, x_2, \ldots, x_n$ are observations from normal distribution with variance 1 and mean μ. The log likelihood function with respect to unknown μ is*

$$\begin{aligned} l(\mu|\boldsymbol{x}) &= \sum_{i=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2 + \right] \\ &= \ln n - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

Differentiating the log likelihood with respect to μ and setting equal to 0 gives the

likelihood equation

$$\begin{aligned} -\frac{1}{2}\sum_{i=1}^{n} 2(-1)(x_i-\mu) &= 0 \\ \sum_{i=1}^{n} x_i - n\mu &= 0 \\ n\mu &= \sum_{i=1}^{n} x_i \\ \mu &= \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \end{aligned}$$

Thus, the sample mean is a candidate for MLE, and the only solution for the ML equation. To verify that it is maximum, not a minimum, we need to study the second derivate of the log likelihood, $\frac{\partial}{\partial\mu} l(\mu, \boldsymbol{x}) = \frac{\partial}{\partial\mu}\sum_{i=1}^{n} x_i - n\mu = -n$*. The second derivative is a negative constant, so the the first derivative is decreasing for* $infty < \mu < \infty$*. Thus, the first derivative has to be positive for* $\mu < \bar{x}$ *and negative for* $\mu > \bar{x}$*, which implies that* $\mu = \bar{x}$ *is a maximum. Furthermore, since the first derivative is continuous and unique, it is also a global maximum. Thus, we deduce that being within the domain of the parameter value, the global maximum* $\bar{x}$ *is the ML-estimate for* μ*. Had we a restriction for* μ *to be positive, then the ML estimator would be* $\mu = \bar{x}$ *if* $\bar{x} > 0$ *and* $\mu = 0$ *if* $\bar{x} \leq 0$

As shown above, the second derivative called Fishers information is $\frac{\partial^2 l(\mu|\boldsymbol{x})}{\partial\mu^2} = -n$*. Writing this to the Rao-Cramer lower bound gives*

$$\begin{aligned} \mathrm{var}(\hat{\mu}) &= \frac{1}{-\,\mathrm{E}(-n)} \\ &= 1/n\,. \end{aligned}$$

Thus, the standard error of estimate is $1/\sqrt{n}$*. It is a special case of the well-known standard error of mean* $\sigma^2/\sqrt{n}$*. Even though we will not show it here, we conclude by saying that the standard error of ML-estimate for a population with other variance than 1 would be just* $\sigma/\sqrt{n}$*.*

Example 1.29 *A lazy forester does not much care on exact results, but uses numerical easy-to-use algorithms always for maximum likelihood estimation instead. Even though this approach should not be suggested for a true scientist, at least for such a simple distribution as normal, we do that for demonstration purposes.*

Thus, we have the sample from normal distribution below, with known variance of 1.

```
> library(stats4)
> y10<-c(4.99, 4.42, 5.95, 4.49, 5.75, 3.92, 7.74, 5.77, 4.75, 5.61)
> y20<-c(5.80, 6.00, 5.35, 3.98, 4.99, 5.34, 5.14, 5.36, 3.19, 4.62,
+        4.57, 3.23, 5.57, 4.18, 4.66, 6.76, 5.83, 4.74, 6.91, 4.44)
```

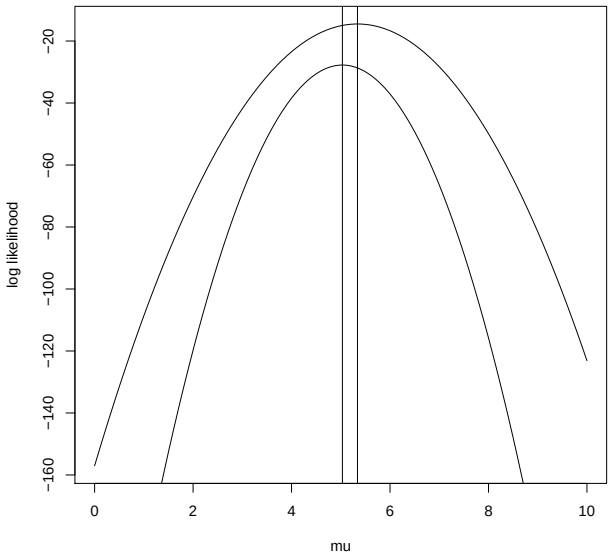


Figure 1.8: The normal likelihood based on the data of 10 observations (upper) and that based on 20 observations (lower). The vertical lines show the ML estimates.

Next, we define the log likelihoods as the sum over logarithmic densities. The likelihood is a function of μ. The mle package needs the negative log likelihood, so that is why we return value $-l$.

```
> # define a function giving the negative log likelihood for y10
> nll10<-function(mu) {
+    l<-sum(log(dnorm(y10,mu)))
+    -l
+    }
> # The same function for the larger sample
> nll20<-function(mu) {
+    l<-sum(log(dnorm(y20,mu)))
+    -l
+    }
```

For demonstration purposes, we plot the likelihoods for both datasets. These plots are shown in Figure 1.8. We see that the likelihood based on larger dataset is narrower and has a sharper peak than the likelihood based on the smaller dataset.

```
> mu<-seq(0,10,0.1)
> plot(mu,-sapply(mu,function(x) nll10(x)),type="l",xlab="mu",ylab="log likelihood")
> lines(mu,-sapply(mu,function(x) nll20(x)),type="l")
```

The ML estimates are obtained numerically using function `mle`.

```
> sol10<-mle(minuslogl=nll10,start=list(mu=0))
> sol20<-mle(minuslogl=nll20,start=list(mu=0))
>
> summary(sol10)
Maximum likelihood estimation

Call:
```

```
mle(minuslogl = nll10, start = list(mu = 0))

Coefficients:
   Estimate Std. Error
mu    5.339  0.3162278

-2 log L: 28.99266
>
> summary(sol20)
Maximum likelihood estimation

Call:
mle(minuslogl = nll20, start = list(mu = 0))

Coefficients:
   Estimate Std. Error
mu    5.033  0.2236068

-2 log L: 55.47056

> lines(rep(coef(sol10),2),c(-500,0))
> lines(rep(coef(sol20),2),c(-500,0))
```

We see that the estimates are 5.339 *for the smaller data and* 5.033 *for larger data. The standard errors are* 0.316 *and* 0.223*, respectively. The vertical lines added to the plot of likelihoods show that they are at the maximums of the respective likelihoods. The higher sample size leads to more accurate estimates, as indicated by the standard errors. This can also be seen from the plot of likelihoods: standard errors are inversely related to the second derivative of the likelihood, and the second derivative is the higher the more peaked the likelihood is.*

Finally, we use the exact results from the previous example to compute the MLE:s. The estimate is just the sample mean, and standard error is $1/n$*.*

```
>
> mean(y10)
[1] 5.339
> sqrt(1/10)
[1] 0.3162278
> mean(y20)
[1] 5.033
> sqrt(1/20)
[1] 0.2236068
```

The numerically obtained estimates and standard errors are the same as the exact values up to the default reporting accuracy of R.

Casella and Berger (2002) presents a number of other examples on finding ML estimators for binomial and normal distribution under different assumptions. Below are two examples on finding ML-estimators in R.

Example 1.30 *Assume that tree diameters follow Weibull*(α, β) *distribution, and an independent sample of diameters* $\boldsymbol{x} = (15.1, 17.1, 10.0, 13.5, 16.6, 18.1, 11.0, 13.4, 15.2, 17.1, 15.9, 17.7, 14.4, 13$ *has been taken from a stand. Find numerically ML-estimates for parameters* α *and* β*.*

We use R-function `mle` *in package* `stats4`*. For this purpose, we need to define a function that evaluates the negative log likelihood.*

```
library(stats4)

# the measured diameters in a vector
DBH<-c(15.1, 17.1, 10.0, 13.5, 16.6, 18.1, 11.0, 13.4, 15.2, 17.1,
       15.9, 17.7, 14.4, 13.8, 14.8, 11.9, 13.0, 14.7, 16.5, 11.0)

# negative of Weibull log likelihood
nll<-function(shape,scale) {
   cat(shape,scale," ")
   value<--sum(log(dweibull(DBH,shape,scale)))
   cat(value,"\n")
   value
   }

# save ML-estimate into object 'solution'
solution<-mle(minuslogl=nll,start=list(shape=10,scale=10))
10 10  1806.286
10.001 10  1807.236
9.999 10  1805.337
10 10.001  1804.436
10 9.999  1808.139
-939.2394 1861.296  NaN
-179.8479 380.2591  NaN
-27.96957 84.05182  NaN
2.406085 24.81036  67.82072
2.407085 24.81036  67.82053
2.405085 24.81036  67.8209
2.406085 24.81136  67.8221
2.406085 24.80936  67.81933
...
<part of the output was removed>
...
7.603536 15.50216  44.4477
7.604536 15.50116  44.44770
7.602536 15.50116  44.4477
7.603536 15.50216  44.4477
7.603536 15.50016  44.44771
Warning messages:
1: NaNs produced in: dweibull(x, shape, scale, log)
2: NaNs produced in: dweibull(x, shape, scale, log)
3: NaNs produced in: dweibull(x, shape, scale, log)
>
> # Take the summary and asymptotic variance-covariance matrix
> summary(solution)
Maximum likelihood estimation

Call:
mle(minuslogl = nll, start = list(shape = 10, scale = 10))

Coefficients:
       Estimate Std. Error
shape  7.603536  1.3646050
scale 15.502161  0.4802482

-2 log L: 88.8954
> cov2cor(vcov(solution))
          shape     scale
shape 1.0000000 0.3144117
scale 0.3144117 1.0000000
```

The above output shows that the ML estimates for shape and scale are $\hat{\alpha} = 7.60$ *and* $\hat{\beta} = 15.50$*, with estimation errors of 1.36 and 0.48, respectively. The correlation of estimation errors is 0.314.*

The estimation gave three warnings. To see the causes for these warnings, the function giving the negative log likelihood was set to print the evaluated values of α *and* β*, as well as the value of the negative log likelihood onto the screen using function* `cat()`*. We see that for three cases, the Negative likelihood got the value NaN (Not*

a number). This resulted from that the algorithm used in finding estimators tried to evaluate the likelihood with a negative value of shape parameter. However, the shape can only be positive, and function `dweibull()` *returns value NaN for negative values of shape. However, the algorithm did not fail to converge for this problem.*

A link function can be used to ensure that only values within the parameter range would be evaluated. Defining $\alpha = \exp(\theta_1)$ *and* $\beta = \exp(\theta_2)$*, ensures that* α *and* β *are always positive, just because the exponential function is defined for any real-valued* x*, and it gives only values above zero. Thus, we are applying link function* $\tau(\theta) = \ln(\theta)$ *The invariance property of the MLE implies that such approach should lead to same point estimates than the previous approach.*

```
nll2<-function(theta1,theta2) {
    shape<-exp(theta1)
    scale<-exp(theta2)
    value<--sum(log(dweibull(DBH,shape,scale)))
    cat(shape,scale,value,"\n")
    value
    }

# save ML-estimate into object 'solution'
solution2<-mle(minuslogl=nll,start=list(theta1=log(10),theta2=log(10)))
10 10 1806.286
10.01001 10 1815.809
9.990005 10 1796.824
10 10.01001 1787.867
10 9.990005 1824.893
0 Inf NaN
0 Inf NaN
1.256009e-164 Inf NaN
...
<part of the output was removed>
...
7.603944 15.53322 44.45
7.603944 15.50218 44.4477
7.611552 15.48669 44.44836
7.596344 15.48669 44.44824
7.603944 15.50218 44.4477
7.603944 15.47121 44.45003
There were 14 warnings (use warnings() to see them)

exp(coef(solution2))
   theta1    theta2
 7.603944 15.502181
```

We see that the evaluation still results warnings, because the algorithm evaluates the likelihood with ultimately high or small positive values. The resulting point estimates are equal to the one of the previous approach up to three significant digits. The small differences in the estimates arise from the numerical accuracy of the algorithm.

Example 1.31 *In Example 1.12, we derived the distribution of canopy height observations, assuming that the crown shape seen from the side is an ellipsoid centered at* (x_0, y_0)*. Assume that we have observed the canopy height at random points within the crown with sampling density 4 observations per m*^{2} *from within a tree crown using laser scanner. We know that the tree height is necessarily larger than equal to the maximum observed height, where the equality corresponds to the improbable situation that one of the observations has hit the tree top. Assuming that the crown shape can be well described with the ellipsoid, we want to estimate the crown shape, and especially, the tree height. We proceed by fitting the density of canopy height observations, by maximizing a profile likelihood, where one of the parameters is profiled out of the likelihood using the information on the sampling density. We first reparameterize the density* (1.7) *using the maximum crown radius, i.e., the relative radius (as a fragment of total height) at relative height* y_0 *by defining* $maxr = y0 + b$*. Then, based on the assumption about*

circular shape of the cross-sectional crown, the maximum crown radius can be solved from equation $n/4 = \pi * (h * maxr)^2$ *as* $maxr = \frac{1}{h}\sqrt{\frac{n}{4\pi}}$*, where* n *is the number of laser observations we have on the tree crown. Thus we can eliminate* b *from our density, resulting in that the likelihood has only three parameters to be estimated, namely* h*,* x_0*, and* y_0*. The likelihood is defined in the R-code below.*

```
# Read the data
> thistree<-read.table("c:/laurim/biometria/lasertree.txt",header=TRUE)
# save laser heights to x
> x<-thistree$z
# initial guesses for the three parameters
> hinit<-max(x+0.01)
> x0init<-min(x-0.01)/hinit
# the profile likelihood
> minuslogl<-function(h=hinit,x0=x0init,y0=-0.015) {
+            maxr<-sqrt(length(x)/(4*pi))/h
+            -sum(log(pdf.laser(x,maxr-y0,x0,y0,h)))
+            }
```

The assumed model was fitted to data. Note that we have constraints for the parameters: hx_0 *should be less than the lowest laser observation,* h *should be greater than the highest laser observation, and* y_0 *should be below zero. These constraints could be implemented using log links, as we did in an earlier example on Weibull distribution. However, this time we use constrained optimization in finding the parameter estimation, by using algorithm* `L-BFGs-B` *in estimation. The simple bounds for the parameters are given using parameters* `lower` *and* `upper`*. We need to use slightly lower maximum for* x_0 *than* $\min(x)$ *and a slightly higher minimum for* h *than* $\max(x)$*, because equality would lead to infinite profile likelihood. The summary of fit is given below.*

```
> fit<-mle(minuslogl,method="L-BFGS-B",lower=c(hinit,-Inf,-Inf),upper=c(Inf,x0init,0))
> summary(fit)
Maximum likelihood estimation

Call:
mle(minuslogl = minuslogl, method = "L-BFGS-B", lower = c(hinit,
    -Inf, -Inf), upper = c(Inf, x0init, 0))

Coefficients:
     Estimate Std. Error
h  15.1564548 0.14582940
x0  0.4385392 0.02250756
y0 -0.1114142 0.09192179

-2 log L: 345.7253
> windows(width=3,height=5)
> maxr<-sqrt(length(x)/(4*pi))/coefs[1]
> par(mfcol=c(2,1),mai=c(0.6,0.5,0.1,0.1),mgp=c(2,0.7,0),cex=0.8)
> hist(x,freq=FALSE,xlim=c(6,16),main=NA,xlab="Height,m")
> coefs<-coef(fit)
> y<-seq(0,16,0.1)
> lines(y,pdf.laser(y,maxr-coefs[3],coefs[2],coefs[3],coefs[1]))
> plot(y,radius.ellipse(y,maxr-coefs[3],coefs[2],coefs[3],coefs[1]),
       type="l",xlab="Height,m",ylab="Crown radius, m")
```

The estimated tree height was 15.16 meters, being 0.17 meters higher than the maximum of laser observations, 14.99 meters. The standard error of estimate was 0.15 meters. The upper plot of figure 1.9 shows a histogram of the observations and the fitted distribution. The lower plot shows the corresponding crown shape. The utilized functions `pdf.laser` *and* `radius.ellipse` *were defined in example 1.12.*

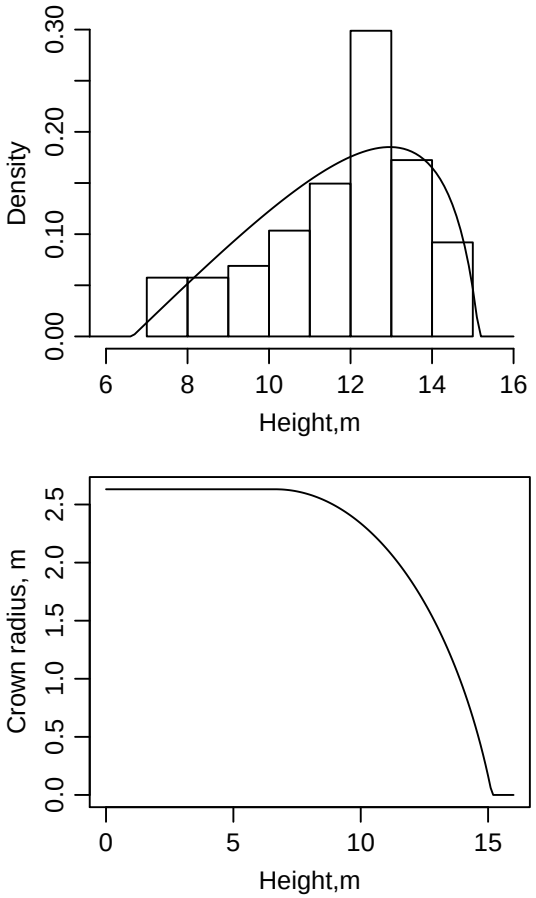


Figure 1.9: Illustration of example 1.31

1.5.2 Other methods

The principle of Maximum likelihood is the best justified approach for finding point estimates for unknown parameters. However, there are also other methods, that have been used in forestry, especially in fitting diameter distributions to data. Such methods are the method of moments, the method of percentiles, and cdf regression.

In the method of moments, the first k moments $\mathrm{E}(X^1), \ldots, \mathrm{E}(X^k)$ of an assumed distribution function are set equal to the sample moments, and parameters are solved from the resulting system of k equations. In the case of a two-parameter distribution, the method of moments gives a distribution having the same expectation and variance as are the sample estimators.

The method of percentiles is analogous to the method of moments, but it utilizes sample percentiles, which are set equal to the theoretical quantiles of the assumed distribution function. In addition to moments and percentiles, also other quantities can be used for finding equations for a system of k equations. For example, the values of basal area, mean diameter, and the number of stems could be set equal to those derived from the assumed distribution. For those equations, analytical solutions do not usually exist, and the resulting equations need to be solved numerically. Methods based on percentiles or other stand characteristics are commonly called parameter recovery methods in the literature of diameter distributions. In some approaches, the number of equations is higher than k, and a subset of the equations is used for each parameter. In such case, the solution fulfills all the equations only in a rare special case.

Some forestry studies have used the method of cdf regression to fit an assumed distribution function to tree diameter data. In that approach, the diameter observations are ordered, and the ith smallest observation of the sample of size n is interpreted as $i/(n+1)$th quantile of the distribution. The estimated values are obtained by fitting the cdf of the assumed distribution function to the data of diameters and the corresponding values of the empirical cdf. I cannot see any reason to favor this method over the theoretically better justified methods, such as the method of maximum likelihood.

The bayesian approach is still one more method of finding estimators. In the bayesian approach, the parameters are treated as random variables. The prior belief of the value of a parameter are parameterized using a prior distribution, which is combined with the distribution of the data. This yields a posterior distribution of the parameter of interest. A point estimate of the parameter could be obtained as the mean of the posterior distribution. The Posterior distribution can also be used to make more detailed inference on the parameter of interest, such as computing different confidence intervals.

1.6 Linear prediction

The general case from (Lappi et al. 2006).

Assume that a random vector $\boldsymbol{h}$ of length k can be divided into two parts

$$\boldsymbol{h} = \begin{pmatrix} \boldsymbol{h_1} \\ \boldsymbol{h_2} \end{pmatrix}$$

where $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are random vectors, or scalars in the special case that the length of random vector is 1. It is assumed that $\mathrm{E}(\boldsymbol{h}_1) = \boldsymbol{\mu}_1$, $\mathrm{E}(\boldsymbol{h}_2) = \boldsymbol{\mu}_2$, $\mathrm{var}(\boldsymbol{h}_1) = \boldsymbol{V}_1$, $\mathrm{var}(\boldsymbol{h}_1) = \boldsymbol{V}_1$, and $\mathrm{cov}(\boldsymbol{h}_,\boldsymbol{h}_2) = \boldsymbol{V}_{12}$. Using the notation of McCulloch and Searle (2001, p. 247), this can be written as

$$\begin{pmatrix} \boldsymbol{h_1} \\ \boldsymbol{h_2} \end{pmatrix} \sim \left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_{12} \\ \boldsymbol{V}'_{12} & \boldsymbol{V}_2 \end{pmatrix} \right]$$

Assume that we have observed the random vector $\boldsymbol{h}_2$ and want to predict vector h_1. The Best Linear Predictor (BLP) of $\boldsymbol{h}_1$ is

$$BLP(\boldsymbol{h}_1) = \widehat{\boldsymbol{h}_1} = \boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_2^{-1}(\boldsymbol{h}_2 - \boldsymbol{\mu}_2) \tag{1.39}$$

with a prediction variance of

$$\mathrm{var}(\widehat{\boldsymbol{h}_1} - \boldsymbol{h}_1) = \boldsymbol{V}_1 - \boldsymbol{V}_{12}\boldsymbol{V}_2^{-1}\boldsymbol{V}'_{12} \tag{1.40}$$

(McCulloch and Searle 2001, p. 250). This result means that if the expectations and variance-covariance matrices of two random vectors are known and either one of them is observed, the other one can be predicted using (1.39). Furthermore, the variance of the prediction error can be calculated using Equation (1.40).

If $\boldsymbol{h}$ follows the multinormal distribution, i.e.,

$$\begin{pmatrix} \boldsymbol{h_1} \\ \boldsymbol{h_2} \end{pmatrix} \sim N_k \left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_{12} \\ \boldsymbol{V}'_{12} & \boldsymbol{V}_2 \end{pmatrix} \right],$$

then BLP is the Best Predictor. This results from that with multivariate normal distribution, all covariances are linear, implying that no nonlinear predictor can be better than the best linear one.

In practice, matrices $\boldsymbol{V}_1$, $\boldsymbol{V}_2$, and $\boldsymbol{V}_{12}$ as well as vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{mu}_2$ are unknown, and will be replaced with their estimates. The resulting predictor is the Estimated Best Linear Predictor (EBLP).

Example 1.32 *Assume that a model system is available for prediction of tree volume and tree height using tree diameter only. Assuming that the form of the model system is correct and the parameters do not include errors, such a model system would give the expected values for height and volume for trees with known diameter. Furthermore,*

the residuals of the model are assumed to have constant variances so that the residual variance of height would be $\boldsymbol{V}_2 = 0.5^2$ *and that of volume would be* $\boldsymbol{V}_1 = 2^2$*, and the covariance between them is* $\boldsymbol{V}_{12} = 0.5$.

Assume that a sample tree has been measured for diameter and height, the observed height being $\boldsymbol{h}_2 = 14$ *meters, respectively. The volume model gives the expected volume as* $\boldsymbol{\mu}_1' = 70$ *liters, and expected height as* $\boldsymbol{\mu}_2' = 13$ *meters. In this case, both vectors are scalars. However, we write them using boldface font in order to avoid introduction of new notations, and to show that they can just be treated as vectors of length one. The BLP is*

$$\begin{aligned}\widehat{\boldsymbol{h}_1} &= \boldsymbol{\mu} + \boldsymbol{V}_1\boldsymbol{V}_{12}^{-1}(\boldsymbol{h}_2 - \boldsymbol{\mu}_2) \\ &= 70 + 0.5 \times 0.25^{-1}(14 - 13) \\ &= 70 + 2 * 1 = 72\mathrm{dm}^3\end{aligned}$$

Further examples on linear prediction will be given when dealing the linear mixed models and models systems. With mixed-effects models, the BLP is used to predict stand effects using observations of the response. For example, we can predict stand effects of a height-diameter curve using measured heights and diameters from the stand of interest, resulting into a local models for that particular stand. With models systems, the BLP could be utilized to carry information from one model to another, as we did in the above example.

1.7 Exercises

Try to do exercises 1-5, the rest are additional excercises if you have time left.

1. Tree diameter in a stand follows *Weibull*(8,15) distribution. Stand density is 500 trees per ha. Tree height h depends on tree diameter d according to the Korf function $h(d) = 25 * exp(-3/d)$.

   (a) Plot the distribution function and density of tree diameter.

   (b) Compute the expected value and standard deviation of tree diameter and add a vertical line to the plots of density and distribution function at the mean diameter. Interpret the standard deviation.

   (c) Compute the number of trees with diameter above 18 cm.

   (d) Plot the height-diameter curve.

   (e) Plot the inverse function of the H-D curve.

   (f) Plot the distribution function of tree diameter ($F_H(h)$)and the corresponding density ($f_H(h) = F_H'(h)$).

(g) Numerically compute the mean and standard deviation of tree heights as follows. Define first R-functions for $hf_H(h)$ and $h^2 f_H(h)$. Then use `integrate(fn,lower,upper)` for numerically evaluate $\mathrm{E}(H) = \int_0^\infty hf_H(h)dh$ and $\mathrm{E}(H^2) = \int_0^\infty h^2 f_H(h)dh$. In `integrate(fn,lower,upper)`, `fn` is the function to be integrated and `lower` and `upper` are the bounds of integration. Finally, use rule the rule 1.17 to compute the variance.

(h) Dominant trees include 100 tallest trees of a stand. Plot the distibution function and density of the diameter of dominant trees.

(i) Plot the distibution function and density of the height of dominant trees.

(j) Compute the dominant height, i.e., the mean of dominant trees as the expected value of the height distribbution of dominant trees.

2. Simulate square 20 by 20 meter sample plot from a forest stand, where the stand density is 1000 trees per ha, trees are located randomly, and tree diameter follows *Weibull(*$5, 15$*)* distribution as follows

(a) Generate the realized number of sample trees for the plot from *Poisson*(λA) distribution, where λ is stand density and A is the plot area.

(b) Generate x and y coordinates for tree locations from the *Uniform* distribution.

(c) Generate tree diameters from the *Weibull* distribution.

(d) Plot the trees using `plot(x,y,cex)`, where x and y are coordinates and cex is the size of symbol which is prportional to tree diameter.

(e) Fit Weibull distribution to the simulated data using the method of maximum likelihood.

3. In a forest stand, tree joint distribution of tree diameter and height is characterized with the Multinormal distribution with $\mathrm{E}\begin{bmatrix} D \\ H \end{bmatrix} = \begin{bmatrix} 20 \\ 18 \end{bmatrix}$ and $\mathrm{cov}\begin{bmatrix} D \\ H \end{bmatrix} = \begin{bmatrix} 9 & 3 \\ 3 & 4 \end{bmatrix}$.

(a) Predict tree heights for a tree with known diameter of 10 cm using BLP.

(b) List the properties of your estimate assuming that the multivariate normality holds.

(c) Plot the implicitly assumed H-D curve. Hint, predict heights for several diameters with regular intervals and plot the results.

(d) Write down the mathematical expression of the assumed H-D relationship.

4. Using equation on equation (1.19), show that covariance is a special case of variance.

5. Derive the expected value and variance of the uniform distribution.

6. Let Y have the *Binomial(n, p)* distribution. Using rules (1.15) and (1.24), show that $\mathrm{E}(Y) = np$ and $\mathrm{var}(Y) = np(1-p)$.

7. Let X have *Poisson(λ)* distribution. Using (1.11) and (1.17), show that $\mathrm{E}(X) = \mathrm{var}(X) = \lambda$.

8. Show that if X follows the two-parameter version of *beta* distribution, then $Y = a + (b-a)X$ follows the four-parametric version of the *beta* distribution. Based on this relationship, derive the expected value and variance of Y.

9. The density of random variable $Y_{r:n}$, that is, the rth smallest tree in a sample of size n, from a population with underlying cdf and pdf of $F_Y(y)$ and $f_Y(y)$, respectively, is given by

$$f_{r:n}(y) = \frac{n!}{(r-1)!(n-r)!} f_Y(y) \left[F_Y(y)\right]^{r-1} \left[1 - F_Y(y)\right]^{n-r} .$$

   Assume that trees have been sampled from a stand with the percentile-based diameter distribution of example 1.6?

   (a) What is the density of $Y_{r:n}$?

   (b) Compute the experted value of the minimum diameter of a sample of 12 trees, $Y_{1:12}$.

   (c) Compute the variance and standard deviation of $Y_{1:12}$.

   (d) Interpret the computed values for expected value and standard deviation.

10. The joint density of $Y_{r1:n}$ and $Y_{r2:n}$ ($r1 < r2$), i.e., the $r1$th smallest and $r2$th smallest trees in a sample of size n from a population with underlying cdf and pdf of $F_Y(y)$ and $f_Y(y)$, respectively, is given by

$$\begin{aligned} f(y_1, y_2) \quad = \quad & \frac{n!}{(n-r2)!(r_2 - r_1 - 1)!(r_1 - 1)} f_Y(y_1) f_Y(y_2) \\ & \left[F_Y(y)\right]^{r_1 - 1} \left[F_Y(y_2) - F_Y(y_1)\right]^{r_2 - r_1 - 1} \left[1 - F_Y(y)\right]^{n - r_2} . \end{aligned}$$

   Assume that the underlying population distribution is the percentile-based distribution of example 1.6.

   (a) Derive the joint density of $Y_{r1:n}$ and $Y_{r2:n}$.

   (b) Derive the conditional density of $Y_{1:12}$ given that $Y_{2:12} = 10$.

(c) Compute the expected value $\mathrm{E}(Y_{1:12}Y_{2:12})$.

(d) Compute the expected values $\mathrm{E}(Y_{1:12})$, $\mathrm{E}(Y_{2:12})$, $\mathrm{E}(Y_{1:12}^2)$, and $\mathrm{E}(Y_{2:12}^2)$ using the results from the earlier exercise on a univariate order statistic.

(e) Using the previous results, compute $\mathrm{cov}(Y_{1:12}Y_{2:12})$ and $\mathrm{corr}(Y_{1:12}Y_{2:12})$.

11. Bivariate distribution for diameter distribution at two points in time (the paper published by xxx in Biometrics).

Chapter 2

Linear model

Linear regression, or the linear model, is maybe the most widely used statistical tool. It is used to analyze the dependence of variable y, e.g., tree volume, on other variables x, e.g., tree diameter and height. The terminology of regression modeling depends much on the field of study, and also on the purpose of the modeling. The different names used for variable y include terms dependent variable, response variable, the regressand, the measured variable, the responding variable, the explained variable, and the outcome variable. For variable x, terms independent variable, predictor variable, regressor, controlled variable, manipulated variable, and explanatory variable are used. In the basic regression analysis, it is often assumed that x-variables are set by the researcher, and the effect of changes in x on the value of y is analyzed. However, in many cases, the values of x cannot be controlled. This does not cause problems into the analysis, but the inference on the model is valid for the dataset used in analysis.

In the case of single predictor regression, we will have one response and one predictor. For example, we may predict tree height on tree diameter. In multiple regression, the number of predictors may is higher than 1. For example, Standing tree volume may be predicted on tree diameter, height and upper diameter. In most cases, the number of responses is one. However, we may also have several responses, for example, we may wish to simultaneously predict tree volume and height. In this situation, a system of models is fitted. Those situations are handled in chapter 5.

The regression may be either nonlinear or linear. In this context, the linearity means that variable y, or any transformation of it, is linear in predictors x or any transformations of it. Thus, the linear regression applies in all instances where the relationship between $x_1, \ldots, x_n$ and y can be written as

$$f_y(y) = \beta_0 + \beta_1 f_1(x_1) + \beta_2 f_2(x_1) + , \ldots , + \beta_n f_n(x_n) \, ,$$

where $\beta_1, \ldots, \beta_n$ are parameters to be estimated from the data.

In nonlinear regression, the relationship may be expressed with any function

$$y = f(x_1, \ldots, x_n, \beta_1, \ldots, \beta_m) .$$

For example, we could assume relationship

$$y = \beta_0 x^{\beta_1} ,$$

between x and y, which is no more linear. However, the relationship could be linearized by taking logarithms from both sides to get

$$\ln y = \ln \beta_0 + \beta_1 \ln(x) .$$

The above example shows a very common way to linearize nonlinear relationships, namely taking logarithms. However, in many cases the relationship cannot be linearized. The widely used and flexible Champman-Richards function (Richards 1959) is an example of such a function

$$y = \beta_1 \left(1 - e^{-\beta_2 x}\right)^{\beta_3} .$$

2.1 Single-predictor regression

2.1.1 Model formulation

In single predictor regression, we have one response and one predictor. Assume that variables y_i and x_i are measured for individuals i, $i = 1, \ldots, n$. The simple linear regression model for the ith individual can be written as

$$y_i = b_0 + b_1 x_i + e_i . \tag{2.1}$$

In equation (2.1), we assume that variable y_i comprises of two parts: of a systematic part that depends on the values of x_i through the assumed relationship, and of a random part caller residual or random error. Parameters b_0 and b_1 are parameters of the assumed regression model, which specify the systematic dependence on x_i, and e_i is the residual of individual i. Part $b_0 + b_1 x_i$ expresses the average or mean relationship between variables x and y. It can also be interpreted as the expected value of y for an individual with given value of x. The residual e_i in turn expresses how much the value of y of the particular tree i differs from that expected value or mean in the horizontal direction. It is usually assumed that $\mathrm{E}(e_i = 0$, $\mathrm{var}(e_i) = \sigma^2$, and $\mathrm{cov}(e_i, e_j) = 0$ when $i \neq j$.

Example 2.1 *20 Trees were measured for diameter and height in a stand. The observations have been printed in table 2.1 plotted in figure 2.1. It seems a linear model of form*

$$h_i = b_0 + b_1 d_i + e_i$$

Table 2.1: Twenty measured diameter-height pairs from a Scots pine - Norway spruce mixture.

d	h	pl	d	h	pl	d	h	pl	d	h	pl
31	22.8	1	28.3	19.7	1	26.9	21.3	2	26.2	19.3	1
29.9	21.6	2	28.7	21.3	1	25.7	20.7	2	25.3	20.3	1
30	20.5	2	27.9	21.3	1	26.3	19.3	1	25.5	21	2
29.6	22.5	1	27.6	20.8	2	26.2	20.2	2	25.6	19.2	1
29.4	20.4	1	27.3	24	2	25.4	20.5	2	24.3	19.3	2

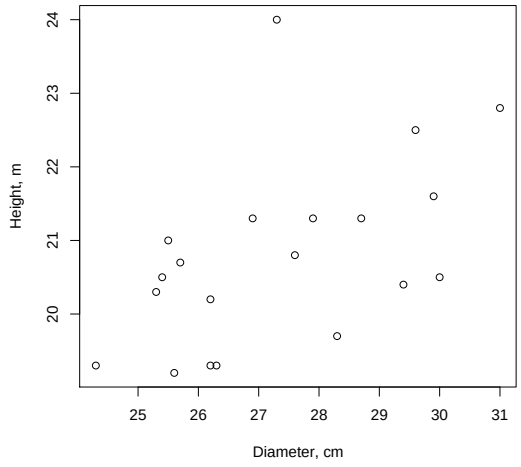


Figure 2.1: Twenty measured diameter-height pairs from a Scots pine - Norway spruce mixture.

could be a good starting point for modeling the height-diameter relationship in the example stand.

Example 2.2 *A total of 200 Scots pine trees were measured for diameter and height from 44 plots in North Carelia. The data were read to R using*

```
> hddata<-read.table("c:/laurim/biometria/hddata.txt",header=TRUE)
```

To see if the relationship between diameter and height is linear, we start by plotting the data.

```
> windows(width=3,height=6)
> par(mfcol=c(2,1),mai=c(0.6,0.5,0.1,0.1),mgp=c(2,0.7,0),cex=0.8)
> plot(hddata$d,hddata$h)
```

The plot shows slight curvature. We try to linearize the relationship by making transformations to diameter. After some trials, we find that the relationship between $\ln(d+10)$ *and* h *is quite close to linear.*

```
> plot(log(hddata$d+10),hddata$h)
```

Thus, a good starting point for our modeling purposes is the following simple linear model

$$h_i = b_0 + b_1 \ln(d_i + 10) + e_i$$

where h_i*,* d_i*, and* e_i *are the height, diameter and residual of tree* i*,* $i = 1, \ldots, 200$*. The residual is assumed to have expectation of zero, i.e., the average difference between observed height and the fitted model is assumed to be zero. Furthermore, it is assumed that the variance of true height around the regression line is constant, i.e., it does not vary according to the predicted height.*

2.1.2 Estimation with least squares

In the method of least squares, values for parameter vector $\boldsymbol{b}$ are estimated by searching the value that minimize the sum of squared residuals. With single predictor regression (2.1), the sum of squared residuals can be written as

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

The values of b_0 and b_1 that minimize the squared residuals is found at a point where the first derivative is zero, given that that point is the global minimum. Differentiating the sum of squared residuals with respect to b_0 and b_1 (we believe that they are global minimums), and setting the derivates equal to 0 gives equations

$$-2\sum_{i=1}^{n} y_i + 2nb_0 - 2b_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} 2y_i x_i - 2b_0 \sum_{i=1}^{n} x_i - 2b_1 \sum_{i=1}^{n} x_i^2 = 0$$

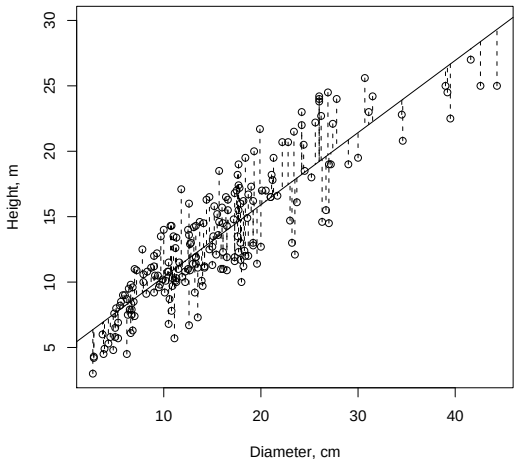


Figure 2.2: The plot of tree height against diameter in the data of example 2.2.

Solving the upper equation for b_0 yields $b_0 = \bar{y} - b_1\bar{x}$. Writing the solution into the latter equation and solving for b_1 yields $b_1 = \frac{\frac{1}{n}\sum_i x_i \sum_i y_i - \sum_i xy}{\frac{1}{n}\left(\sum_i x_i\right)^2 - \sum_i x^2}$.

The residual variance is estimated by

$$\widehat{\sigma^2} = \frac{RSS}{n-2}\,, \tag{2.2}$$

where $RSS = \sum_i (y_i - \widehat{b_0} - \widehat{b_1}x_i)^2$ is the residual sum of squares.

Example 2.3 *In example 2.5, the required sums are* $\sum_i d_i = 547.1$, $\sum_i h_i = 416$, $\sum_i d_i h_i = 11403.52$, *and* $\sum_i d_i^2 = 15035.59$. *The OLS estimates for regression coefficients become*

$$\begin{aligned}\widehat{b_1} &= \frac{\frac{1}{n}\sum_i d_i \sum_i h_i - \sum_i d_i h_i}{\frac{1}{n}\left(\sum_i d_i\right)^2 - \sum_i d^2} \\ &= \frac{1/20 \times 547.1 \times 416 - 11403.52}{1/20 \times 547.1^2 - 15035.59} = 0.3421870\end{aligned}$$

and

$$\begin{aligned}\widehat{b_0} &= \bar{h} - b_1\bar{d} \\ &= 1/20 \times 416 - 0.3421870 \times 1/20 \times 547.1 = 11.43947\end{aligned}$$

The estimate of residual variance is

$$\begin{aligned}\widehat{\sigma^2} &= \frac{\sum_i (h_i - \widehat{b_0} - \widehat{b_1}d_i)^2}{n-2} \\ &= 1.215681\end{aligned}$$

The above computations were carried out with the following code:

```
> sumd<-sum(onestand$d)
> sumh<-sum(onestand$h)
> sumdh<-sum(onestand$d*onestand$h)
> sumd2<-sum(onestand$d^2)
> n<-dim(onestand)[1]
> b1<-(1/n*sumd*sumh-sumdh)/(1/n*sumd^2-sumd2)
> b1
[1] 0.3421870
> b0<-sumh/n-b1*sumd/n
> b0
[1] 11.43947
> RSS<-sum((onestand$h-b0-b1*onestand$d)^2)
> sigma2<-RSS/(n-2)
> sigma2
[1] 1.215681
> sqrt(sigma2)
[1] 1.102579
```

The same results are obtained using the R-function `lm`

```
fm1<-lm(h~d,data=onestand)
fm1
```

A little bit more information on the model is obtained using function `summary()`

```
> summary(fm1)

Call:
lm(formula = h ~ d, data = onestand)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42337 -1.02454 -0.01555  0.51366  3.21882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.4395     3.6219   3.158  0.00544 **
d             0.3422     0.1321   2.590  0.01847 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.103 on 18 degrees of freedom
Multiple R-squared: 0.2716,     Adjusted R-squared: 0.2311
F-statistic:  6.71 on 1 and 18 DF,  p-value: 0.01847
```

At this stage, we just note that our manually computed figures agree with those given by function `lm`*. The fitted line can be added to the plot using*

```
> abline(b0,b1) # try also abline(fm1)
```

Furthermore, we illustrate the residuals using vertical dashed lines

```
> fn1<-function(x) lines(rep(onestand$d[x],2),c(onestand$h[x],predict(fm1)[x]),lty="dashed")
> sapply(1:dim(onestand)[1],fn1)
```

2.2 Multiple regression

2.2.1 Model formulation

The situation does not change very much from that of the previous subsection if we have more than one predictor. The linear model is

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i}, \ldots, + b_p x_{pi} + e_i \,. \tag{2.3}$$

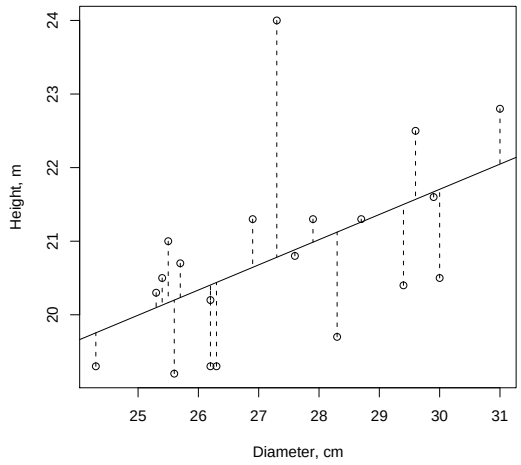


Figure 2.3: The fitted regression line and the residuals of the twenty measured diameter-height pairs from a Scots pine - Norway spruce mixture.

Now coefficicients $b_0, b_1, \ldots, b_p$ specify the expected dependence of y_i on the predictors $x_{1i}, \ldots, x_{pi}$, or the systematic part of the model. Again, the residual e_i in expresses how much the value of y of the particular tree i differs from that expected value or mean in the horizontal direction, the assumptions are the same as with single-predictor regression, namely that $\mathrm{E}(e_i = 0$, $\mathrm{var}(e_i) = \sigma^2$, and $\mathrm{cov}(e_i, e_j) = 0$ when $i \neq j$. Note that no specific distribution is assumed for e_i at this stage. Assumptions $\mathrm{var}(e_i) = \sigma^2$ and $\mathrm{cov}(e_i, e_j) = 0$ can be relaxed so, that a specific structure may be assumed for variances and covariances. These issues are discussed later in chapters 3 and 2.1.2

Example 2.4 *As an alternative to model of example 2.2, one could assume that tree height depends not only on tree diameter, but on some properties of the stand, such as the basal area, basal area median diameter, and geographical northern co-ordinate of the plot location. The assumed model would be*

$$h_i = b_0 + b_1 \ln(d_i + 10) + b_2 G_i + b_3 DGM_i + b_4 YK_i + e_i$$

where G_i, DGM_i, and YK_i are the basal area (m^3/ha), basal area median diameter (cm,), and the geographical north co-ordinate of i, $i = 1, \ldots, 200$. A practical difference to the single-predictor regression is immediately noticed, namely that the predictions are represented by points that do not form a line in the $h - d$-plane. Plot 2.4 shows the observed heights with open circles, and the fitted values of a multiple regression model with black triangles.

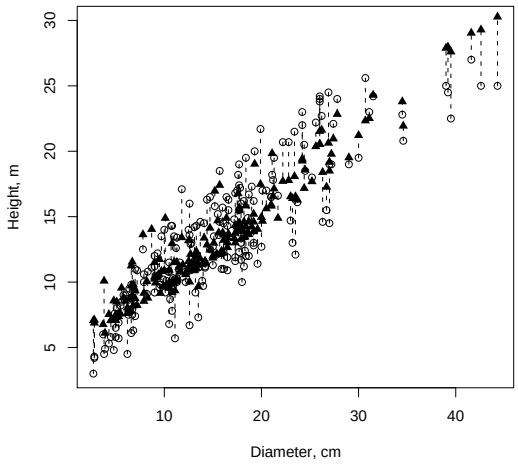


Figure 2.4: The plot of tree height against diameter in the data of example 2.4.

According to the type of predictors, the linear model can be called regression model, analysis of variance model, and analysis of covariance model. In the **regression analysis** model, predictors are usually continuous random variables, such as tree diameter, or altitude above sea level. In **variance analysis** (ANOVA) model, the predictors are so called *dummy variables* or *indicator variables*, which only can get values 0 and 1, indicating either presence or absence of a certain property or feature. They may indicate, for example, the origin of a seedling (e.g., 1 means planted and 0 natural), or if a certain treatment was applied or not. For example, value 1 may indicate the new silvicultural treatment and value 0 the old treatment. Furthermore, indicator variables can be used to parameterize categorical variables, such as site fertility class, to a regression model. Classifications with more than two classes require several dummy variables. The **covariance analysis** model is a combination of these two models, including both continuous variables and indicator variables as predictors.

It is often convenient to write the linear model (2.3) in a matrix form. In the matrix form, all observations of y are written into a column vector $\boldsymbol{y}_{n\times 1}$, the predictors into matrix $\boldsymbol{X}_{n\times(p+1)}$, the coefficients into column vector $\boldsymbol{b}_{p\times 1}$, and the residuals into a column vector $\boldsymbol{e}_{n\times 1}$. The model becomes

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

or simply

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \tag{2.4}$$

Matrix $\boldsymbol{X}$ is called the *design matrix*. Note that the first column of the design matrix has ones. This is to include the constatnt to the model, i.e, the constant of (2.3) can be written as $b_i \times 1$ The assumptions about the residual vector $\boldsymbol{e}$ can be stated as

$$\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$$
$$\mathrm{var}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$$

Furthermore, because the fixed part does not include randomness (design matrix X and coefficients b are thought as fixed), these assumptions imply that

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$$
$$\mathrm{var}(\boldsymbol{y}) = \mathrm{var}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$$

Example 2.5 *On one sample plot, 20 trees were measured for diameter and height. The data are shown in table 2.1. The simple linear model of Example 2.2 can be presented in a matrix form as*

$$\boldsymbol{h} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}\,,$$

where

$$\boldsymbol{h} = \begin{pmatrix} 22.8 \\ 21.6 \\ 20.5 \\ 22.5 \\ 20.4 \\ 19.7 \\ 21.3 \\ 21.3 \\ 20.8 \\ 24.0 \\ 21.3 \\ 20.7 \\ 19.3 \\ 20.2 \\ 20.5 \\ 19.3 \\ 20.3 \\ 21.0 \\ 19.2 \\ 19.3 \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & 31 \\ 1 & 29.9 \\ 1 & 30 \\ 1 & 29.6 \\ 1 & 29.4 \\ 1 & 28.3 \\ 1 & 28.7 \\ 1 & 27.9 \\ 1 & 27.6 \\ 1 & 27.3 \\ 1 & 26.9 \\ 1 & 25.7 \\ 1 & 26.3 \\ 1 & 26.2 \\ 1 & 25.4 \\ 1 & 26.2 \\ 1 & 25.3 \\ 1 & 25.5 \\ 1 & 25.6 \\ 1 & 24.3 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \quad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_1 0 \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \\ e_{17} \\ e_{18} \\ e_{19} \\ e_{20} \end{pmatrix},$$

2.2.2 Estimation with least squares

Solution of the model (2.3) becomes easily tedious to compute, because the number of equations needed equals to the number of parameters in $\boldsymbol{b}$. That is why linear models

and the least squares solutions are usually presented in the matrix form. The sum of squared residuals is expressed in the matrix form as $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$. The OLS estimator of $\boldsymbol{b} = (b_0, b_1)'$ is in the matrix form

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

It can be shown that the least squares estimator is unbiased and has variance-covariance matrix $\sigma^2\boldsymbol{X}'\boldsymbol{X}^{-1}$:

$$\begin{aligned} \mathrm{E}(\widehat{\boldsymbol{b}}) &= \mathrm{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{E}\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} \\ &= \boldsymbol{b} \end{aligned}$$

$$\begin{aligned} \mathrm{cov}(\widehat{\boldsymbol{b}}) &= \mathrm{cov}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\left(\mathrm{cov}\boldsymbol{y}\right)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2\boldsymbol{I}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right] \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

For estimation of residual variance, let us define matrix H as

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

An unbiased estimator for the residual variance is

$$\widehat{\sigma^2} = \frac{\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}}{n - p}\,, \tag{2.5}$$

where p is the number of fixed parameters in the model, i.e., the length of vector $\boldsymbol{b}$.

Note that we did not need any assumption on the distribution of residuals for these proofs. It can be shown that the LS estimator is the Best Linear Unbiased Estimator of the parameters of model (2.4). This means that from among estimators that are unbiased and linear with respect to observations, (i.e., are of form $\boldsymbol{a}\boldsymbol{y}$), the OLS-estimator has the smallest variance. However, this holds only when the assumptions (constant variance and no correlation) on residuals hold.

Example 2.6 *Using $\boldsymbol{X}$ and $\boldsymbol{h}$ as defined in example 2.5, the OLS estimate for $\boldsymbol{b}$ is*

$$\begin{aligned} \widehat{\boldsymbol{b}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{h} \\ &= \begin{pmatrix} 11.4394735 \\ 0.3421870 \end{pmatrix} \end{aligned}$$

The residual variance is

$$\widehat{\sigma^2} = \frac{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}}{n-p} = 1.215681$$

The R-code for Example 2.6

```
> onestand<-read.table("c:/laurim/biometria/onestand.txt",header=TRUE)
> X<-cbind(1,onestand$d)
> h<-onestand$h
> b<-solve(t(X)%*%X)%*%t(X)%*%h
> H<-X%*%solve(t(X)%*%X)%*%t(X)
> I<-diag(rep(1,20))
> sigma2<-t(h)%*%(I-H)%*%h/18
> b
           [,1]
[1,] 11.4394735
[2,]  0.3421870
> sigma2
         [,1]
[1,] 1.215681
```

The estimates are, again, exactly the same as those obtained in example 2.3

Example 2.7 *Let us fit the model of example 2.2 to the data of 200 height-diameter pairs from different stands. To study the trends in residuals, we plot the mean of variable y as a function of x in a given number of classes. In addition, we add two vertical, centered at the class mean. The first one is proportional to the class-specific standard deviation (multiplied by 2×1.96 to yield a 95% confidence interval based on assumption of normality) at. The second one is proportional to the class-specific standard errors of means is a similar way. Adding tehse lines to a scatterplot of residuals, gives a good plot for analyzing the trends in residuals. The lines based on class-specific standard deviations can be used to analyse if the residuals have constant variance ort not. The lines based on the standard error of meas can be used to assess if the model fits well enough over the whole range of predictions.*

Function mywhiskers can be used to make such a plot. Using `se=TRUE` produces the first version and using `se=FALSE` produces the second version.

```
> mywhiskers<-function(x, y, nclass=10, limits=NA, add=FALSE, se=TRUE, main="", xlab="x",
 ylab="y", ylim=NA, lwd=1) {
+     away<-is.na(x+y)
+     x<-x[!away]
+     y<-y[!away]
+     if (is.na(limits[1]))
+        limits<-seq(min(x),max(x)+1e-10,length=nclass+1)
+        else
+        nclass=length(limits)-1
+     means<-sapply(1:nclass,function(i) mean(y[x>=limits[i]&x<limits[i+1]]))
+     if (se) {
+        ses<-sapply(1:nclass,function(i) sd(y[x>=limits[i]&x<limits[i+1]])/
+                                          sqrt(sum(x>=limits[i]&x<limits[i+1])))
+        } else {
+        ses<-sapply(1:nclass,function(i) sd(y[x>=limits[i]&x<limits[i+1]]))
+        }
+     lb<-means-1.96*ses
+     ub<-means+1.96*ses
+     xclass<-1/2*(limits[-1]+limits[-nclass-1])
+     if (add) {
+        points(xclass,means)
```

```
+        } else {
+        if (is.na(ylim[1])) ylim<-c(min(lb),max(ub))
+           plot(xclass,means,ylim=ylim,main=main,xlab=xlab,ylab=ylab,xlim=range(x))
+        }
+     sapply(1:nclass,function(i) lines(xclass[c(i,i)],c(lb[i],ub[i]),lwd=lwd))
+     }
```

The following code wasused t ofit the model and plot the residuals:

```
> fm3<-lm(h~log(d+10),data=hddata)
> fm3

Call:
lm(formula = h ~ log(d + 10), data = hddata)

Coefficients:
(Intercept)  log(d + 10)
     -33.78        14.80
> windows()
> plot(predict(fm3),resid(fm3),xlab="Predicted value",ylab="Residual",ylim=c(-8,8))
> mywhiskers(predict(fm3),resid(fm3),se=FALSE,add=TRUE)
> mywhiskers(predict(fm3),resid(fm3),add=TRUE,lwd=3)
> abline(0,0)
```

All 95% confidence intervals for mean (the thick vertical lines) overlap the the x-axis (Figure 2.6). This means that the model of form $h = b_0 + b_1 \ln(d + 10)$ *wits rther well to the data. However, the class-specific standard deviations (the thin vertical lines) indicate increasing variance as a function of prediction.*

2.2.3 The design matrix

The design matrix specifies the structure of the model. It is usually a big matrix: the number of rows equals to the number of observations, and it has at least as many columns than there are predictors in your model. Some important points about the structure and construction of the design matrix are given here. Statistical packages such as R, SPSS or SAS automatically construct the design matrix, but it is still important for a modeler to understand what is done in the computer memory.

Continuous predictors are included in the original form, as we did in example 2.5. Including factors is a little different. Factors having only two levels are included as one dummy variable, which gets values 0 or 1. The coefficient of that variable expresses the difference of the level coded as 1 to the level coded as 0. Factors having $p > 2$ levels are coded as $p - 1$ dummy variables, each of them indicating if a specific level was present (dummy=1) or not (dummy=0). One level is selected as the default or reference level, which the other levels are compared to. From mathematical point of view, the selection of default level is unimportant. However, the selection of the default level affects the interpretation of the coefficients: the coefficients of other levels specify the difference to the default level. Also the t-tests on the individual coefficients test the hypothesis "do this level significantly differ from the default level'.' That is why the level can, and should be selected so that the interpretation makes sense.

Example 2.8 *In example 2.5, half of the trees were Scots pines, the rest 10 sample trees being Norway spruces. In addition to tree diameter, we could include a dummy variable into our model, indicating whether the tree is spruce or pine. There is no good reason to select which of the tree species would be coded as 0 (the default level). We just select Scots pine as the default.*

The model would be

$$h_i = b_0 + b_1 d_i + b_2 SPRUCE_i + e_i \tag{2.6}$$

where $SPRUCE_i$ is an indicator variable getting value 1 when tree i is spruce, and 0 elsewhere. The design matrix and parameter vector $\mathbf{b}$ become

$$\boldsymbol{X} = \begin{pmatrix} 1 & 31 & 0 \\ 1 & 29.9 & 1 \\ 1 & 30 & 1 \\ 1 & 29.6 & 0 \\ 1 & 29.4 & 0 \\ 1 & 28.3 & 0 \\ 1 & 28.7 & 0 \\ 1 & 27.9 & 0 \\ 1 & 27.6 & 1 \\ 1 & 27.3 & 1 \\ 1 & 26.9 & 1 \\ 1 & 25.7 & 1 \\ 1 & 26.3 & 0 \\ 1 & 26.2 & 1 \\ 1 & 25.4 & 1 \\ 1 & 26.2 & 0 \\ 1 & 25.3 & 0 \\ 1 & 25.5 & 1 \\ 1 & 25.6 & 0 \\ 1 & 24.3 & 1 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$$

The parameter estimates are obtained as

```
> X<-cbind(1,onestand$d,onestand$pl-1)
> h<-onestand$h
> b<-solve(t(X)%*%X)%*%t(X)%*%h
> H<-X%*%solve(t(X)%*%X)%*%t(X)
> I<-diag(rep(1,20))
> sigma2<-t(h)%*%(I-H)%*%h/18
> b
          [,1]
[1,] 9.6564516
[2,] 0.3935878
[3,] 0.7539084
> sigma2
          [,1]
[1,] 1.068024
```

or using

```
> onestand$pl<-as.factor(onestand$pl)
> fm2<-lm(h~d+as.factor(pl),data=onestand)
> summary(fm2)
```

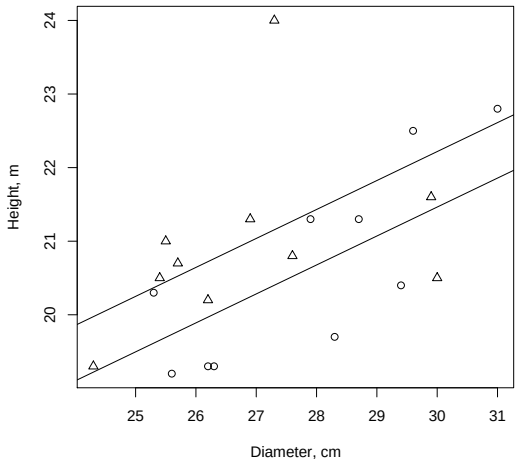


Figure 2.5: Observations of the sample stand (circles for pine and triangles for spruce), and the fitted line of model (2.6) for spruce (upper) and pine (lower).

```
Call:
lm(formula = h ~ d + as.factor(pl), data = onestand)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7180 -0.6700 -0.1904  0.5805  2.8447

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.6565     3.6817   2.623  0.01782 *
d                0.3936     0.1317   2.988  0.00827 **
as.factor(pl)2   0.7539     0.4918   1.533  0.14366
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.063 on 17 degrees of freedom
Multiple R-squared:  0.36,      Adjusted R-squared: 0.2847
F-statistic: 4.782 on 2 and 17 DF,  p-value: 0.02251
```

Note that we used function `as.factor` *to coerce the tree species to a factor. The output shows that tree species 1 was set as the default level, as we did also with our design matrix.*

```
> plot(onestand$d,onestand$h,pch=as.numeric(onestand$pl),xlab="Diameter, cm",ylab="Height, m")
> abline(coef(fm2)[1],coef(fm2)[2])
> abline(sum(coef(fm2)[c(1,3)]),coef(fm2)[2])
```

The above design matrix specified a model where the dependence of height on diameter is similar for both tree species, except for a horizontal shift in the level of the model. Thus, the assumed model for Scots pine is $h_i = b_0 + b_1 d_i + e_i$. For Norway spruce, the assumed model is $h_i = b_0 + b_2 + b_1 d_i + e_i$. Those are illustrated in plot 2.5.

One could also be interested in fitting a models where also the slope varies between the tree species. Such a model is obtained by including an interaction term into the model, as demonstrated in the following example.

Example 2.9 *Assume that we want to have also different slopes for Scots pine and Norway spruce by including an interaction term into our model. Such a model is defined as*

$$h_i = b_0 + b_1 d_i + b_2 SPRUCE_i + b_3 * SPRUCE_i d_i + e_i$$

The design matrix corresponding to this model is defined as

$$\boldsymbol{X} = \begin{pmatrix} 1 & 31 & 0 & 0 \\ 1 & 29.9 & 1 & 29.9 \\ 1 & 30 & 1 & 30 \\ 1 & 29.6 & 0 & 0 \\ 1 & 29.4 & 0 & 0 \\ 1 & 28.3 & 0 & 0 \\ 1 & 28.7 & 0 & 0 \\ 1 & 27.9 & 0 & 0 \\ 1 & 27.6 & 1 & 27.6 \\ 1 & 27.3 & 1 & 27.3 \\ 1 & 26.9 & 1 & 26.9 \\ 1 & 25.7 & 1 & 25.7 \\ 1 & 26.3 & 0 & 0 \\ 1 & 26.2 & 1 & 26.2 \\ 1 & 25.4 & 1 & 25.4 \\ 1 & 26.2 & 0 & 0 \\ 1 & 25.3 & 0 & 0 \\ 1 & 25.5 & 1 & 25.5 \\ 1 & 25.6 & 0 & 0 \\ 1 & 24.3 & 1 & 24.3 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

The assumed model for Scots pine is now $h_i = b_0 + b_1 d_i + e_i$, and that for Norway spruce is $h_i = b_0 + b_2 + (b_1 + b_3) d_i + e_i$. Thus, we actually assume separate models for the tree species. An equivalent model would be obtained by splitting the data according to tree species, and fitting separate models for both sub-dataset.

Another important question about the design matrix is if the constant is included or not. Usually the constant is included, i.e., the first column of the design matrix consists of ones. However, sometimes the phenomenon to be modeled itself is such that constant does not have a practical meaning, or the model without constant would be more realistic. In such cases, the constant can be left out of the model.

Example 2.10 *As diameters have been measured at breast height, it could be realistic to assume that height for diameter $d = 0$ is 1.3. Thus, a realistic assumption would be*

to assume the constant to be 1.3. Such an assumption can be implemented by subtracting the value 1.3 meters from the measured heights, and assuming that the constant b_0 is zero in the resulting model

$$d_i - 1.3 = b_0 + b_1 d_i + e_i.$$

Vectors and matrices $\boldsymbol{h}$, $\boldsymbol{X}$, and $\boldsymbol{b}$ become

$$\boldsymbol{h} = \begin{pmatrix} 21.5 \\ 20.3 \\ 19.2 \\ 21.2 \\ 19.1 \\ 18.4 \\ 20 \\ 20 \\ 19.5 \\ 22.7 \\ 20 \\ 19.4 \\ 18 \\ 18.9 \\ 19.2 \\ 18 \\ 19 \\ 19.7 \\ 17.9 \\ 18 \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 31 \\ 29.9 \\ 30 \\ 29.6 \\ 29.4 \\ 28.3 \\ 28.7 \\ 27.9 \\ 27.6 \\ 27.3 \\ 26.9 \\ 25.7 \\ 26.3 \\ 26.2 \\ 25.4 \\ 26.2 \\ 25.3 \\ 25.5 \\ 25.6 \\ 24.3 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} b_1 \end{pmatrix}$$

2.2.4 Least squares for LM with a general residual variance structure

If the assumptions on the variance and correlation of error terms do not hold, OLS estimation is still unbiased. However, the OLS estimator is no more the minimum-variance-estimator. With such data, Generalized least squares estimation can be used. In GLS, the variance among residuals may vary, end residuals may also be correlated. These assumptions are parameterized into the variance-covariance matrix of residuals, which is then used to obtain GLS estimates for the parameters of the linear model. A WLS estimator is also sometimes used. It is a special case of GLS, where residuals are assumed to be uncorrelated, but the variance is assumed to vary among observations. The different methods are consistent so, that GLS with zero correlations leads to WLS. Furthermore, WLS with constant variance yields OLS. We will go directly to GLS principle, is just a simple special case of the more general GLS approach.

For presentation of the GLS estimation, we assume that the response $\boldsymbol{y}$ depends on

the design matrix $\boldsymbol{X}$ in the similar fashion than before.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

The model differs from model (2.4) in the assumptions about the residual vector. We assume that

$$\begin{aligned} \mathrm{E}(\boldsymbol{e}) &= \boldsymbol{0} \\ \mathrm{var}(\boldsymbol{e}) &= \sigma^2 \boldsymbol{V}\,, \end{aligned}$$

where V is any symmetric positive definite matrix.

We are still assuming that design matrix X and coefficients b are fixed, and these assumptions imply that the following results still hold

$$\begin{aligned} \mathrm{E}(\boldsymbol{y}) &= \boldsymbol{X}\boldsymbol{b} \\ \mathrm{var}(\boldsymbol{y}) = \mathrm{var}(\boldsymbol{e}) &= \sigma^2 \boldsymbol{I} \end{aligned}$$

An interesting special case of GLS results if we know only means of y in p classes of variable x, and classes have unequal number of observations. Assumption of constant variance for raw unobserved data implies that classes have unequal variances, variance of class i being σ^2/n_i. The variance-covariance matrix of the classified data is then $\sigma^2\boldsymbol{W}$, where

$$\boldsymbol{W} = \begin{pmatrix} 1/n_1 & 0 & \dots & 0 \\ 0 & 1/n_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/n_p \end{pmatrix}$$

The OLS estimator generalizes to the case. The Estimator that properly accounts for the relaxed assumptions on the residual is the estimator that minimizes the sum of squares $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$. The Generalized Least Squares (GLS) estimator is

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{2.7}$$

For model (2.2.4), this is the BLUE, i.e, has the smallest is unbiased and has the smallest variance.

It can be shown that the GLS estimator is unbiased and has variance-covariance matrix $\sigma^2\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}^{-1}$:

$$\begin{aligned} \mathrm{E}(\widehat{\boldsymbol{b}}) &= \mathrm{E}\left[(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}\right] \\ &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\mathrm{E}\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{b} \\ &= \boldsymbol{b} \end{aligned}$$

$$\begin{aligned}
\mathrm{cov}(\widehat{\boldsymbol{b}}) &= \mathrm{cov}\left[(\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}y}\right] \\
&= (\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}}\,(\mathrm{cov}\boldsymbol{y})\left[(\boldsymbol{X'V^{-1}X})^{-1}X'V^{-1}\right]' \\
&= \sigma^2\left[(\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}VV^{-1}X}(\boldsymbol{X'V^{-1}X})^{-1}\right] \\
&= \sigma^2\left[(\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}X}(\boldsymbol{X'V^{-1}X})^{-1}\right] \\
&= \sigma^2(\boldsymbol{X'V^{-1}X})^{-1}
\end{aligned}$$

For estimation of residual variance, let us define matrix $\boldsymbol{G}$ as

$$\boldsymbol{G} = \boldsymbol{X}(\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}}$$

An unbiased estimator for the residual variance is

$$\widehat{\sigma^2} = \frac{\boldsymbol{y'}(\boldsymbol{I}-\boldsymbol{G})'\boldsymbol{V^{-1}}(\boldsymbol{I}-\boldsymbol{G})\boldsymbol{y}}{n-p}\,, \tag{2.8}$$

where p is the number of fixed parameters in the model, i.e., the length of vector $\boldsymbol{b}$.

In R, function `lm` can be used in fitting OLS and WLS models, i.e., it can be used for modeling when all off-diagonal elements of matrix $\boldsymbol{V}$ are zeros. For fitting GLS models with any positive definite $\boldsymbol{V}$, one alternative is to use function `gls` in package `nlme`.

Example 2.11 *In example 2.7, a good form was found for the fixed part of the model, but the residuals showed an increasing variance as a function of prediction. In order to make a better assumption on the residual variance, we model the squared residuals of the OLS model on prediction. The following code plots the squared residuals and fits a linear model to them.*

```
> plot(predict(fm3),resid(fm3)^2)
> lm.var<-lm(I(resid(fm3)^2)~predict(fm3)-1)
> lm.var

Call:
lm(formula = I(resid(fm3)^2) ~ predict(fm3) - 1)

Coefficients:
predict(fm3)
      0.4568
> abline(lm.var)
```

The model of squared residuals has a positive coefficient, implying that the residual variance is increasing as a function of predicted value. A model assuming a heterogeneous residual variance according to the variance model is fitted by using the `weight=1/predict(lm.var)` *in function* `lm`.

```
> fm4<-update(fm3,weight=1/predict(lm.var))
> fm4

Call:
lm(formula = h ~ log(d + 10), data = hddata, weights = 1/predict(lm.var))
```

```
Coefficients:
(Intercept)  log(d + 10)
     -32.59        14.42

> summary(fm4)$sigma
[1] 0.9774896
```

The coefficients are slightly different from those obtained using OLS. With weighted fit, the standardized residuals should be homogeneous, indicating that the assumed variance function properly takes into account the heteroscedastic variance. Those residuals are obtained by using `type="pearson"` *in function* `resid`*. The plot of standardized residuals shows that the residuals are not that much heteroscedastic any more.*

```
> plot(predict(fm4),
+      resid(fm4,type="pearson"),
+      xlab="Predicted value",
+      ylab="Standardized residual")
> mywhiskers(predict(fm4),resid(fm4,type="pearson"),se=FALSE,add=TRUE)
> mywhiskers(predict(fm4),resid(fm4,type="pearson"),add=TRUE,lwd=3)
> abline(0,0)
```

Even though estimates seldom need to be computed manually using the matrix equations, it is useful to be able to do it in order to better understand the general applicability of the GLS-estimator 2.7. Furthermore, it may be sometimes useful to be able to compute the estimates. For example, it is not always very clearly reported if the weights needed in a function fitting a wls model should be given as variances or standard deviations, and if they need to be inverted or not. To test which alternative should be used, one option is to fit a small model manually, and to test which definitions leads to the estimates obtained using the function used for model fitting.

Example 2.12 *To show that the estimates are based on the estimator 2.7, we fit the model of example 2.11 manually.*

```
> V<-diag(predict(lm.var))
> X<-cbind(1,log(hddata$d+10))
> y<-hddata$h
> est<-solve(t(X)%*%solve(V)%*%X)%*%t(X)%*%solve(V)%*%y
> est
          [,1]
[1,] -32.58795
[2,]  14.42168
> G<-X%*%solve(t(X)%*%solve(V)%*%X)%*%t(X)%*%solve(V)
> I<-diag(rep(1,length(y)))
> sigma2<-1/(dim(X)[1]-dim(X)[2])*(t(y)%*%t(I-G)%*%solve(V)%*%(I-G)%*%y)
> sqrt(sigma2)
          [,1]
[1,] 0.9774896
```

We see that the estimates are exactly equal to those obtained using `lm`*.*

Example 2.13 *Let us fit the model of example 2.11 using function* `gls`*. In* `gls`*, we have several options for the form of the variance function. Using* `weights=varFixed( x)`*, where* `x` *is a variable of our data, the variance is assumed to depend on the given covariate according to* $var(e) = \sigma^2 x$*. To specify a model where variance depends on the predicted value, we first save the predicted value into our data, and then use it as a covariate in the variance function. The numerical values of fixed parameters are very similar to those of the previous example.*

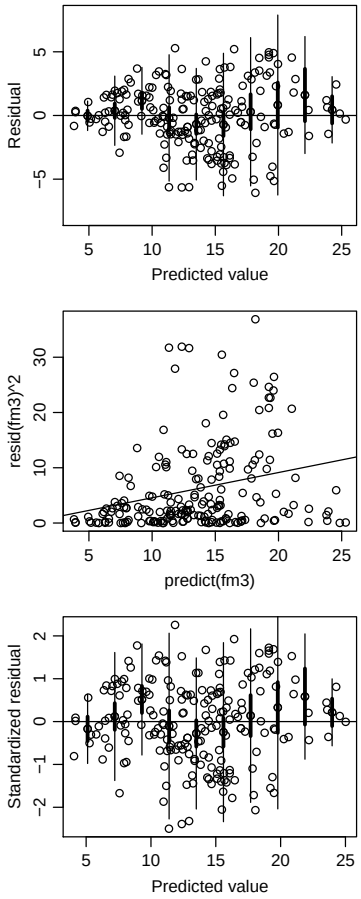


Figure 2.6: Plot for example 2.11

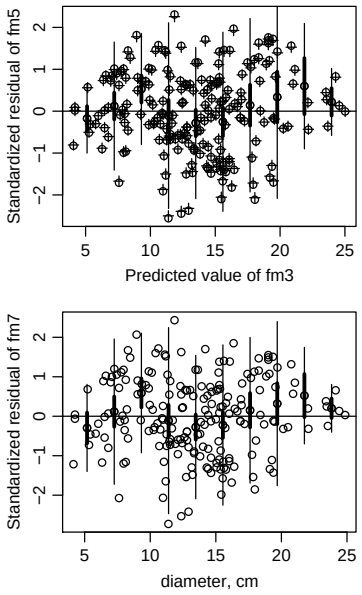


Figure 2.7: Plot for example 2.13. The crosses of teh upper plot show the standardized residuals of model `fm4` and the circles those of `fm5`.

```
> library(nlme)
> hddata$fitted.fm3<-predict(fm3)
> fm5<-gls(h~log(d+10),data=hddata,weights=varFixed(~fitted.fm3))
> fm5
Generalized least squares fit by REML
  Model: h ~ log(d + 10)
  Data: hddata
  Log-restricted-likelihood: -454.3684

Coefficients:
(Intercept) log(d + 10)
  -32.58795    14.42168

Variance function:
 Structure: fixed weights
 Formula: ~fitted.fm3
Degrees of freedom: 200 total; 198 residual
Residual standard error: 0.660629

> windows(width=3,height=2.5)
> par(mai=c(0.6,0.5,0.1,0.1),mgp=c(2,0.7,0),cex=0.8)
> plot(predict(fm5),
+      resid(fm5,type="pearson"),
+      xlab="Predicted value",
+      ylab="Standardized residual")
> points(predict(fm4),
+      resid(fm4,type="pearson"),pch=3)
> mywhiskers(predict(fm5),resid(fm5,type="pearson"),se=FALSE,add=TRUE)
> mywhiskers(predict(fm5),resid(fm5,type="pearson"),add=TRUE,lwd=3)
```

Plot 2.7 shows that, the residuals differ slightly, and estimates for residual standard error are very different. The slight differences result from the use of REML for estimating the residual variance. The difference in standard error is an artifact, resulting from that we assumed the variance to be proportional to predicted value from fm4, whereas in the previous example it was assumed to be proportional to the predicted variance. The following code illustrates a manually computed model that corresponds to the REML-model. The squared ratio of standard errors $(0.6606/0.9775)^2$ *is equal*

to the coefficient of the variance function, 0.4568.

```
> fm6<-update(fm3,weight=1/predict(fm3))
> summary(fm6)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.5880     1.5801  -20.62   <2e-16 ***
log(d + 10)  14.4217     0.5115   28.19   <2e-16 ***
...
Residual standard error: 0.6606 on 198 degrees of freedom
...
```

Another alternative for the variance function could be $var(e_i) = \sigma^2 |d_i|^{2\delta}$. *A model with such a variance function can be fitted using*

```
> fm7<-gls(h~log(d+10),data=hddata,weights=varPower(0.5,~d))
> fm7
Generalized least squares fit by REML
  Model: h ~ log(d + 10)
  Data: hddata
  Log-restricted-likelihood: -452.6542

Coefficients:
(Intercept) log(d + 10)
  -32.08281    14.25532

Variance function:
 Structure: Power of variance covariate
 Formula: ~d
 Parameter estimates:
    power
0.5590031
Degrees of freedom: 200 total; 198 residual
Residual standard error: 0.5427085
```

The lower plot of figure2.7 shows the standardized residuals from this model. It is impossible to say which of the two alternative variance functions was better. Slight differences are seen in the estimates of the fixed parameters.

It is easy to see, how all the results on OLS can be easily derived by replacing $\boldsymbol{V}$ with $\boldsymbol{I}$. Furthermore, we still rely only on assumptions on expectation, variance and covariance of residuals, and no specific assumption on the distribution was made.

2.3 Tests of the regression relationship

Until now, we only have assumed that $\boldsymbol{\epsilon}$ is independent and has constant variance. In order to be able to test the significance of predictors, we additionally assume that ϵ_i are independent observations from normal distribution

$$\epsilon_i \sim N(0, \sigma^2)$$

. This implies that

$$\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

and

$$\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2 \boldsymbol{I})$$

It results from the normality of the residuals (or of the response) that

$$\begin{aligned} \widehat{\boldsymbol{b}} &\sim N_p(\boldsymbol{b}, \sigma^2(\boldsymbol{X'X})^{-1}) \\ \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}})}{\sigma^2} &\sim \chi^2(n-p) \end{aligned}$$

Furthermore, the fit $\boldsymbol{X}\hat{\boldsymbol{b}}$ and residual $\boldsymbol{\epsilon}$ are uncorrelated. Because their joint distribution is multinormal, they are also independent. Thus, also $\widehat{\boldsymbol{b}}$ and $\widehat{\sigma^2}$ are independent.

2.3.1 Sums of squares

The overall variation of response around its mean is defined as

$$SS_{tot} = \sum (y_i - \bar{y})^2 .$$

It can be can be divided into two components: the unexplained variation (residual sum of squares)

$$RSS = \sum (y_i - \hat{y_i})$$

and explained variation

$$SS_{reg} = SS_{tot} - RSS .$$

A widely used figure for evaluation of the goodness of fit of the regression relationship is the degree of determination

$$R^2 = 1 - \frac{RSS}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

which tells which portion of the total variation in y was explained by the estimated regression relationship. However, even though R^2 is a useful figure for comparison, the model evaluations should never be based on it alone. It does not tell if the assumptions of the linear model were met or not. In addition, one should note that R^2 is different in different scale. With the single-predictor regression, the coefficient of determination is the squared correlation coefficient between x and y.

The adjusted R^2 takes into account the degrees of freedom that are used for model fitting. it is defined as

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} .$$

Example 2.14 *The following code manually computes R^2 for the model of example 2.1.*

```
> SStot<-sum((onestand$h-mean(onestand$h))^2)
> RSS<-sum(resid(fm1)^2)
> SSreg<-SStot-RSS
> R2<-SSreg/SStot
> R2
[1] 0.2715626
> 1-(1-R2)*(20-1)/(20-1-1)
[1] 0.2310938
```

The same value is printed by summary:

```
> summary(fm1)
...
Multiple R-squared: 0.2716,     Adjusted R-squared: 0.2311
...
```

Example 2.15 *Two models, where one has $log(V)$ as the response and the other has V. The value of R^2 is completely different.*

2.3.2 F-test for the significance of regression

A very basic question about the regression relationship is, whether the explanatory variables significantly explain the variation of the response. This question can be answered by testing if the explained variation is big when compared to the unexplained variation. The Null and alternative hypotheses are

$$\begin{aligned} H_0 &: b_0 = b_1 = \ldots = b_p = 0 \\ H_1 &: someb_i \neq 0 \end{aligned}$$

The test statistic is

$$F_{obs} = \frac{SS_{reg}/p}{RSS/(n-p-1)}$$

It results from the normality of residuals that the the two sums of squares in the test statistic are distributed according to χ^2 distribution with p and $n-p-1$ degrees of freedom for numerator and denominator, respectively. Thus, we remember from section 1.4.3 that the test statistic is distributed according to $F(p, n-p-1)$ degrees of freedom. The p-value is

$$p = P(F > F_{obs}),$$

i.e., the value of cdf of F-distribution with p and $n-p-1$ degrees of freedom at F_{obs}.

Example 2.16 *To do tests, the first step is to check if the assumptions on the model were met. To do this, we plot the Normal quantile-quantile plots, which plot the realized residual quantiles against the theoretical quantiles based on Normal distribution. If normality is met, these observations should be close to a line having constant of 0 and coefficient of 1.*

```
qqnorm(resid(fm1),main="Q-Q-plot for fm1")
abline(0,1)
qqnorm(resid(fm2),main="Q-Q-plot for fm2")
abline(0,1)
```

The plots are shown in figure 2.8. The largest observed residual is overly large to be from a normal distribution, especially for model fm1. However, we regard this plot good enough, and trust on normality in our tests.

Example 2.17 *For model 2.1, the F test statistic is computed as follows*

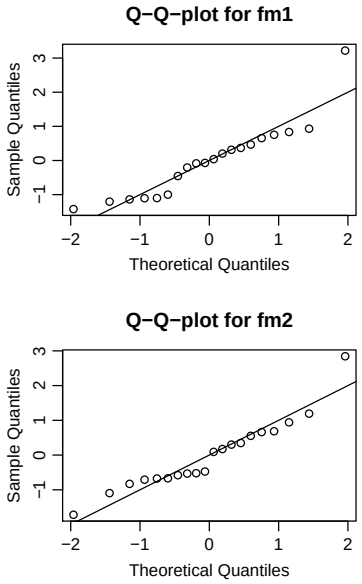


Figure 2.8: Normal Q-Q plots for model fm1 and fm2.

```
> Fobs<-SSreg/1/(RSS/(20-1-1))
> Fobs
[1] 6.710426
> 1-pf(Fobs,1,18)
[1] 0.01846628
```

Compare also to

```
> summary(fm1)
...
F-statistic:  6.71 on 1 and 18 DF,  p-value: 0.01847
```

The results show that there is a statistically significant (p=0.018) dependence between tree height and diameter in the example stand.

2.3.3 t-test for coefficients

Because σ^2 is replaced with the estimated error variance, $\widehat{\sigma^2}$, the confidence intervals and tests for parameter vector $\boldsymbol{b}$ can be derived using result

$$\frac{\boldsymbol{p'}\widehat{\boldsymbol{b}} - \boldsymbol{p'b}}{\hat{\sigma}\sqrt{\boldsymbol{p'(X'X)^{-1}p}}} \sim N(0,1)$$

where $\boldsymbol{p}$ is any vector of length same length as $\boldsymbol{b}$. The distribution of only one element of $\boldsymbol{b}$say b_i is obtained by having all elements of p as zeros, except for the ith element, which is given the value of 1. Based on the above equation, the $100(1-\alpha)\%$ confidence interval for $\boldsymbol{p'}\widehat{\boldsymbol{b}}$ is

$$\boldsymbol{p'}\widehat{\boldsymbol{b}} \pm t_{\alpha/2;n-p}\hat{\sigma}\sqrt{\boldsymbol{p'(X'X)^{-1}p}}$$

The same results can also be used for construction of tests. The null and alternative hypotheses are

$$H_0 : \boldsymbol{p'b} = 0$$
$$H_1 : \boldsymbol{p'b} \neq 0$$

The test statistic is

$$t_{obs} = \frac{\boldsymbol{p'\hat{b}}}{\hat{\sigma}\sqrt{\boldsymbol{p'(X'X)^{-1}p}}}$$

Under the null hypothesis, $t \sim t(n-p)$. The null hypothesis is rejected if $|t| > t_{\alpha/2;n-p}$. A more informative approach is to compute and report the p-value. The p-value is computed as $p = P(|t| > |t_{obs}|) = 2P(t > t_{obs})$, i.e, twice the the value of the cumulative distribution function of t-distribution with $n-p$ degrees of freedom. The t-tests and p-values reported bu statistical model-fitting procedures are applications of this test.

Example 2.18 *A researcher may be interested whether the level of the H-D curve could be assumed to be the same for both tree species. A test on such a null hypothesis is seen directly from the t-test of the tree species dummy variable. In this case, null hypothesis on different constant for different tree species would be rejected (p-value = 0.14).*

```
> summary(fm2)
...
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.6565     3.6817   2.623  0.01782 *
d                 0.3936     0.1317   2.988  0.00827 **
as.factor(pl)2    0.7539     0.4918   1.533  0.14366
...
```

Example 2.19 *We want to test whether the constant of the model of example 2.1 is 1.3 meters. For this purpose, we define*

$$\boldsymbol{p} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} 1.3 \\ 0 \end{bmatrix}$$

The test statistic is computed as follows.

```
> p<-c(1,0)                   # t-value for constant
> p
[1] 1 0
> b<-c(1.3,0)
> bhat<-coef(fm1)               # estimate of b
> bhat
(Intercept)           d
 11.4394735   0.3421870
> sigma<-summary(fm1)$sigma  # residual s.e.
> sigma
[1] 1.102579
> X<-cbind(1,onestand$d)
> se<-(sigma*sqrt(p%*%solve(t(X)%*%X)%*%p))
> se
         [,1]
[1,] 3.621877
> t<-(p%*%bhat-p%*%b)/se
> t
         [,1]
[1,] 2.799508
> 2*(1-pt(t,20-2))
           [,1]
[1,] 0.01184922
```

The p-value is 0.01, which shows that the null hypothesis about constant of 1.3 meters is rejected. Our model predicts height of 11 meters for trees of diameter 0. Maybe a better model specification is needed.

The same result is obtained by fitting a restricted model using `offset` *in our model. As the restriction is made for the constant, we first need to add a variable including ones into our dataset. Then we fit a model without constant using '-1 in* `h=d-1`*, but specify a predefined constant of* 1.3 *by using argument* `offset=I(1.3*ones)`.

```
> onestand$ones<-1
> fm2<-lm(h~d-1,data=onestand,offset=I(1.3*ones))
> anova(fm1,fm2)
Analysis of Variance Table

Model 1: h ~ d
Model 2: h ~ d - 1
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     18 21.8823
2     19 31.4099 -1   -9.5276 7.8372 0.01185 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The same result can be seen from the following table by comparing the standard error of the constant to the difference between assumed and estimated value of the constant.

```
> summary(fm1)

Call:
lm(formula = h ~ d, data = onestand)
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.4395     3.6219   3.158  0.00544 **
d             0.3422     0.1321   2.590  0.01847 *
...
```

2.3.4 Testing several coefficients at the same time

The t-tests produced by statistical packages usually perform two kinds of standard tests: the F-test for significance of the regression relationship, and a t-test for testing if individual coefficients significantly differ from 0. The t-test approach of the previous section can also be used for testing several individual coefficients at once. Another, perhaps more easily applicable approach is to use F-test.

Assume that we have regression relationship $\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e}$, and we want to test a reduced model with different constraints for the model. Those constraints may be whatever, including dropping some predictors, giving a constant value for a coefficient, or setting some coefficients to be equal. Whatever constraints are made they result to **a constrained model**, which is a special case of the more general model. A test for testing if removing the constraints significantly improved the fit is formulated as follows. Define RSS_{NH} as the sum of squares from the reduced (constrained) model, having q predictors, and RSS_{AH} as the sum of squares from the full (unconstrained) model having p predictors. These sums of squares are independent, χ^2-distributed random variables. Furthermore, also the difference between two χ^2-distributed random

variables is distributed according to χ_2 distribution. Thus, the test statistic

$$F_{obs} = \frac{(RSS_{NH} - RSS_{AH})/(p-q)}{RSS_{AH}/(n-p-1)}$$

is distributed according to F-distribution with $p-q$ and $n-p-1$ degrees of freedom if the null hypothesis is true. The p-value is obtained from F distribution in a similar way as earlier. Exactly the same test would be obtained also using the likelihood ratio principle.

In R, these comparisons can be carried out easily by using the `anova` function, as demonstrated in the following examples. The constrained model can be fitted by dropping predictors. A constant coefficient is assumed for a specific pedictor can be given by using the `offset` argument in `lm(formula,data,offset)`.

Example 2.20 *Assume we want to test if the model with separate constants and slopes for the two species was significantly different from the model with common slope and constant. The full model is the one with species-specific coefficients and the constrained model is the one with common coefficients. The models are fitted and test statistics computed as follows. Note that all main effects and interactions are included by having model equation `h d*pl`.*

```
> full<-lm(h~d*pl,data=onestand)
> constr<-lm(h~d,data=onestand)
> RSSnh<-sum(resid(constr)^2)
> RSSah<-sum(resid(full)^2)
> Fobs<-(RSSnh-RSSah)/(4-2)/(RSSah/(20-3-1))
> Fobs
[1] 1.988027
> 1-pf(Fobs,2,16)
[1] 0.1693878
```

The full model is not significantly better than the restricted model. Using function `anova` results in an identical result:

```
> anova(constr,full)
Analysis of Variance Table

Model 1: h ~ d
Model 2: h ~ d * pl
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     18 21.8823
2     16 17.5268  2    4.3555 1.988 0.1694
```

The t-test and F test lead to similar inference if the test is such that it can be performed with both methods. This is demonstrated by the following example.

Example 2.21 *In example 2.18, model fm2 included separate coefficients for both tree species. A constrained version is the model fm1, which includes no species-specific parameters. It was shown that the p-value for the species dummy was 0.14366, which indicates that the level does not vary between tree species. The same hypothesis can be tested by making F-test between models fm2 and fm1. We get*

```
> anova(fm1,fm2)
Analysis of Variance Table

Model 1: h ~ d
Model 2: h ~ d + as.factor(pl)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     18 21.8823
2     17 19.2244  1    2.6578 2.3503 0.1437
```

The p-value and resulting inference are exactly the same as we got from the t-test.

2.4 Checking the validity of assumptions

The basic assumptions of a liner model are that the model is correct, observations are independent and have equal variance. In addition, for testing hypotheses on the model estimated by least squares, an assumption about normality of residuals is needed. For ML-estimation, the normality needs to be assumed already at the model fitting stage. In this section, strategies for situations where those assumptions are not met are discussed. R has method `plot()` for producing some standard plots (such as residual and qq-plots) from a fitted model.

2.4.1 Model shape

The assumption that the model is correct means that the assumed function for the modeled relationship is able to describe the true relationship between the variables. In practice, we seldom have any well-justified theory for the assumed relationship, and the model form is many times a result of a trial-and-error procedure. The simplest way to account for the observed non-linearities in the relationships is to make transformations **to the predictors**. The relationship could also be linearized by making transformations to the response, but this should be avoided if possible.

With single-predictor regression, an initial guess for the model shape can be found by plotting the response against predictor. If the relationship is concave, then a concave transformation to the predictor would linearize the relationship. With convex relationship, a convex transformation can be used. Commonly used convex transformations are e^x and x^2. Commonly used concave transformations are $ln(x)$, $\sqrt{x}$ and $(-)1/x$. With more complex relationships, the same predictor can be included twice, with different transformations. If the relationship is good, a plotting model residuals against the predictors should not show any trend. To recognize the trends, one can use lowess smoothers. One alternative is to use the approach that was presented in example 2.7, i.e., plotting the residuals against the predictor, and computing class means of the residuals, and plotting their confidence intervals. If roughly 95% of the confidence intervals overlap x-axis and no clear trend can be found in class means, I would believe that the model form is good. Function `mywhiskers()` of example 2.7 with option `se=TRUE` can be used to perform such analysis.

The above rules can also be used with multiple regression. In multiple regression, the response should be first plotted against all potential predictors to get a view on the shape of the relationship. The fitted model should be further analyzed by plotting the residuals against all the predictors, and checking if there are nonlinear trends in the residuals. Lowess smoothers or function `mywhiskers` can be used to detect also

those trends.

The nonlinearity can also be tested through a formal test. In such a test, two alternative regressions are fitted: The null model with the predictor x included as a continuous predictor and the alternative model in which all k distinct values of x are included as an indicator (dummy variable). Testing the performance of alternative model against the null model using a F-test gives a test for nonlinearity.

$$\begin{aligned} H_0 &: \text{Modelislinear} \\ H_1 &: \text{Modelisnonlinear} \end{aligned}$$

$$F_{obs} = \frac{(RSS_0 - RSS_1)/(k-2)}{RSS_1/(n-k)}$$

where RSS_0 is the residual sum of squares from the null model and RSS_1 that of the alternative model. The p-value is obtained from F-distribution with $k-2$ and $n-k$ degrees of freedom.

It is quite common that none of the commonly used transformations linearizes the relationship. In such case, one could try Box-Cox transformations (Box and Cox 1962), which is a generalized transformation giving the most commonly used transformations as special cases. Another approach is spline regression. Harrell (2001, p. 20-34) presents an easy-to-use restricted cubic spine approach, which is able to produce various shapes of relationship. A third approach is the nonlinear regression, where no restrictions are made for the function of the relationship.

Example 2.22 *Harrells spline regression*

2.4.2 Independence of observations

The independence of observations should be thought about when collecting the data. Maybe the most important sources causing dependence among observations are spatial and temporal autocorrelation.

In forestry, the spatial autocorrelation often arises from that trees that are close to each other tend to be more similar to each other than distant trees. If the distances between trees are known, the spatial autocorrelation can be modeled through a correlogram, variogram and covariogram, which are related to each others in a similar way than variance, covariance and correlation are related to each other. The covariogram is constructed by fitting an OLS regression to the data, and then plotting a the cross-products of all possible pairs of residuals against the geographical distance between these two observations. The covariogram is then obtained by fitting an appropriate (decreasing) function to these data. If such a trend is found, then spatial autocorrelation exists among the observations. The covariagram can be used to estimate the covariance

between any two observations with known distance, which are then used in GLS approach to account for the autocorrelation. The estimated covariances (up to a scaling constant) can be written to matrix $\boldsymbol{V}$, and the model can be then fitted using generalized least squares. This approach, which is clearly one special case of the linear model, is known as **kriging** in spatial statistics (e.g., Cressie 1993).

The temporal autocorrelation can be analyzed and taken into account in naanalogous way than the spatial autocorrelation. We just have the geographical distance replaced with the time difference between the observations. Such approach leads to analysis of lagged residuals, which are used to fit an appropriate model for the temporal autocorrelation. After the autocorrelation model has been estimated, GLS approach is used to fit a linear model that properly accounts for the temporal autocorrelation.

Another cause of interdependencies among observations is hierarchy of the data. For example, branches may be sampled from different trees for determination of biomass, the branches of one tree being somehow similar to each other. As an another example, trees may be measured from different stands, trees of one stand being similar and different from trees of other stands. Furthermore, repeated measurements may be taken from several plots, observations of one and the same plot at different points in time being similar and different to observations of other plots. All these examples lead to a situation where the data consists of several groups, and observations within the same group are correlated. This kind of structure can be taken into account through variance component modeling, which is dealt in more detail in the next chapter. In the variance component approach, a constant correlation is assumed among observations from different groups.

2.4.3 Constant residual variance

In ordinary least squares, the variance of residuals is assumed to be constant. The validity of this assumption can be taken into account by plotting the residuals of a fitted OLS model against the fitted values. The obtained plot should not show any trend in the variance. However, it is not always easy to see if there is trend in the variance or not, especially if the observations are not distributed evenly along the x-axis, having the fitted values. To help with this problem, one approach is to compute standard deviations of the residuals in different classes of fitted value, and compare them to each other. If the standard deviations are equal and do not show any trend with respect to the fitted values, the assumption about constant variance seems to be met. An example of such analysis was domne in example 2.7 using the function `mywhiskers` with option `se=FALSE`.

A formal test can als be conducted for testing the heteroscedasticity (see Weis-

berg2005).

The violation of the constant variance can be corrected either by making a transformation to the response or through the use of a variance function. A very usually applied transformation is the logarithmic transformation, which is justified by the log-normal distribution for variables getting only positive values. A straightforward way for using a variance function is is to fit OLS model, then model the variance using the squared residuals, and refit the model with GLS using the estimated variance function. An example of such procedure was shown in example 2.11. To check if the utilized variance function properly accounts for the heterogeneous variance, a plot of standardized residuals should express a constant variance against the response.

A problem with the transformation approach is that it also has an effect on the relationship between predictors and response. Thus, a transformation should be found that both homogenizes the variance and results in linear relationship between response and predictor. Furthermore, making transformation causes that the model gives unbiased predictions for the transformed variables (e.g., for the logarithmic height), implying that the prediction for the back transformed variable (e.g., for the total height) is biased. For these reasons, I would recommend using variance functions rather than making transformations into the response to homogenize the residual variance. If the transformation approach is used and normality of residuals of the fitted model are assumed, the bias can be corrected by adding half of the residual error to the prediction before backward transformation. This bias correction is based on the expected value of lognormal random variable $\mathrm{E}X = e^{\mu+\sigma^2/2}$.

2.4.4 Normality

As noted several times before, the normality of residuals is needed only for testing purposes. Thus, if the previously stated assumptions are met, normality is not needed e.g. for the OLS or GLS estimates $\hat{\boldsymbol{b}}$ to be best linear unbiased estimators for $\boldsymbol{b}$. However, all t- and F-test on the regression relationship require normality. Thus, the normality becomes important for example, in studies where the linear model is used for testing the impacts of predictors to the response.

The very general statistical results (central limit theorem and law of large numbers) can be used as a justification to assume that distribution of any random variable is normal, unless it is clearly violated. In regression analysis, the normality can be checked using quantile-to-quantile plots (qq-plots), which plot the ordered residuals against corresponding quantiles of standard normal distribution. If the qq-plot are close to a straight line, the residuals follow the normal distribution.

2.5 Prediction

With the linear model, prediction for a new individual with predictors $\boldsymbol{X}^*$ is obtained as $\tilde{y} = \hat{b}\boldsymbol{X}^*$.

The prediction includes errors caused by

- model uncertainty,
- estimation errors of the parameters, $\mathrm{var}(\hat{\boldsymbol{b}})$, and
- residual variance.

The variance-covariance matrix of prediction, that accounts for the residual and parameter uncertainty, is given as

$$\mathrm{var}(\tilde{\boldsymbol{y}}) = \sigma^2 + \sigma^2 \boldsymbol{X}_0 \boldsymbol{X'} \boldsymbol{X}^{-1} \boldsymbol{X}_0' = \sigma^2 + \boldsymbol{X}_0 \,\mathrm{var}(\hat{\boldsymbol{b}}) \boldsymbol{X}_0$$

The confidence interval for the prediction can be formulated based on the prediction variance and t-distribution.

Example 2.23 *We want to make predictions from model* `fm1` *for diameters* $25, \ldots, 30$. *The code below defines the matrix* $\boldsymbol{X}_0$ *for these predictions.*

```
> Xstar<-cbind(1,25:30)
> attributes(fm1)  # look for attributes that include the desired matrices and vectors
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

$class
[1] "lm"

> attributes(summary(fm1))  # cov unscaled includes the matrix solve(X%*%X)
$names
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"

$class
[1] "summary.lm"

> sigma<-summary(fm1)$sigma
> varb<-sigma^2*summary(fm1)$cov.unscaled
>
> varh<-diag(rep(sigma^2,dim(Xstar)[1]))+Xstar%*%varb%*%t(Xstar)
> varh
             [,1]         [,2]       [,3]       [,4]         [,5]         [,6]
[1,]  1.373239251  0.116465081 0.07537208 0.03427907 -0.006813936 -0.047906941
[2,]  0.116465081  1.308502499 0.06917759 0.04553384  0.021890096 -0.001753651
[3,]  0.075372075  0.069177588 1.27866427 0.05678861  0.050594127  0.044399640
[4,]  0.034279070  0.045533842 0.05678861 1.28372455  0.079298158  0.090552931
[5,] -0.006813936  0.021890096 0.05059413 0.07929816  1.323683354  0.136706221
[6,] -0.047906941 -0.001753651 0.04439964 0.09055293  0.136706221  1.398540676
> cov2cor(varh)
             [,1]         [,2]       [,3]       [,4]         [,5]         [,6]
[1,]  1.000000000  0.086883108 0.05687994 0.02581785 -0.005053973 -0.034569077
[2,]  0.086883108  1.000000000 0.05348104 0.03513266  0.016632913 -0.001296338
[3,]  0.056879937  0.053481044 1.00000000 0.04432483  0.038889270  0.033201953
[4,]  0.025817846  0.035132665 0.04432483 1.00000000  0.060832424  0.067581688
[5,] -0.005053973  0.016632913 0.03888927 0.06083242  1.000000000  0.100475153
[6,] -0.034569077 -0.001296338 0.03320195 0.06758169  0.100475153  1.000000000
```

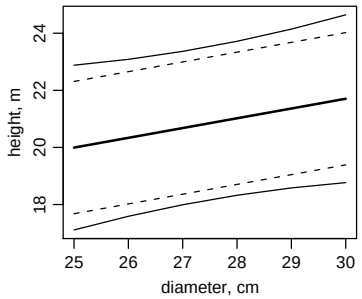


Figure 2.9: Predictions (thick solid line) and approximate 95% prediction intervals for model fm1 (thin solid line). The dashed line show intervals that do not take into account the estimation errors of the coefficients.

The variance-covariance matrix is not a diagonal, but the predictions are slightly correlated. These correlations arise from the estimation errors of the parameters: all predictions are based on the same estimates of $\mathbf{b}$*, and thus are correlated due to this. However, the correlations are not very strong, as indicated by the correlation matrix shown in the lower matrix. Thus, we ignore these correlations and plot the confidence intervals as if the predictions were uncorrelated. The confidence intervals are shown in Figure 2.9. The narrow intervals plotted using dashed line do not take into account the estimation errors of parameters. The wider intervals shown by the solid lines are the prediction intervals that also take into account the estimation errors.*

```
> pred<-Xstar%*%coef(fm1)
> tquantile<-qt(0.975,20-2)
> ub1<-pred+sigma*tquantile
> lb1<-pred-sigma*tquantile
> ub2<-pred+diag(varh)*tquantile
> lb2<-pred-diag(varh)*tquantile
>
> windows(width=3,height=3)
> par(mfcol=c(1,1),mai=c(0.6,0.5,0.5,0.1),mgp=c(2,0.7,0),cex=0.8)
> plot(Xstar[,2],pred,type="l",ylim=range(c(ub2,lb2)),xlab="diameter, cm",ylab="height, m",lwd=2)
> lines(Xstar[,2],ub1,lty="dashed")
> lines(Xstar[,2],lb1,lty="dashed")
> lines(Xstar[,2],ub2)
> lines(Xstar[,2],lb2)
```

2.6 Estimation with maximum likelihood

The previous section presented how the parameters of the linear model can be estimated using methods based on least squares. In this section, we make an assumption about the distribution of the random variable y, and estimate the parameters using the method of maximum likelihood. Section 1.5.1 presented the general idea of maximum likelihood. In the ML estimation of the linear model, the response $\boldsymbol{y}$ is assumed to be normally distributed. Furthermore, parameter μ is not assumed to be constant, but it is is written as a linear function of predictors as $\mu = \mu(\boldsymbol{Xb})$. The likelihood (or log-likelihood)

will then be a function of $\boldsymbol{b}$ and σ^2, and the ML estimate of them is the combination that maximizes the (log-)likelihood.

Letting also $\exp(\sigma^2)$ to be a function of predictors (say, $\exp(\sigma^2) = f(\boldsymbol{X}, \boldsymbol{c})$), we can fit a model with heteroscedastic residuals. The exponential transformation is used to constraint the σ^2 to be positive. This is based om the invariance property of ML-estimator, and was used also in example 1.30. Both the parameters $\boldsymbol{b}$ and $\boldsymbol{c}$ can be estimated by maximizing the (log-)likelihood on $\boldsymbol{b}$ and $\boldsymbol{c}$. Furthermore, we may define any positive definite structure for $\boldsymbol{V}$, parameterize it and estimate the parameters by maximizing the likelihood on all the specified parameters.

We could also assume some other distribution for $\boldsymbol{y}$ than the normal. For example, y_i may be a binary random variables having the *Bernoulli(p)* distribution. Such a variable would get only values 0 and 1, the value 0 indicating absence and value 1 presence of the characteristic being modeled. Furthermore, if $\boldsymbol{y}$ are counts, a realistic assumption could be the *Poisson(λ)* distribution. In those cases, the parameters of the assumed distribution are just written as a linear combination of predictors, (e.g., $\text{logit}^{-1}(p) = f(\boldsymbol{Xb})$ in the case of bernoulli distribution and $\exp(\lambda) = \lambda\boldsymbol{Xb}$ in the case of Poisson distribution. The functions logit^{-1} and exp are link functions which are used to ensure that the parameter is within the parameter space, i.e., $0 \leq p \leq 1$ and $\lambda \geq 0$. The parameters are then estimated by maximizing the log-likelihood on $\boldsymbol{b}$. Non-normal models where a function of parameters is assumed to be linear in $\boldsymbol{b}$ are called **generalized linear models**, and are dealt more in detail in chapter 4.

2.6.1 ML for the single predictor regression

Let us state model (2.1) in a bit different form:

$$\begin{aligned} \mathrm{E}(y_i) &= \mu(x_i) = b_0 + b_1 x_i \\ y_i &\sim \text{indep.} N[\mu(x_i), \sigma^2] \quad i = 1, 2, \ldots, n \end{aligned}$$

where $\mu(x_i)$ indicates that the expectation of y is a function of x. The four important assumptions incorporated in the models are (i) The y_i follow a normal distribution, (ii) are independent, (iii) have equal variance, and (iv) the mean of y is linear in x. Compared to the least squares assumptions, the only difference is the assumption on normality of y.

The likelihood based on normal distribution is

$$\begin{aligned} L &= \prod_i f_N(\mu(x_i), \sigma) \\ &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - \mu(x_i)\right)^2\right] \\ &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - b_0 - b_1 x_i\right)^2\right] \end{aligned}$$

Taking logarithms yields the log likelihood as

$$\begin{aligned} l &= \ln L \\ &= \sum_i \left[-\ln\left(2\pi\sigma^2\right)^{1/2} - \frac{1}{2\sigma^2}\left(y_i - b_0 - b_1 x_i\right)^2 - \right] \\ &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_i \left(y_i - b_0 - b_1 x_i\right)^2 . \end{aligned}$$

Differentiating the log likelihood with respect to the three parameters (b_0, b_1m σ^2) gives the following equations

$$\begin{aligned} \frac{\partial l}{\partial b_0} &= \frac{1}{\sigma^2}\sum_i \left(y_i - b_0 - b_1 x_i\right) \\ \frac{\partial l}{\partial b_1} &= \frac{1}{\sigma^2}\sum_i \left(x_i y_i - b_0 x_i - b_1 x_i^2\right) \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}\sum_i \left(y_i - b_0 - b_1 x_i\right)^2 \end{aligned}$$

Setting these equations to 0 and solving for the three parameters gives the ML estimators. Start with the lowest equation by solving it for σ^2. Setting it equal to 0 gives

$$\widehat{\sigma^2} = \frac{1}{n}\sum_i \left(y_i - \widehat{b_0} - \widehat{b_i} x_i\right),$$

which shows the ML estimator for the residual variance, given the estimates for b_0 and b_1. It also shows that $\widehat{\sigma^2}$ is always positive and greater than zero, unless predictions and observations are exactly the same. Thus, the equations setting the two upper equations to zero can be multiplied with σ^2 which yields exactly same system of equations that was solved to get the OLS estimates. Thus, the ML-estimators

$$\begin{aligned} \widehat{b_0} &= \bar{y} - b_1\bar{x} \\ \widehat{b_1} &= \frac{\frac{1}{n}\sum_i x_i \sum_i y_i - \sum_i xy}{\frac{1}{n}\left(\sum_i x_i\right)^2 - \sum_i x^2} \end{aligned}$$

are exactly same as the OLS-estimators. It is important to note that the estimator for the residual variance, $\widehat{sigma^2} = \frac{RSS}{N}$ differs from (2.2) in that nominator $n - 2$ is replaced by n. Thus, the ML estimate of residual variance is biased downwards.

2.6.2 ML for the linear model with uncorrelated errors and constant variances

As with LS methods, the results on simple linear regression generalize to the multivariate case. For estimation of the linear model with the method of maximum likelihood, the model with uncorrelated errors and constant variance is defined as

$$\boldsymbol{y} \sim N[\boldsymbol{Xb}, \sigma^2 \boldsymbol{I}]$$

The likelihood is based on the multivariate normal distribution:

$$L = L(\boldsymbol{b}, \sigma^2 | \boldsymbol{y}) = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{Xb})'\boldsymbol{I}/\sigma^2(\boldsymbol{y}-\boldsymbol{Xb})\right]}{(a\pi\sigma^2)^{1/2n}}$$

The log likelihood is

$$l = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{Xb})'(\boldsymbol{y}-\boldsymbol{Xb})/\boldsymbol{\sigma^2}\,.$$

The partial derivatives with respect to $\boldsymbol{b}$ and σ^2 are

$$\frac{\partial l}{\partial \boldsymbol{b}} = \frac{\boldsymbol{X'y} - \boldsymbol{X'Xb}}{\sigma^2} \frac{\partial l}{\partial \sigma^2} = \frac{(\boldsymbol{y}-\boldsymbol{Xb})'(\boldsymbol{y}-\boldsymbol{Xb})}{2\sigma^2} - \frac{n}{a\sigma^2}$$

Setting the upper equation to zero gives the ML estimator for $\boldsymbol{b}$ as $\widehat{\boldsymbol{b}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$, which is identical to the OLS estimator. From equating the lower derivative to zero and solving for σ^2 yields the ML estimator for residual variance, which is $\widehat{\sigma^2} = \frac{1}{n}(\boldsymbol{y}-\boldsymbol{Xb})'(\boldsymbol{y}-\boldsymbol{Xb})$. This is identical to the estimator (2.5), except for the important replacement of $n-p$ by n in the denominator.

2.6.3 ML for LM with a general residual variance structure

The more general model, corresponding to the GLS model, is

$$\boldsymbol{y} \sim N[\boldsymbol{Xb}, V]$$

The log likelihood is

$$l = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{Xb})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{Xb})\,,$$

The derivation for the ML estimators can be found from McCulloch and Searle (2001, p. 178-181). As one could guess, the results are identical to the results on GLS, except for that the denominator of (2.8) is replaced with n.

As a summary, the method of maximum likelihood is equivalent to OLS and GLS approaches in estimating the parameters of a linear model, except that it results in downward biased estimates for the residual variance. The bias is higher with smaller datasets, and vanishes with very large datasets.

2.6.4 Restricted maximum likelihood

It has be seen that the LS and ML principles lead to identical results for estimation of $\boldsymbol{b}$, except for that LS principle is based on more relaxed assumptions about the distribution of residuals (or distribution of $\boldsymbol{y}$). However, with the variance of the residual error, these two methods differ in the denominator of the estimator for residual variance. It was shown that the estimator obtained from OLS was unbiased, whereas the ML estimator was an underestimate. The Restricted or Residual ML (REML) approach has not this problem, and that is why it has become one of the most widely applied estimation methods for linear mixed models and variance components models.

In restricted maximum likelihood, a specific linear combination of the original data is modeled using maximum likelihood. The assumed model is specified as

$$\boldsymbol{K'y} \sim N[0, \boldsymbol{K'V'K}]\,,$$

where $\boldsymbol{K}$ is a $n \times p$ matrix having full column rank. Furthermore, it has the specific property that $\boldsymbol{K'X} = \boldsymbol{0}$. The matrix $\boldsymbol{K}$ is specified as $\boldsymbol{K'} = \boldsymbol{C'}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-}\boldsymbol{X'})$ with some matrix $\boldsymbol{C}$. The ML estimators can thus be derived from ML-estimators by making replacements

$$\boldsymbol{y} = \boldsymbol{K'yX} = \boldsymbol{K'XV} = \boldsymbol{K'VK}$$

Essentially, we see that the property $\boldsymbol{K'X} = \boldsymbol{0}$ causes the fixed effect part of REML model to be zero, implying that it does not involve $\boldsymbol{b}$ at all. Thus REML can be used only for estimating the parameters specifying matrix V. With balanced data, these estimators are unbiased and equivalent to the LS counterparts. The estimation of $\boldsymbol{b}$ involves first estimating $\boldsymbol{V}$ using ML with the REML version of the model, and then estimating b using the generalized least squares. The REML approach has become popular in mixed-effect-modeling for the following reasons (McCulloch and Searle 2001):

1. REML is sensible for balanced data, where REML solutions are the GLS and ANOVA estimators, which are minimal variance unbiased estimators under normality. However these proiperties may not apply for unbalanced data.

2. REML takes into account the degrees of freedom used in estimating the fixed effects $\boldsymbol{b}$. This is important when the length of $\boldsymbol{b}$ is large in relation to the number of observations. Presumably this feature occurs for unbalanced data, too.

3. REML estimates of the parameters of $\boldsymbol{V}$ are invariant to the estimates of $\boldsymbol{b}$.

4. REML estimates may not be as sensitive to outliers as ML estimates.

2.7 About modeling strategies

2.7.1 Selection of predictors

There are several ways for prediction of predictors to a regression model. The researcher should always use his/her own expertise in selecting the predictors. If a theory exists on the relationship between the variables, the theory should be used in selecting the predictors and also the shape of the model. Automaticed procedures, such as stepwise backward or forward algorithms should be avoided, and reasons for having the selected predictors in the model should be searched for. Plotting residuals, predictors and responses in all possible ways always helps, for example, in detecting peculiar observations that may the cause cause for unexpected predictors to be significant.

2.7.2 The purpose of modeling

It is important for a researcher to recognize the aim of the modeling. With different research problems, different assumptions of the model become fundamental. As an example let us consider two distinct uses of regression model: explaining and prediction.

Explaining is used, for example, when the effect of a new silvicultural method on the growth of a sapling is evaluated. With such a model, the interesting research questions are (i) does the method have an impact on the growth and (b) how big is the impact. In this situation, the regression model is used in a similar fashion as the analysis of variance or analysis of covariance. When regression model is used for explanation or for testing interesting effects, study of the assumptions behind the test of regression coefficients is very crucial. It is, for example, quite common to make transformations to the response in order to make the residuals normal and homogeneous. If the assumptions are not met, also the test may be biased. Also high correlation among the predictors may cause problems. Of two highly correlated predictors, it may be impossible to say which one has an effect on the response or which not. It is also a quite common situation that the highly correlated predictors are insignificant predictors when used alone, but including them both may give very high significance to them.

The way of selecting the predictors is also important with models used for explanation and testing purposes, as well as documenting all utilized predictors; also those that were found insignificant. For example, suppose we have response variable y and data where we have 100 potential x-variables, of which none is a significant predictor in reality. However, by testing the significance of all these predictors, one at a time, at the risk level of 0.05, the expected number of predictors with p-value of 0.05 or less is 5 just due to the definition of p-value. Reporting only those 5 variables without mentioning the 95 insignificant predictors would lead to very different inference than

reporting of all the tests.

In prediction, the aim is to find a function that as accurately as possible predicts the expected value of the response with given values of the predictors. With this model, different assumptions become crucial. With prediction, the model form needs to be chosen with care. Especially, the behavior of the selected function needs to be analyzed so that it would not lead to unrealistic results when applied beyond the range of the modeling data. Such a risk is especially high if the model includes polynomial terms. Furthermore, with predictive models, making transformations to the response causes biased predictions in the original scale, correction of which requires assumptions on the distribution. The bias correction is also quite vulnerable to a violation in the assumption.

On the other hand, violation of the assumptions about the normality and homogeneity of residuals is not that big a problem with predictive models: the estimates are still unbiased, even though not the best possible. The violations in these assumptions may just lead to problems in testing, leading to including predictors that are insignificant in reality. Including insignificant predictors is possible especially if the predictors are selected from among a large set of potential predictors through trial-and-error procedure, or using a automatic rule, such as backward or forward procedure using large datasets. However, these predictors usually have very small coefficient, meaning that they do not have a big effect on the predictions. However, including extra predictors may not lead to bad predictions, even though it may cause extra work in applications, where the values of all predictors need to be known for the prediction to be possible.

2.8 Exercises

1. Data `cherry.txt` give the volume (cubic feet), height (feet) and diameter (inches) (at 54 inches above ground) for a sample of 31 black cherry trees in the Allegheny National Forest, Pennsylvania. The data were collected in order to find an estimate for the volume of a tree (and therefore the timber yield), given its height and diameter (Hand et al. 1994). Start your model development from relationship $V_i = b_0 D_i^{b_1} H_i^{b_2}$.

   (a) Linearize the model using an appropriate transformation. Specify the relationship between original parameters and the parameters of the linearized model.

   (b) Fit the model using `lm`.

   (c) Define the model matrix $\boldsymbol{X}$ and response vector $\boldsymbol{y}$ for this model and estimate the parameters using matrix operations. Confirm that you got the

same values as returned by function `lm`.

(d) Compute the hat matrix $\boldsymbol{H}$, and estimate the residual variance $\hat{\sigma}^2$. Confirm that you got the same values as returned by function `lm`.

(e) Compute the standard erros of parameter estimates as $\hat{\sigma}^2(\boldsymbol{X'X})^{-1}$, t-statistics, and p-values. Compare to coef(summary(model)).

(f) Compute the model residuals as $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb}$. Compute R^2 for the model, F-statistic and corresponding p-value.

(g) Based on that volume is determined by three dimensions, and diameter and height by one dimension, specify a justified null hypothesis about the the values of b_1 and b_2. Using argument `offset` in function `lm`, fit a null model. Using function `anova` test whether the full model is significantly better than the null model.

2. Data `plot48.txt` includes data from one measured plot. The variables in the data are: puunro=tree id, pl=tree species (1=pine), xk=x-coordinate of the tree location in the stand, yk=y-coordinate of tree location in the satnd, d=tree diameter (cm), h=tree height (m), t=tree age, ig1=basal area growth during the coming 5 year period, id1=diameter growth in the coming 5 year perioid, ig2=past basal area growth, id2=past diameter growth. The aim is to model the coming basal are growth (ig1) of a tree using information on past growth, and current age, diameter and height of the tree.

(a) Graphically explore the relationships of potential explanatory variables and ig1. What would you do with the one outlier of the data.

(b) Start from model $ig1_i = b_0 + b_1 * ig2_i + e_i)$. Check the validity of your assumptions, and find a model that best fulfills the assumptions.

(c) Include other predictors and test whether they improved the model or not. Report your final model.

(d) Extra. Using different models for spatial autocorrelation structures, test whether the model could be improved by taking into account the spatial autocorrelation.

3. Solve LS estimators $b_0 = \bar{y} - b_1\bar{x}$ and $b_1 = \frac{\frac{1}{n}\sum_i x_i \sum_i y_i - \sum_i xy}{\frac{1}{n}\left(\sum_i x_i\right)^2 - \sum_i x^2}$ from equations

$$-2\sum_{i=1}^{n} y_i + 2nb_0 - 2b_1\sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} 2y_i x_i - 2b_0\sum_{i=1}^{n} x_i - 2b_1\sum_{i=1}^{n} x_i^2 = 0$$

Furthermore, show that the solution for b_1 is equivalent to the more commonly used form $b_1 = \frac{S_{xy}}{S_{xx}}$ where $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$ and $S_{xx} = \sum_i (x_i - \bar{x})^2$

Chapter 3

Linear mixed models

The mixed effect model is a model for modeling hierarchical populations, consisting of groups and and individuals within groups. For example, the group may bee a sample plot, and the individual may be tree within the plot. As another example, individuals might be stands within a municipality, municipalities representing the groups. Such population could be modeled either by using mixed models or by using fixed effect models, i.e., we could fit a model including a binary indicator variable describing if the individual belongs to a certain group or not. The difference between these models arises from the type of the groups. With the fixed modeling approach, coefficients would be estimated for each group in the data. Those effects would be fixed, describing the difference between the group in hands from the default group. However, we would mot be able to make inference on groups that are not included in or data. Thus, the data should present all groups that are available, e.g., data should be available from all municipalities of the study area.

If the groups represented in our data represent only a sample from a population of groups, then the mixed model approach would be preferred. Such situation appears, for example, when the data consists of trees within sample plots. All possible plots would not be included in the data, an it is thus realistic to assume that the data only represents a sample from an probably infinite population of sample plots. In the mixed model, we are rather interested in the variation within plots and between plots. In a mixed model, the total residual variation of the observations is divided to within-plot and between-plot variation.

3.1 Model formulation

3.1.1 The variance component model

A variance component model for variable y of individual i in group k is

$$y_{ki} = \mu + a_k + e_{ki} \tag{3.1}$$

where μ is a fixed population mean and part $a_k + e_{ki}$ includes the random parameters (Pinheiro and Bates 2000). It divides the residual error of linear model into two independent parts, to a random group effect for group k and into a random residual for individual i of group k. It is assumed that the group-effects and residual are independent, normally distributed random variables with

$$a_k \sim NID(0, \sigma_a^2)$$

and

$$e_{ki} \sim NID(0, \sigma^2)$$

There may be also other random effects, due to a more complicated structure of the data. For example, individual $i = 1, \ldots, n_k$ of group $k = 1, \ldots, K$ may have been observed at different points in time $t = 1, \ldots, T$. A candidate model for such data would be

$$y_{kti} = \mu + a_k + b_t + e_{kti}$$

where $a_k \sim NID(0, \sigma_a^2)$, $b_t \sim NID(0, \sigma_b^2)$, and $e_{kti} \sim NID(0, \sigma^2)$. In this model, we assume that the observations of different individuals have been taken in the same point in time, thus assuming each observation to include the same effect for time t. In this model, the group and time effects are crossed. And individuals are nested both within groups and points in time.

If the measurements are taken at different points in time, it would be better to assume them to be nested, i.e., measurements taken for individual i at different points are somehow similar, but the tth measurement occasion of individual i is not similar to the tth measurement occasion for other individuals. The nested model would be

$$y_{ki} = \mu + a_k + b_{ki} + e_{kti}$$

where $a_k \sim NID(0, \sigma_a^2)$, $b_{kt} \sim NID(0, \sigma_b^2)$, and $e_{kti} \sim NID(0, \sigma^2)$. Now points in time are nested within individuals, which are further nested within groups. This kind of structure arises, for example, in an analysis of permanent plot data, where the plots have been established at different years and remeasured with fixed (e.g., 5 year) intervals. It would also be possible to define a model that is a combination of these two models, i.e., has both crossed and nested time effects. In most cases, the hierarchical populations lead to mixed models having one or more nested effects.

3.1.2 Mixed model

The linear mixed model is a generalization of the variance component model. It includes both fixed and random parameters. With only one level of grouping and k fixed predictors, it is defined as

$$y_{ki} = b_0 + b_1 x_{1ki} + \ldots + b_p x_{pki} + a_k + e_{ki}\,. \tag{3.2}$$

Part $b_0 + b_1 x_1 + \ldots + b_p x_p$ is the fixed part, which has exactly the same meaning and interpretation as the fixed part of the linear model of the previous section had. Part $a_k + e_{ki}$ is the fixed part, which has exactly the same assumption and interpretation as the random part of a variance component model had. Compared to the variance component model (3.1), mixed model (3.2) differs only in that the fixed population mean has been replaced with a linear function of fixed predictors and parameters. Compared to the fixed-effects model, the residual of model (2.3), is partitioned into two parts in model (3.2). This is such a special case of the multiple regression model, where observations of the same group are correlated. The covariance between observations i and i' of group k is

$$\begin{aligned}
\operatorname{cov}(a_k + e_{ki}, a_k + e_{ki'}) &= \operatorname{cov}(a_k, a_k) + \operatorname{cov}(a_k, e_k i') + \operatorname{cov}(a_k, e_{ki}) + \operatorname{cov}(e_{ki}, e_{ki'}) \\
&= var(a_k) \\
&= \sigma_a^2
\end{aligned}$$

As in the variance component model (3.1), additional crossed and nested levels of grouping can be included into the mixed-effect model.

By reorganizing terms, the mixed-effect model 3.2 can be written as

$$y_{ki} = (b_0 + a_k) + b_1 x_{1ki} + \ldots + b_p x_{pki} + e_{ki}\,.$$

which shows that we are actually assuming that the constant of the model varies between groups, whereas the other coefficients are fixed. However, also other parameters can be assumed to be random as well. This yields a special case of the mixed-effects model that has sometimes been called the random coefficient model. With the single predictor regression, the model becomes

$$y_{ki} = (b_0 + a_k) + (b_1 + c_k) x_{ki} + e_{ki}\,, \tag{3.3}$$

where a_k and c_k are random, group effects which have bivariate normal distribution with mean 0 and variance-covariance matrix

$$\boldsymbol{D} = \operatorname{var}\begin{pmatrix} a_k \\ c_k \end{pmatrix} = \begin{pmatrix} \operatorname{var}(a_k) & \operatorname{cov}(a_k, c_k) \\ \operatorname{cov}(a_k, c_k) & \operatorname{var}(a_k) \end{pmatrix}. \tag{3.4}$$

The other assumptians are as they were before. It is also possible to make restrictions to the above structure, for example, by assuming that the covariance between random effects is 0.

Presentation 3.3 is a good way to show that we are actually assuming random constants and coefficients in a mixed modeling. However, it is often better to organize terms into fixed and random parts. furthermore, we can add also other predictors to the model to get a general mixed-effects model for a population with one level of grouping

$$y_{ki} = b_0 + b_1 x_{1ki} + \ldots + b_p x_{pki} + a_{0k} + a_{1k} x_{1ki} + \ldots + a_{qk} x_{qki} + e_{ki}\,, \quad (3.5)$$

where $b_0 + b_1 x_{1ki} + \ldots + b_p x_{pki}$ is the fixed part and $a_{0k} + a_{1k} x_{1ki} + \ldots + a_{qk} x_{qki} + e_{ki}$ is the random part. The assumptions about them are as stated earlier.

3.1.3 Multiple levels of grouping

As the variance component model, also the mixed model can have multiple levels of grouping. For example, we may have tree diameters and heights measured from permanent plots. Assuming a linear relationship between height and diameter, the height for tree i at time t on plot k would follow model

$$h_{kti} = A_{kt} + B_{kt} d_{kti} + e_k ti$$

This model could be fitted separately for each plot and measurement occasion, leading to estimation of fixed parameters A_{kt} and B_{kt} separately for each measurement occasion within a plot.

Assuming that the measurement occasions in the data are a random sample of possible measurement occasions for that plot, and assuming that the plots represented in the data are a sample from a population of sample plot, we could assume a mixed-effects model having random plot- and time-effects both in the constant and slope,

$$h_{kti} = \alpha + a_k + a_{kt} + (\beta + b_k + b_{kt}) d_{kti} + e_k ti,$$

which can be reorganized to

$$h_{kti} = \alpha + \beta d_{kti} + a_k + b_k d_{kti} + a_{kt} + b_{kt} d_{kti} + e_k ti,$$

where α and β are the population-level means of the constant and slope, a_k and b_k the plot-level random effects with

$$\boldsymbol{D_k} = \begin{pmatrix} a_k \\ b_k \end{pmatrix} \sim \begin{pmatrix} \mathrm{var}(a_k) & \mathrm{cov}(a_k, b_k) \\ \mathrm{cov}(a_k, b_k) & \mathrm{var}(b_k) \end{pmatrix}$$

and

$$\boldsymbol{D_{kt}} \begin{pmatrix} a_{kt} \\ b_{kt} \end{pmatrix} \sim \begin{pmatrix} \mathrm{var}(a_{kt}) & \mathrm{cov}(a_{kt}, b_{kt}) \\ \mathrm{cov}(a_{kt}, b_{kt}) & \mathrm{var}(b_{kt}) \end{pmatrix}$$

with any positive definite $\boldsymbol{D_k}$ and $\boldsymbol{D_{ki}}$. The random effects at different levels, as well as the random errors are multinormally distributed and uncorrelated across the levels. Again, the first two terms in the latter form are the fixed part, and the rest of the equation is the random part.

3.1.4 Matrix formulation

Single level of grouping

Mixed-effects model (3.5) can be stated in a matrix form as follows

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{c} + \boldsymbol{e}\,. \tag{3.6}$$

where $\boldsymbol{X}b$ is the fixed part and $Zc + e$ the random part. The fixed part is similar to that of the fixed-effects model.

For the random part, $\boldsymbol{Z}$ is the design matrix of the random part and $\boldsymbol{c}$ is the vector of random parameters. The length of c id $k \times q$, where k is the number of groups, and q the number of group-specific random parameters. Correspondingly, matrix $\boldsymbol{Z}$ is an has n rows and $x \times p$ columns. Thus, the number of columns in $\boldsymbol{Z}$ and the length of $\boldsymbol{c}$ depends on the number of groups in the data. Matrix $\boldsymbol{Z}$ is usually organized so that all parameters of the first group are included first, then those of the second group. This organization leads to a block-diagonal structure of $var(\boldsymbol{c})$ and $\boldsymbol{Z}$, which eases the computations. The structure is probably best described by the following example.

Example 3.1 *Assume model $y_{ki} = \alpha + \beta + a_k + b_k x_{ki} + e_{ki}$, where $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$. The model can be written as*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{c} + \boldsymbol{e}$$

by defining

$$\boldsymbol{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{K1} \\ y_{K2} \\ \vdots \\ y_{Kn_K} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & y_{1n_1} \\ 1 & y_{21} \\ 1 & y_{22} \\ \vdots & \vdots \\ 1 & y_{2n_2} \\ \vdots & \vdots \\ 1 & y_{K1} \\ 1 & y_{K2} \\ \vdots & \vdots \\ 1 & y_{Kn_K} \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \boldsymbol{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n_1} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2n_2} \\ \vdots \\ e_{K1} \\ e_{K2} \\ \vdots \\ e_{Kn_K} \end{bmatrix},$$

$$\boldsymbol{Z} = \begin{bmatrix} 1 & x_{11} & 0 & 0 & \cdots & 0 & 0 \\ 1 & x_{12} & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & y_{1n_1} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & x_{21} & \cdots & 0 & 0 \\ 0 & 0 & 1 & x_{22} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 1 & y_{2n_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & x_{K1} \\ 0 & 0 & 0 & 0 & \cdots & 1 & x_{K2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & x_{Kn_K} \end{bmatrix}, \boldsymbol{c} = \begin{bmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \\ \cdots \\ a_K \\ b_K \end{bmatrix}.$$

The above example showed the matrices for the whole data. Matrices $\boldsymbol{Z}$ and $\boldsymbol{D}$ consist of blocks, each of which presents a single group. The model for each group can also be written in matrix a form, as shown in the following example. The model for a single group is needed in prediction of random effects.

Example 3.2 *In example 3.1, the model for a single group is written in matrix form as*

$$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{b} + \boldsymbol{Z}_k\boldsymbol{c}_k + \boldsymbol{e}_k$$

where

$$\boldsymbol{y}_k = \begin{bmatrix} y_{k1} \\ y_{k2} \\ \vdots \\ y_{kn_k} \end{bmatrix}, \boldsymbol{X}_k = \begin{bmatrix} 1 & x_{k1} \\ 1 & x_{k2} \\ \vdots & \vdots \\ 1 & y_{kn_k} \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix},$$

$$\boldsymbol{e}_k = \begin{bmatrix} e_{k1} \\ e_{k2} \\ \vdots \\ e_{kn_k} \end{bmatrix}, \boldsymbol{Z}_k = \begin{bmatrix} 1 & x_{k1} \\ 1 & x_{k2} \\ \vdots & \vdots \\ 1 & y_{kn_k} \end{bmatrix}, \boldsymbol{c}_k = \begin{bmatrix} a_k \\ b_k \end{bmatrix}.$$

Thus, to write a mixed model in a matrix form, the essential task is to be able to write the model for a single group. The other matrices and vectors for the model of whole data are then obtained as follows

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ X_K \end{bmatrix} \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \cdots & 0 \\ 0 & \boldsymbol{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_K \end{bmatrix} \boldsymbol{c} = \begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \\ \vdots \\ \boldsymbol{c}_K \end{bmatrix} \boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_K \end{bmatrix} \quad (3.7)$$

The vector of fixed parameters, $\boldsymbol{b}$, is the same for whole data and a single group.

In the model for a single group,

$$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{b} + \boldsymbol{Z}_k\boldsymbol{c}_k + \boldsymbol{e}_k, \quad (3.8)$$

$\boldsymbol{c}_k$, and $\boldsymbol{e}_k$ are random vectors. Of the vector of random effects, it is assumed that $\boldsymbol{c}_k \sim N_q(\boldsymbol{0}, \boldsymbol{D}_k)$, where D is a positive definite matrix. With model (3.3) such an assumption was specified in (3.4). Of $\boldsymbol{e}_k$, the very general assumption is that $\boldsymbol{e}_k \sim N_{n_k}(\boldsymbol{0}, \boldsymbol{R}_k)$, where $\boldsymbol{R}_k$ parameterizes the assumed trends in the conditional residual variance and dependencies in the conditional residuals within a group. The usual (starting) hypothesis is that $\boldsymbol{R}_k = \sigma^2 \boldsymbol{I}$, which implies an assumption on uncorrelated conditional residuals with a constant variance. Relaxing this assumption is possible, for example assumptions on heteroscedastic variance and different autocorrelation structures are possible.

The assumptions about the residual error and random effects lead to that

$$\boldsymbol{y}_k \sim N_{n_k}(\boldsymbol{X}_k\boldsymbol{b}, \boldsymbol{Z}_k\boldsymbol{D}\boldsymbol{Z}_k' + R)$$

and

$$\mathrm{cov}(\mathrm{c_k}, \mathrm{y_k'}) = \mathrm{DZ_k'}$$

These variance-covariance matrices can be easily derived by writing the assumed expressions for $\boldsymbol{u}_k$ into the equations of variance and covariance, and then applying the rules of chapter 1.

Multiple levels of grouping

Assume that we have two nested levels of grouping, the first (outer) level of grouping being indexed by $k = 1, \ldots, K$, and the innermost grouping being indexed by $l = 1, \ldots, n_k$. The linear mixed-effects model for group l within group k is written as

$$\boldsymbol{y}_{kl} = \boldsymbol{X}_{kl}\boldsymbol{b} + \boldsymbol{Z}_{k,l}\boldsymbol{c}_k' + \boldsymbol{Z}_{kl}\boldsymbol{c}_{kl} + \boldsymbol{e}_{kl}$$

where $\boldsymbol{c}_k'$ includes the first (outer) level random effects for group k, and $\boldsymbol{c}_{kl}$ the second (innermost) level random effects for level l within k. Matrices $\boldsymbol{Z}_{k,l}$ and bmZ_{kl} are the corresponding design matrices.

The model for the whole group k is

$$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{b} + \boldsymbol{Z}_k\boldsymbol{c}_k + \boldsymbol{e}_k$$

where

$$\boldsymbol{y}_k = \begin{bmatrix} \boldsymbol{y}_{k1} \\ \boldsymbol{y}_{k2} \\ \vdots \\ \boldsymbol{y}_{kn_k} \end{bmatrix}, \boldsymbol{X}_k = \begin{bmatrix} \boldsymbol{X}_{k1} \\ \boldsymbol{X}_{k2} \\ \vdots \\ \boldsymbol{X}_{kn_k} \end{bmatrix},$$

$$\boldsymbol{Z}_k = \begin{bmatrix} \boldsymbol{Z}_{k,1} & \boldsymbol{Z}_{k1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{Z}_{k,2} & \boldsymbol{0} & \boldsymbol{Z}_{k1} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Z}_{k,n_k} & \boldsymbol{0} & \boldsymbol{0} & \vdots & \boldsymbol{Z}_{kn_k} \end{bmatrix}, \boldsymbol{c}_k = \begin{bmatrix} \boldsymbol{c}_k' \\ \boldsymbol{c}_{k1} \\ \boldsymbol{c}_{k2} \\ \vdots \\ \boldsymbol{c}_{kn_k} \end{bmatrix}$$

and the vector of fixed parameters is the same as in the previous model, $\boldsymbol{b}$. The model for the whole data can be written as described in equation (3.7)

Example 3.3 *Model*

$$h_{kti} = \alpha + \beta d_{kti} + a_k + b_k d_{kti} + a_{kt} + b_{kt} d_{kti} + e_k ti,$$

for time point t within stand k can be written as

$$\boldsymbol{h}_{kt} = \boldsymbol{X}_{kt}\boldsymbol{b} + \boldsymbol{Z}_{k,t}\boldsymbol{c}'_k + \boldsymbol{Z}_{kt}\boldsymbol{c}_{kt} + \boldsymbol{e}_{kt}$$

by defining

$$\boldsymbol{h}_{kt} = \begin{bmatrix} h_{kt1} \\ h_{kt2} \\ \vdots \\ h_{ktn_{kt}} \end{bmatrix}, \boldsymbol{X}_{kt} = \boldsymbol{Z}_{k,t} = \boldsymbol{Z}_{kt} = \begin{bmatrix} 1 & d_{kt1} \\ 1 & d_{kt2} \\ \vdots & \vdots \\ 1 & d_{ktn_{kt}} \end{bmatrix},$$

$$\boldsymbol{c}'_k = \begin{bmatrix} a_k \\ b_k \end{bmatrix}, \boldsymbol{c}_{kt} = \begin{bmatrix} a_{kt} \\ b_{kt} \end{bmatrix}, \boldsymbol{e}_{kt} = \begin{bmatrix} e_{kt1} \\ e_{kt2} \\ \vdots \\ e_{ktn_{kt}} \end{bmatrix}$$

The model for plot k becomes

$$\boldsymbol{h}_k = \boldsymbol{X}_k\boldsymbol{b} + \boldsymbol{Z}_k\boldsymbol{c}_k + \boldsymbol{e}_k$$

where the design matrix for random part and the vector of random effects are

$$\boldsymbol{Z}_k = \begin{bmatrix} 1 & d_{k11} & 1 & d_{k11} & 0 & 0 & \cdots & 0 & 0 \\ 1 & d_{k12} & 1 & d_{k12} & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & d_{k1n_{k1}} & 1 & d_{k1n_{k1}} & 0 & 0 & \cdots & 0 & 0 \\ 1 & d_{k21} & 0 & 0 & 1 & d_{k21} & \cdots & 0 & 0 \\ 1 & d_{k22} & 0 & 0 & 1 & d_{k22} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & d_{k2n_{k2}} & 0 & 0 & 1 & d_{k2n_{k2}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & d_{kn_k1} & 0 & 0 & 0 & 0 & \cdots & 1 & d_{kn_k1} \\ 1 & d_{kn_k2} & 0 & 0 & 0 & 0 & \cdots & 1 & d_{kn_k2} \\ 1 & d_{kn_kn_{kn_k}} & 0 & 0 & 0 & 0 & \cdots & 1 & d_{kn_kn_{kn_k}} \end{bmatrix}, \boldsymbol{b}_k = \begin{bmatrix} a_k \\ b_k \\ a_{k1} \\ b_{k1} \\ a_{k2} \\ b_{k2} \\ \vdots \\ a_{kn_k} \\ b_{kn_k} \end{bmatrix},$$

The general matrix formulation

In all the above cases, the mixed-effects model for the whole data can be written in form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{c} + \boldsymbol{e}' \tag{3.9}$$

where $\boldsymbol{Xb}$ is the fixed part and $\boldsymbol{Zc} + \boldsymbol{e'}$ is the random part. By defining that $\boldsymbol{e} = \boldsymbol{Zc}+\boldsymbol{e'}$, we see that the mixed-effects model is a special case of the linear model (2.4). In the mixed-effects model, the residual of the linear model is partitioned into several terms. Some of these terms are common for for some observations. For example, the group effects are common to all observations of a specific group. These shared parts of the residual cause correlation among the residuals of the linear model, $\boldsymbol{e}$. This correlation is parameterized into the matrix $\boldsymbol{V}$ through the variance-covariance matrix of the random effects, D. The variance-covariance matrix of the residual of the linear model can be computed as

$$\begin{aligned} \boldsymbol{V} &= \mathrm{var}(\boldsymbol{e}) \\ &= \mathrm{var}(\boldsymbol{Zc} + \boldsymbol{e'}) \\ &= \boldsymbol{ZDZ'} + \boldsymbol{R} \end{aligned}$$

Thus, model (3.9) is such a special case of model (2.4), where the variance covariance-matrix of residuals is specified as

$$\mathrm{var}(e) = \boldsymbol{V} = \boldsymbol{ZDZ'} + \boldsymbol{R}\,.$$

3.2 Estimation

3.2.1 Analysis of variance

The earliest estimation methods of a mixed-effects model are based on an ANOVA-approach. In the anova approach, the group-effects are first included into the model as dummy variables, and then the estimates for the variance components were based on the residual and regression sums of squares, which are divided with the number of observations and corrected for the degrees of freedom used. The ANOVA approach leads to unique, unbiased estimators for datasets with equal number of observations in each group and no missing data. However, problems arise with unbalanced data and missing observations, and no agreement exists in which estimators should be used in such cases. Thus, currently the most often applied and most widely accepted methods are those based on maximum likelihood and restricted (or residual) maximum likelihood.

3.2.2 Maximum Likelihood

The method of maximum likelihood is a straightforward application of what was presented in section 2.6. One should be aware that it requires the assumption on normality already in the estimation stage, as does the REML method, too. In the most common case, an unrestricted positive definite structure is given to the variance-covariance

matrix for random effects at each level of grouping, and matrix $\boldsymbol{R}$ is defined as the diagonal matrix $\sigma^2\boldsymbol{I}$. The ML yields estimates for the variance-covariance parameters for the random parameters at each level of grouping, for the residual error variance, and for vector $\boldsymbol{b}$. Furthermore, the dependence of residuals or heteroscedastic variance can be modeled by giving a more sophisticated structure for $\boldsymbol{R}$. An important result for the ML-method is, that whatever structure $\boldsymbol{V}$ has, the ML-estimates of b are the GLS estimates, where the required variance-covariance matrix is replaced with the ML-estimate. A widely known problem with the ML-estimates is that it is downward biased. The extent of bias can be computed only for some special cases.

3.2.3 Restricted maximum likelihood

The method of restricted maximum likelihood (REML) was presented in section 2.6. In the REML method, the ML equation includes only the parameters specifying the random part of the LMM. Thus, REML itself does not involve estimation of $\boldsymbol{b}$. Instead, the estimation of $\boldsymbol{b}$ is conducted by direct application of GLS-method, after the estimates for the parameters of $\boldsymbol{V}$ are first obtained by REML. In general, REML is usually preferred over ML because it yields the minimum-variance unbiased estimators for the variance components in the case of balanced data, which are also the ANOVA-estimators. However, it is not known if these estimators are unbiased for unbalanced data, but at least they are less biased than the ML-estimators. Other reasons for preferring REML over ML were listed in section 2.6. However, as REML itself does not involve estimation of the fixed parameters, tests for finding appropriate structure **for the fixed part** should not be based on REML-likelihood ratio; the ML-estimation can be used instead. The final model can then be estimated using REML, after appropriate structure has been found for the fixed part.

3.3 Prediction of random effects

Model estimation yields the estimates for the fixed parameters, those for the variances and covariances at different levels of grouping, and for the variance of the residual. However, estimation of the LMM does not yield estimates for the group effects, $\boldsymbol{c}_k$ for groups $k = 1, \ldots, K$. These effects can, however, be predicted using the Best Linear Predictor, which was presented for a general case in section 1.6 and is taken back below. If

$$\begin{pmatrix} \boldsymbol{h_1} \\ \boldsymbol{h_2} \end{pmatrix} \sim \left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{12}' & \boldsymbol{V}_2 \end{pmatrix} \right],$$

then

$$BLP(\boldsymbol{h}_1) = \widehat{\boldsymbol{h}_1} = \boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_2^{-1}\left(\boldsymbol{h}_2 - \boldsymbol{\mu}_2\right)$$

with a prediction variance of

$$\mathrm{var}(\widehat{\boldsymbol{h}_1} - \boldsymbol{h}_1) = \boldsymbol{V}_1 - \boldsymbol{V}_{12}\boldsymbol{V}_2^{-1}\boldsymbol{V}_{12}'$$

Consider the model (3.8) for a single group k. We want to predict the random effects $\boldsymbol{b}_k$ using the observations of y_k. Thus, $\boldsymbol{h}_1 = \boldsymbol{b}_k$ and $\boldsymbol{h}_2 = \boldsymbol{y}_k$ The required variances and covariances were given before

$$\begin{aligned} \boldsymbol{V}_1 &= \boldsymbol{D} \\ \boldsymbol{V}_2 &= \boldsymbol{Z}_k\boldsymbol{D}\boldsymbol{Z}_k' + \boldsymbol{R} \\ \boldsymbol{V}_{12} &= \boldsymbol{D}\boldsymbol{Z}_k' \end{aligned}$$

The BLP of random effects, and their prediction error variance become

$$\begin{aligned} \widehat{\boldsymbol{b}_k} &= \boldsymbol{D}\boldsymbol{Z}_k' \left(\boldsymbol{Z}_k\boldsymbol{D}\boldsymbol{Z}_k' + \boldsymbol{R}\right)^{-1} (\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{b}) \\ \mathrm{var}(\widehat{\boldsymbol{b}_k} - \boldsymbol{b}_k) &= \boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_k' \left(\boldsymbol{Z}_k\boldsymbol{D}\boldsymbol{Z}_k' + \boldsymbol{R}\right)^{-1} \boldsymbol{Z}_k\boldsymbol{D} \end{aligned}$$

Henderson mixed model equations give equivalent equations

$$\begin{aligned} \widehat{\boldsymbol{b}_k} &= \left(\boldsymbol{Z}_k'\boldsymbol{R}^{-1}\boldsymbol{Z}_k + \boldsymbol{D}^{-1}\right)^{-1} \boldsymbol{Z}'\boldsymbol{R}^{-1} (\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{b}) \\ \mathrm{var}(\widehat{\boldsymbol{b}_k} - \boldsymbol{b}_k) &= \left(\boldsymbol{Z}_k'\boldsymbol{R}^{-1}\boldsymbol{Z}_k + \boldsymbol{D}^{-1}\right)^{-1} \end{aligned}$$

(Searle et al. 1992, Lappi 1991, see). The lower form requires inversion of a matrix of dimension $p \times p$ in addition to the inversion of $\boldsymbol{R}$, which is of dimension $n_k \times n_k$, but is usually diagonal. The upper requires the inversion of a non-diagonal matrix of dimension $n \times n$. Thus, the lower form is computationally better if $\boldsymbol{R}$ is diagonal and the number of observations is high compared to the number of random parameters.

3.4 Inference and tests

3.4.1 Checking of the assumptions of the mixed model

The model form

With the mixed model, checking that the assumed model form fits to the data is as important as with the simpler models, too. However, checking the assumptions is not as simple as it was before. The plot of residuals on all the predictors is a good starting point to see if the model has a good shape. However, there may also other ways of checking the fit. an example of such a case is presented in the following example. For correcting the model shape, the same rules apply as in the case of linear models (see section 2.4)

Example 3.4 *Compare three functions for H-D modeling*

Other assumptions to be checked are, if the assumption of random parameters are realistic. For example, in a variance component model, only the level of the assumed model is assumed to vary among groups, whereas the model shape is assumed to be the same. Plots should be used to see if this is a realistic assumption.

Example 3.5 *Longitudinal model for H-D curves*

Assumptions on the residual variance

A plot of standardized (conditional) residuals on the predicted value should express a constant variance with no trends. A variance function should be used to homogenize the residuals, or alternatively, a transformation could be made to the response. The normality of residuals should be checked, for example, by using q-q plots, as the ML and REML methods are based on normality. Slight discrepancy from normality is usually allowed, as no very good methods for normalizing the data have been presented. Transformations to the response could be used to get the data better met the normality. However, this results into the bias problem, which was discussed in section 2.4.3.

In a repeated measurements data, lagged residuals could be used to analyze the possible temporal autocorrelation. In spatial data, covariagrams could be used to analyze the possible spatial autocorrelation.

Assumptions on the random effects

Also the ransom effects are assumed to have constant variance and normal distribution. These assumptions should be checked by plotting the BLUPs of random effects against the predicted value. The normality of random effects should be checked by using q-q-plots. If several random effects are included at a certain level, the linearity of the correlation can be checked by plotting the predicted random effects against each other. The nonlinearity is usually linked with discrepancy from (multi)normality; this results from that with multinormal distribution, all marginal and joint distributions are normal and all correlations are linear.

3.4.2 Tests of the model

Tests on the fixed part can be based on the same principles as in the case of linear model. However, one should keep in mind that LR test for two models with different fixed effects should be carried out with models estimated by ML, not wit models based on REML. Assumptions on the random part can be tested with either of the models.

3.5 Extending the linear mixed model

The assumption of constant variance and independence of residuals can be relaxed also with the linear mixed-effects-model. In R, the same functions can be used for specifying the structure for the residual term than wee used with the linear model and function `glm`.

3.6 An analysis of H-D curve

3.6.1 Selection of model form

Data on 1678 height-diameter observations for Scots Pine trees from 56 plots in North Carelia, Finland is analyzed. The data is cross-sectional, so the only cause for hierarchy is the organization of data into plots and trees within plots. The analysis has the following steps

1. Selecting appropriate model form from among four alternatives: Power equation, Meyer equation, Korf Curve, and Näslunds equation.

2. Fitting a first mixed-effects model, and checking the assumptions on model form and residual errors, and development of the model to better meet the assumptions.

3. Including additional stand-specific covariates to capture part of the between-stand variation.

4. Demonstrate the use of these two models for prediction when one sample tree has been measured for diameter and height.

5. Demonstrate the use of these two models for prediction when three sample tree has been measured for diameter and height.

The R-code of in file `examples_ch3.R` includes the analysis.

First, define the possible functions for the relationship, and find initial estimates for the parameter values.

```
> spati<-read.table("c:/laurim/biometria/spati.txt",header=TRUE)
> # Include only pine sample trees
> spati<-spati[spati$pl==1,]
> spati$h2<-spati$h-1.3
>
> # Select appropriate model form from among alternatives:
> # power H=aD^b
> HDpower<-function(d,a,b) {
+          a*d^b
+          }
> # Meyer: H=a(1-exp{-bD})
> HDmeyer<-function(d,a,b) {
+          a*(1-exp(-b*d))
+          }
```

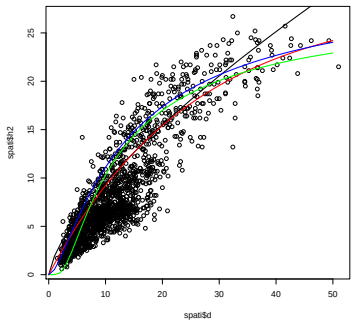


Figure 3.1: Plot of the data, and the four alternative models with their initial parameter values.

```
> # Korf: H=a exp{-bD^{-c}}
> HDkorf<-function(d,a,b) {
+          a*exp(-b*d^(-1))
+          }
> # Nslund: H=1.3+d^2/(a+bd)^2
> HDnaslund<-function(d,a,b) {
+             d^2/(a+b*d)^2
+             }
> # We will fit each of these models into the data of each plot using nls and study the residuals
> # to select the best fitting model.
>
>
> # plot the data
> plot(spati$d,spati$h2)
> # manually find starting values for the parameters
> d<-seq(0,50)
> lines(d,HDpower(d,1.9,0.7),lwd=2)
> lines(d,HDmeyer(d,28,0.04),col="red",lwd=2)
> lines(d,HDkorf(d,28,10),col="green",lwd=2)
> lines(d,HDnaslund(d,1.2,0.18),col="blue",lwd=2)
```

The following code fits each of the four alternative models for each plot. It also plots the observations of each plot, and adds the fitted curves to the plots.

```
> # A vector including plot numbers
> plots<-unique(spati$plot)
>
> # do these computations for each plot
> # and plot the models
> korfa<-korfb<-rep(NA,length(plots))
> for (i in 1:length(plots)) {
+     # open a new window at every fourth iteration
+     if (((i-1)%%4)==0) {
+        windows()
+        par(mfcol=c(2,2))
+        }
+     # thisplot will include observations only from i:th plot
+     thisplot<-spati[spati$plot==plots[i],]
+
+     # fit each of the models for each plot using nls
+     fmpower<-nls(h2~HDpower(d,a,b),
+                  data=thisplot,
+                  start=list(a=1.9,b=0.7))
+     fmmeyer<-nls(h2~HDmeyer(d,a,b),
+                  data=thisplot,
+                  start=list(a=30,b=0.04))
+     # common start values for each plot did not converge for each plot.
+     # That is why startong values are estimated separately for each plot
+     # by fitting a linearized form of the model.
+     # This could be automatized by defining a self-start function.
```

```
+     korfstart<-lm(log(h2)~I(1/d),data=thisplot)
+     fmkorf<-nls(h2~HDkorf(d,a,b),
+                 data=thisplot,
+                 start=list(a=exp(coef(korfstart)[1]),b=-coef(korfstart)[2]))
+     fmnaslund<-nls(h2~HDnaslund(d,a,b),
+                    data=thisplot,
+                    start=list(a=1.2,b=0.18))
+
+     # Make a new plot that shows the h-d data of thisparticular plot
+     # and add the fitted lines from each of the four fits into the plot.
+     plot(thisplot$d,thisplot$h2)
+     lines(d,HDpower(d,coef(fmpower)[1],coef(fmpower)[2]),                    lwd=2)
+     lines(d,HDmeyer(d,coef(fmmeyer)[1],coef(fmmeyer)[2]),       col="red",   lwd=2)
+     lines(d,HDkorf(d,coef(fmkorf)[1],coef(fmkorf)[2]),          col="green",lwd=2)
+     lines(d,HDnaslund(d,coef(fmnaslund)[1],coef(fmnaslund)[2]), col="blue",  lwd=2)
+     mtext(c("Power","Meyer","Korf","Nslund"),
+           col=c("black","red","green","blue"),
+           at=seq(min(thisplot$d),max(thisplot$d),length=4))
+
+     # Save the residuals from each of these fits
+     # Note: the columns are automatically created in the first iteration.
+     spati$respow[spati$plot==plots[i]]<-resid(fmpower)
+     spati$resmey[spati$plot==plots[i]]<-resid(fmmeyer)
+     spati$reskorf[spati$plot==plots[i]]<-resid(fmkorf)
+     spati$resnas[spati$plot==plots[i]]<-resid(fmnaslund)
+
+     # add also a column including a plotwise standardized diameter
+     spati$dstd[spati$plot==plots[i]]<-(thisplot$d-mean(thisplot$d))/sd(thisplot$d)
+     }
```

The following code computes means, standard deviations and standard errors of residuals for each model. Näslunds model seems to be the best, followed by Korf, Meyer and Power functions.

```
> rbind(sd=sd(spati[,c("respow","resmey","reskorf","resnas")]),
+       mean=mean(spati[,c("respow","resmey","reskorf","resnas")]),
+       se=sd(spati[,c("respow","resmey","reskorf","resnas")])/sqrt(dim(spati)[1]))
         respow      resmey     reskorf       resnas
sd    1.09239050  0.97568785 0.972231759  0.970056688
mean -0.01628642 -0.01793159 0.009149283 -0.003784412
se    0.02666748  0.02381853 0.023734158  0.023681060
```

Next, plot the residuals against diameter and standardized diameter to see assess the fit of each of these models.

```
# plot the residuals in the standard way
# against tree diameter
       windows()
       par(mfcol=c(2,2))
       plot(spati$d,spati$respow)
       abline(0,0)
       plot(spati$d,spati$resmey)
       abline(0,0)
       plot(spati$d,spati$reskorf)
       abline(0,0)
       plot(spati$d,spati$resnas)
       abline(0,0)

# and against standardized d
       windows()
       par(mfcol=c(2,2))
       plot(spati$dstd,spati$respow)
       abline(0,0)
       plot(spati$dstd,spati$resmey)
       abline(0,0)
       plot(spati$dstd,spati$reskorf)
       abline(0,0)
       plot(spati$dstd,spati$resnas)
       abline(0,0)

# Residuals against standardized diameter
```

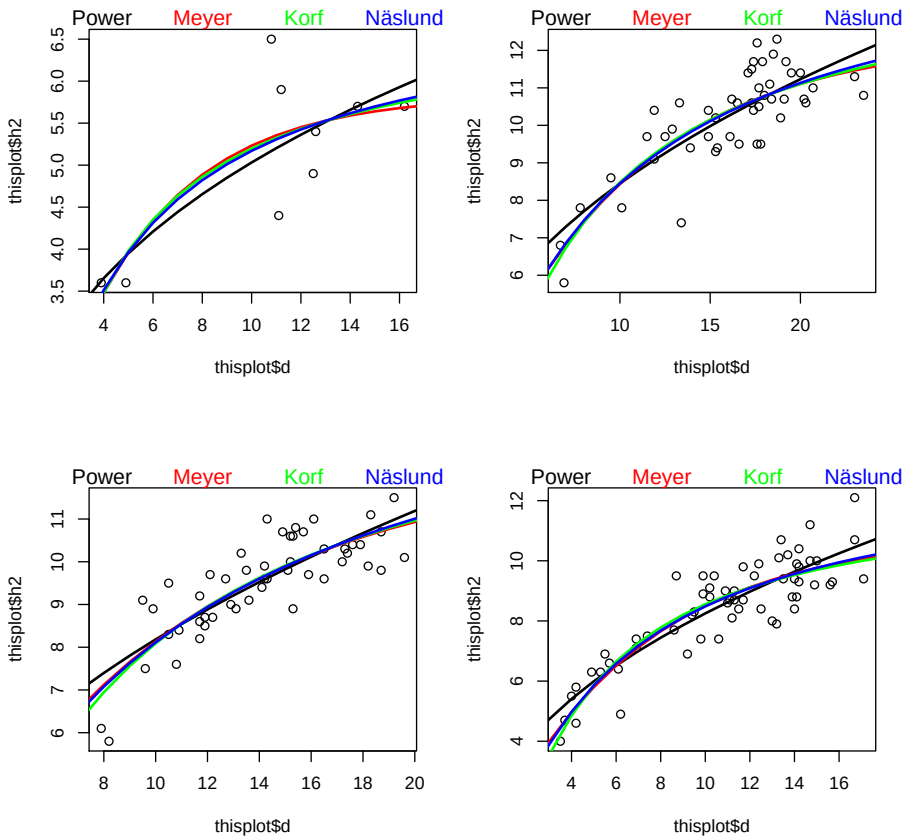


Figure 3.2: Data and fitted models for four example plots.

```
limits<-quantile(spati$dstd,probs=seq(0,1,0.1))
windows()
par(mfcol=c(2,2))
mywhiskers(spati$dstd,spati$respow,limits=limits,se=FALSE,main="Power")
mywhiskers(spati$dstd,spati$respow,limits=limits,add=TRUE,lwd=2)
abline(0,0)
mywhiskers(spati$dstd,spati$resmey,limits=limits,se=FALSE,main="Meyer")
mywhiskers(spati$dstd,spati$resmey,limits=limits,add=TRUE,lwd=2)
abline(0,0)
mywhiskers(spati$dstd,spati$reskorf,limits=limits,se=FALSE,main="Korf")
mywhiskers(spati$dstd,spati$reskorf,limits=limits,add=TRUE,lwd=2)
abline(0,0)
mywhiskers(spati$dstd,spati$resnas,limits=limits,se=FALSE,main="Nslund")
mywhiskers(spati$dstd,spati$resnas,limits=limits,add=TRUE,lwd=2)
abline(0,0)

# Residuals against raw diameter
limits<-quantile(spati$d,probs=seq(0,1,0.1))
windows()
par(mfcol=c(2,2))
mywhiskers(spati$d,spati$respow,limits=limits,se=FALSE,main="Power")
mywhiskers(spati$d,spati$respow,limits=limits,add=TRUE,lwd=2)
abline(0,0)
mywhiskers(spati$d,spati$resmey,limits=limits,se=FALSE,main="Meyer")
mywhiskers(spati$d,spati$resmey,limits=limits,add=TRUE,lwd=2)
abline(0,0)
mywhiskers(spati$d,spati$reskorf,limits=limits,se=FALSE,main="Korf")
mywhiskers(spati$d,spati$reskorf,limits=limits,add=TRUE,lwd=2)
abline(0,0)
mywhiskers(spati$d,spati$resnas,limits=limits,se=FALSE,main="Nslund")
mywhiskers(spati$d,spati$resnas,limits=limits,add=TRUE,lwd=2)
abline(0,0)
```

The plots are shown in figures 3.3 and 3.4. Residuals against diameter do not show any trend, whereas plotting against standardized diameter shows that Power and Meyer equations are not flexible enough for our purposes. They lead underestimation of height for medium sized trees of a stand, and overestimation of height for largest and smallest trees. No difference can be seen between Korf and Näslunds curves. Plots 3.5 and 3.6 show the corresponding plots from mywhiskers. I selected Korf curve, but Näslunds curve could have been as good or even a better alternative.

3.6.2 Fitting the linear mixed-effects model

The previously shown computations were carried out in the original scale using the nonlinear regression approach. However, many modeling tasks become much easier in the linear scale, even though it leads to biased predictions in nonlinear scale. However, we select a linearized model and then do the normality-based bias correction. The linearized Korf model with random effects for both a and be is obtained using logarithmic transformation as

$$\ln(h_{ki} - 1.3) = \alpha + a_k - (\beta + b_k)\frac{1}{d}$$

The constant of the linearized model, $\alpha + a_k$, is obtained by making the logarithmic transformation to the original scale parameter. The following code fits the linearized korf curve to each plot, and plots the obtained parameters. The plots are shown in figure 3.7. The distribution of the stand-specific parameters seems not to be normal, and the correlation shows a nonlinear trend. However, we do not have good tools to fix

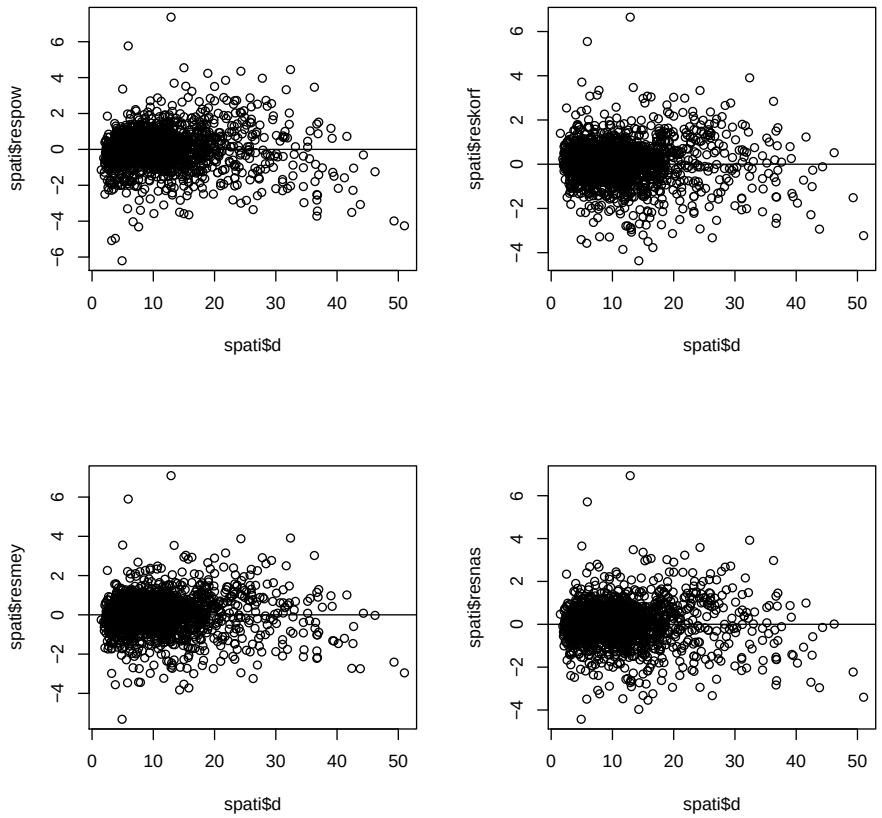


Figure 3.3: Residuals against diameter.

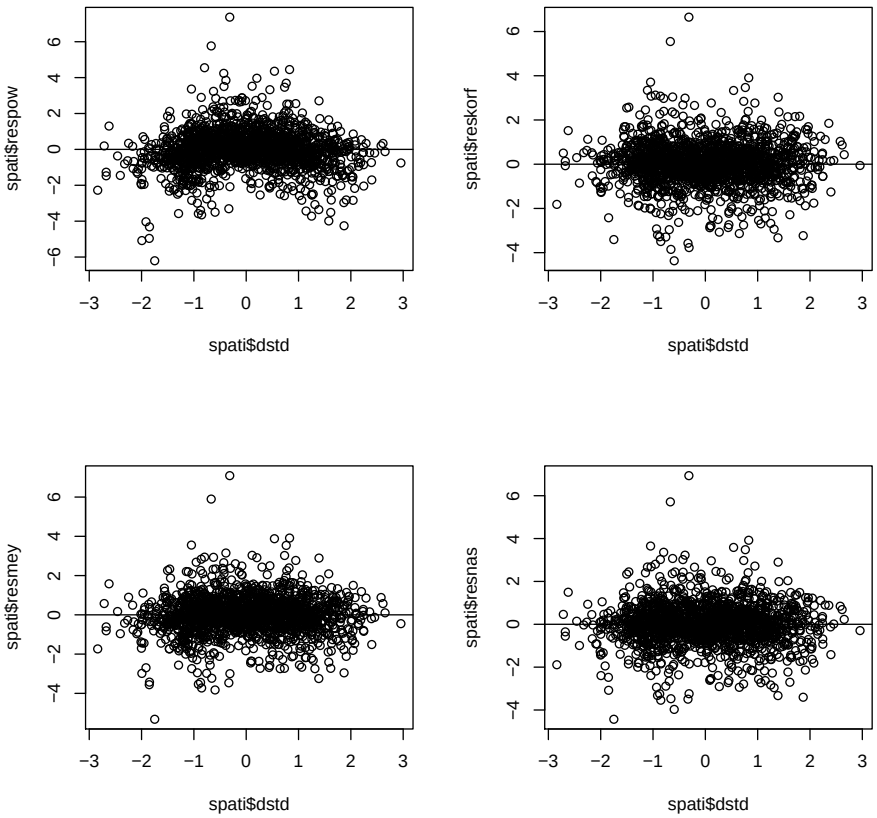


Figure 3.4: Residuals against standardized diameter.

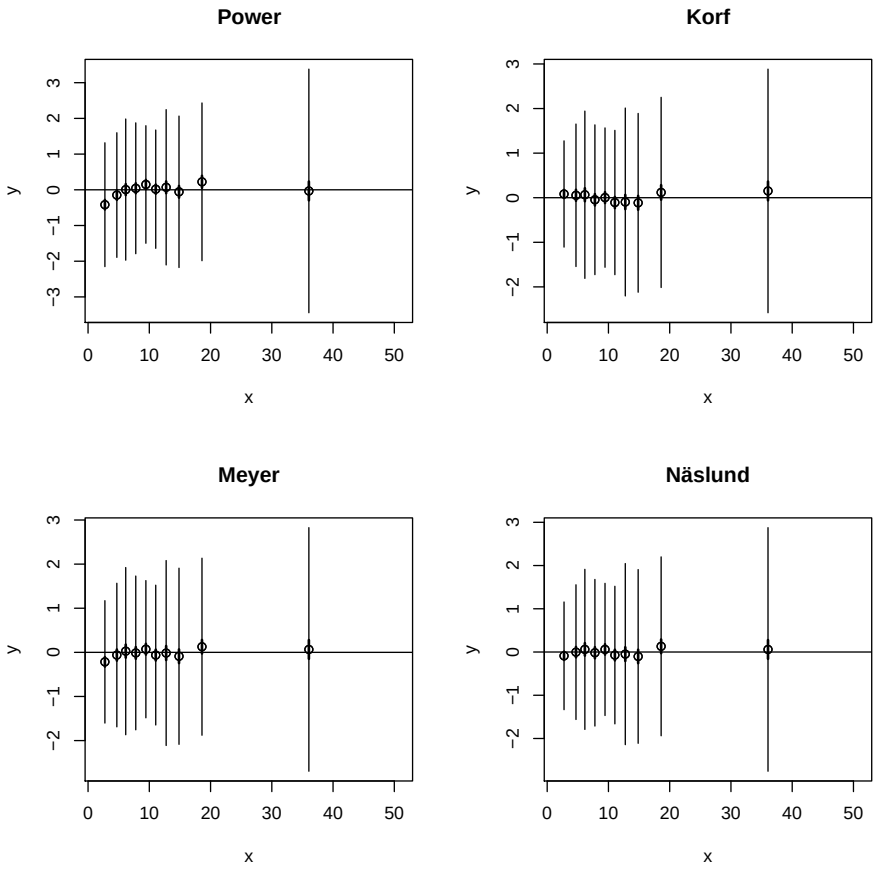


Figure 3.5: Residuals against diameter.

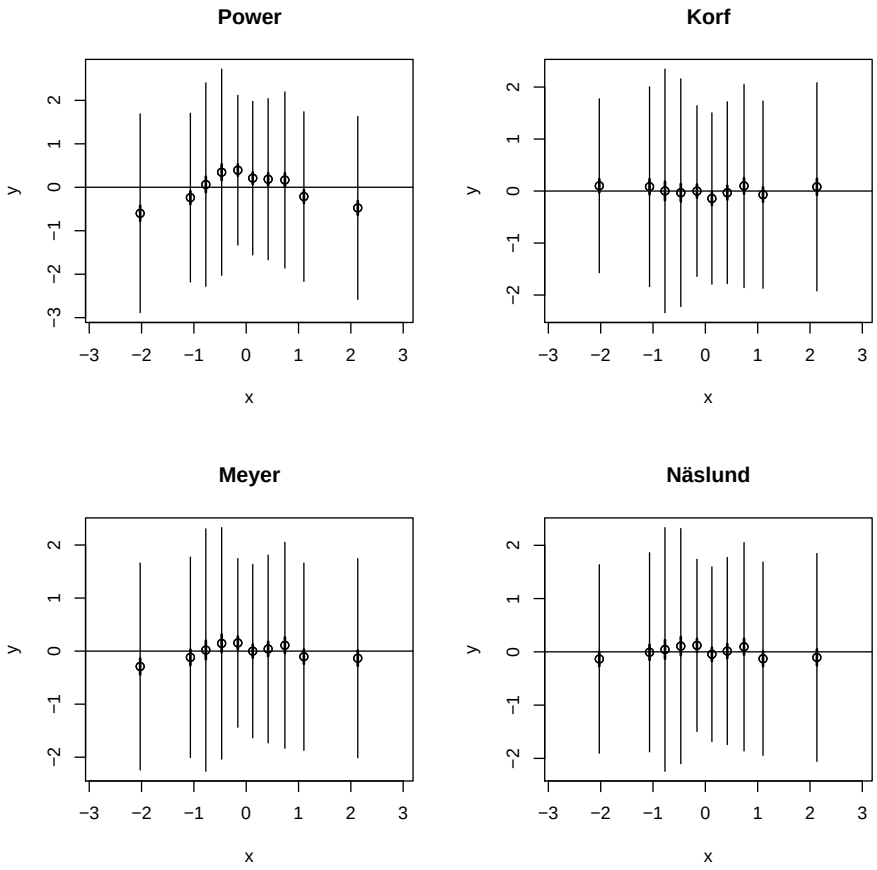


Figure 3.6: Residuals against standardized diameter.

these departures, and we will proceed with this model, even though we are aware that the assumptions are not very well met.

```
> korfpar<-data.frame(a=rep(NA,length(plots)),b=rep(NA,length(plots)))
> for (i in 1:length(plots)) {
+
+     # thisplot will include observations only from i:th plot
+     thisplot<-spati[spati$plot==plots[i],]
+
+     lmkorf<-lm(log(h2)~I(1/d),data=thisplot)
+     korfpar[i,]<-c(1,-1)*coef(lmkorf)
+     }
> head(korfpar)
         a        b
1 3.107455 8.633308
2 2.981028 6.266065
3 2.753103 4.404059
4 2.964051 6.243494
5 3.051922 5.531834
6 3.226972 7.581148
> # Plot the plot-specific estimates
> windows(4,7)
> par(mfcol=c(3,2))
> hist(korfpar$a)
> qqnorm(korfpar$a,main="a")
> hist(korfpar$b)
> qqnorm(korfpar$b,main="b")
> plot(korfpar)
```

Next, we will fit the mixed-effects model into the data.

```
> # This model can be fitted using function lme in package nlme (OLD)
> library(nlme)
> lmm1<-lme(log(h2)~I(1/d),random=~1+I(1/d)|plot,data=spati)
> lmm1
Linear mixed-effects model fit by REML
  Data: spati
  Log-restricted-likelihood: 639.0039
  Fixed: log(h2) ~ I(1/d)
(Intercept)      I(1/d)
   2.807888   -6.279249

Random effects:
 Formula: ~1 + I(1/d) | plot
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 0.4922689 (Intr)
I(1/d)      2.2729989 -0.875
Residual    0.1472408

Number of Observations: 1678
Number of Groups: 56
>
> # Explore object lmm1
> attributes(lmm1)
$names
 [1] "modelStruct"  "dims"         "contrasts"    "coefficients"
 [5] "varFix"       "sigma"        "apVar"        "logLik"
 [9] "numIter"      "groups"       "call"         "terms"
[13] "method"       "fitted"       "residuals"    "fixDF"
[17] "na.action"    "data"

$class
[1] "lme"
```

Compared to the residual variation, the between-stand variation is high, and the need for random effects is obvious. We could also make a formal test by fitting a restricted model and test if the full model is significantly better than the restricted model.

Next, study if the assumptions of the model are met. We assumed

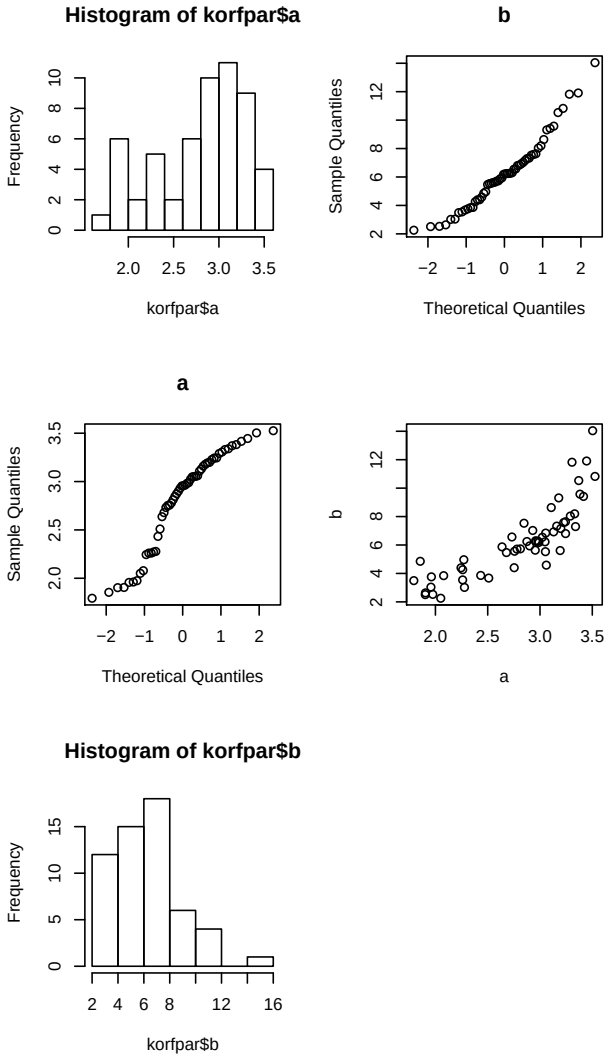


Figure 3.7: Plots of stand-specific parameter estimates.

1. The dependence of log(H-1.3) on D can be described using korf curve
2. Residuals have constant variance
3. Residuals are normally distributed
4. Random effects are multinormally distributed

The validity of these assumptions are studied in the following code. The resulting plots are shown in figure 3.8. The plots show

```
# Assumption 1: model form
windows(4,8)
par(mfcol=c(4,2))
plot(spati$d,resid(lmm1),col="red")
mywhiskers(spati$d,resid(lmm1),add=TRUE,lwd=2)
mywhiskers(spati$d,resid(lmm1),se=FALSE,add=TRUE)
abline(0,0)
# Model seems to fit fairly well.


# Residual variance has a decreasing trend:
plot(fitted(lmm1),resid(lmm1))
mywhiskers(fitted(lmm1),resid(lmm1),add=TRUE,se=FALSE)

# Normality of residuals
#qqnorm(resid(lmm2,type="pearson"))
#abline(0,1)
qqnorm(resid(lmm1,type="pearson"))
abline(0,1)
# The residuals have heavy tails!

# Normality of random effects
qqnorm(unlist(random.effects(lmm1)[1]))
qqnorm(unlist(random.effects(lmm1)[2]))
plot(unlist(random.effects(lmm1)[1]),unlist(random.effects(lmm1)[2]))
```

The following code makes a plot of the fixed part and plot-specific models. The plot is shown in Figure 3.9.

```
> windows()
> plot(spati$d,log(spati$h2),col="red")
> for (i in 1:56) {
+     thisplot<-spati[spati$plot==plots[i],]
+     d<-seq(min(thisplot$d),max(thisplot$d),length=20)
+     lines(d,coef(lmm1)[i,1]+coef(lmm1)[i,2]/d)
+     }
> d<-seq(min(spati$d),max(spati$d),length=20)
> lines(d,fixef(lmm1)[1]+fixef(lmm1)[2]/d,lwd=3,col="blue")
```

We fit a new model with heteroscedastic variances. The plot looks better (Figure 3.10), and the model fits significantly better.

```
> lmm1b<-lme(log(h2)~I(1/d),
+            random=~1+I(1/d)|plot,
+            data=spati,
+            weights=varPower(-0.5,~d))
>
>
> # Residual variance looks better
> plot(fitted(lmm1b),resid(lmm1b,type="pearson"))
> mywhiskers(fitted(lmm1b),resid(lmm1b,type="pearson"),add=TRUE,se=FALSE)
>
> # The model also fits better
> anova(lmm1,lmm1b)
      Model df       AIC       BIC   logLik   Test  L.Ratio p-value
lmm1      1  6 -1266.008 -1233.463 639.0039
lmm1b     2  7 -1619.514 -1581.545 816.7572 1 vs 2 355.5065  <.0001
```

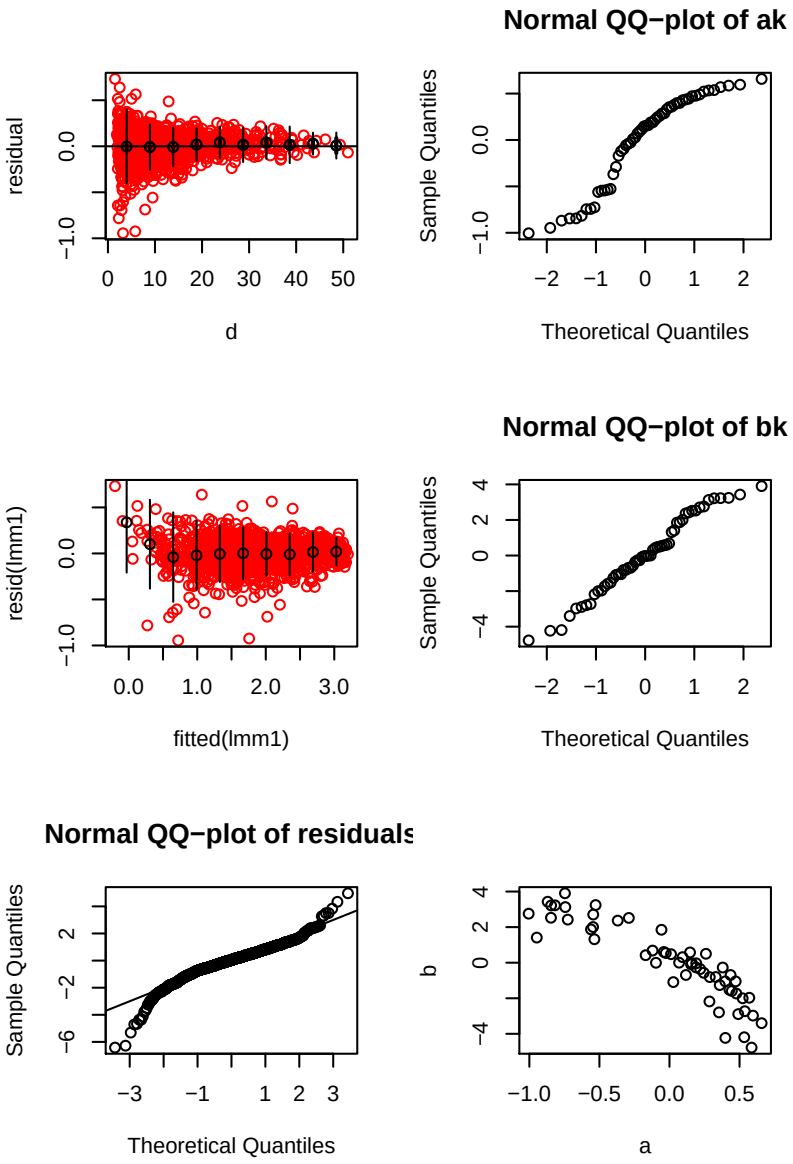


Figure 3.8: Diagnostoic plots on lmm1.

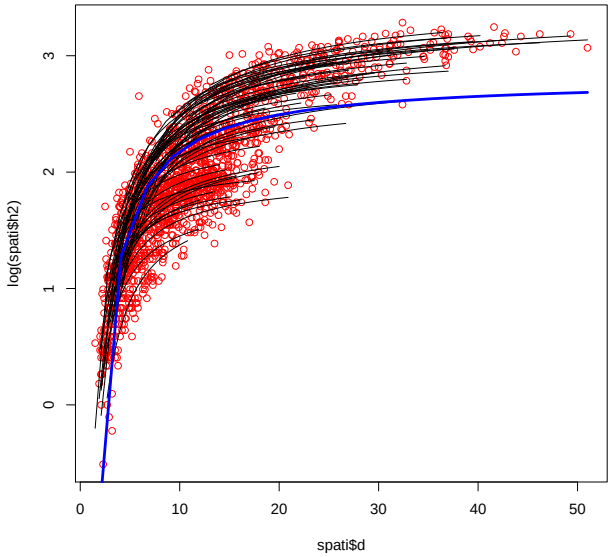


Figure 3.9: Data (red), marginal (blue) and conditional (black) curves.

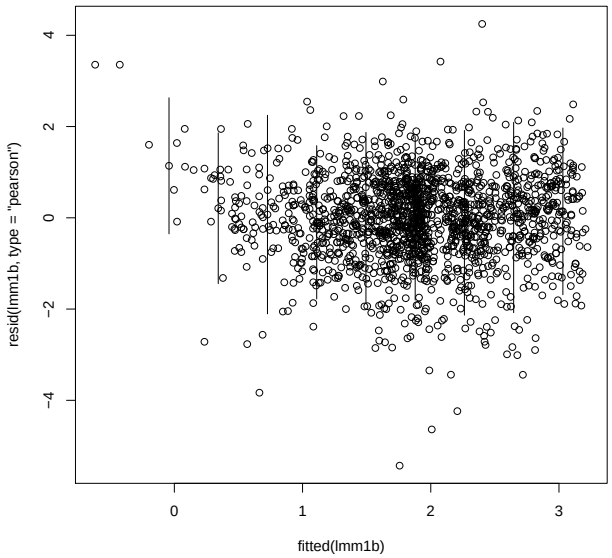


Figure 3.10: Residuals of model fmm1b.

Assume we want to model the dependence in stand specific values of a and by using some predictors. First take a look on how these parameters depend on stand characteristics. These plots are made using the following commands. The plots are shown in figures 3.11 and 3.12. We see that the random parameters clearly depend on stand characteristics. The dependence seems to be strongest for mean diameter Dg. For this covariate, the relationship between a is nonlinear, but the relationship with b looks fairly linear. A logarithmic transformation in Dg seems to linearize the relationship fairly well (Figure 3.13)

```
plotdata<-cbind(unique(spati[,c("plot","X","Y","N","G","Dg","Tg")]),korfpar)
windows(4,7)
par(mfcol=c(3,2))
plot(plotdata$X,plotdata$a)
plot(plotdata$Y,plotdata$a)
plot(plotdata$N,plotdata$a)
plot(plotdata$G,plotdata$a)
plot(plotdata$Dg,plotdata$a)
plot(plotdata$Tg,plotdata$a)

windows(4,7)
par(mfcol=c(3,2))
plot(plotdata$X,plotdata$b)
plot(plotdata$Y,plotdata$b)
plot(plotdata$N,plotdata$b)
plot(plotdata$G,plotdata$b)
plot(plotdata$Dg,plotdata$b)
plot(plotdata$Tg,plotdata$b)

> windows()
> plot(log(plotdata$Dg),plotdata$a)
> abline(lm(a~log(Dg),data=plotdata))
```

Thus, our new model would be

$$\ln(h_{ki}) = \alpha_0 + \alpha_1 * \ln(Dg) + a_k + (\beta_0 + \beta D_g + b_k) * (1/d_{ki}) + e_{ki}$$

This model is fitted using the following code. Note that we refit also the restricted model using ML, because we are going to make a LR-test on the fixed effects. The new relaxed model is significantly better than the old one.

```
> lmm4<-lme(log(h2)~I(1/d),
+           random=~1+I(1/d)|plot,
+           data=spati,
+           weights=varPower(-0.5,~d),
+           method="ML")
>
>
> lmm4b<-lme(log(h2)~I(1/d)+log(Dg)+I(1/d):Dg,
+            random=~1+I(1/d)|plot,
+            data=spati,
+            weights=varPower(-0.5,~d),
+            method="ML")
>
> # Make the LR test
> anova(lmm4,lmm4b)
      Model df       AIC       BIC   logLik   Test  L.Ratio p-value
lmm4      1  7 -1624.754 -1586.776 819.3768
lmm4b     2  9 -1728.881 -1680.053 873.4406 1 vs 2 108.1276  <.0001
```

The next question is, whether including additional predictors would improve the model. Let us add stand age as an additional predictor and do the test. The additional predictor provides no improvement to the model. The explanation can be seen by plotting the stand effects of model lmm4b against stand-specific predictors (Figures

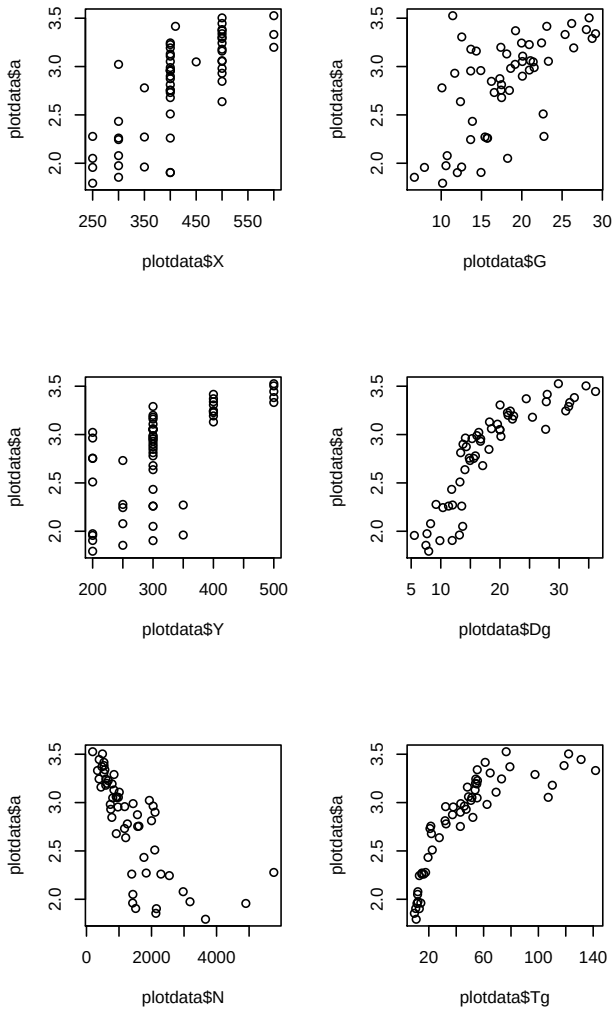


Figure 3.11: Parameter a against stand variables.

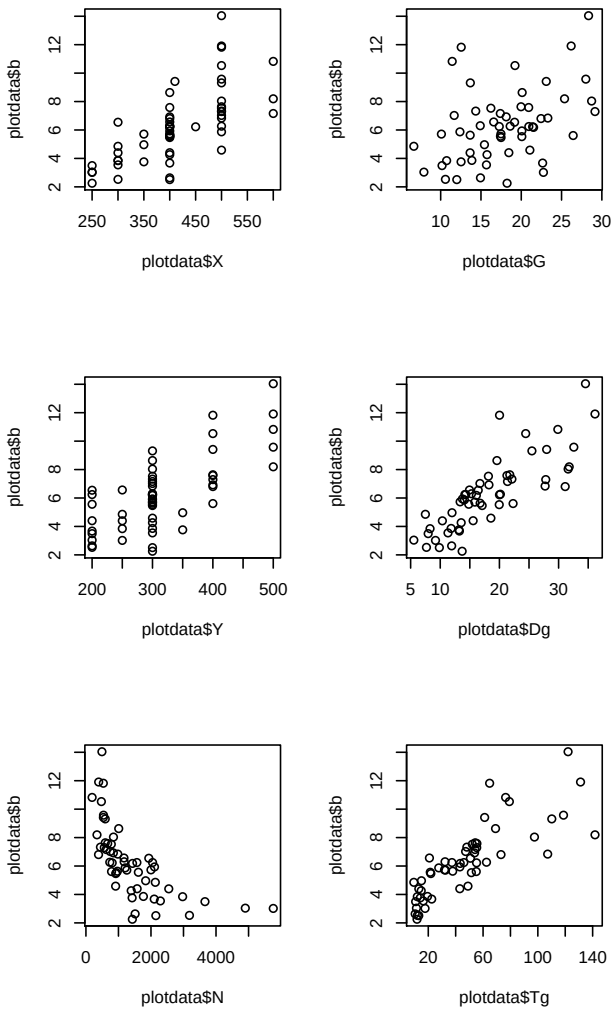


Figure 3.12: Parameter b against stand variables.

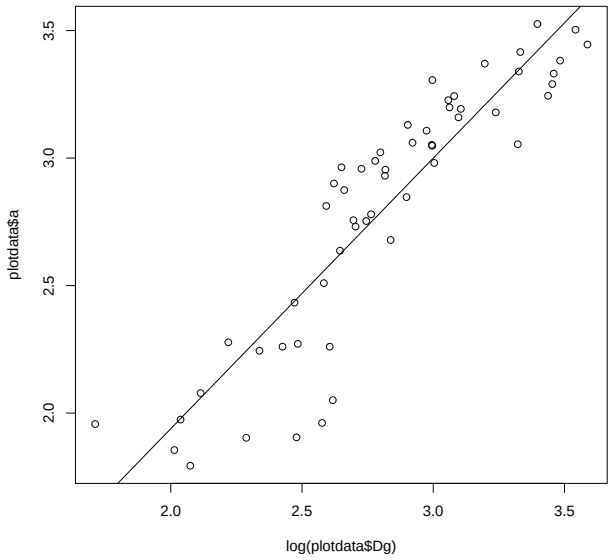


Figure 3.13: Dependence of $\log(b)$ on Dg.

3.14 and 3.15). There seems not to be any unexplained trend in them, and the trends that were seen in plots 3.11 and 3.12 were caused by correlations among stand variables.

```
> lmm4c<-lme(log(h2)~I(1/d)+log(Dg)+Tg+I(1/d):Dg,
+            random=~1+I(1/d)|plot,
+            data=spati,
+            weights=varPower(-0.5,~d),
+            method="ML")
>
> anova(lmm4b,lmm4c)
      Model df       AIC       BIC   logLik   Test     L.Ratio p-value
lmm4b     1  9 -1728.881 -1680.053 873.4406
lmm4c     2 10 -1726.890 -1672.637 873.4452 1 vs 2 0.009259103  0.9233
```

Thus, we refit the final model using REML and save the two final models into objects fm1 and fm2. Furthermore, we plot the fixed parts of the model (Figure 3.17), as well as the modeled dependence of a and b on Dg (Figure 3.16).

```
# Refit lmm4b with REML

lmm4b<-lme(log(h2)~I(1/d)+log(Dg)+I(1/d):Dg,
           random=~1+I(1/d)|plot,
           data=spati,
           weights=varPower(-0.5,~d),
           method="REML")

# Final models are lmm1b and lmm4b
fm1<-lmm1b
fm2<-lmm4b


# Plot the plot-specific estimates of a, and their estimates
windows(3,5)
par(mfcol=c(2,1))
plot(plotdata$Dg,plotdata$a)
```

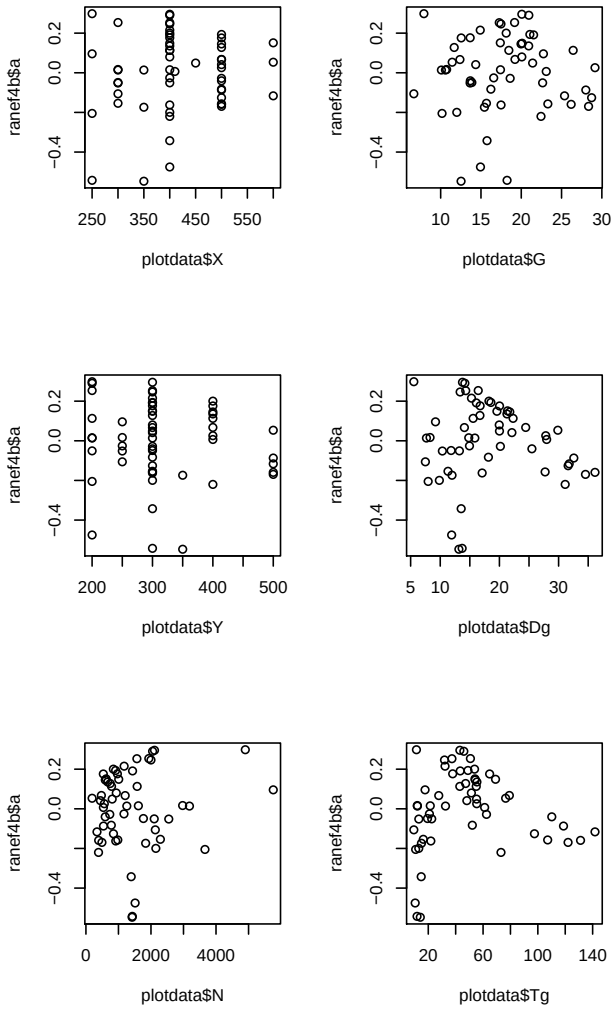


Figure 3.14: Stand effects a of lmm4b against stand variables.

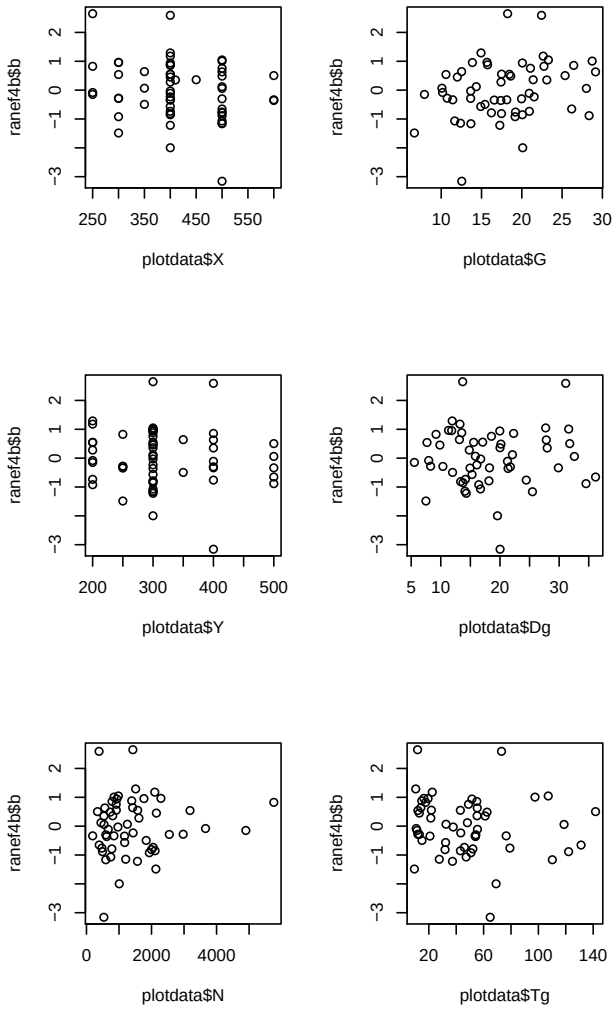
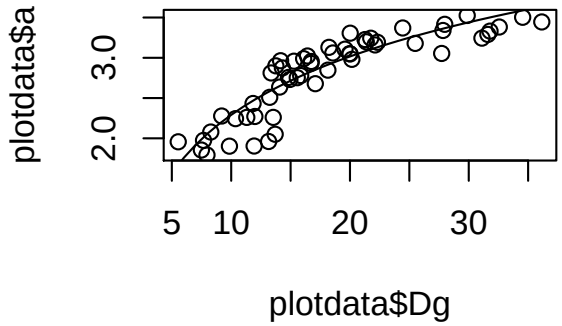


Figure 3.15: Stand effects b of lmm4b against stand variables.



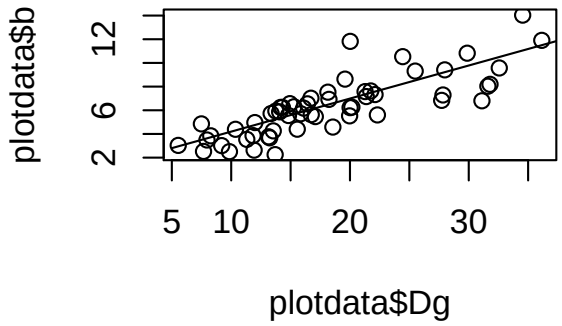


Figure 3.16: Modeled dependence of a and b on Dg.

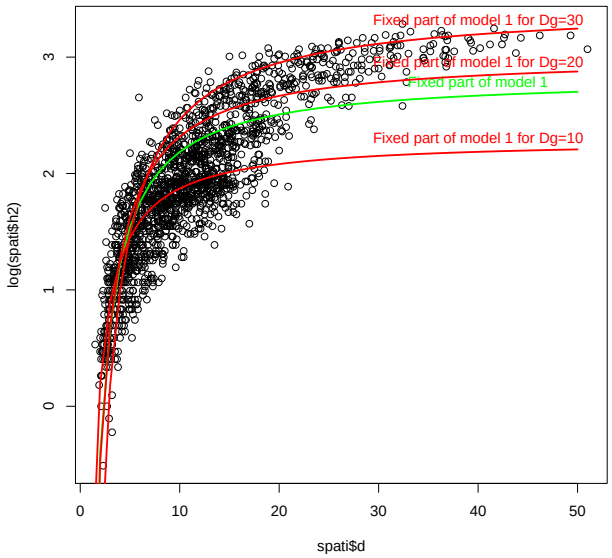


Figure 3.17: The fixed part of fm1 and that of fm2 with three different values of Dg

```
Dg<-seq(5,40)
lines(Dg,fixed.effects(fm2)[1]+fixed.effects(fm2)[3]*log(Dg))
plot(plotdata$Dg,plotdata$b)
lines(Dg,-fixed.effects(fm2)[2]-fixed.effects(fm2)[4]*Dg)
```

```
# Plot the data and the fixed part of the model
windows()
plot(spati$d,log(spati$h2))
d<-seq(0,50)
lines(d,cbind(1,1/d)%*%fixed.effects(fm1),
      type="l",col="green",lwd=2)
text(40,cbind(1,1/40)%*%fixed.effects(fm1),
     "Fixed part of model 1",pos=3,col="green")
Dg<-10
lines(d,cbind(1,1/d,log(Dg),Dg/d)%*%fixed.effects(fm2),
      type="l",col="red",lwd=2)
text(40,cbind(1,1/40,log(Dg),Dg/40)%*%fixed.effects(fm2),
     "Fixed part of model 1 for Dg=10",pos=3,col="red")
Dg<-20
lines(d,cbind(1,1/d,log(Dg),Dg/d)%*%fixed.effects(fm2),
      type="l",col="red",lwd=2)
text(40,cbind(1,1/40,log(Dg),Dg/40)%*%fixed.effects(fm2),
     "Fixed part of model 1 for Dg=20",pos=3,col="red")
Dg<-30
lines(d,cbind(1,1/d,log(Dg),Dg/d)%*%fixed.effects(fm2),
      type="l",col="red",lwd=2)
text(40,cbind(1,1/40,log(Dg),Dg/40)%*%fixed.effects(fm2),
     "Fixed part of model 1 for Dg=30",pos=3,col="red")
```

The same diagnostic plots were plotted for the model including Dg as a predictor (Figure 3.18) than for the model without Dg (Figure 3.8). The assumptions about normality seem to be better met when the nonlinear dependence of a on Dg has been modeled.

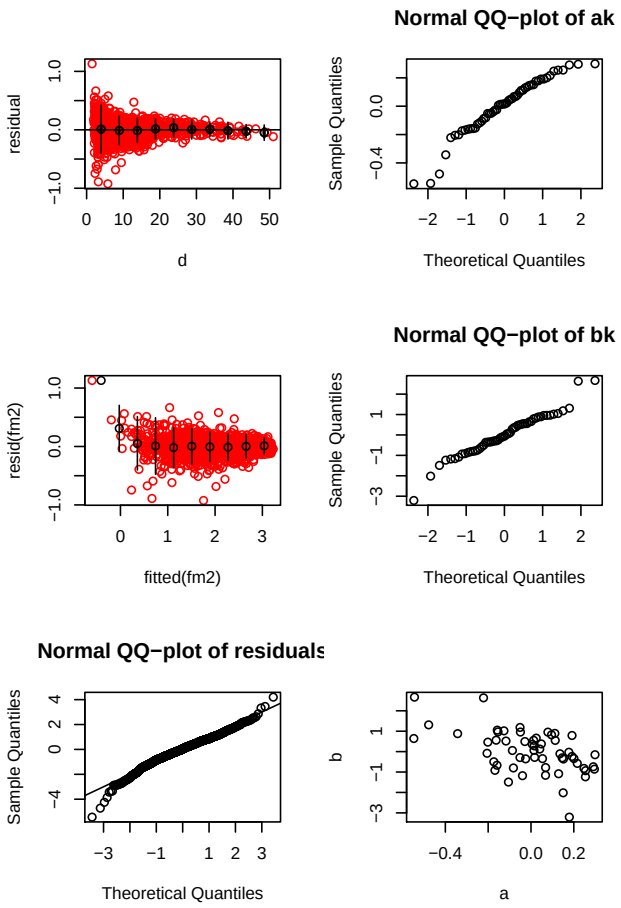


Figure 3.18: Diagnostic plots from model fm2.

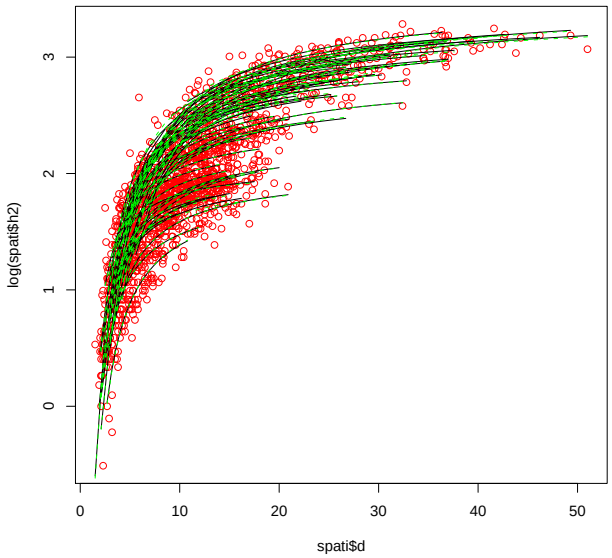


Figure 3.19: Conditional predictions from model fm2 (black) and fm2 (green).

As a final plot of the model, Figure 3.19 shows conditional predictions (i.e., those with stand effects predicted for the modeling data) for both models. The conditional models are very close to each other. Further analysis could be done on which of the models is better.

Figure 3.19 was generated using code

```
windows()
plot(spati$d,log(spati$h2),col="red")
for (i in 1:56) {
    thisplot<-spati[spati$plot==plots[i],]
    d<-seq(min(thisplot$d),max(thisplot$d),length=20)
    lines(d,cbind(1,1/d,log(thisplot$Dg[1]),thisplot$Dg[1]/d)%*%t(coef(fm2)[i,]))
    }

for (i in 1:56) {
    thisplot<-spati[spati$plot==plots[i],]
    d<-seq(min(thisplot$d),max(thisplot$d),length=20)
    lines(d,coef(fm1)[i,1]+coef(fm1)[i,2]/d,col="green",lty="dashed")
    }
```

3.6.3 Model application

A very useful application of the linear mixed-effects model is using it for prediction. In such case, we can use either marginal prediction, i.e., the fixed part only. However, if we have measurements of tree height and diameter available from even one sample tree, we can make conditional, stand specific prediction by predicting the stand effects of our model for the stand in hand. In the following code, we make predictions

- for fm1 using one measured sample tree

- for fm2 using one measured sample tree
- for fm1 using three measured sample trees
- for fm1 using three measured sample trees

The predictions of $\ln(h - 1.3)$ are a straightforward application of the best linear predictor. Also the conditional predictions plotted in Figure 3.19 utilize the same predictors, but they are carried out in R.

In mixed model, we estimated the fixed parameters and variance components. We are often interested also in the random effects. They are predicted using the BLP or BLUP. For the plots of the data, R has calculated them. They can be studied using random.effects() or ranef(). However, a useful situation is that we want to predict the random effects for a plot that was not included in the data. For example, we may have one tree height measured for one stand, and we want to utilize that information in predicting the HD-curve for that stand. First, we assume that from a stand with $Dg = 10$, one tree with $d = 12$ and $h = 10$ has been observed. We first specify the vectors and matrices for BLP, and then compute the BLP.

```
> hobs<-10
> dobs<-12
> # To predict random effects of model 1, we use BLP with
> # h1=c(a_k,b_k)
> # h2=the observed ln(height-1.3)=ln(8.7)= 2.163323
> h2<-log(hobs-1.3)
> # the expectations are
> # mu1=c(0,0)
> mu1<-c(0,0)
> # mu2=the fixed part of the model for d=12
> mu2<-c(1,1/dobs)%*%fixed.effects(fm1)
> # V1 = the variance-covariance matrix of random effects
> V1<-D<-getVarCov(fm1)
> # V2 = the variance-covariance matrix of heights, i.e., ZDZ'+R,
> # where D = V1
> Z<-cbind(1,1/dobs)
> rho<-attributes(fm1$apVar)$Pars[4]
> sigma<-fm1$sigma
> R<-sigma^2*dobs^(2*rho)
> V2<-Z%*%V1%*%t(Z)+R
> V12<-V1%*%t(Z)
> h2
[1] 2.163323
> mu1
[1] 0 0
> mu2
         [,1]
[1,] 2.292262
> V1
Random effects variance covariance matrix
            (Intercept)   I(1/d)
(Intercept)     0.23954 -0.98661
I(1/d)         -0.98661  5.49210
  Standard Deviations: 0.48942 2.3435
> V2
          [,1]
[1,] 0.1268992
> V12
                 [,1]
(Intercept)  0.1573194
I(1/d)      -0.5289329
```

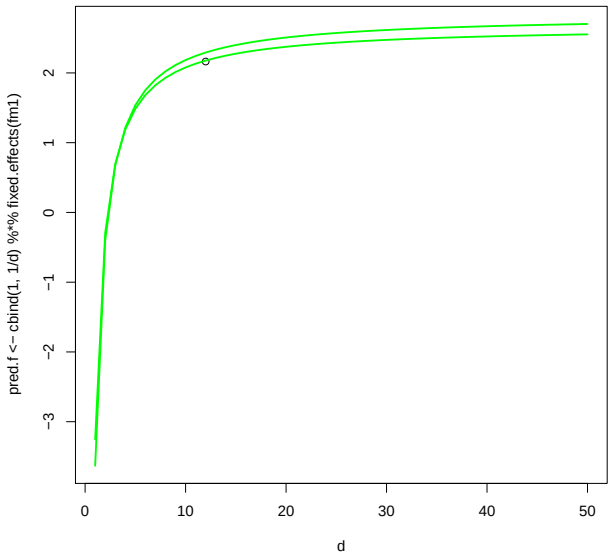


Figure 3.20: Local and population-level curve for the example stand in logarithmic scale.

```
>
> # The blup of random effects is
> h1<-bhat<-V12%*%solve(V2)%*%(h2-mu2)
> h1
                  [,1]
(Intercept) -0.1598475
I(1/d)       0.5374327
> fixef(fm1)+h1
                 [,1]
(Intercept)  2.670914
I(1/d)      -5.924571
```

Thus, the intercept (a) for this particular stand is $\alpha + 0.157 = 2.97$ and the shape parameter is $\beta + 0.529 = 5.925$. The localized, (or calibrated) conditional curve for this plot is shown in Figure 3.20.

```
# plot the localized or calibrated H-D curve.
# Save the predictions to vectors
# pred.f  - logarithmic prediction based on the fixed part only
# pred.fr - logarithmic prediction based on fixed and random parts
d<-seq(1,50)
plot(d,pred.f<-cbind(1,1/d)%*%fixed.effects(fm1),type="l",col="green",lwd=2)
points(dobs,h2)
lines(d,pred.fr<-cbind(1,1/d)%*%(fixed.effects(fm1)+bhat),type="l",col="green",lwd=2)

# The local curve almost passes through the observed height. Why?
> cbind(1,1/12)%*%(fixed.effects(fm1)+bhat)
       [,1]
[1,] 2.1772
> h2
[1] 2.163323
```

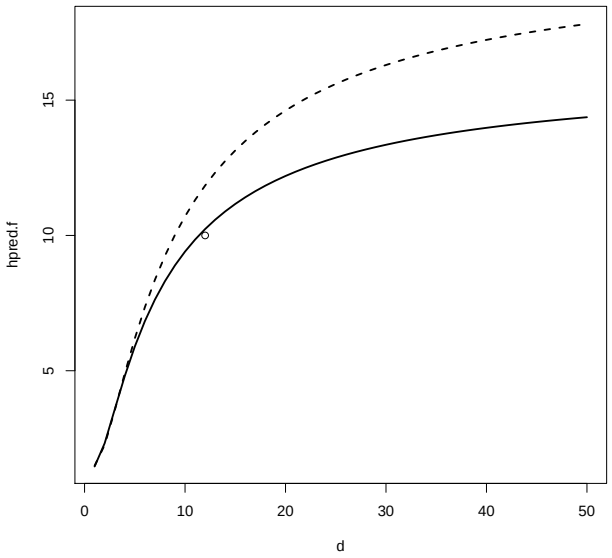


Figure 3.21: Local (solid) and population-level (dashed) curve for the example stand in original scale.

Next, we want to make predictions in the original scale. A bias correction based on normality of logarithmic residuals is applied by adding half of the prediction error variance into the predictions before applying the exponential transformation. The following code makes the computations and produces the plot shown in figure 3.21.

```
# Transformation to the original scale.
# Half of the prediction variance needs to be added before back-transformation.
# 1. The prediction based on fixed and random parts
# The prediction errors of random parameters are
> # Transformation to the original scale.
> # Half of the prediction variance needs to be added before back-transformation.
> # 1. The prediction based on fixed and random parts
> # The prediction errors of random parameters are
> varb<-V1-V12%*%solve(V2)%*%t(V12)
> varb
Random effects variance covariance matrix
             (Intercept)   I(1/d)
(Intercept)    0.044505 -0.33088
I(1/d)        -0.330880  3.28740
  Standard Deviations: 0.21096 1.8131
> V1
Random effects variance covariance matrix
             (Intercept)   I(1/d)
(Intercept)     0.23954 -0.98661
I(1/d)         -0.98661  5.49210
  Standard Deviations: 0.48942 2.3435
> Z0<-cbind(1,1/d)
> R0<-sigma^2*d^(2*rho)
> var.lnh2<-Z0%*%varb%*%t(Z0)+diag(R0)
> hpred.fr<-1.3+exp(pred.fr+0.5*diag(var.lnh2))
>
> # 2. The prediction based on fixed part only
> var.lnh2<-Z0%*%D%*%t(Z0)+diag(R0)
> hpred.f<-1.3+exp(pred.f+0.5*diag(var.lnh2))
```

```
>
> plot(d,hpred.f,type="l",lwd=2,lty="dashed")
> lines(d,hpred.fr,lwd=2)
> points(dobs,hobs)
```

The same computations were made using the model with Dg as a predictor. The resulting figure in original scale is shown in Figure 3.22. Even though there is a huge difference in the fixed parts of the model, the one sample tree moves the local curves fairly close to each other.

```
> # Let us make the corresponding predictions using fm2
> Dg<-10
> fixef(fm2)
(Intercept)      I(1/d)     log(Dg)   I(1/d):Dg
 -0.1187020  -1.4058238   1.0464768  -0.2791521
> mu2<-c(1,1/dobs,log(Dg),Dg/dobs)%*%fixed.effects(fm2)
>
> # V1 = the variance-covariance matrix of random effects
> V1<-D<-getVarCov(fm2)
> # V2 = the variance-covariance matrix of heights, i.e., ZDZ'+R,
> # where D = V1
> Z<-cbind(1,1/dobs)
> rho<-attributes(fm2$apVar)$Pars[4]
> sigma<-fm2$sigma
> R<-sigma^2*dobs^(2*rho)
> V2<-Z%*%V1%*%t(Z)+R
> V12<-V1%*%t(Z)
>
> # The blup of random effects is
> bhat<-V12%*%solve(V2)%*%(h2-mu2)
>
> # plot the localized or calibrated H-D curve.
> # Save the predictions to vectors
> # pred.f  - logarithmic prediction based on the fixed part only
> # pred.fr - logarithmic prediction based on fixed and random parts
> d<-seq(1,50)
> plot(d,pred.f2<-cbind(1,1/d,log(Dg),Dg/d)%*%fixed.effects(fm2),type="l",col="green",lwd=2)
> points(dobs,h2)
> lines(d,pred.fr2<-cbind(1,1/d,log(Dg),Dg/d)%*%(fixed.effects(fm2)+c(bhat,0,0)),type="l",col="green",lwd=
>
> # The local curve does not pass that close to the observed height. Why?
> cbind(1,1/12,log(Dg),Dg/12)%*%(fixed.effects(fm2)+c(bhat,0,0))
         [,1]
[1,] 2.093866
> h2
[1] 2.163323
>
> # Transformation to the original scale.
> # Half of the prediction variance needs to be added before back-transformation.
> # 1. The prediction based on fixed and random parts
> # The prediction errors of random parameters are
> varb<-V1-V12%*%solve(V2)%*%t(V12)
> varb
Random effects variance covariance matrix
            (Intercept)   I(1/d)
(Intercept)    0.020079 -0.12404
I(1/d)        -0.124040  1.44110
  Standard Deviations: 0.1417 1.2005
> V1
Random effects variance covariance matrix
            (Intercept)   I(1/d)
(Intercept)    0.042251 -0.13301
I(1/d)        -0.133010  1.44470
  Standard Deviations: 0.20555 1.202
> Z0<-cbind(1,1/d)
> R0<-sigma^2*d^(2*rho)
> var.lnh2<-Z0%*%varb%*%t(Z0)+diag(R0)
> hpred.fr2<-1.3+exp(pred.fr2+0.5*diag(var.lnh2))
>
> # 2. The prediction based on fixed part only
> var.lnh2<-Z0%*%D%*%t(Z0)+diag(R0)
> hpred.f2<-1.3+exp(pred.f2+0.5*diag(var.lnh2))
>
```

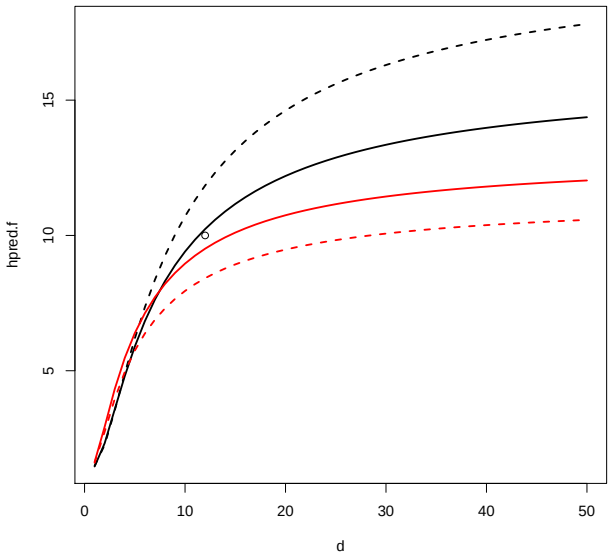


Figure 3.22: Local (solid) and population-level (dashed) curve for the example stand in original scale for fm1 (black) and fm2 (red).

```
> plot(d,hpred.f,type="l",lwd=2,lty="dashed")
> lines(d,hpred.fr,lwd=2)
> lines(d,hpred.f2,type="l",lwd=2,lty="dashed",col="red")
> lines(d,hpred.fr2,lwd=2,col="red")
> points(dobs,hobs)
```

Finally, we repeat the prediction using three sample trees. The commands are shown below. The plot of local curves from both models is shown in figure 3.23

```
> # EXAMPLE 2 Three trees
> # Assume that from a stand with Dg=10, three trees with d=c(5,9,12), and  and h=10 have been observed.
> hobs<-c(6,8,10)
> dobs<-c(5,9,12)
> # To predict random effects of model 1, we use BLP with
> # h1=c(a_k,b_k)
> # h2=the observed ln(height-1.3)=ln(8.7)= 2.163323
> h2<-log(hobs-1.3)
> # the expectations are
> # mu1=c(0,0)
> mu1<-c(0,0)
> # mu2=the fixed part of the model for d=12
> mu2<-cbind(1,1/dobs)%*%fixed.effects(fm1)
> # V1 = the variance-covariance matrix of random effects
> V1<-D<-getVarCov(fm1)
> # V2 = the variance-covariance matrix of heights, i.e., ZDZ'+R,
> # where D = V1
> Z<-cbind(1,1/dobs)
> rho<-attributes(fm1$apVar)$Pars[4]
> sigma<-fm1$sigma
> R<-diag(sigma^2*dobs^(2*rho))
> V2<-Z%*%V1%*%t(Z)+R
> V12<-V1%*%t(Z)
>
> # The blup of random effects is
> bhat<-V12%*%solve(V2)%*%(h2-mu2)
```

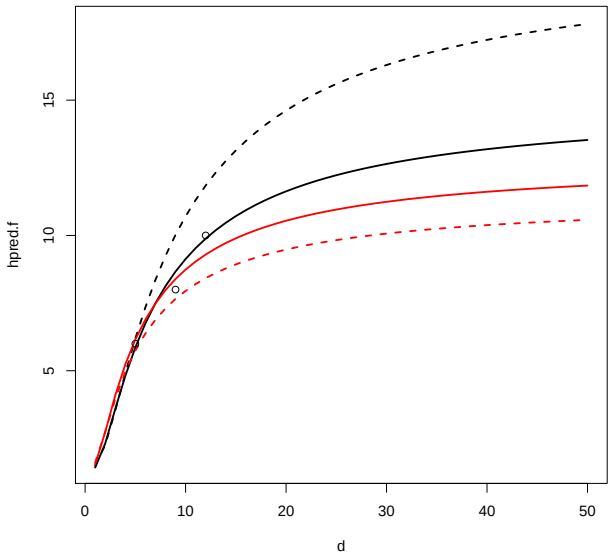


Figure 3.23: Local (solid) and population-level (dashed) curve for the example stand in original scale for fm1 (black) and fm2 (red).

```
>
> # plot the localized or calibrated H-D curve.
> # Save the prediictions to vectors
> # pred.f  - logarithmic prediction based on the fixed part only
> # pred.fr - logarithmic prediction based on fixed and random parts
> d<-seq(1,50)
> plot(d,pred.f<-cbind(1,1/d)%*%fixed.effects(fm1),type="l",col="green",lwd=2)
> points(dobs,h2)
> lines(d,pred.fr<-cbind(1,1/d)%*%(fixed.effects(fm1)+bhat),type="l",col="green",lwd=2)
>
>
> # Transformation to the original scale.
> # Half of the prediction variance needs to be added before back-transformation.
> # 1. The prediction based on fixed and random parts
> # The prediction errors of random parameters are
> varb<-V1-V12%*%solve(V2)%*%t(V12)
> varb
Random effects variance covariance matrix
            (Intercept)   I(1/d)
(Intercept)    0.033286 -0.22755
I(1/d)        -0.227550  1.88180
  Standard Deviations: 0.18244 1.3718
> V1
Random effects variance covariance matrix
            (Intercept)   I(1/d)
(Intercept)     0.23954 -0.98661
I(1/d)         -0.98661  5.49210
  Standard Deviations: 0.48942 2.3435
> Z0<-cbind(1,1/d)
> R0<-sigma^2*d^(2*rho)
> var.lnh2<-Z0%*%varb%*%t(Z0)+diag(R0)
> hpred.fr<-1.3+exp(pred.fr+0.5*diag(var.lnh2))
>
> # 2. The prediction based on fixed part only
> var.lnh2<-Z0%*%D%*%t(Z0)+diag(R0)
> hpred.f<-1.3+exp(pred.f+0.5*diag(var.lnh2))
>
```

```
> plot(d,hpred.f,type="l",lwd=2,lty="dashed")
> lines(d,hpred.fr,lwd=2)
> points(dobs,hobs)
>
> # Let us make the corresponding predictions using fm2
> Dg<-10
> mu2<-cbind(1,1/dobs,log(Dg),Dg/dobs)%*%fixed.effects(fm2)
>
> # V1 = the variance-covariance matrix of random effects
> V1<-D<-getVarCov(fm2)
> # V2 = the variance-covariance matrix of heights, i.e., ZDZ'+R,
> # where D = V1
> Z<-cbind(1,1/dobs)
> rho<-attributes(fm2$apVar)$Pars[4]
> sigma<-fm2$sigma
> R<-diag(sigma^2*dobs^(2*rho))
> V2<-Z%*%V1%*%t(Z)+R
> V12<-V1%*%t(Z)
>
> # The blup of random effects is
> bhat<-V12%*%solve(V2)%*%(h2-mu2)
>
> # plot the localized or calibrated H-D curve.
> # Save the prediictions to vectors
> # pred.f  - logarithmic prediction based on the fixed part only
> # pred.fr - logarithmic prediction based on fixed and random parts
> d<-seq(1,50)
> plot(d,pred.f2<-cbind(1,1/d,log(Dg),Dg/d)%*%fixed.effects(fm2),type="l",col="green",lwd=2)
> points(dobs,h2)
> lines(d,pred.fr2<-cbind(1,1/d,log(Dg),Dg/d)%*%(fixed.effects(fm2)+c(bhat,0,0)),type="l",col="green",lwd=2)
>
> # Transformation to the original scale.
> # Half of the prediction variance needs to be added before back-transformation.
> # 1. The prediction based on fixed and random parts
> # The prediction errors of random parameters are
> varb<-V1-V12%*%solve(V2)%*%t(V12)
> varb
Random effects variance covariance matrix
             (Intercept)   I(1/d)
(Intercept)    0.017626 -0.11147
I(1/d)        -0.111470  1.00390
  Standard Deviations: 0.13276 1.0019
> V1
Random effects variance covariance matrix
             (Intercept)   I(1/d)
(Intercept)    0.042251 -0.13301
I(1/d)        -0.133010  1.44470
  Standard Deviations: 0.20555 1.202
> Z0<-cbind(1,1/d)
> R0<-sigma^2*d^(2*rho)
> var.lnh2<-Z0%*%varb%*%t(Z0)+diag(R0)
> hpred.fr2<-1.3+exp(pred.fr2+0.5*diag(var.lnh2))
>
> # 2. The prediction based on fixed part only
> var.lnh2<-Z0%*%D%*%t(Z0)+diag(R0)
> hpred.f2<-1.3+exp(pred.f2+0.5*diag(var.lnh2))
>
> plot(d,hpred.f,type="l",lwd=2,lty="dashed")
> lines(d,hpred.fr,lwd=2)
> lines(d,hpred.f2,type="l",lwd=2,lty="dashed",col="red")
> lines(d,hpred.fr2,lwd=2,col="red")
> points(dobs,hobs)
```

3.7 Exercises

1. Show that in the linear mixed-effects model $\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{b} + \boldsymbol{Z}_k\boldsymbol{c}_k + \boldsymbol{e}_k$, $\mathrm{var}(\boldsymbol{y}_k) = Z_k\boldsymbol{D}\boldsymbol{Z}_k' + \boldsymbol{R}$ and $\mathrm{cov}(\boldsymbol{c}_k, \boldsymbol{y}_k') = \boldsymbol{D}\boldsymbol{Z}_k'$.

2. Perform similar analysis as we did in the last section of this chapter using spruce

dataset from file `spruce.txt`. The data structure is similar to the pine data used in the analysis.

(a) Analyse possible model shapes.

(b) Develop and fit appropriate linear mixed-effects model. Report different stages of the modeling work and justify the choices you have made.

(c) File `sprucestand.txt` includes 60 measured height diameter pairs from a stand that was not included in the modeling data. Select randomly (i) one sample tree and (ii) five sample trees from the example stand, and compute the stand effects of your model. Plot all the data of the example stand. Add to the plot (i) the prediction of the fixed part (ii) conditional (localized, calibrated) prediction based on the one sample tree, (iii) the conditional (localized, calibrated) prediction based on the five sample trees, and (iv) mark the sample trees utilized in prediction.

3. File `hdmod.R` includes functions that can be used for easy application of the longitudinal height models of (Mehtätalo 2004) and (Mehtätalo 2005).

   (a) Run the code using the measurements given in the example at the end of the file.

   (b) Use the model for prediction for a pine stand stand with DGM=30 cm, and two sample trees measured 10 years ago. At that time, DGM was 22 cm, and the measured diameters were 15 and 22 cm, and the corresponding heights were 14 and 19 meters.

   (c) Plot the fixed part predictions, and localized H-D curves 10 years before and now.

4. Using plots 1-16 of data `pinevol`, analyze the dependence of tree volume on diameter. Use the fitted model for prediction for plot 51. Some help for this data can be found in Lappi et al. (2006).

Chapter 4

Generalized linear models

The linear mixed-effects model assume normality of the response. With the linear (fixed-effects) model, normality is needed for the test. However, the response may also have some other distribution. For example, it may be binary: dead or alive, success or failure. Or it my be count: e.g., number of dead trees within a plot. Furthermore, the diameter of a tree within a stand or plot may be assumed to have a Weibull distribution. Even in those cases, the linear model may lead to fairly good model. However, a better justified model is obtained through generalized linear model (GLM). Note the distinction between terms GLS and GLM. They have nothing in common, except for the two first letters.

The generalized linear models are closely related to fitting assumed distributions to data using the method of maximum likelihood. Generalized linear model actually means just fitting the assumed distribution to the data using the maximum likelihood. The only difference to the examples of section 1.5.1 is that in GLM, the parameter(s) of the assumed distribution, or some functions of them, are written as a linear function of predictors, instead of assuming a common value for the whole data. To ensure that the predicted parameter is always within the support (e.g, probability is between 0 and 12, or the expected number of stems is positive), a link function is used to link a linear function of predictors to the parameter to be modeled. This link has exactly the same idea as we had in example 1.30, where the parameters of a Weibull-distribution were restricted to be above 0 by using a log link.

4.1 Formulation of the LM as a GLM

With the linear model, the usual way of thinking is that there is a regression line which is usually of interest to the modeler. In addition, there is a normally distributed error term, which is added to the regression line to get the realized observations. This way of

thinking is used in development of the least squares estimation methods, thus we call it the LS-formulation of the linear model. It can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \boldsymbol{V})$.

Another way of thinking is the formulation we used in section 2.6 to derive the ML-estimators for the model parameters. This formulation, called ML-formulation, is specified as

$$\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \boldsymbol{V})$$

Thus, we do not explicitly write the error term, but the assumed linear relationship between response and the predictors has been written to the mean of the assumed normal distribution.

These two formulations, LS and ML-formulation, are equivalent, but arise from two different ways of thinking. The lower alternative represents such a way of thinking that is used in generalized linear models. The most essential difference is that **the generalized linear model does not have an explicitly written error term**. Furthermore, the LS-way of thinking is not necessarily possible for GLM with other distributions than the normal. Thus, the GLM's are based on thinking the model through the the ML-formulation.

The ML estimation of the linear model, which was presented in section 2.6, is the generalized linear model formulation for normally distributed response. The generalized linear model can be developed also for other types of response in a similar manner. We just write the parameters (or functions of them) as a linear function of predictors, and fit the assumed distribution into the data using the method of maximum likelihood.

With distributions belonging to the exponential family, certain good properties hold for the estimators, and statistical packages generally do not provide tools for estimation of GLM for other distributions than those belonging the exponential family.

4.2 GLM for exponential family

4.2.1 Model formulation

The generalized linear model for a random variable that follows a distribution of exponential family is (McCulloch and Searle 2001)

$$\begin{aligned} y_i &\sim \text{indep.} f_{Y_i|\gamma_i}(y_i) \\ f_{Y_i|\gamma_i} &= \exp\left(\frac{y_i\gamma_i - b(\gamma_i)}{\tau^2} - c(y_i, \tau)\right) \end{aligned} \tag{4.1}$$

The first expression states the family of the distribution. It is assumed that the data, y_i, $i = 1, \ldots, n$, constitute an independent, random sample form a distribution with density f_{Y_i}. The second expression states a general formulation to that density, i.e., the so called canonical formulation of the exponential family, where γ_i and τ are the parameters of the distribution. Especially, τ is related to the variance of the distribution and γ_i to the mean. Functions b and c are assumed to be known. Parameter γ is called the canonical parameter and is related to the mean through

$$\mu_i = \mathrm{E}(Y_i) = b'(\gamma_i)$$

and the variance is

$$\begin{aligned} \mathrm{Var}(Y_i) &= \tau^2 b''(\gamma_i) \\ &= \tau^2 b''(b'^{-1}(\mu_i)) \\ &\equiv \tau^2 v(\mu_i) \end{aligned}$$

In addition to the above assumptions, we assume that a function of the mean is linear in predictors,

$$g(\mu_i) = \boldsymbol{x'b}$$

which is equivalent to

$$g(-b'(\gamma_i)) = \boldsymbol{x'b}\,.$$

Function g is the applied link function, which is usually selected so that it retains the estimated parameters within its support.

For example, if the expected value of the random variable, $\mu = -b'(\gamma_i)$ has to be between 0 and 1, the link $g(\mu)$ should be a monotone function between $\mu \in [0, 1]$ that gets all possible values between $]-\infty, \infty[$. Such a function is obtained as the inverse cdf (i.e., quantile function) of any continuous distribution. Another typical example restricts the parameter $\mu = -b'(gamma_i)$ to be positive. An appropriate link for this case is a monotone function that gets values $]-\infty, \infty[$ for $\mu \in [0, \infty]$. The idea of the link is, that function $\boldsymbol{x'b}$ may give any values within $]-\infty, \infty[$. Thus, by applying the inverse link $\mu = g^{-1}(\boldsymbol{x'b})$, the fitted value of μ is always within its allowed range.

In addition, for each density function f, the so called canonical link ensures that $\boldsymbol{X'y}$ is a sufficient statistic for γ_i. The canonical links for some distributions are given in table 4.1.

4.2.2 Estimation

The estimation of GLM is based on the maximum likelihood principle. The log likelihood of a random variable that is distributed according to the exponential family is

Table 4.1: Canonical link functions for the most common distributions of the exponential family (adopted from Wikipedia 28.4.2008)

Distribution	Name	Link Function	Mean Function
Normal	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential	Inverse	$\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma	Inverse	$\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Binomial	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}$
Multinomial	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}$

obtained from (4.1) as

$$l = \frac{1}{\tau^2}\sum_{i=1}^{n}[y_i\gamma_i - b(\gamma_i)] - \sum_{1=1}^{n} c(y_i, \tau)$$

It can be shown (see McCulloch and Searle (2001) for details) that

$$\frac{\partial l}{\partial \boldsymbol{b}} = \frac{1}{\tau^2}\sum (y_i - \mu_i) w_i g(\mu_i) \boldsymbol{x}_i'$$

where $w_i = [v(\mu_i)g^2(\mu_i)]^{-1}$, and v is the variance function, which was defined earlier.

By defining $\boldsymbol{W}$ and $\boldsymbol{\Delta}$ as

$$\boldsymbol{W}_{n\times n} = \begin{bmatrix} w_1 & \dots & 0 \\ \dots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix} \quad \boldsymbol{\Delta}_{n\times n} = \begin{bmatrix} g(\mu_1) & \dots & 0 \\ \dots & \ddots & \vdots \\ 0 & \cdots & g(\mu_n) \end{bmatrix}$$

we can write the likelihood equation in the matrix form as

$$\frac{\partial l}{\partial \boldsymbol{b}} = \frac{1}{\tau^2}\boldsymbol{X'W\Delta(y-\mu)}.$$

Equating this to 0 gives the ML equations as

$$\boldsymbol{X'W\Delta y = X'W\Delta\mu}.$$

In the equations, both $\boldsymbol{W}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\mu}$ are functions of the parameter of interest, $\boldsymbol{\beta}$. These functions are usually nonlinear, and cannot be solved analytically. Thus, iterative numerical methods, such as Fisher scoring and iterative weighted least squares methods, are usually needed for solving these equations and estimating the parameters.

There are also other methods of estimation for the GLM, such as the maximum quasi-likelihood approach (MQL, suurin näennäis uskottavuus). As the name expresses, it maximizes something that is similar to the likelihood, but is not the likelihood exactly. This method has been developed for such situations, where no other assumptions about the distribution of the data are, except how the variance depends on the mean in

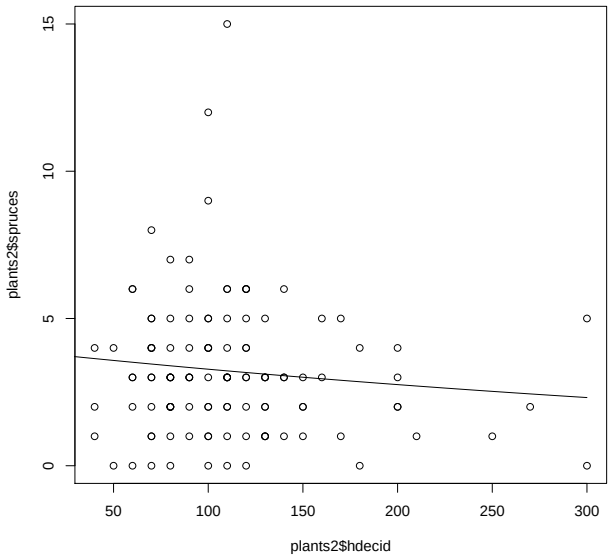


Figure 4.1: Plot of the count data and the fitted regression line.

the data. Thus, this method is somehow looking for a counterpart of the LS approach in the case of GLMS. For some distributions, such as Poisson the ML leads to equivalent results to those of MQL with corresponding assumptions on mean-variance relationship. The MQL also has some advantages in the case of overdispersion (McCulloch and Searle 2001).

Example 4.1 *The number of spruce saplings has been recorded from 123 sample plots. Each of the plots has been taken from different regeneration area. For count data, the Poisson regression would be appropriate methodology. We want to explain the number of spruce saplings per plot using mean height of deciduous trees. We first plot the data to see that the number of spruces decreases as the height of deciduous trees increases (Figure 4.1).*

```
> plants2<-read.table("c:/laurim/biometria/plants2.txt",header=TRUE)
> plants2<-plants2[!is.na(plants2$hdecid),] # remove observations with unknown hdecid
> # plants2$spruces will include both planted and natural spruces
> plants2$spruces<-plants2$planted+plants2$spruces
> plot(plants2$hdecid,plants2$spruces)
```

The model is fitted using

```
> glm1 <- glm(spruces ~ hdecid, family=poisson(), data=plants2)
> summary(glm1)

> summary(glm1)

> summary(glm1)
```

```
Call:
glm(formula = spruces ~ hdecid, family = poisson(), data = plants2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6740  -0.8421  -0.1711   0.5929   4.7531

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.360962   0.135115  10.073   <2e-16 ***
hdecid      -0.001739   0.001149  -1.513    0.130
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 189.85  on 113  degrees of freedom
Residual deviance: 187.46  on 112  degrees of freedom
AIC: 503.69

Number of Fisher Scoring iterations: 5
```

We add the fitted line into the plot of our data

```
> x<-seq(0,300,1)
> lines(x,exp(cbind(1,x)%*%coef(glm1)))
```

Example 4.2 *To show that GLM is just an application of fitting a distribution function to data using the method of maximum likelihood, we fit the model of previous example using mle package. For this purpose, we first define a function that computes the negative log likelihood as a function of our coefficients. Then we use function* `mle` *to find such values for the coefficients that maximize the log likelihood.*

```
> library(stats4)
> nll<-function(b0,b1) {
+      cat(b0," ",b1," ")
+      value<--sum(log(dpois(plants2$spruces,exp(b0+b1*plants2$hdecid))),na.rm=TRUE)
+      cat(value,"\n")
+      value
+      }
>
> solution<-mle(minuslogl=nll,start=list(b0=0.5,b1=0))
0.5   0  317.5801
0.501   0  317.4012
0.499   0  317.7593
0.5   0.001  300.3687
0.5   -0.001  337.6111
179.5457   18621.23  Inf
...
1.380250   -0.002904336  252.3297
1.381250   -0.001904336  249.8583
1.381250   -0.003904336  259.0869
>
> # Estimates, their standard errors, and AIC
> coef(solution)
          b0           b1
 1.381249811 -0.001904336
> sqrt(diag(vcov(solution)))
         b0          b1
0.134270822 0.001140901
> AIC(solution)
[1] 503.7166
> logLik(solution)
'log Lik.' -249.8583 (df=2)
>
> # The same figures from glm-fit
> summary(glm1)$coefficients
                Estimate  Std. Error   z value     Pr(>|z|)
(Intercept)  1.360962060 0.135115156 10.072608 7.301548e-24
hdecid      -0.001738545 0.001149177 -1.512861 1.303151e-01
> AIC(glm1)
[1] 503.6938
```

```
> logLik(glm1)
'log Lik.' -249.8469 (df=2)
```

Our own code results in slightly different estimates. To check if the differences result from different likelihoods, we compute the likelihood at the solution of `glm` *using our own function* `nll`.

```
> nll(coef(glm1)[1],coef(glm1)[2])
1.360962   -0.001738545  249.8469
[1] 249.8469
```

The resulting value is exactly same as the negative log-likelihood of the GLM model. Thus, the differences result from that glm uses a better algorithm for maximization than our general-purpose function mle, which just calls R function optim.

4.2.3 Inference and tests

We remember from section 1.5.1, that the asymptotic variance-covariance -matrix of a ML-estimate can be obtained as

$$\mathrm{var}(\hat{\boldsymbol{\theta}}) \approx [\boldsymbol{I}(\boldsymbol{\theta})]^{-1}$$

where

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\,\mathrm{E}\left[\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]$$

In this case, it can be shown that applying the above equations would give the asymptotic variance as (McCulloch and Searle 2001)

$$\mathrm{var}(\hat{\boldsymbol{b}}) = \tau^2(\boldsymbol{X'WX}^{-1})$$

Thus, the asymptotic distribution of the ML-estimate of $\boldsymbol{b}$, $\hat{\boldsymbol{b}}$ is Normally distributed with mean $\boldsymbol{b}$ and variance-covariance matrix $\hat{\tau}^2(\boldsymbol{X'WX}^{-1})$. However, note that these results are asymptotic, and are not necessarily valid for small samples. However, exact results do not exist for generalized linear models.

To test if the predictors of the model significantly explain the variance in the data, a Likelihood Ratio (LR) test can be formulated. More specifically, assume a null hypothesis

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0.$$

where $\boldsymbol{\theta}_0$ is a specified value we assume for $\boldsymbol{\theta}$ under the null hypothesis. If the null hypothesis is true, the likelihood ratio test statistic

$$-2\ln\Lambda = -2\left[l(\boldsymbol{\theta}_0) - l(\hat{\boldsymbol{\theta}})\right] \tag{4.2}$$

follows the $\chi^2(q)$ distribution with where q is the number of restrictions.

The parameter vector can also be partitioned by making specified assumptions only for some of the parameters. This can be formulated as a LR test as follows.

1. Fit the full model and save the value of the maximum log likelihood.
2. Fit a restricted model, where the parameters of interest are restricted according to the null hypothesis, and other parameters are estimated using the method of maximum likelihood.
3. Compute the LRT test statistic (4.2).
4. Compare the test statistic to $\chi^2(q)$ distribution, where q is the number of restrictions in the restricted model.

Another test, called Wald test, can be used for the same purpose. According to McCulloch and Searle (2001), it is computationally simpler, but may lead to worse approximations of the p-value with moderate or small sample sizes and extreme deviations. To test two models against each other, a Likelihood Ratio test can be formulated.

Example 4.3 *We want to test if dropping the only predictor from our model would cause no remarkable reduction of fit. The restricted model is fitted using*

```
> glm2 <- glm(spruces ~ 1, family=poisson(), data=plants2)
> summary(glm2)

> summary(glm2)

Call:
glm(formula = spruces ~ 1, family = poisson(), data = plants2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5374  -0.7311  -0.1237   0.4191   4.7545

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1692     0.0522    22.4   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 189.85  on 113  degrees of freedom
Residual deviance: 189.85  on 113  degrees of freedom
AIC: 504.09

Number of Fisher Scoring iterations: 5
```

The test statistic is computed as

```
> chiobs<--2*(logLik(glm2)-logLik(glm1))
> chiobs
[1] 2.392032
```

The same statistics is obtained also using anova

```
> anova(glm2,glm1)
Analysis of Deviance Table

Model 1: spruces ~ 1
Model 2: spruces ~ hdecid
  Resid. Df Resid. Dev  Df Deviance
1       113    189.853
2       112    187.461   1    2.392
```

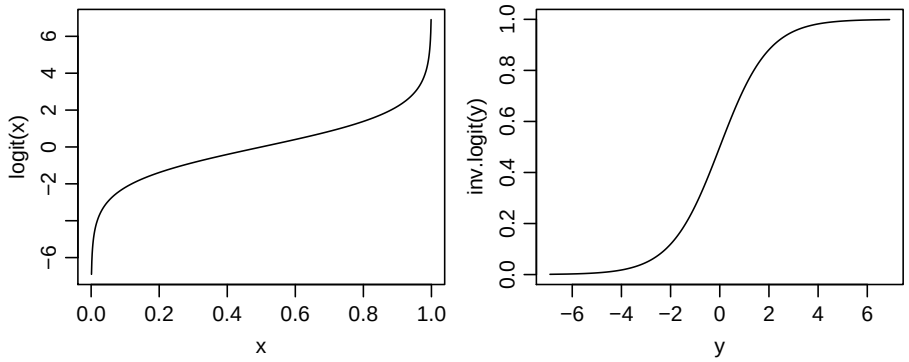


Figure 4.2: An illustration of the logit- (left) and inverse logit (right) transformations.

However, anova does not perform the test for glm, so we need to do it by ourselves. Comparison of the test statistic to χ^2 distribution with one degree of freedom gives

```
> 1-pchisq(chiobs,1)
[1] 0.121955
```

The p-value of 0.12 indicates that the trend seen in the number of spruce saplings could be a result of chance, as obtaining such a trend or stronger would have a probability of 0.12 if null hypothesis is true. The p-value of the coefficient of glm1 would have led to the same inference, even though the p-values differ slightly.

4.3 Logistic regression

In addition to the Linear model under normality, one common application of the generalized linear model is the logistic regression. It is the best-justified method for data of binary observations, such as success or failure, or presence or absence of a certain feature. Let us define the logit-transformation as

$$g(\mu) = \text{logit}(\mu) = \frac{\mu}{1-\mu}$$

Where μ is the expected value of the binomial variable, and is always between 0 and 1. The inverse logit is obtained by solving the above equation for μ. We get

$$\mu = g^{-1}(y) = \frac{1}{1+e^{-y}}\,.$$

These functions are illustrated in figure 4.2

In the Logistic regression model, we assume

$$\begin{aligned} y_i &\sim \text{indep.}Bernoulli(\mu(\boldsymbol{x}_i)) \\ g(\mu_i) &= \text{logit}(\mu_i) \\ &= \boldsymbol{x}_i'\boldsymbol{b} \end{aligned}$$

The estimation and inference follows the procedure outlined for the general model.

4.4 Poisson regression

The Poisson regression is an appropriate model for count data. In the Poisson regression model, we assume

$$\begin{aligned} y_i &\sim \text{indep.} Poisson(\mu(\boldsymbol{x}_i)) \\ g(\mu_i) &= \ln(\mu_i) \\ &= \boldsymbol{x}_i'\boldsymbol{b} \end{aligned}$$

The log link is used to ensure that the expected density, μ_i, is greater than zero. The estimation and inference follows the procedure outlined for the general model, and details are omitted.

As we remember from Chapter 1, the mean and variance of Poisson distribution are the same. However, a very common situation is that this relationship does not hold. In this case, an additional parameter is introduced that specifies the relation of the variance and expectation. It should be noted that this means that the Poisson distribution is not any more assumed. However, this model is commonly referred as extra-poisson model. In reality, no specific distribution is assumed, but the distribution is approximated by a discrete distribution that allows different mean and variance. Such distribution is, for example, the negative binomial distribution.

Example 4.4 *In the previous examples, we fitted Poisson GLM to the count data of spruce saplings. The assumption on the distribution was not evaluated, even though we carried out tests etc for the data. In Poisson situation, a crude test on the assumption can be carried out by comparing variances to the mean in discrete classes of the predictor. Thus, we classified the data from 123 plots to 5 classes of hdecid, with each class having equal number of observations. The class means and variances were computed, and are added to the previously shown plot of data in Figure 4.3.*

We see that the variances are higher than mean for all classes except for the last one, for which mean is slightly higher than the variance. However, the differences are fairly small for three classes. For the rest two classes, variances are about two times as high as the means. This implies that the assumption on Poisson distribution did not hold very well, but the number of seedlings varies more than one could expect. This is related to the spatial pattern of tree locations. We remember that the number of trees within a plot is Poisson distributed if tree locations are independent. Thus, this results shows that this independence is not met. The higher variance than the mean indicates clustering of spruce saplings. The lower-than-the-mean variance would mean regular

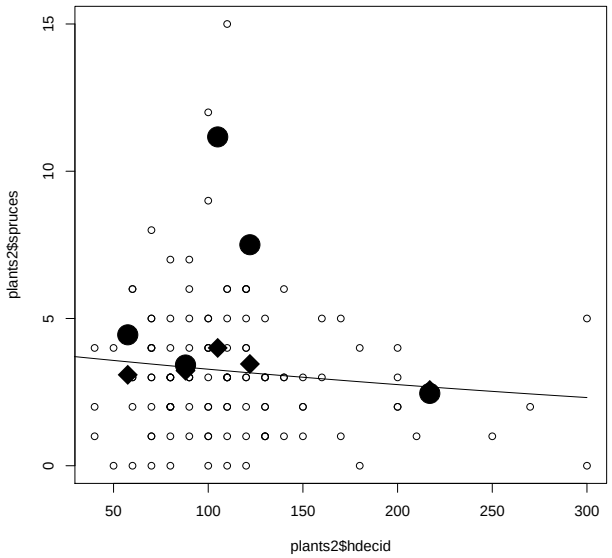


Figure 4.3: Means (diamonds) and variances (filled circles) in five classes of hdecid.

spatial pattern. Thus, we conclude that our data indicates clustering, and some other assumption than Poisson could be better. This is a very common situation with real-life datasets.

A better assumption would be an "overdispersed Poisson distribution. It is important to recognize that this is a misleading term, as we are actually not any more assuming a Poisson distributon. Instead, we assume an unspecified distribution having two parameters: one for mean, and another for the ratio of variance and mean. As no specific assumption on distribution is not made, the likelihood cannot be written, and MLE:s cannot be computed. However, we can write a quasi-likelihood that is something similar to the likelihood, but uses only information on mean and variance, not on the distribution itself. This would lead to Quasi-likelihood estimation. More information on these approaches can can be found e.g., in McCulloch and Searle (2001).

```
limits<-quantile(plants2$hdecid,na.rm=TRUE,probs=seq(0,1,0.2))-c(1,0,0,0,0,0)
means<-sapply(1:5,function(x) mean(plants2$spruce[plants2$hdecid>=limits[x]&plants2$hdecid<limits[x+1]],na.rm=TRUE))
vars<-sapply(1:5,function(x) var(plants2$spruce[plants2$hdecid>=limits[x]&plants2$hdecid<limits[x+1]],na.rm=TRUE))
points((limits[-1]+limits[-6])/2,means,pch=18,cex=3)
points((limits[-1]+limits[-6])/2,vars,pch=19,cex=3)
```

4.5 Weibull regression

The Weibull distribution is not a member of the exponential family. However, as an approach that is very similar to the GLM was presented for modeling diameter distri-

butions with Weibull distribution by Cao (2004), I will present such GLM-type model here.

Assume we have observed tree diameters from several stands, and we want to model the stand-specific diameter distribution using the Weibull distribution. Furthermore, we want to to relate the Weibull parameters to some stand-specific variables, such as stand age, or mean diameter. The ultimate aim is to fit a model that could be used for predicting the diameter distribution for a stand where only the utilized stand-specific predictors are known. This approach is commonly referred as the parameter prediction method (PPM) in the forestry literature.

A commonly used approach for such a modeling task is to first fit a Weibull distribution to the measured diameters of each stand using the method of maximum likelihood. After this first step, a dataset is obtained where each stand is one observation, and Weibull parameters and values of the potential stand-specific predictors are known for each stand. In the second step, the ML-estimates are modeled using the stand-specific predictors for each stand to estimate general relationships between the Weibull parameters and stand characteristics.

The GLM-type approach of Cao (2004) does these two steps at once, which is theoretically better justified and may also result to better fit. The assumed model for the diameter of tree i, $i = 1, \ldots, n$ is stated as follows

$$\begin{aligned} y_i &\sim \text{indep.} Weibull(\alpha_i, \beta_i) \\ g_\alpha(\alpha_i) &= \ln(\alpha_i) \\ &= \boldsymbol{x}'_{\alpha i}\boldsymbol{b}_\alpha \\ g_\beta(\beta_i) &= \ln(\beta_i) \\ &= \boldsymbol{x}'_{\beta i}\boldsymbol{b}_\beta \end{aligned}$$

Thus, we assume that tree diameter follows the $Weibull(\alpha, \beta)$ distribution. Furthermore, we use the log link for both parameters, as they are restricted to be greater than zero. The logarithmic parameters are assumed to be linear functions of the predictors, as specified in the last two equations. The predictors and coefficients may be different for the shape (α) and scale (β) parameters, that is why indices α and β are in vectors $\boldsymbol{x}$ and $\boldsymbol{b}$.

This specification is based on a similar way of thinking as is the GLM. However, there are some differences. First, Weibull distribution is not a member of the exponential family, so the results based on this assumption do not hold. Second, we are not modeling the mean of the distribution, but the parameters directly. Third, two parameters are written as a function of predictors, instead of having only the mean to

be modeled. However, if one manages to fit the model, the asymptotic properties of ML-estimates (consistency, unbiasedness, efficiency and normality) hold. These properties are, however, only large sample properties, and may be badly violated for small samples.

Example 4.5 *Dataset spati includes 10255 measured tree diameters from 66 plots in North Carelia. To relate the diameter distribution parameters on stand characteristics, we first use the well-known PPM approach (Parameter Prediction Method). In that method, the two-parameter Weibull function is first fitted int the data of each plot. Then the obtained estimates are saved into a plot-specific data, and the estimates are related to stand characteristics using a linear model. We omit all details on model diagnostics and assumptions made, as the aim is just to demonstrate the glm approach.*

The following code was used to fit Weibull for each plot using ML, and to save the estimated into data frame plots, which has 66 rows, one for each plot. Then we plot the selected stand variables and parameter estimates to study the relationships between stand variables and distribution parameters (Figure 4.4).

```
> library(stats4)
>
> # 2 parameter Weibull -logL
> nLLweibull<-function(x, shape=5,scale=20) {
+             -sum(dweibull(x,shape=shape,scale=scale,log=TRUE))
+             }
>
> # Fits two-parameter weibull distribution to tree diameter data using MLE starting from values 5 and 20
> # for shape and scale, respectively
> fitw2<-function(d) {
+        est<-mle(function(shape=5,scale=20) nLLweibull(d,shape,scale))
+        if (class(est)=="try-error") list(par=rep(NA,2),neg2LL=NA,conv=NA)
+        else list(par=coef(est),neg2LL=2*attributes(est)$min,conv=attributes(est)$details$convergence)
+        }
> # Fit weibull for each plot
> for (i in 1:dim(plots)[1]) {
+     d<-spati$d[spati$plot==plots$plot[i]]
+     weibullfit<-fitw2(d)
+     plots[i,c("shape","scale")]<-weibullfit$par
+     }
There were 50 or more warnings (use warnings() to see the first 50)
>
> plot(plots[,-1]) # do not include the first column
```

The plot shows no correlation between shape and stand variables. The scale parameter is quite highly correlated with mean diameter and mean height. Based on these observations, we use a constant value for shape, and relate logarithmic scale on Dg and Hg. The logarithmic scale is used for compatibility with the used Weibull regression with log link. The relationships are clearly not linear, but we assume linearity for simplicity. The parameter estimates for the PPM models are shown below.

```
> lmsc<-lm(log(scale)~Dg+Hg,data=plots)
> lmsh<-lm(log(shape)~1,data=plots)
> coef(summary(lmsh))
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 1.067284 0.04159837 25.65688 1.013757e-35
> coef(summary(lmsc))
```

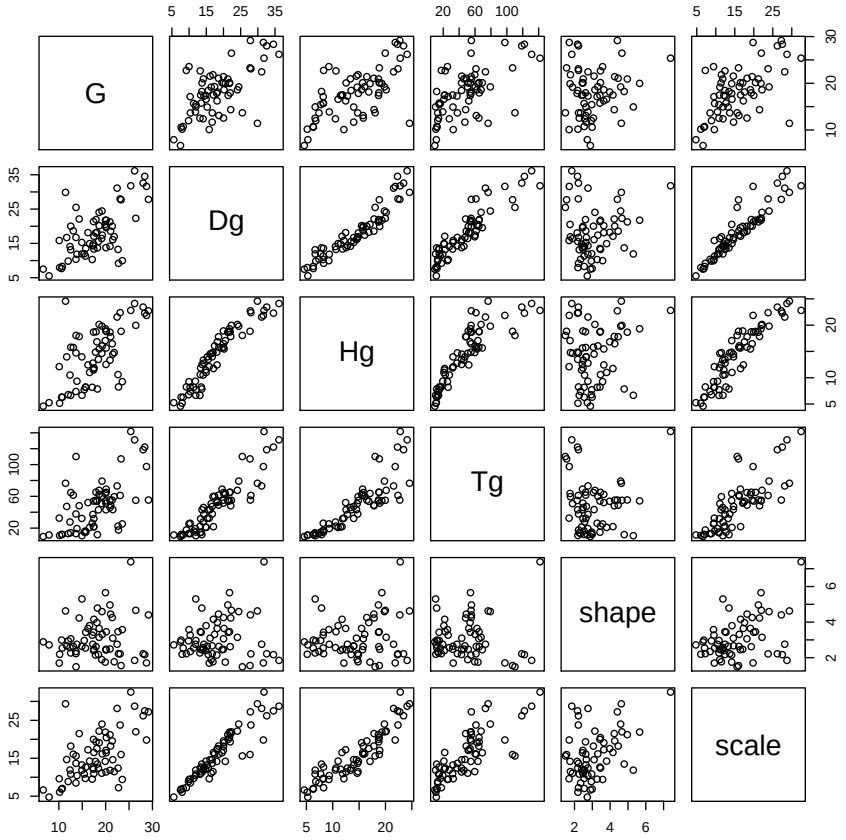


Figure 4.4: Plot of stand-specific Weibull-parameters on stand characteristics.

```
             Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 1.64465659 0.055449860 29.660248 1.002560e-38
Dg          0.02537290 0.008300818  3.056674 3.281326e-03
Hg          0.03956507 0.010854219  3.645133 5.428510e-04
```

Next, we formulate the Weibull GLM, where the ML fit and model fitting are carried out in one and the same stage. The coefficients from the PPM model are used as initial estimates of the parameters.

```
> coef(lmsh)->csh
> coef(lmsc)->csc
> names(csc)<-names(csh)<-c() # Having names in the vector, mle tries to use them as parameter names and gets lost.
>
> # Refit using glm-type approach
>
> nLL<-function(b0sh=csh,b0sc=csc[1],b1sc=csc[2],b2sc=csc[3]) {
+             shape<-exp(b0sh)
+             scale<-exp(b0sc+b1sc*spati$Dg+b2sc*spati$Hg)
+             -sum(dweibull(spati$d,shape=shape,scale=scale,log=TRUE))
+             }
>
> weibull.glm<-mle(minuslogl=nLL)
Warning messages:
1: In dweibull(x, shape, scale, log) : NaNs produced
2: In dweibull(x, shape, scale, log) : NaNs produced
3: In dweibull(x, shape, scale, log) : NaNs produced
4: In dweibull(x, shape, scale, log) : NaNs produced
5: In dweibull(x, shape, scale, log) : NaNs produced
6: In dweibull(x, shape, scale, log) : NaNs produced
7: In dweibull(x, shape, scale, log) : NaNs produced
8: In dweibull(x, shape, scale, log) : NaNs produced
9: In dweibull(x, shape, scale, log) : NaNs produced
> summary(weibull.glm)
Maximum likelihood estimation

Call:
mle(minuslogl = nLL)

Coefficients:
       Estimate  Std. Error
b0sh 0.94660490 0.007817644
b0sc 1.60010673 0.011130143
b1sc 0.04103064 0.001823252
b2sc 0.02342298 0.002213285

-2 log L: 60985.8
```

The estimates from GLM approach are quite different from those of PPM, but are of the same order of magnitude. Cao (2004) reported that the estimates from GLM approach lead to highly more accurate prediction of diameter distribution than the traditional PPM estimates. The improvement is due to that the GLM does the whole procedure in a single stage, whereas in PPM is carried in two stages, and parameter estimates from the first stage include estimation errors.

4.6 Generalized linear mixed models

As one might guess, the generalized linear mixed model is a generalized linear model with random effects. Such a model is appropriate for modeling hierarchical datasets, as was the linear mixed-effects model, too. The random effects are incorporated into the linear part of the model. Let $\boldsymbol{c}$ include the random effects of the model. The generalized

linear mixed model is specified as,

$$\begin{aligned} y_i|\boldsymbol{c} &\sim \text{indep.} f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{c}) \\ f_{Y_i|\boldsymbol{c}} &= \exp\left(\frac{y_i\gamma_i - b(\gamma_i)}{\tau^2} - c(y_i,\tau)\right) \\ g(\mu_i) &= \boldsymbol{x_i'b} + \boldsymbol{z'c} \end{aligned}$$

The model specification is very similar to the GLM, but now we have random effects c added to the linear predictor. The random effects are assumed to have a multinormal distribution, and effects of different groups in the data.

It is important to note that the generalized linear model assumes distribution of the exponential family **for the conditional observations**. Thus, for example, we might assume that tree species within a given stand is bernoulli distributed, or the number of trees of a given species from plots of one stand follows a Poisson distribution. However, the marginal distribution of the data may not be of that family. In linear mixed-effects model, where normality was assumed for the response, the general results on the normal distribution state that all conditional and marginal distributions of a multinormal distribution is normal. Thus, for LMM, marginal distributions are also normal. However, these result do not hold generally, and marginal distribution usually cannot be specified. However, the marginal expectation, variance, and covariance can be computed. The general formulas for marginal expectation, variance, and covariance are as follows.

$$\begin{aligned} \mathrm{E}(y_i) &= \mathrm{E}\left[g^{-1}(\boldsymbol{x_i'b} + \boldsymbol{z_i'c})\right] \\ \mathrm{var}(y_i) &= \mathrm{var}\left[g^{-1}\left(\boldsymbol{x_i'b} + \boldsymbol{z_i'c}\right)\right] + \mathrm{E}(\tau^2 v\left(g^{-1}[\boldsymbol{x_i'b} + \boldsymbol{z_i'c}]\right)) \\ \mathrm{cov}(y_i, y_j) &= \mathrm{cov}\left[g^{-1}(\boldsymbol{x_i'b} + \boldsymbol{z_i'c}), g^{-1}(\boldsymbol{x_j'b} + \boldsymbol{z_j'c})\right] \end{aligned}$$

As one can see, the values of these depend on the utilized link function, and on the assumed distribution of the data. Thus, no easily applicable formulas can be presented for these moments of the marginal distribution. In general, random effects cause dependence among the observations so, that observations that share a random effect are correlated. However, computing the correlation is not that simple as it was with LMM.

The estimation of the parameters of a GLMM is based on the principle of maximum likelihood. Each possible value of the random effect vector c would give different value for the likelihood. Thus, we can think the value of the likelihood as a transformation of the random variable, $\boldsymbol{c}$. The expected value of the likelihood is then the likelihood of the data. The likelihood becomes

$$L = \int_{\boldsymbol{c}} \prod_i f_{Y_i|\boldsymbol{c}}(y_i|\boldsymbol{c}) f_{\boldsymbol{C}}(\boldsymbol{c}) d\boldsymbol{c}$$

where $f_{Y_i|\boldsymbol{c}}(y_i|\boldsymbol{c})$ is the conditional density of the data (e.g., Poisson density), and $f_{\boldsymbol{C}}(\boldsymbol{c})$ is the q-dimensional joint distribution of random effects $c_{q\times 1}$.

Example 4.6 *Assume that n_i sample plots of equal area have been measured from stand i, $i = 1, \ldots, m$. Let y_{ij} denote the number of Scots pines observed on the j th plot of stand i. A model for the number of Scots pines within a plot is specifies as*

$$\begin{aligned} y_{ij}|\boldsymbol{c} &\sim \text{indep.}Poisson(\mu_{ij}) \\ \ln \mu_{ij} &= \boldsymbol{x'_{ij}b} + c_i \\ c_i &= \text{i.i.d.}N(0, \sigma_c^2) \end{aligned}$$

This model assumes only a random constant for each stand, so that the logarithmic density (and variance, as they are equal with Poisson distribution), behaves in a similar way with respect to the predictors, but the level may change.

The likelihood can be written as

$$\begin{aligned} l &= \ln\left(\prod_{i=1}^{m}\int_{-\infty}^{\infty}\prod_{j=1}^{n_i}\frac{\mu_{ij}^{y_{ij}}e^{-\mu_{ij}}}{y_{ij}!}\frac{1}{\sqrt{2\pi\sigma^2}e^{-\frac{1}{2\sigma_c^2}c_i^2}}dc_i\right) \\ &= \boldsymbol{y'Xb} - \sum_{i,j}\ln y_{ij}! + \\ &\quad \sum_i \ln\int_{-\infty}^{\infty}\exp\left[y_{i\cdot}u_i - \sum_j e^{\boldsymbol{x'_{ij}b}+u_i}\right]\frac{1}{\sqrt{2\pi\sigma_c^2}}e^{-\frac{1}{2\sigma_c^2}c_i^2}dc_i \end{aligned}$$

This integral cannot be expressed in closed form. Thus, we cannot express the likelihood equations in closed form, and all the computations need to be carried out numerically. However, numerical methods are so well developed nowadays that modern computers are able to find estimators for GLMMs.

The inference and tests on GLMMs are all based on the asymptotic properties of the ML-estimator. Thus, the inference and diagnostics on GLM:s applies also for GLMM:s.

The prediction of random effects with GLMMs is not as simple as for linear models.

Example 4.7 *A more extensive data of spruce saplings can be found in file* `plants.txt`. *The data includes measurements from 1926 plots from 123 regeneration areas within one county. The data is read using the following code.*

```
> plants<-read.table("c:/laurim/biometria/plants.txt",header=TRUE)
> plants<-plants[!is.na(plants$hdecid),]
> plants$spruces<-plants$planted+plants$spruces
> hist(plants$spruces)
```

A good starting point for a model for spruce sapling counts is a Poisson mixed model with log link and random constant. We fit the model using the default PQL method for approximating the likelihood.

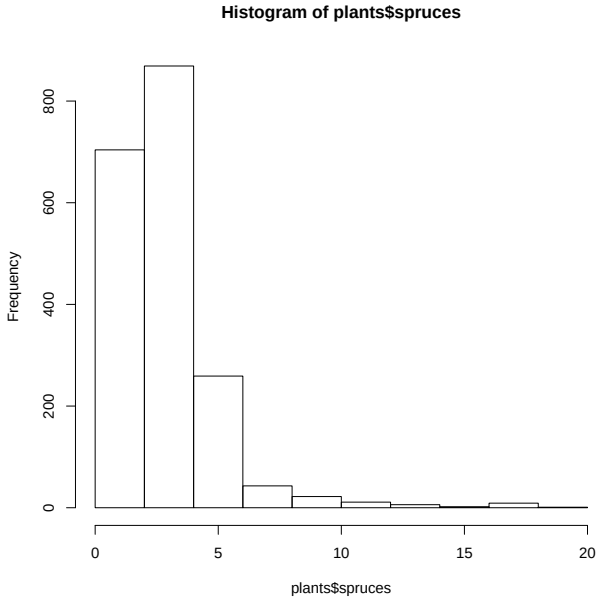


Figure 4.5: Histogram of the number of spruce saplings in dataset plants.

```
> glmm1.PQL<-lmer(spruces ~ (1|stand)+hdecid, family=poisson(), data=plants)
Generalized linear mixed model fit using PQL
Formula: spruces ~ (1 | stand) + hdecid
   Data: plants
 Family: poisson(log link)
  AIC  BIC logLik deviance
 2373 2390  -1184     2367
Random effects:
 Groups Name        Variance Std.Dev.
 stand  (Intercept) 0.065014 0.25498
number of obs: 1926, groups: stand, 123

Estimated scale (compare to  1 )  1.058734

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.1713541  0.0462620  25.320   <2e-16 ***
hdecid      -0.0001229  0.0003464  -0.355    0.723
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
       (Intr)
hdecid -0.820
```

Hdecid seems to be insignificant predictor, and could probably be dropped. Thus, we fit a restricted model and make a likelihood ratio test.

```
> glmm2.PQL<-lmer(spruces ~ (1|stand), family=poisson(), data=plants, method="PQL")
> anova(glmm2.PQL,glmm1.PQL)
Data: plants
Models:
glmm2.PQL: spruces ~ (1 | stand)
glmm1.PQL: spruces ~ (1 | stand) + hdecid
          Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
glmm2.PQL  2  2371.5  2382.7 -1183.8
glmm1.PQL  3  2373.4  2390.1 -1183.7 0.1622      1     0.6871
```

```
> summary(glmm2.PQL)
Generalized linear mixed model fit using PQL
Formula: spruces ~ (1 | stand)
   Data: plants
 Family: poisson(log link)
  AIC  BIC logLik deviance
 2372 2383  -1184     2368
Random effects:
 Groups Name        Variance Std.Dev.
 stand  (Intercept) 0.064826 0.25461
number of obs: 1926, groups: stand, 123

Estimated scale (compare to  1 )  1.058644

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.15792    0.02644    43.8   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The p-value from anova (0.6871) leads to rejection of glmm1, and we conclude that there is no evidence on that the number of spruce samplings would depend on the height of deciduous trees in this data. The next question could be whether the random effects are significant. The summary above showed, that the estimated between-stand variation in the logarithmic number of saplings is 0.25, which shows to be quite high. A test on this needs a restricted model, i.e., a model without random effects. Such a model is fitted using glm, and the likelihood ratio test is then carried out.

```
> glm1<-glm(spruces ~ 1, family=poisson(), data=plants)
> chiobs<--2*(logLik(glm1)-logLik(glmm2.PQL))
> chiobs
[1] 5738.658
> 1-pchisq(chiobs,1)
[1] 0
```

The value of test statistic is extremely high, indicating significant variation in stand density among stands. Thus, we accept the model with random effect for stand and no fixed effects as the final model.

The estimation methods for GLMM:s have not yet been found, and this area is under development. In GLMM:s, the problem is that the likelihood needs to be approximated numerically. Different approximation methods work in different situations, and that is why it is important to test different methods and even different implementations of them to see whether the estimates are the true maximums the likelihood. In R, there are three alternatives for GLMM:s: PQL, Laplace, and AGQ. However, AGQ is available only in lmer2, which is a development version of lmer. The authors of lmer write of the estimation methods as follows. ”The PQL method is fastest but least accurate. The Laplace method is intermediate in speed and accuracy. The AGQ method is the most accurate but can be considerably slower than the others. However, it appears that AGQ does not work for all situations.

Example 4.8 *The following code fits the model of previous example using Laplace method.*

```
> glmm2.Laplace<-lmer(spruces ~ (1|stand), family=poisson(), data=plants, method="Laplace")
> glmm2.Laplace
Generalized linear mixed model fit using Laplace
Formula: spruces ~ (1 | stand)
   Data: plants
 Family: poisson(log link)
  AIC  BIC logLik deviance
 2371 2382  -1183     2367
Random effects:
 Groups Name        Variance Std.Dev.
 stand  (Intercept) 0.0758   0.27532
number of obs: 1926, groups: stand, 123

Estimated scale (compare to  1 )  1.055265

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1482     0.0281   40.86   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

There are slight differences in the estimates of coefficients. According to the suggestion of the authors of lmer package, I would select the model estimated using Laplace instead of the one estimated using PQL.

The AQL method does not work for this data.

```
> glmm2.AGQ<-lmer(spruces ~ (1|stand), family=poisson(), data=plants, method="AGQ")
Error in lmer(spruces ~ (1 | stand), family = poisson(), data = plants,  :
  method = "AGQ" not yet implemented for supernodal representation
```

4.7 exercises

- Model the dependence of the number of planted spruces on the covariates of dataset plants2 using a GLM.
- Model the dependence of the number of planted spruces on the covariates of dataset plants using a GLMM.

Chapter 5

Model systems

5.1 Types of model systems

Models systems arise from a need to simultaneously model several responses. Typically, methods developed for simultaneous model estimation are needed when a model system, comprised of several individual models, is estimated. In fitting the model system, it may not be enough to model the behavior of individual models of the system. Instead, a model system may bee needed that realistically describes the system as a whole. For this reason, the interrelationships between models need to be taken into account.

Model systems have been used a lot of in econometrics for modeling the markets. For example, the so called Kleins small macroeconomic model includes equations for consumption, demand and investment (Greene 1997). The individual equations of this system are directly related, for example the demand is a function of consumption and investment. In forestry, model systems are needed, for example, in developing a forest simulator. Such a simulator may include models for stand structure, tree growth, mortality, and silvicultural operations. The stand structure has affects stand growth, which in turn determines the timing of the silvicultural practice, e.g., thinning. Furthermore, the thinning has an effect on the stand structure etc. Thus, different model components should formulate a model system that describes the development of a forest stands in a realistic way.

The model systems may be directly related, seemingly unrelated, or unrelated. In directly related models, response of one model may be the predictor in another model. In seemingly unrelated model, direct relationships do not exist between the individual models. However, the models may be related in that the residuals are correlated. In unrelated models, no direct relationship exists between the models, and the residuals are uncorrelated.

5.2 Estimation of directly related models: 2SLS

5.2.1 Illustration through height and volume models

In directly related models, a predictor of one component model is the response of another component. For example, we may have a model system for height h_i and volume v_i for tree i:

$$\ln h_i = a_h + b_h \ln d_i + e_{hi} \quad (5.1)$$

$$\ln v_i = a_v + b_v \ln d + c_v \ln_h + e_v \quad (5.2)$$

where d is tree diameter, a_h, b_h, a_v, b_v, and c_v are parameters to be estimated, and e_h and e_v are the residuals of the models, which are assumed to be uncorrelated and have expectation 0.

The two models of the above system are directly related, because one of the predictors of the volume model, the tree height, is the response of the height model. Assume that we have a model fitting data including observed diameters, heights, and volumes for trees from one stand. We could fit the models separately, i.e, first fit the model for tree height and then for tree volume. This would be a rather good approach in many situations.

Problems arise if we use the model for prediction for trees with unknown height. In such case we would first predict tree height using model (5.1), and further write the **predicted height** into the lower model to predict tree volume. However the predicted height is a different random variable then the true tree height. Thus, we are replacing a required predictor with something else, which is an unbiased estimator of the random variable we would need.

An estimation method that accounts for this problem is obtained by replacing the true height in the volume model with the predicted height also in the estimation stage. Then using the predicted height utilized in the application stage would be exactly the same random variable which was used in the model fitting stage. Thus, the estimation would be carried out in two stages:

1. Fit the height model (5.1) to the data, and save the fitted values of the model
2. Fit the volume model (5.2) to the data, but replace the height in the RHS by the predicted height from step 1.

This two-stage approach yields a model system that could be used in the application described below. Note that the volume model fitted in the second stage usually fits much worse to the data than a model with the true height as a predictor would fit. This is just a result of the fact that we are able to utilize only that part of the variation in

tree height that could be explained with variation in tree diameter. Thus, the fitted volume model is actually based on tree diameter only, and it does not account for the variation in tree height. However, this is what we actually need, as we do not have that information on tree height available in applications.

5.2.2 General formulation

In general case, the variables appearing in a model system can be classified into two classes: endogenous and exogenous. The exogenous variables are called instrumental variables, and they are the variables that bring the information into the system. In the previous example, the only instrumental variable was the tree diameter. The instrumental variables are those variables we will actually have known when the model system is applied in practice.

The direct relationships among the components of a model system can be taken into account through a two-stage least squares approach. In that approach, models for all the predictors are first fitted using the instruments as predictors, and the fitted values of these models are saved. In the second stage, the actual values of the predictors are replaced with the predictions from the first stage, to estimate the coefficients of the model system.

Let $\boldsymbol{X}$ be the model matrix of the model component to be estimated, and let $\boldsymbol{Z}$ be the matrix of instruments. The two-stage least squares approach can be written as:

1. $\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}$
2. $\widehat{\boldsymbol{B}_{\boldsymbol{IV}}} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y}$

Note that in the case of unrelated models, this approach would lead to OLS as a special case. This is because Z and X would be identical, giving $\hat{\boldsymbol{X}} = \boldsymbol{X}$.

5.3 Estimation of unrelated models with correlated residuals: SUR

5.3.1 Model formulation

The seemingly unrelated models do not have such a direct relationship as the directly related models do. However, they are related to each other because the residuals are correlated. Let us consider an example where tree height and volume are both modeled on tree diameter only

$$\ln h_i = a_h + b_h \ln d_i + e_{hi} \tag{5.3}$$

$$\ln v_i = a_v + b_v \ln d + e_{vi} \tag{5.4}$$

where the notations are the same as we used in models (5.1) and (5.2).

If the models are estimated separately, we would assume e_{hi} are independent, identically distributed random variables with constant variance, and the same assumptions would also be done for e_{vi}. However, the residuals for different models may be correlated for same individuals. For example, it might be realistic to assume that if height is overestimated for a given tree i, then also the volume might be overestimated. The separate estimation does not provide information on this kind of correlation. In a SUR model, we allow correlation among responses by specifying $\mathrm{cov}(e_{hi}, e_{vi})$.

The SUR model can also be expressed in a matrix form. Assume that we have m individual models,

$$\begin{aligned} \boldsymbol{y}_1 &= \boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{e}_1 \\ \boldsymbol{y}_2 &= \boldsymbol{X}_2\boldsymbol{b}_1 + \boldsymbol{e}_2 \\ &\vdots \\ \boldsymbol{y}_m &= \boldsymbol{X}_m\boldsymbol{b}_1 + \boldsymbol{e}_m \end{aligned}$$

By defining

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{X}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{X}_m \end{bmatrix}$$

$$\boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_m \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_m \end{bmatrix}$$

we can write the model system in the form of linear model (2.4) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

Thus, this model is just a special case of the (very!) general matrix formulation of the linear model. In the above formulation, $\boldsymbol{X}$ is a block-diagonal matrix. Matrix $\boldsymbol{D} = \mathrm{var}(\boldsymbol{e})$ has quite a peculiar structure

$$\boldsymbol{D} = \begin{bmatrix} \sigma_1^2\boldsymbol{I} & \varphi_{12}\boldsymbol{I} & \cdots & \varphi_{1m}\boldsymbol{I} \\ \varphi_{12}\boldsymbol{I} & \sigma_2^2\boldsymbol{I} & \cdots & \varphi_{2m}\boldsymbol{I} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{1m}\boldsymbol{I} & \varphi_{2m}\boldsymbol{I} & \cdots & \sigma_m^2\boldsymbol{I} \end{bmatrix}.$$

If the observations are ordered so that observations of different response of one observation are one after another, then the structure of $\boldsymbol{X}$ would change to something peculiar, but matrix $\boldsymbol{D}$ would then be block-diagonal.

5.3.2 Estimation

As the above specification showed that the SUR model is just a special case of the LM, the estimation methods do not change from those of the Linear model. Thus, the estimation involves first estimating the parameters specifying matrix $\boldsymbol{D}$, and then estimating the model coefficients using GLS.

Matrix $\boldsymbol{D}$ is specified with response-specific variances of residual error, and the between-response covariances. For estimating these parameters, two alternative methods are available. The first one, which we call Zellnerrs method, is very easy to implement even manually. In that method, the individual models are first fitted to the data separately using OLS. The variance-covariance matrix of residuals is then estimated using the residuals from these fits. Another approach is to assume multinormality of residuals and estimate the parameters of matrix $\boldsymbol{D}$ by using the method of maximum likelihood. In both methods, the estimated variances and covariances are written to matrix $\boldsymbol{D}$, and the parameter vector $\boldsymbol{b}$ is fitted using GLS.

5.3.3 Prediction

Prediction from a SUR model is very similar to prediction from any other linear model. We just write the values of the predictors to the model and make prediction. The prediction intervals are calculated in very similar manner than with Linear models.

However, an interesting application arises if the value of one or more responses of the simultaneous model system has been observed, and other responses are being predicted for that individual. In that case, the estimated between-model correlation can be used to carry information from one model to another. The prediction is a straightforward application of The linear predictor of section 1.6. For example, assume that the response of model 1 has been observed, and others are to be predicted. We just write the observed value of y_1 into vector $\boldsymbol{h}_2$, and the other variables for the same sampling unit into vector $\boldsymbol{h}_1$. The predictions is then direct application of the ideas presented in Section 1.6

5.3.4 Why simultaneous estimation

The across-model correlation may improve the analysis because

1. The fixed parameters of the model may be better estimated.

2. Prediction of the models will be more efficient, if the response of one of the component models have been observed.

3. A more realistic simulations could be obtained from the models system.

The first item of the list is maybe the most well-known result of simultaneous estimation. However, it seldom causes any big improvement to the efficiency. SUR leads to more efficient estimation, if correlations among the residuals exist, and if the predictors of the component models are not the same. However, if a predictor is dropped from one component model because of its statistical insignificance, and the predictors are otherwise the same, the gain of SUR is negligible. I would see simultaneous estimation as an alternative to using all the predictors in all the component models. This being the case, it seldom causes any remarkable gain in the efficiency of predictors. Thus, I would regard the two later points as more important from a practical point of view.

However, even though the information on between-model correlation would not improve the model at all, utilizing a known correlation would lead to much better and more realistic results in prediction or simulation. However, in these cases it is not necessary to fit the model simultaneously, but it may be enough just to estimate the between model covariances using the residuals of individual models. Furthermore, such an approach could lead to easy implementation of more sophisticated models for the interdependencies than just a constant covariance (Lappi 2006b).

The prediction from SUR model may be highly more efficient than from separately estimated models if the residuals of the models are correlated, and if observed response of one component model can be used in prediction. For example, assume that simultaneously fitted models (5.3) and (5.4) are available, and they are used for prediction for a tree with known height and diameter. In this case, we can compute the realized residual of the height model for that tree. Furthermore, the between-model correlation among residuals can be utilized to predict the residual of the volume model for that tree, which would lead to significantly better prediction of total volume than using tree diameter alone.

Simulation from a model system is a demanding task, where omitting important interrelationships between models can lead to drastic results. Thus, knowledge of the interrelationships between-models is extremely important in simulation. The simulation very often faces to a situation where an assumed linear, constant correlation is not sufficiently realistic and detailed assumption, and more is needed.

5.4 Estimation of directly related models with correlated residuals: 3SLS

The three-stage least squares is an approach combining 2SLS and SUR approaches. In 3SLS approach, The initial models are first estimated using the instrumental variable method, i.e., using 2SLS. Residuals of the estimated 2SLS models are used to estimate

matrix D for the whole models system. This matrix includes also non-zero cross-model covariances. Finally, the final model is estimated by GLS, using matrix $\hat{\boldsymbol{D}}$ in the estimation.

5.5 Simultaneous mixed models

Simultaneous mixed models are obtained, if linear mixed models are fitted for several responses at the same time. Assume that we have m individual mixed models,

$$\begin{aligned} \boldsymbol{y}_1 &= \boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{Z}_1\boldsymbol{c}_1 + \boldsymbol{e}_1 \\ \boldsymbol{y}_2 &= \boldsymbol{X}_2\boldsymbol{b}_1 + \boldsymbol{Z}_2\boldsymbol{c}_2 + \boldsymbol{e}_2 \\ &\vdots \\ \boldsymbol{y}_m &= \boldsymbol{X}_m\boldsymbol{b}_1 + \boldsymbol{Z}_m\boldsymbol{c}_m + \boldsymbol{e}_m \end{aligned}$$

where $\mathrm{cov}(\boldsymbol{e}_i, \boldsymbol{e}_j)$ are nonzero. This means that the residuals of the individual observations are correlated. If only this kind of correlation is assumed, then an easy way to implement the model is to write all the responses in the same column of the data, and add an additional categorical variable that specifies which response is used at that row. In model fitting, the categorical variable indicating the response would be used to specify the innermost level of grouping for the data.

In addition, we may assume also the random effects to be correlated, i.e., we would allow $\mathrm{cov}(\boldsymbol{c}_i, \boldsymbol{c}_j)$ be nonzero. Unfortunately, such models are not that easy to estimate with R. One could surely rely on estimating the models separately, and then computing the cross-model covariances using the residuals and predicted random effects. However, SAS has quite good capabilities for fitting such models.

An application of a simultaneous mixed model was presented by Lappi (1991), and this model was further demonstrated in Lappi et al. (2006). The following pages present an example from Lappi et al. (2006).

Example 6.4. The multivariate case

Lappi (1991) constructed the following multivariate model for the logarithmic height and logarithmic volume of tree i in stand k from stem analysis data (Laasasenaho 1982):

$$\ln H_{ki} = 3.410 - 18.58\frac{1}{D_{ki}} + a_{0k} - a_{1k}\frac{1}{D_{ki}} + e_{ki} \text{ and}$$

$$\ln V_{ki} = 2.704 - 48.93\frac{1}{D_{ki}} + 1.387\ln D_{ki} + c_{0k} - c_{1k}\frac{1}{D_{ki}} + u_{ki},$$

where D_{ki} is DBH+7 cm, parameters a_{0k}, a_{1k}, c_{0k} and c_{1k} are stand-specific random parameters and e_{ki} and u_{ki} are residuals with estimated variances of var(e_{ki})=0.01113, var(u_{ki})=0.01540 and covariance cov(e_{ki},u_{ki})=0.01040. Let us write the random parameters as vectors $\mathbf{a}_k = \begin{pmatrix} a_{0k} & a_{1k} \end{pmatrix}'$ $\mathbf{c}_k = \begin{pmatrix} c_{0k} & c_{1k} \end{pmatrix}'$ and

define $\mathbf{b}_k = (\mathbf{a}_k' \quad \mathbf{c}_k')'$. The estimated dispersion matrix of $\mathbf{b}_k$ is (Lappi 1991)

$$\mathrm{var}(\mathbf{b}_k) = \mathbf{D} = \left[\begin{array}{cc|cc} 0.04739 & -0.3887 & 0.05082 & -0.4772 \\ -0.3887 & 20.64 & -0.6036 & 24.88 \\ \hline 0.05082 & -0.6036 & 0.05988 & -0.7876 \\ -0.4772 & 24.88 & -0.7876 & 31.11 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{D}_1 & \mathbf{D}_{12}' \\ \mathbf{D}_{12} & \mathbf{D}_2 \end{array}\right] = [\mathbf{C} \quad \mathbf{H}]$$

.

The last two parts define a partition of matrix **D** that is needed in the following calculations. The measured height of a sample tree will be used below to predict the random parameters of the volume function. Assume that two sample trees of diameters 20 and 30 cm and heights 20 and 26 m have been measured. The measured heights follow the model

$$\mathbf{y}_k = \boldsymbol{\mu} + \mathbf{Z}\mathbf{a}_k + \mathbf{e}_k ,$$

where vector $\mathbf{y}_k$ includes the measured logarithmic heights, $\mathbf{y}_k = \begin{bmatrix} \ln 20 \\ \ln 26 \end{bmatrix} = \begin{bmatrix} 3.00 \\ 3.26 \end{bmatrix}$, and $\boldsymbol{\mu}$ their expectations, which are obtained using the first two terms of the height model as $\boldsymbol{\mu} = \begin{bmatrix} 2.72 \\ 2.91 \end{bmatrix}$. Matrix **Z** is the design matrix of the random part, i.e., $\mathbf{Z} = \begin{bmatrix} 1 & 1/(20+7) \\ 1 & 1/(30+7) \end{bmatrix}$, and $\mathbf{a}_k$ and $\mathbf{e}_k$ are unknown vectors of random parameters and random residuals with variances $\mathrm{var}(\mathbf{a}_k) = \mathbf{D}_1$ and $\mathrm{var}(\mathbf{e}_k) = \mathbf{R} = 0.01113 \cdot \mathbf{I}$. Using the height and volume models, equation (6.1) can be written as

$$\begin{bmatrix} \mathbf{b}_k \\ \mathbf{y}_k \end{bmatrix} \sim \left(\begin{bmatrix} \mathbf{0} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{C}\mathbf{Z}' \\ \mathbf{Z}\mathbf{C}' & \mathbf{Z}\mathbf{D}_1\mathbf{Z}' + \mathbf{R} \end{bmatrix} \right)$$

and the BLUP of $\mathbf{b}_k$ is (Equation 6.2)

$$\hat{\mathbf{b}}_k = \mathbf{C}\mathbf{Z}'(\mathbf{Z}\mathbf{D}_1\mathbf{Z}' + \mathbf{R})^{-1}(\mathbf{y}_k - \boldsymbol{\mu}) = \begin{pmatrix} 0.244 \\ 0.985 \\ 0.230 \\ 1.131 \end{pmatrix} ,$$

i.e. the predicted random parameters are a_{0k}=0.244, a_{1k}=0.985, c_{0k}=0.230 and c_{1k}=1.131. The predicted logarithmic heights and volumes are obtained by writing these estimates into the height and volume models.

In order to arrive at unbiased predictions of volumes and heights, half of the prediction variance was added to the predicted logarithmic heights and volumes before applying the exponential transformation. The prediction variance of random parameters was first calculated to be

$$\mathrm{var}\left(\hat{\mathbf{b}}_k - \mathbf{b}_k\right) = \mathbf{D} - \mathbf{ZC}'\left(\mathbf{ZD}_1\mathbf{Z}' + \mathbf{R}\right)^{-1}\mathbf{CZ}' = \begin{pmatrix} 0.0212 & -0.536 & 0.0266 & -0.648 \\ -0.536 & 17.6 & -0.716 & 21.3 \\ 0.0266 & -0.716 & 0.0372 & -0.918 \\ -0.648 & 21.3 & -0.918 & 26.8 \end{pmatrix}$$

.

Ignoring the estimation errors in the fixed parameters, the prediction variances of the predicted logarithmic heights were then obtained from the diagonal of

$$\mathrm{var}\left(\hat{\mathbf{y}}_k * - \mathbf{y}_k *\right) = \mathbf{Z} * \mathrm{var}\left(\hat{\mathbf{a}}_k - \mathbf{a}_k\right)\mathbf{Z}*' + 0.01113\mathbf{I}\,,$$

where $\mathbf{y}_k$* denotes the heights of the tally trees, **Z*** the design matrix of tally trees and $\mathrm{var}\left(\hat{\mathbf{a}}_k - \mathbf{a}_k\right)$ includes the first two rows and columns of $\mathrm{var}\left(\hat{\mathbf{b}}_k - \mathbf{b}_k\right)$ (see the definition of $\mathbf{b}_k$). The height and volume models corrected for population level and local bias are shown in Figure 6.4.

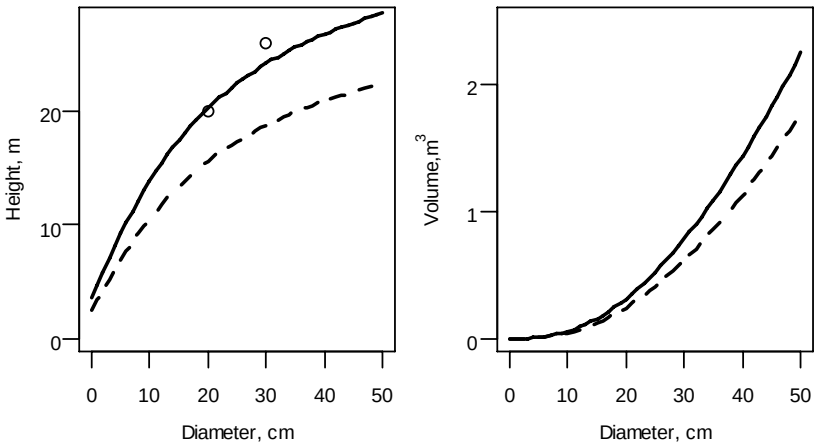


Figure 6.4. Predicted height and volume models when random parameters are 0 (dashed lines) and are predicted using the two observed heights shown in the plot on the left.

5.6 Excercises

1. Show that 2SLS will lead to OLS if the models are unrelated. Explain how this would happen in practice.

Appendix A

Matrix algebra

These pages have been taken from lecture notes on Forest biometrics by professor Annika Kangas.

Matrix calculus

Matrix	**Vector**
$\mathbf{X}_{(n\times p)} = \begin{bmatrix} x_{11} & x_{12} & .. & .. & x_{1p} \\ x_{21} & x_{22} & .. & .. & x_{2p} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ x_{n1} & .. & .. & .. & x_{np} \end{bmatrix}$	$\mathbf{x}_{(p\times 1)} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ .. \\ x_p \end{bmatrix}$
Element x_{ij} means the jth element of row i in matrix **X**. R: `X[i,j]`	A matrix with only one column is called vector. x_j is the jth element of vector **x**. R: `x[i]`

A single number is called **scalar.**

Transpose of a vector	**Transpose of a matrix**
$\mathbf{x}' = \mathbf{x}^T$ Rows and columns interchange. R: `t(x)`	$X_{ij}=X'_{ji}$ R: `t(X)`
$\mathbf{x}_{(1\times p)}{}' = \begin{bmatrix} x_1 & x_2 & x_3 & .. & x_p \end{bmatrix}$	$\mathbf{X}^T_{(p\times n)} = \begin{bmatrix} x_{11} & x_{12} & .. & .. & x_{1n} \\ x_{21} & x_{22} & .. & .. & x_{2n} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ x_{p1} & .. & .. & .. & x_{pn} \end{bmatrix}$

Example

$$\mathbf{X}_{(4\times3)} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}$$

```
X<-matrix(seq(1,12),
          ncol=3,byrow=TRUE)
X<-rbind(1:3,4:6,7:9,10:12)
```

element x_{23}=6 `X[2,3]`

element x_{32}=8

$$\mathbf{Y}_{(3\times4)} = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}$$

```
Y<-t(X)
```

element x_{23} =8

element x_{32}=6

$$\mathbf{x}_{(4\times1)} = \begin{bmatrix} 1 \\ 4 \\ 7 \\ 10 \end{bmatrix}$$

```
x<-X[,1]
```

element x_3 = 7

$$\mathbf{x}_{(1\times4)} = \begin{bmatrix} 1 & 4 & 7 & 10 \end{bmatrix}$$

Square matrix

- has equal number of rows and columns

$$\mathbf{X}_{(n\times n)} = \begin{bmatrix} x_{11} & x_{12} & .. & .. & x_{1n} \\ x_{21} & x_{22} & .. & .. & x_{2n} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ x_{n1} & .. & .. & .. & x_{nn} \end{bmatrix}$$

Symmetric (square) matrix

- $x_{ij}=x_{ji}$
- for example,

$$\mathbf{X}_{(4\times4)} = \begin{bmatrix} 1 & 2 & 3 & 5 \\ 2 & 1 & 4 & 6 \\ 3 & 4 & 2 & 7 \\ 5 & 6 & 7 & 3 \end{bmatrix}$$

Diagonal	**Diagonal matrix**
The diagonal of **X** includes elements x_{ii} `x<-diag(X)`	Only diagonal has non-zero values $\mathbf{X}_{(n\times n)} = \begin{bmatrix} x_{11} & 0 & 0 & .. & 0 \\ 0 & x_{22} & 0 & .. & 0 \\ .. & .. & .. & .. & .. \\ 0 & 0 & 0 & .. & 0 \\ 0 & 0 & 0 & .. & x_{nn} \end{bmatrix}$ `X<-diag(x)`

Identity matrix	**Block diagonal matrix**
A diagonal matrix where all diagonal elements are ones. $\mathbf{I}_n = \mathbf{I}_{(n\times n)} = \begin{bmatrix} 1 & 0 & 0 & .. & 0 \\ 0 & 1 & 0 & .. & 0 \\ .. & .. & .. & .. & .. \\ 0 & 0 & 0 & .. & 0 \\ 0 & 0 & 0 & .. & 1 \end{bmatrix}$ `I<-diag(rep(1,n))`	A block diagonal matrix has m (n_m by n_m) square matrices on the diagonal. $\mathbf{X}_{(n\times n)} = \begin{bmatrix} x_{11} & x_{21} & .. & 0 & 0 \\ x_{12} & x_{22} & .. & 0 & 0 \\ .. & .. & .. & .. & .. \\ 0 & 0 & .. & x_{n-1n-1} & x_{n-1n} \\ 0 & 0 & .. & x_{nn-1} & x_{nn} \end{bmatrix}$ $= \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & .. & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_m \end{bmatrix}$

Example

$$\mathbf{X}_{(6\times6)} = \begin{bmatrix} 1 & 2 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 4 & 0 & 0 \\ 0 & 0 & 4 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 6 \\ 0 & 0 & 0 & 0 & 6 & 5 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{(2\times2)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{(2\times2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_{(2\times2)} \end{bmatrix}$$

Product of a matrix and a scalar

$$a\mathbf{X}_{(n\times p)} = \begin{bmatrix} ax_{11} & ax_{12} & .. & .. & ax_{1p} \\ ax_{21} & ax_{22} & .. & .. & ax_{2p} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ ax_{n1} & .. & .. & .. & ax_{np} \end{bmatrix}$$

Example

$$3\mathbf{X}_{(3\times4)} = 3\begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix} = \begin{bmatrix} 3 & 12 & 21 & 30 \\ 6 & 15 & 24 & 33 \\ 9 & 18 & 27 & 36 \end{bmatrix}$$

Sum of matrices

The dimensions of the matrices need to be equal. For each element of the sum matrix

$z_{ij}=x_{ij}+y_{ij}$

$$\mathbf{X}_{(n\times p)} + \mathbf{Y}_{(n\times p)} = \begin{bmatrix} x_{11} & x_{12} & .. & .. & x_{1p} \\ x_{21} & x_{22} & .. & .. & x_{2p} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ x_{n1} & .. & .. & .. & x_{np} \end{bmatrix} + \begin{bmatrix} y_{11} & y_{12} & .. & .. & y_{1p} \\ y_{21} & y_{22} & .. & .. & y_{2p} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ y_{n1} & .. & .. & .. & y_{np} \end{bmatrix}$$

$$= \begin{bmatrix} x_{11}+y_{11} & x_{12}+y_{12} & .. & .. & x_{1p}+y_{1p} \\ x_{21}+y_{21} & x_{22}+y_{22} & .. & .. & x_{2p}+y_{2p} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ x_{n1}+y_{n1} & .. & .. & .. & x_{np}+y_{np} \end{bmatrix}$$

Example

$$\mathbf{X}_{(3\times 4)} + \mathbf{Y}_{(3\times 4)} = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix} = \begin{bmatrix} 2 & 6 & 10 & 14 \\ 7 & 11 & 15 & 19 \\ 12 & 16 & 20 & 24 \end{bmatrix}$$

Inner product of vectors

$$\mathbf{z}_{1\times 1} = \mathbf{x}_{1\times p}'\mathbf{y}_{p\times 1} = \begin{bmatrix} x_1 & x_2 & x_3 & .. & x_p \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ .. \\ y_p \end{bmatrix} = \sum_{j=1}^{p} x_j y_j$$

Example

$$\mathbf{z}_{1\times 1} = \mathbf{x}_{1\times 4}'\mathbf{y}_{4\times 1} = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \\ 7 \\ 8 \end{bmatrix} = 1\cdot 5 + 2\cdot 6 + 3\cdot 7 + 4\cdot 8$$

$$= 5 + 12 + 21 + 32 = 70$$

```
z<-x%*%y
```

Outer product of vectors

$$\mathbf{Z}_{p\times p} = \mathbf{y}_{p\times 1}\mathbf{x}_{1\times p}' = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ .. \\ y_p \end{bmatrix} \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} = \begin{bmatrix} y_1x_1 & y_1x_2 & .. & .. & y_1x_p \\ y_2x_1 & y_2x_2 & .. & .. & y_2x_p \\ y_3x_1 & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ y_px_1 & y_px_2 & .. & .. & y_px_p \end{bmatrix}$$

Example

$$\mathbf{Z}_{4\times 4} = \mathbf{y}_{4\times 1}\mathbf{x}_{1\times 4}{}' = \begin{bmatrix} 5 \\ 6 \\ 7 \\ 8 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 5\cdot 1 & 5\cdot 2 & 5\cdot 3 & 5\cdot 4 \\ 6\cdot 1 & 6\cdot 2 & 6\cdot 3 & 6\cdot 4 \\ 7\cdot 1 & 7\cdot 2 & 7\cdot 3 & 7\cdot 4 \\ 8\cdot 1 & 8\cdot 2 & 8\cdot 3 & 8\cdot 4 \end{bmatrix} = \begin{bmatrix} 5 & 10 & 15 & 20 \\ 6 & 12 & 18 & 24 \\ 7 & 14 & 21 & 28 \\ 8 & 16 & 24 & 32 \end{bmatrix}$$

```
Z<-y%o%x
```

Product of matrices

$$\mathbf{Z}_{(n\times q)} = \mathbf{X}_{(n\times p)}\mathbf{Y}_{(p\times q)} = \begin{bmatrix} x_{11} & x_{12} & .. & .. & x_{1p} \\ x_{21} & x_{22} & .. & .. & x_{2p} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ x_{n1} & .. & .. & .. & x_{np} \end{bmatrix} \begin{bmatrix} y_{11} & y_{12} & .. & .. & y_{1q} \\ y_{21} & y_{22} & .. & .. & y_{2q} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ y_{p1} & .. & .. & .. & y_{pq} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{p} x_{1j}y_{j1} & \sum_{j=1}^{p} x_{2j}y_{j1} & .. & .. & \sum_{j=1}^{p} x_{nj}y_{j1} \\ \sum_{j=1}^{p} x_{1j}y_{j1} & \sum_{j=1}^{p} x_{2j}y_{j2} & .. & .. & \sum_{j=1}^{p} x_{nj}y_{j2} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ \sum_{j=1}^{p} x_{nj}y_{j1} & .. & .. & .. & \sum_{j=1}^{p} x_{nj}y_{jq} \end{bmatrix}$$

Example

$$\mathbf{Z}_{(2\times3)} = \mathbf{X}_{(2\times3)}\mathbf{Y}_{(3\times2)} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 10 \\ 8 & 11 \\ 9 & 12 \end{bmatrix} = \begin{bmatrix} 1\cdot7+2\cdot8+3\cdot9 & 1\cdot10+2\cdot11+3\cdot12 \\ 4\cdot7+5\cdot8+6\cdot9 & 4\cdot10+5\cdot11+6\cdot12 \end{bmatrix}$$

$$= \begin{bmatrix} 7+16+27 & 10+22+36 \\ 28+40+54 & 40+55+72 \end{bmatrix} = \begin{bmatrix} 50 & 68 \\ 122 & 167 \end{bmatrix}$$

```
Z=X%*%Y
```

Multiplying a matrix by a diagonal matrix

$$\mathbf{Z}_{(n\times m)} = \mathbf{X}_{(n\times n)}\mathbf{Y}_{(n\times m)} = \begin{bmatrix} x_{11} & 0 & 0 & .. & 0 \\ 0 & x_{22} & 0 & .. & 0 \\ 0 & 0 & x_{33} & .. & 0 \\ .. & .. & .. & .. & .. \\ 0 & .. & .. & .. & x_{nn} \end{bmatrix} \begin{bmatrix} y_{11} & y_{12} & .. & .. & y_{1m} \\ y_{21} & y_{22} & .. & .. & y_{2m} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ y_{n1} & .. & .. & .. & y_{nm} \end{bmatrix}$$

$$= \begin{bmatrix} x_{11}y_{11} & x_{11}y_{12} & .. & .. & x_{11}y_{1m} \\ x_{22}y_{21} & x_{22}y_{22} & .. & .. & x_{22}y_{2m} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ x_{nn}y_{n1} & .. & .. & .. & x_{nn}y_{nm} \end{bmatrix}$$

Matrix inverse

Inverse matrix $\mathbf{X}^{-1}$ is a square matrix which fulfills $\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}$

Inverse of a diagonal matrix is obtained by inverting the diagonal elements.

$$\mathbf{X}^{-1}_{(n\times n)} = \begin{bmatrix} x_{11} & 0 & 0 & .. & 0 \\ 0 & x_{22} & 0 & .. & 0 \\ 0 & 0 & x_{33} & .. & 0 \\ .. & .. & .. & .. & .. \\ 0 & .. & .. & .. & x_{nn} \end{bmatrix}^{-1} = \begin{bmatrix} x_{11}^{-1} & 0 & .. & .. & 0 \\ 0 & x_{22}^{-1} & .. & .. & 0 \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ 0 & .. & .. & .. & x_{nn}^{-1} \end{bmatrix}$$

Inverting a matrix is hard work even for small matrices.

For a 2 by 2 matrix, $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ the inverse is

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{d}{ad-cb} & \frac{-c}{ad-cb} \\ \frac{-b}{ad-cb} & \frac{a}{ad-cb} \end{bmatrix}$$

```
Xinv<-solve(X)
```

Rules for matrix calculations

- A(BC)=(AB)C
- (A+B)C=AC+BC
- AB≠BA
- (AB)'=B'A'
- IA=AI=A
- $(AB)^{-1}=B^{-1}A^{-1}$

Appendix B

A short story on R

R is a programming environment that includes lot of ready functions for statistical computing. Most users just use one or some of the most common functions, such as `lm` or `plot`.

The following page includes a R-reference card by Jonathan Baron.

R reference card, by Jonathan Baron

Parentheses are for functions, brackets are for indicating the position of items in a vector or matrix. (Here, items with numbers like x1 are user-supplied variables.)

Miscellaneous

q(): quit
<-: assign
INSTALL package1: install package1
m1[,2]: column 2 of matrix m1
m1[,2:5] or m1[,c(2,3,4,5)]: columns 2–5
m1$a1: variable a1 in data frame m1
NA: missing data
is.na: true if data missing
library(mva): load (e.g.) the mva package

Help

help(command1): get help with command1 (NOTE: USE THIS FOR MORE DETAIL THAN THIS CARD CAN PROVIDE.)
help.start(): start browser help
help(package=mva): help with (e.g.) package mva
apropos("topic1"): commands relevant to topic1
example(command1): examples of command1

Input and output

source("file1"): run the commands in file1.
read.table("file1"): read in data from file1
data.entry(): spreadsheet
scan(x1): read a vector x1
download.file(url1): from internet
url.show(url1), read.table.url(url1): remote input
sink("file1"): output to file1, until sink()
write(object, "file1"): writes an object to file1
write.table(dataframe1,"file1"): writes a table

Managing variables and objects

attach(x1): put variables in x1 in search path
detach(x1): remove from search path
ls(): lists all the active objects.
rm(object1): removes object1
dim(matrix1): dimensions of matrix1
dimnames(x1): names of dimensions of x1
length(vector1): length of vector1
1:3: the vector 1,2,3
c(1,2,3): creates the same vector
rep(x1,n1): repeats the vector x1 n1 times
cbind(a1,b1,c1), rbind(a1,b1,c1): binds columns or rows into a matrix
merge(df1,df2): merge data frames
matrix(vector1,r1,c1): make vector1 into a matrix with r1 rows and c1 columns
data.frame(v1,v2): make a data frame from vectors v1 and v2
as.factor(), as.matrix(), as.vector(): conversion
is.factor(), is.matrix(), is.vector(): what it is
t(): switch rows and columns
which(x1==a1): returns indices of x1 where x1==a1

Control flow

for (i1 in vector1): repeat what follows
if (condition1) ...else ...: conditional

Arithmetic

%*%: matrix multiplication
%/%, ^, %%, sqrt(): integer division, power, modulus, square root

Statistics

max(), min(), mean(), median(), sum(), var(): as named
summary(data.frame): prints statistics
rank(), sort() rank and sort
ave(x1,y1): averages of x1 grouped by factor y1
by(): apply function to data frame by factor
apply(x1,n1,function1): apply function1 (e.g. mean) to x by rows (n1=1) or columns (n2=2)
tapply(x1,list1,function1): apply function to x1 by list1
table(): make a table
tabulate(): tabulate a vector

basic statistical analysis

aov(), anova(), lm(), glm(): linear and nonlinear models, anova
t.test(): t test
prop.test(), binom.test(): sign test
chisq.test(x1): chi-square test on matrix x1
fisher.test(): Fisher exact test
cor(a): show correlations
cor.test(a,b): test correlation
friedman.test(): Friedman test

some statistics in mva package

prcomp(): principal components
kmeans(): kmeans cluster analysis
factanal(): factor analysis
cancor(): canonical correlation

Graphics

plot(), barplot(), boxplot(), stem(), hist(): basic plots
matplot(): matrix plot
pairs(matrix): scatterplots
coplot(): conditional plot
stripplot(): strip plot
qqplot(): quantile-quantile plot
qqnorm(), qqline(): fit normal distribution

Bibliography

Bailey, R., and T. Dell. 1973. Quantifying diameter distributions with the weibull function. Forest Science 19:97–104.

Box, G.E.P., and D.R. Cox. 1962. An analysis of transformations. Journal of the royal statistical society, Series B 26:211–252.

Cao, Q.V. 2004. Predicting parameters of a Weibull function for modeling diameter distribution. Forest Science 50(5):682–685.

Casella, G., and R.L. Berger. 2002. Statistical Inference. 2nd edition. Duxbury, Pacific Grove, USA. 660 p.

Cressie, N.A. 1993. Statistics for spatial data. Wiley, New York, USA.

Faraway, J. 2004. Linear models with R. Chapman and Hall, London, UK.

Faraway, J. 2006. Extending the linear model with R. Generalized linear, mixed effects and nonparametric regression models. Chapman and Hall, London, UK.

Greene, W.H. 1997. Econometric Analysis. 3rd edition. Prentice Hall, Upper Saddle River, New Jersey, USA.

Hand, D.J., F. Daly, A. Lunn, K.J. McConway, and E. Ostrowski. 1994. A Handbook of Small Data Sets. Chapman and Hall, London, UK.

Harrell, F.J. 2001. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. Springer, New York, USA.

Lappi, J. 1991. Calibration of height and volume equations with random parameters. Forest Science 37:781–801.

Lappi, J. 2006a. Metsäbiometrian menetelmiä. Silva Carelica 24, University of Joensuu, Faculty of Forest Sciences. 18 p. In Finnish.

Lappi, J. 2006b. A multivariate, nonparametric stem-curve prediction method. Canadian Journal of Forest Research 36:1017–1027.

Lappi, J., L. Mehtätalo, and K.T. Korhonen. 2006. Generalizing sample tree information. P. 85–106, *in* Forest inventory - methodology and application, Kangas, A. and Maltamo, M., (ed.). Springer.

McCulloch, C.E., and S.R. Searle. 2001. Generalized, linear, and mixed models. Wiley-Interscience, New York, USA. 325 p.

Mehtätalo, L. 2004. A longitudinal height-diameter model for norway spruce in finland. Canadian Journal of Forest Research 34(1):131–140.

Mehtätalo, L. 2005. Height-diameter models for scots pine and birch in finland. Silva Fennica 39(1):55–66.

Muller, K.E., and P.W. Stewart. 2006. Linear model theory. Univariate, multivariate, and mixed models. Wiley, Hoboken, New Jersey, USA.

Pinheiro, J.C., and D.M. Bates. 2000. Mixed-effects models in S and Splus. Springer-Verlag, New York, USA. 528 p.

Richards, F.J. 1959. A flexible growth function for empirical use. J. Exp. Bot. 10(29):290–300.

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. Variance components. Wiley, New York, USA. 501 p.

Siipilehto, J. 1999. Improving the accuracy of predicted basal-area diameter distribution in advanced stands by determining stem number. Silva Fennica 33:281–301.

Tadikamalla, P.P., , and N.L. Johnson. 1982. Systems of frequency curves generated by transformations of logistic variables. Biometrika 69:461–465.

Venables, W.N., and B.D. Ripley. 2002. Modern applied statistics with S. 4th edition. Springer, New York, USA. 495 p.

Wang, M., and K. Rennolls. 2005. Tree diameter distribution modeling: introducing a logit-logistic distribution. Canadian Journal of Forest Research 35:1305–1313.