Introduction to Nonlinear Regression

Andreas Ruckstuhl
IDP Institut für Datenanalyse und Prozessdesign
ZHAW Zürcher Hochschule für Angewandte Wissenschaften

October 2010*†

Contents

*The author thanks Werner Stahel for his valuable comments.

†E-Mail Address: Andreas.Ruckstuhl@zhaw.ch; Internet: http://www.idp.zhaw.ch

Goals

The *nonlinear regression model* block in the Weiterbildungslehrgang (WBL) in angewandter Statistik at the ETH Zurich should

1. introduce problems that are relevant to the fitting of nonlinear regression functions,
2. present graphical representations for assessing the quality of approximate confidence intervals, and
3. introduce some parts of the statistics software R that can help with solving concrete problems.

1. The Nonlinear Regression Model

a **The Regression Model.** Regression studies the relationship between a **variable of interest** Y and one or more **explanatory or predictor variables** $x^{(j)}$. The general model is

$$Y_i = h\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)} \; ; \; \theta_1, \theta_2, \ldots, \theta_p \rangle + E_i \; .$$

Here, h is an appropriate function that depends on the explanatory variables and parameters, that we want to summarize with vectors $\underline{x} = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}]^T$ and $\underline{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^T$. The unstructured deviations from the function h are described via the random errors E_i. The normal distribution is assumed for the distribution of this random error, so

$$E_i \sim \mathcal{N}\left\langle 0, \sigma^2 \right\rangle \; , \quad \text{independent.}$$

b **The Linear Regression Model.** In (multiple) linear regression, functions h are considered that are linear in the parameters θ_j,

$$h\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)} \; ; \; \theta_1, \theta_2, \ldots, \theta_p \rangle = \theta_1 \widetilde{x}_i^{(1)} + \theta_2 \widetilde{x}_i^{(2)} + \ldots + \theta_p \widetilde{x}_i^{(p)} \; ,$$

where the $\widetilde{x}^{(j)}$ can be arbitrary functions of the original explanatory variables $x^{(j)}$. (Here the parameters are usually denoted as β_j instead of θ_j.)

c **The Nonlinear Regression Model** In nonlinear regression, functions h are considered that can not be written as linear in the parameters. Often such a function is derived from theory. In principle, there are unlimited possibilities for describing the deterministic part of the model. As we will see, this flexibility often means a greater effort to make statistical statements.

Example d **Puromycin.** The speed with which an enzymatic reaction occurs depends on the concentration of a substrate. According to the information from Bates and Watts (1988), it was examined how a treatment of the enzyme with an additional substance called Puromycin influences this reaction speed. The initial speed of the reaction is chosen as the variable of interest, which is measured via radioactivity. (The unit of the variable of interest is count/min^2; the number of registrations on a Geiger counter per time period measures the quantity of the substance present, and the reaction speed is proportional to the change per time unit.)

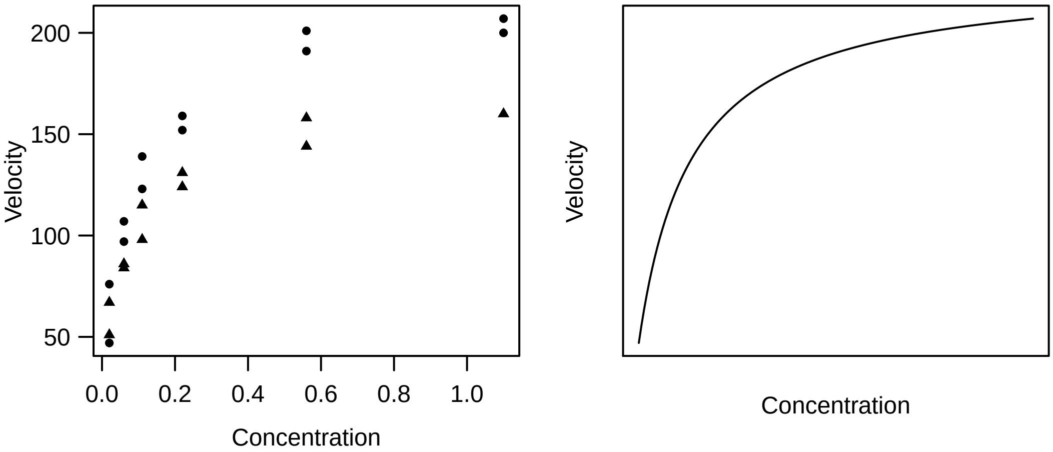

Figure 1.d: Puromycin Example. (a) Data (● treated enzyme; △ untreated enzyme) and (b) typical course of the regression function.

The relationship of the variable of interest with the substrate concentration x (in ppm) is described via the Michaelis-Menten function

$$h\langle x;\underline{\theta}\rangle = \frac{\theta_1 x}{\theta_2 + x} \; .$$

An infinitely large substrate concentration ($x \rightarrow \infty$) results in the ”asymptotic” speed θ_1. It has been suggested that this variable is influenced by the addition of Puromycin. The experiment is therefore carried out once with the enzyme treated with Puromycin and once with the untreated enzyme. Figure 1.d shows the result. In this section the data of the treated enzyme is used.

Example e **Oxygen Consumption.** To determine the biochemical oxygen consumption, river water samples were enriched with dissolved organic nutrients, with inorganic materials, and with dissolved oxygen, and were bottled in different bottles. (Marske, 1967, see Bates and Watts (1988)). Each bottle was then inoculated with a mixed culture of microorganisms and then sealed in a climate chamber with constant temperature. The bottles were periodically opened and their dissolved oxygen content was analyzed. From this the biochemical oxygen consumption [mg/l] was calculated. The model used to connect the cumulative biochemical oxygen consumption Y with the incubation time$x$, is based on exponential growth decay, which leads to

$$h\left\langle x,\underline{\theta}\right\rangle = \theta_1\left(1 - e^{-\theta_2 x}\right)$$

. Figure 1.e shows the data and the regression function to be applied.

Example f **From Membrane Separation Technology** (Rapold-Nydegger (1994)). The ratio of protonated to deprotonated carboxyl groups in the pores of cellulose membranes is dependent on the pH value x of the outer solution. The protonation of the carboxyl carbon atoms can be captured with ^{13}C-NMR. We assume that the relationship can be written with the extended *“Henderson-Hasselbach Equation”* for polyelectrolytes

$$\log_{10}\left\langle \frac{\theta_1 - y}{y - \theta_2}\right\rangle = \theta_3 + \theta_4\, x \; ,$$

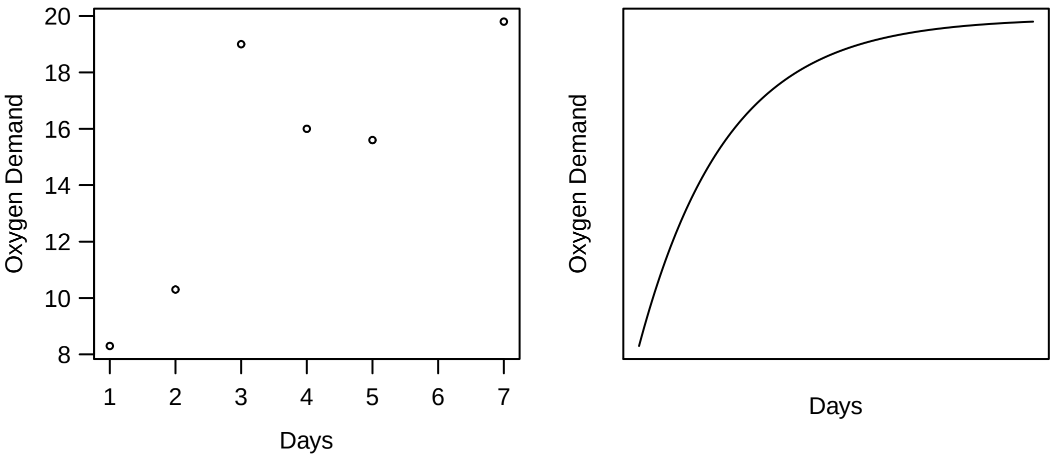

Figure 1.e: Oxygen consumption example. (a) Data and (b) typical shape of the regression function.

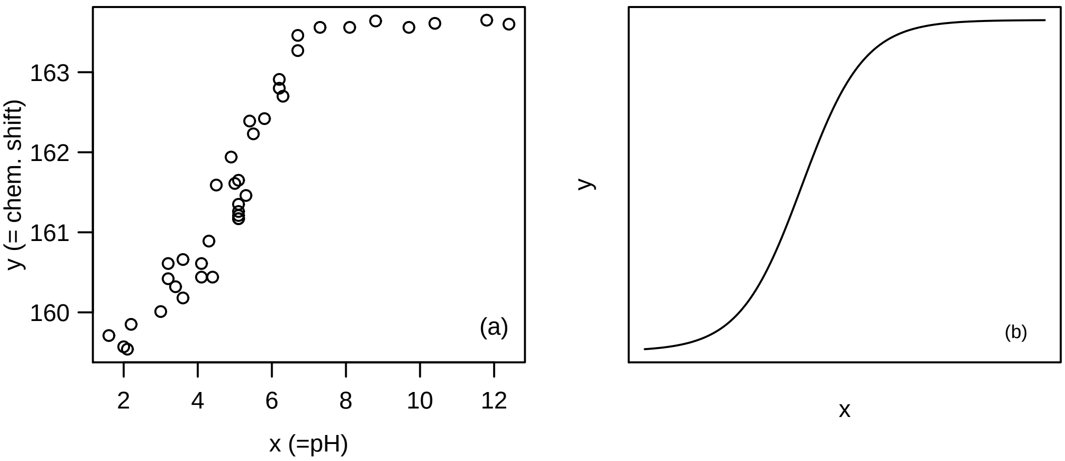

Figure 1.f: Membrane Separation Technology.(a) Data and (b) a typical shape of the regression function.

where the unknown parameters are θ_1, θ_2 and $\theta_3 > 0$ and $\theta_4 < 0$. Solving for y leads to the model

$$Y_i = h\langle x_i; \underline{\theta}\rangle + E_i = \frac{\theta_1 + \theta_2\, 10^{\theta_3+\theta_4 x_i}}{1 + 10^{\theta_3+\theta_4 x_i}} + E_i\ .$$

The regression funtion $h\langle x_i, \underline{\theta}\rangle$ for a reasonably chosen $\underline{\theta}$ is shown in Figure 1.f next to the data.

g A Few Further Examples of Nonlinear Regression Functions:

- Hill Model (Enzyme Kinetics): $h\langle x_i, \underline{\theta}\rangle = \theta_1 x_i^{\theta_3}/(\theta_2 + x_i^{\theta_3})$
  For $\theta_3 = 1$ this is also known as the Michaelis-Menten Model (1.d).
- Mitscherlich Function (Growth Analysis): $h\langle x_i, \underline{\theta}\rangle = \theta_1 + \theta_2 \exp\langle\theta_3 x_i\rangle$.
- From kinetics (chemistry) we get the function

$$h\langle x_i^{(1)}, x_i^{(2)}; \underline{\theta}\rangle = \exp\langle -\theta_1 x_i^{(1)} \exp\langle -\theta_2/x_i^{(2)}\rangle\rangle.$$

- Cobbs-Douglas Production Function

$$h\left\langle x_i^{(1)}, x_i^{(2)}; \underline{\theta}\right\rangle = \theta_1 \left(x_i^{(1)}\right)^{\theta_2} \left(x_i^{(2)}\right)^{\theta_3} .$$

Since useful regression functions are often derived from the theory of the application area in question, a general overview of nonlinear regression functions is of limited benefit. A compilation of functions from publications can be found in Appendix 7 of Bates and Watts (1988).

h **Linearizable Regression Functions.** Some nonlinear regression functions can be **linearized** through transformation of the variable of interest and the explanatory variables.

For example, a power function

$$h\langle x; \underline{\theta}\rangle = \theta_1 x^{\theta_2}$$

can be transformed for a linear (in the parameters) function

$$\ln\langle h\langle x; \underline{\theta}\rangle\rangle = \ln\langle\theta_1\rangle + \theta_2 \ln\langle x\rangle = \beta_0 + \beta_1 \widetilde{x} ,$$

where$\beta_0 = \ln\langle\theta_1\rangle$, $\beta_1 = \theta_2$ and $\widetilde{x} = \ln\langle x\rangle$. We call the regression function h **linearizable**, if we can transform it into a function linear in the (unknown) parameters via transformations of the arguments and a monotone transformation of the result.

Here are some more linearizable functions (also see Daniel and Wood, 1980):

$$\begin{array}{lcl}
h\langle x, \underline{\theta}\rangle = 1/(\theta_1 + \theta_2 \exp\langle -x\rangle) & \longleftrightarrow & 1/h\langle x, \underline{\theta}\rangle = \theta_1 + \theta_2 \exp\langle -x\rangle \\
h\langle x, \underline{\theta}\rangle = \theta_1 x/(\theta_2 + x) & \longleftrightarrow & 1/h\langle x, \underline{\theta}\rangle = 1/\theta_1 + \theta_2/\theta_1 \frac{1}{x} \\
h\langle x, \underline{\theta}\rangle = \theta_1 x^{\theta_2} & \longleftrightarrow & \ln\langle h\langle x, \underline{\theta}\rangle\rangle = \ln\langle\theta_1\rangle + \theta_2 \ln\langle x\rangle \\
h\langle x, \underline{\theta}\rangle = \theta_1 \exp\langle\theta_2 g\langle x\rangle\rangle & \longleftrightarrow & \ln\langle h\langle x, \underline{\theta}\rangle\rangle = \ln\langle\theta_1\rangle + \theta_2 g\langle x\rangle \\
h\langle x, \underline{\theta}\rangle = \exp\langle -\theta_1 x^{(1)} \exp\langle -\theta_2/x^{(2)}\rangle\rangle & \longleftrightarrow & \ln\langle\ln\langle h\langle x, \underline{\theta}\rangle\rangle\rangle = \ln\langle -\theta_1\rangle + \ln\langle x^{(1)}\rangle - \theta_2/x^{(2)} \\
h\langle x, \underline{\theta}\rangle = \theta_1 \left(x^{(1)}\right)^{\theta_2} \left(x^{(2)}\right)^{\theta_3} & \longleftrightarrow & \ln\langle h\langle x, \underline{\theta}\rangle\rangle = \ln\langle\theta_1\rangle + \theta_2 \ln\langle x^{(1)}\rangle + \theta_3 \ln\langle x^{(2)}\rangle .
\end{array}$$

The last one is the Cobbs-Douglas Model from 1.g.

i **The Statistically Complete Model.** A linear regression with the linearized regression function in the referred-to example is based on the model

$$\ln\langle Y_i\rangle = \beta_0 + \beta_1 \widetilde{x}_i + E_i ,$$

where the random errors E_i all have the same normal distribution. We back transform this model and thus get

$$Y_i = \theta_1 \cdot x^{\theta_2} \cdot \widetilde{E}_i$$

with $\widetilde{E}_i = \exp\langle E_i\rangle$. The errors $\widetilde{E}_i$, $i = 1, \ldots, n$ now contribute multiplicatively and are lognormal distributed! The assumptions about the random deviations are thus now drastically different than for a model that is based directily on h,

$$Y_i = \theta_1 \cdot x^{\theta_2} + E_i^*$$

with random deviations E_i^* that, as usual, contribute additively and have a specific normal distribution.

A linearization of the regression function is therefore advisable only if the assumptions about the random deviations can be better satisfied - in our example, if the errors actually act multiplicatively rather than additively and are lognormal rather than normally distributed. These assumptions must be checked with residual analysis.

j * Note: In linear regression it has been shown that the variance can be stabilized with certain transformations (e.g. $\log\langle\cdot\rangle, \sqrt{\cdot}$). If this is not possible, in certain circumstances one can also perform a weighted linear regression . The process is analogous in nonlinear regression.

k The introductory examples so far:

We have spoken almost exclusively of regression functions that only depend on one original variable. This was primarily because it was possible to fully illustrate the model graphically. The ensuing theory also functions well for regression functions $h\left\langle \underline{x}; \underline{\theta}\right\rangle$, that depend on several explanatory variables $\underline{x} = [x^{(1)}, x^{(2)}, \ldots, x^{(m)}]$.

2. Methodology for Parameter Estimation

a **The Principle of Least Squares.** To get estimates for the parameters $\underline{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^T$, one applies, like in linear regression calculations, the principle of least squares. The sum of the squared deviations

$$S(\underline{\theta}) := \sum_{i=1}^{n} (y_i - \eta_i\langle\underline{\theta}\rangle)^2 \qquad \text{mit} \;\; \eta_i\langle\underline{\theta}\rangle := h\langle x_i; \underline{\theta}\rangle$$

should thus be minimized. The notation where $h\langle x_i; \underline{\theta}\rangle$ is replaced by $\eta_i\langle\underline{\theta}\rangle$ is reasonable because $[x_i, y_i]$ is given by the measurement or observation of the data and only the parameters $\underline{\theta}$ remain to be determined.

Unfortunately, the minimum of the squared sum and thus the estimation can not be given explicitly as in linear regression. **Iterative numeric procedures** help further. The basic ideas behind the common algorithm will be sketched out here. They also form the basis for the easiest way to derive tests and confidence intervals.

b **Geometric Illustration.** The observed values $\underline{Y} = [Y_1, Y_2, \ldots, Y_n]^T$ determine a point in n-dimensional space. The same holds for the "model values" $\underline{\eta}\left\langle\underline{\theta}\right\rangle = [\eta_1\left\langle\underline{\theta}\right\rangle, \eta_2\left\langle\underline{\theta}\right\rangle, \ldots, \eta_n\left\langle\underline{\theta}\right\rangle]^T$ for given $\underline{\theta}$.

Take note! The usual geometric representation of data that is standard in, for example, multivariate statistics, considers the observations that are given by m variables $x^{(j)}$, $j = 1, 2, \ldots, m$,as points in m-dimensional space. Here, though, we consider the Y- and η-values of all n observations as points in n-dimensional space.

Unfortunately our idea stops with three dimensions, and thus with three observations. So, we try it for a situation limited in this way, first for simple linear regression. As stated, the observed values $\underline{Y} = [Y_1, Y_2, Y_3]^T$ determine a point in 3-dimensional space. For given parameters $\beta_0 = 5$ and $\beta_1 = 1$ we can calculate the model values $\eta_i\left\langle\underline{\beta}\right\rangle = \beta_0 + \beta_1 x_i$ and represent the corresponding vector $\underline{\eta}\left\langle\underline{\beta}\right\rangle = \beta_0\underline{1} + \beta_1\underline{x}$ as a point. We now ask where all points lie that can be achieved by variation of the parameters. These are the possible linear combinations of the two vectors $\underline{1}$ and $\underline{x}$ and thus form the

plane "spanned by $\underline{1}$ and $\underline{x}$" . In estimating the parameters according to the principle of least squares, geometrically represented, the squared distance between $\underline{Y}$ and $\underline{\eta}\left\langle\underline{\beta}\right\rangle$ is minimized. So, we want the point on the plane that has the least distance to $\underline{Y}$. This is also called the **projection** of $\underline{Y}$ onto the plane. The parameter values that correspond to this point $\underline{\widehat{\eta}}$ are therefore the estimated parameter values $\underline{\widehat{\beta}} = [\widehat{\beta}_0, \widehat{\beta}_1]^T$.
Now a nonlinear function, e.g. $h\langle \underline{x};\underline{\theta}\rangle = \theta_1 \exp\langle 1-\theta_2 x\rangle$, should be fitted on the same three observations. We can again ask ourselves where all points $\underline{\eta}\langle\underline{\theta}\rangle$ lie that can be achieved through variations of the parameters θ_1 and θ_2. They lie on a two-dimensional *curved* surface (called the **model surface** in the following) in three-dimensional space. The estimation problem again consists of finding the point $\underline{\widehat{\eta}}$ on the model surface that lies nearest to $\underline{Y}$. The parameter values that correspond to this point $\underline{\widehat{\eta}}$, are then the estimated parameter values $\underline{\widehat{\theta}} = [\widehat{\theta}_1, \widehat{\theta}_2]^T$.

c **Solution Approach for the Minimization Problem.** The main idea of the usual algorithm for minimizing the sum of squared deviations (see 2.a) goes as follows: If a preliminary best value $\underline{\theta}^{(\ell)}$ exists, we approximate the model surface with the plane that touches the surface at the point $\underline{\eta}\left\langle\underline{\theta}^{(\ell)}\right\rangle = h\left\langle\underline{x};\underline{\theta}^{(\ell)}\right\rangle$. Now we seek the point in this plane that lies closest to $\underline{Y}$. This amounts to the estimation in a linear regression problem. This new point lies on the plane, but not on the surface, that corresponds to the nonlinear problem. However, it determines a parameter vector $\underline{\theta}^{(\ell+1)}$ and with this we go into the next round of iteration.

d **Linear Approximation.** To determine the approximated plane, we need the partial derivative

$$A_i^{(j)}\langle\underline{\theta}\rangle := \frac{\partial\eta_i\langle\underline{\theta}\rangle}{\partial\theta_j} ,$$

which we can summarize with a $n \times p$ matrix $\boldsymbol{A}$. The approximation of the model surface $\underline{\eta}\langle\underline{\theta}\rangle$ by the "tangential plane" in a parameter value $\underline{\theta}^*$ is

$$\eta_i\langle\underline{\theta}\rangle \approx \eta_i\langle\underline{\theta}^*\rangle + A_i^{(1)}\langle\underline{\theta}^*\rangle(\theta_1-\theta_1^*) + ... + A_i^{(p)}\langle\underline{\theta}^*\rangle(\theta_p-\theta_p^*)$$

or, in matrix notation,

$$\underline{\eta}\langle\underline{\theta}\rangle \approx \underline{\eta}\langle\underline{\theta}^*\rangle + \boldsymbol{A}\langle\underline{\theta}^*\rangle(\underline{\theta}-\underline{\theta}^*) .$$

If we now add back in the random error, we get a linear regression model

$$\underline{\widetilde{Y}} = \boldsymbol{A}\langle\underline{\theta}^*\rangle\,\underline{\beta} + \underline{E}$$

with the "preliminary residuals" $\widetilde{Y}_i = Y_i - \eta_i\langle\underline{\theta}^*\rangle$ as variable of interest, the columns of $\boldsymbol{A}$ as regressors and the coefficients $\beta_j = \theta_j - \theta_j^*$ (a model without intercept β_0).

e **Gauss-Newton Algorithm.** The Gauss-Newton algorithm consists of, beginning with a start value $\underline{\theta}^{(0)}$ for $\underline{\theta}$, solving the just introduced linear regression problem for $\underline{\theta}^{*} = \underline{\theta}^{(0)}$ to find a correction $\underline{\beta}$ and from this get an improved value $\underline{\theta}^{(1)} = \underline{\theta}^{(0)} + \underline{\beta}$. For this, again, the approximated model is calculated, and thus the "preliminary residuals" $\underline{Y} - \underline{\eta}\left\langle \underline{\theta}^{(1)} \right\rangle$ and the partial derivatives $\boldsymbol{A}\left\langle \underline{\theta}^{(1)} \right\rangle$ are determined, and this gives us $\underline{\theta}_2$. This iteration step is continued as long as the correction $\underline{\beta}$ is negligible. (Further details can be found in Appendix A.)

It can not be guaranteed that this procedure actually finds the minimum of the squared sum. The chances are better, the better the p-dimensionale model surface at the minimum $\underline{\widehat{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)^T$ can be locally approximated by a p-dimensional "plane" and the closer the start value $\underline{\theta}^{(0)}$ is to the solution being sought.

* Algorithms comfortably determine the derivative matrix $\boldsymbol{A}$ numerically. In more complex problems the numerical approximation can be insufficient and cause convergence problems. It is then advantageous if expressions for the partial derivatives can be arrived at analytically. With these the derivative matrix can be reliably numerically determined and the procedure is more likely to converge (see also Chapter 6).

f **Initial Values.** A iterative procedure requires a starting value in order for it to be applied at all. Good starting values help the iterative procedure to find a solution more quickly and surely. Some possibilities to get these more or less easily are here briefly presented.

g **Initial Value from Prior Knowledge.** As already noted in the introduction, nonlinear models are often based on theoretical considerations from the application area in question. Already existing **prior knowledge** from similar experiments can be used to get an initial value. To be sure that the chosen start value fits, it is advisable to graphically represent the regression function $h\langle x; \underline{\theta}\rangle$ for various possible starting values $\underline{\theta} = \underline{\theta}^0$ together with the data (e.g., as in Figure 2.h, right).

h **Start Values via Linearizable Regression Functions.** Often, because of the distribution of the error, one is forced to remain with the nonlinear form in models with linearizable regression functions. However, the linearized model can deliver starting values.

In the **Puromycin Example** the regression function is linearizable: The reciprocal values of the two variables fulfill

$$\widetilde{y} = \frac{1}{y} \approx \frac{1}{h\langle x; \underline{\theta}\rangle} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1}\frac{1}{x} = \beta_0 + \beta_1 \widetilde{x} \; .$$

The least squares solution for this modified problem is $\underline{\widehat{\beta}} = [\widehat{\beta}_0,\, \widehat{\beta}_1]^T = (0.00511,\, 0.000247)^T$ (Figure 2.h (a)). This gives the initial value

$$\theta_1^{(0)} = 1/\widehat{\beta}_0 = 196 \; , \qquad \theta_2^{(0)} = \widehat{\beta}_1/\widehat{\beta}_0 = 0.048 \; .$$

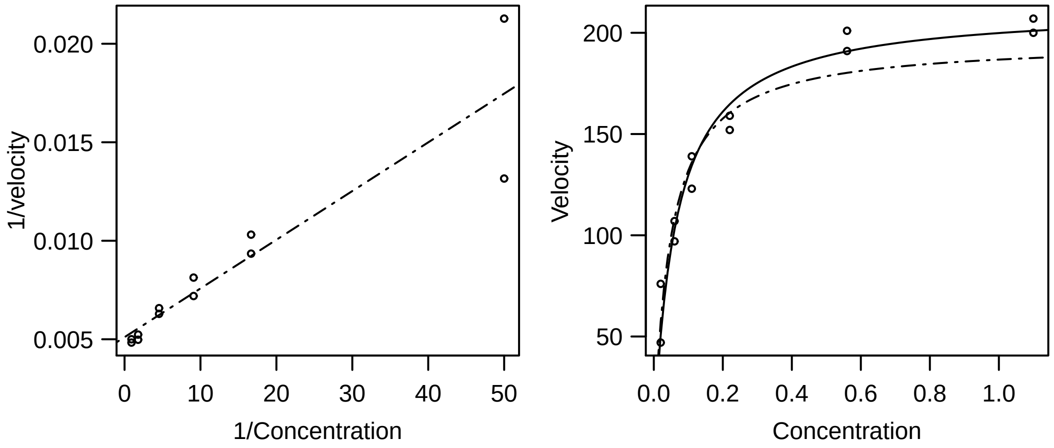

Figure 2.h: Puromycin Example. Left: Regression line in the linearized problem. Right: Regression function $h\langle x;\underline{\theta}\rangle$ for the initial values $\underline{\theta} = \underline{\theta}^{(0)}$ (- - - - -) and for the least squares estimation $\underline{\theta} = \widehat{\underline{\theta}}$ (———).

i **Initial Values via Geometric Meaning of the Parameter.** It is often helpful to consider the geometrical features of the regression function.

In the **Puromycin Example** we can thus arrive at an initial value in another, instructive way: θ_1 is the y value for $x = \infty$. Since the regression function is monotone increasing, we can use the maximal y_i-value or a visually determined "asymptotic value" $\theta_1^0 = 207$ as initial value for θ_1. The parameter θ_2 is the x-value, at which y reaches half of the asymptotic value θ_1. This gives $\theta_2^0 = 0.06$.

The initial values thus result from the geometrical meaning of the parameters and a coarse determination of the corresponding aspects of a curve "fitted by eye."

Example j **Membrane Separation Technology.**In the Membrane Separation example we let $x \to \infty$, so $h\langle x;\underline{\theta}\rangle \to \theta_1$ (since $\theta_4 < 0$); for $x \to -\infty$, $h\langle x;\underline{\theta}\rangle \to \theta_2$. From Figure 1.f(a) along with the data shows $\theta_1 \approx 163.7$ and $\theta_2 \approx 159.5$. We know θ_1 and θ_2, so we can linearize the regression function through

$$\widetilde{y} := \log_{10}\langle \frac{\theta_1^{(0)} - y}{y - \theta_2^{(0)}} \rangle = \theta_3 + \theta_4 x \; .$$

We speak of a **conditional linearizable** function. The linear regression leads to the initial value $\theta_3^{(0)} = 1.83$ and $\theta_4^{(0)} = -0.36$.

With this initial value the algorithm converges to the solution $\widehat{\theta}_1 = 163.7$, $\widehat{\theta}_2 = 159.8$, $\widehat{\theta}_3 = 2.675$ and $\widehat{\theta}_4 = -0.512$. The functions $h\langle \cdot;\underline{\theta}^{(0)}\rangle$ and $h\langle \cdot;\widehat{\underline{\theta}}\rangle$ are shown in Figure 2.j(b).

* The property of conditional linearity of a function can also be useful for developing an algorithm specially suited for this situation (see e.g. Bates and Watts, 1988).

3. Approximate Tests and Confidence Intervals

a The estimator $\widehat{\underline{\theta}}$ gives the value of $\underline{\theta}$ that fits the data optimally. We now ask *which parameter values $\underline{\theta}$ are compatible with the observations.* The **confidence region** is

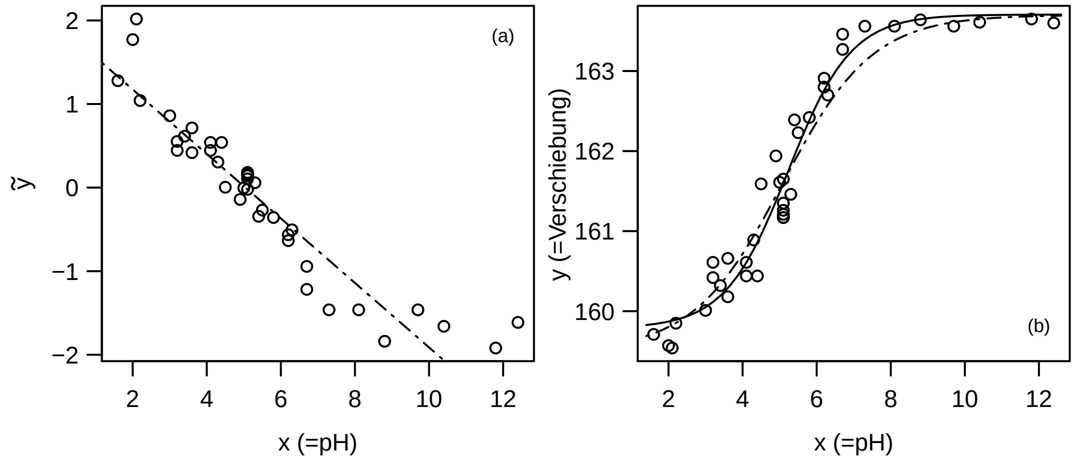

Figure 2.j: Membrane Separation Technology Example. (a) Regression line, which is used for determining the initial values for θ_3 and θ_4. (b) Regression function $h\langle x; \underline{\theta}\rangle$ for the initial value $\underline{\theta} = \underline{\theta}^{(0)}$ (- - - - -) and for the least squares estimation $\underline{\theta} = \widehat{\underline{\theta}}$ (———).

the set of all these values. For an individual parameter θ_j the confidence region is the **confidence interval**.

The results that now follow are based on the fact that the estimator $\widehat{\underline{\theta}}$ is asymptotically multivariate normally distributed. For an individual parameter that leads to a “z-Test” and the corresponding confidence interval; for several parameters the corresponding Chi-Square test works and gives elliptical confidence regions.

b The **asymptotic properties** of the estimator can be derived from the linear approximation. The problem of nonlinear regression is indeed approximately equal to the linear regression problem mentioned in 2.d

$$\underline{\widetilde{Y}} = \boldsymbol{A}\left\langle \underline{\theta}^*\right\rangle \underline{\beta} + \underline{E}\,,$$

if the parameter vector $\underline{\theta}^*$, which is used for the linearization lies near to the solution. If the estimation procedure has converged (i.e. $\underline{\theta}^* = \widehat{\underline{\theta}}$), then $\underline{\beta} = 0$ – otherwise this would not be the solution. The standard error of the coefficients $\underline{\beta}$ – and more generally the covariance matrix of $\widehat{\underline{\beta}}$ – then correspond approximately to the corresponding values for $\widehat{\underline{\theta}}$.

* A bit more precisely: The standard errors characterize the uncertainties that are generated by the random fluctuations in the data. The available data have led to the estimation value $\widehat{\underline{\theta}}$. If the data were somewhat different, then $\widehat{\underline{\theta}}$ would still be approximately correct, thus we accept that it is good enough for the linearization. The estimation of $\underline{\beta}$ for the new data set would thus lie as far from the estimated value for the available data, as this corresponds to the distribution of the parameter in the linearized problem.

c **Asymptotic Distribution of the Least Squares Estimator.** From these considerations it follows: Asymptotically the least squares estimator $\widehat{\underline{\theta}}$ is normally distributed (and consistent) and therefore

$$\widehat{\underline{\theta}} \overset{a}{\sim} \mathcal{N}\left\langle \underline{\theta}, \frac{\boldsymbol{V}\left\langle\underline{\theta}\right\rangle}{n} \right\rangle\,,$$

with asymptotic covariance matrix $\boldsymbol{V}\left\langle\underline{\theta}\right\rangle = \sigma^2(\boldsymbol{A}\left\langle\underline{\theta}\right\rangle^T \boldsymbol{A}\left\langle\underline{\theta}\right\rangle)^{-1}$, where $\boldsymbol{A}\left\langle\underline{\theta}\right\rangle$ is the $n \times p$ matrix of the partial derivatives (see 2.d).

To determine the covariance matrix $V\langle\underline{\theta}\rangle$ explicitly, $\boldsymbol{A}\langle\underline{\theta}\rangle$ is calculated at the point $\widehat{\underline{\theta}}$ instead of the unknown point $\underline{\theta}$, and for the error variance σ^2 the usual estimator is plugged

$$\widehat{\boldsymbol{V}\langle\underline{\theta}\rangle} = \widehat{\sigma}^2 \left(\boldsymbol{A}\left\langle\widehat{\underline{\theta}}\right\rangle^T \boldsymbol{A}\left\langle\widehat{\underline{\theta}}\right\rangle\right)^{-1} \quad \text{mit } \widehat{\sigma}^2 = \frac{S\langle\widehat{\underline{\theta}}\rangle}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \eta_i\left\langle\widehat{\underline{\theta}}\right\rangle\right)^2 .$$

With this the distribution of the estimated parameters is approximately determined, from which, like in linear regression, standard error and confidence intervals can be derived, or confidence ellipses (or ellipsoids) if several variables are considered at once.

The denominator $n-p$ in $\widehat{\sigma}^2$ is introduced in linear regression to make the estimator unbiased. – Tests and confidence intervals are not determined with the normal and chi-square distribution, but with the **t and F distributions**. There it is taken into account that the estimation of σ^2 causes an additional random fluctuation. Even if the distribution is no longer exact, the approximations get more exact if we do this in nonlinear regression. Asymptotically the difference goes to zero.

Example d **Membrane Separation Technology.** A computer output for the Membrane Separation example shows Table 3.d. The estimations of the parameters are in the column ”Value”, followed by the estimated approximate standard error and the test statistics (”t value”), that are approximately t_{n-p} distributed. In the last row the estimated standard deviation $\widehat{\sigma}$ of the random error E_i is given.

From this output, in linear regression the confidence intervals for the parameters can be determined: The approximate 95% confidence interval for the parameter θ_1 is

$$163.706 \pm q_{0.975}^{t_{35}} \cdot 0.1262 = 163.706 \pm 0.256 \ .$$

```
Formula: delta ~ (T1 + T2 * 10^(T3 + T4 * pH)) / (10^(T3 + T4 * pH) + 1)
Parameters:
     Estimate  Std. Error   t value   Pr(> |t|)
 T1  163.7056      0.1262  1297.256   < 2e-16
 T2  159.7846      0.1594  1002.194   < 2e-16
 T3    2.6751      0.3813     7.015  3.65e-08
 T4   -0.5119      0.0703    -7.281  1.66e-08

Residual standard error: 0.2931 on 35 degrees of freedom
Number of iterations to convergence: 7
Achieved convergence tolerance: 5.517e-06
```

Table 3.d: Membrane Separation Technology Example: R summary of the fitting.

Example e **Puromycin.** For checking the influence of treating an enzyme with Puromycin of the postulated form (1.d) a general model for the data with and without the treatment can be formulated as follows:

$$Y_i = \frac{(\theta_1 + \theta_3 z_i) x_i}{\theta_2 + \theta_4 z_i + x_i} + E_i \ .$$

Where z is the indicator variable for the treatment ($z_i = 1$, if treated, otherwise =0). Table 3.e shows that the parameter θ_4 at the 5% level is not significantly different from 0, since the P value of 0.167 is larger then the level (5%). However, the treatment has a clear influence, which is expressed through θ_3; the 95% confidence interval covers the region $52.398 \pm 9.5513 \cdot 2.09 = [32.4, 72.4]$ (the value 2.09 corresponds to the 0.975 quantile of the t_{19} distribution).

```
Formula: velocity ∼ (T1 + T3 * (treated == T)) * conc/(T2 + T4 * (treated
== T) + conc)
Parameters:
     Estimate  Std. Error  t value  Pr(> |t|)
 T1   160.280       6.896   23.242   2.04e-15
 T2     0.048       0.008    5.761   1.50e-05
 T3    52.404       9.551    5.487   2.71e-05
 T4     0.016       0.011    1.436      0.167
Residual standard error: 10.4 on 19 degrees of freedom
Number of iterations to convergence: 6
Achieved convergence tolerance: 4.267e-06
```

Table 3.e: R summary of the fit for the Puromycin example.

f **Confidence Intervals for Function Values.** Besides the parameters, the function value $h\left\langle \underline{x}_0, \underline{\theta} \right\rangle$ for a given $\underline{x}_0$ is of interest. In linear regression the function value $h\left\langle \underline{x}_0, \underline{\beta} \right\rangle = \underline{x}_0^T \underline{\beta} =: \eta_0$ is estimated by $\widehat{\eta}_0 = \underline{x}_0^T \widehat{\underline{\beta}}$ and the estimated $(1-\alpha)$ confidence interval for it is

$$\widehat{\eta}_0 \pm q^{t_{n-p}}_{1-\alpha/2} \cdot \mathrm{se}\left\langle \widehat{\eta}_0 \right\rangle \quad \text{with} \ \ \mathrm{se}\left\langle \widehat{\eta}_0 \right\rangle = \widehat{\sigma} \sqrt{\underline{x}_o^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \underline{x}_o} \ .$$

With analogous considerations and asymptotic approximation we can specify confidence intervals for the function values $h\left\langle \underline{x}_0; \underline{\theta} \right\rangle$ for nonlinear h. If the function $\eta_0\left\langle \widehat{\underline{\theta}} \right\rangle := h\left\langle x_0, \widehat{\underline{\theta}} \right\rangle$ is approximated at the point $\underline{\theta}$, we get

$$\eta_o\langle \widehat{\underline{\theta}} \rangle \approx \eta_o\langle \underline{\theta} \rangle + \underline{a}_o^T \, (\widehat{\underline{\theta}} - \underline{\theta}) \quad \text{mit } \underline{a}_o = \frac{\partial h\langle x_o, \underline{\theta} \rangle}{\partial \underline{\theta}} \ .$$

(If $\underline{x}_0$ is equal to an observed $\underline{x}_i$, then $\underline{a}_0$ equals the corresponding row of the matrix $\boldsymbol{A}$ from 2.d.) The confidence interval for the function value $\eta_0\left\langle \underline{\theta} \right\rangle := h\left\langle \underline{x}_0, \underline{\theta} \right\rangle$ is then approximately

$$\eta_0\left\langle \widehat{\underline{\theta}} \right\rangle \pm q^{t_{n-p}}_{1-\alpha/2} \cdot \mathrm{se}\left\langle \eta_0\left\langle \widehat{\underline{\theta}} \right\rangle \right\rangle \qquad \text{mit } \mathrm{se}\left\langle \eta_0\left\langle \widehat{\underline{\theta}} \right\rangle \right\rangle = \widehat{\sigma} \sqrt{\widehat{\underline{a}}_o^T \left(\boldsymbol{A}\left\langle \widehat{\underline{\theta}} \right\rangle^T \boldsymbol{A}\left\langle \widehat{\underline{\theta}} \right\rangle \right)^{-1} \widehat{\underline{a}}_o}.$$

In this formula, again the unknown values are replaced by their estimations.

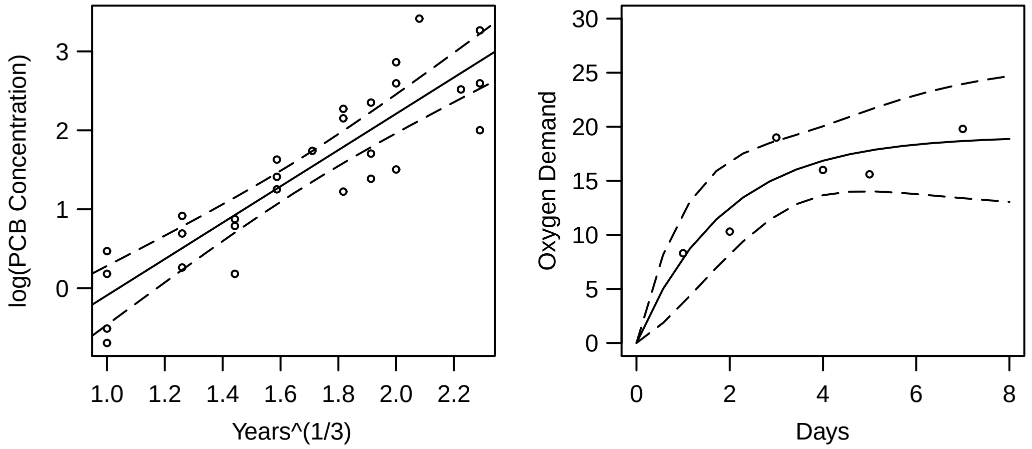

Figure 3.g: Left: Confidence band for an estimated line for a linear problem. Right: Confidence band for the estimated curve $h\langle x, \underline{\theta}\rangle$ in the oxygen consumption example.

g **Confidence Band.** The expression for the $(1-\alpha)$ confidence interval for $\eta_o\langle\underline{\theta}\rangle := h\langle x_o, \underline{\theta}\rangle$ also holds for arbitrary x_o. As in linear regression, it is obvious to represent the limits of these intervals as a "confidence band" that is a function of x_o, as this Figure 3.g shows for the two examples of Puromycin and oxygen consumption.

Confidence bands for linear and nonlinear regression functions behave differently: For linear functions this confidence band is thinnest by the center of gravity of the explanatory variables and gets gradually wider as it move out (see Figure 3.g, left). In the nonlinear case, the bands can be arbitrary. Because the functions in the “Puromycin” and “Oxygen Consumption’ must go through zero, the interval shrinks to a point there. Both models have a horizontal asymptote and therefore the band reaches a constant width for large x (see Figure 3.g, right) .

h **Prediction Interval.** The considered confidence band indicates where the **ideal function values** $h\langle x\rangle$, and thus the expected values of Y for given$x$, lie. The question, in which region **future observations** Y_0 for given $\underline{x}_0$ will lie, is not answered by this. However, this is often more interesting than the question of the ideal function value; for example, we would like to know in which region the measured value of oxygen consumption would lie for an incubation time of 6 days.

Such a statement is a prediction about a **random variable** and is different in principle from a confidence interval, which says something about a **parameter**, which is a fixed but unknown number. Corresponding to the question posed, we call the region we are now seeking a **prediction interval** or **prognosis interval**. More about this in Chapter 7.

i **Variable Selection.** In nonlinear regression, unlike linear regression, variable selection is not an important topic, because

- a variable does not correspond to each parameter, so usually the number of parameters is different than the number of variables,
- there are seldom problems where we need to clarify whether an explanatory variable is necessary or not – the model is derived from the subject theory.

However, there is sometimes a reasonable question of whether a portion of the parameters in the nonlinear regression model can appropriately describe the data (see Beispiel Puromycin).

4. More Precise Tests and Confidence Intervals

a The quality of the approximate confidence region depends strongly on the quality of the linear approximation. Also the convergence properties of the optimization algorithms are influenced by the quality of the linear approximation. With a somewhat larger computational effort, the linearity can be checked graphically and, at the same time, we get a more precise confidence interval.

b **F Test for Model Comparison.** To test a null hypothesis $\underline{\theta} = \underline{\theta}^*$ for the whole parameter vector or also $\theta_j = \theta_j^*$ for an individual component, we can use an **F-Test for model comparison** like in linear regression. Here, we compare the sum of squares $S\langle\underline{\theta}^*\rangle$ that arises under the null hypothesis with the sum of squares $S\langle\widehat{\underline{\theta}}\rangle$. (For $n \to \infty$ the F test is the same as the so-called Likelihood Quotient test, and the sum of squares is, up to a constant, equal to the log likelihood.)

Now we consider the null hypothesis $\underline{\theta} = \underline{\theta}^*$ for the whole parameter vector. The test statistic is

$$T = \frac{n-p}{p} \frac{S\langle\underline{\theta}^*\rangle - S\langle\widehat{\underline{\theta}}\rangle}{S\langle\widehat{\underline{\theta}}\rangle} \overset{a}{\sim} F_{p,n-p} \,.$$

From this we get a confidence region

$$\left\{ \underline{\theta} \;\middle|\; S\langle\underline{\theta}\rangle \le S\langle\widehat{\underline{\theta}}\rangle \left(1 + \tfrac{p}{n-p}\, q\right) \right\}$$

where $q = q_{1-\alpha}^{F_{p,n-p}}$ is the $(1-\alpha)$ quantile of the F distribution with p and $n-p$ degrees of freedom.

In linear regression we get the same exact confidence region if we use the (multivariate) normal distribution of the estimator $\widehat{\underline{\beta}}$. In the nonlinear case the results are different. The region that is based on the F tests is not based on the linear approximation in 2.d and is thus (much) more exact.

c **Exact Confidence Regions for p=2.** If $p = 2$, we can find the exact confidence region by calculating $S\langle\underline{\theta}\rangle$ on a grid of $\underline{\theta}$ values and determine the borders of the region through interpolation, as is familiar for contour plots. In Figure 4.c are given the contours together with the elliptical regions that result from linear approximation for the **Puromycin** example (left) and the **oxygen consumption** example (right).

For $p > 2$ contour plots do not exist. In the next chapter we will be introduced to graphical tools that also work for higher dimensions. They depend on the following concepts.

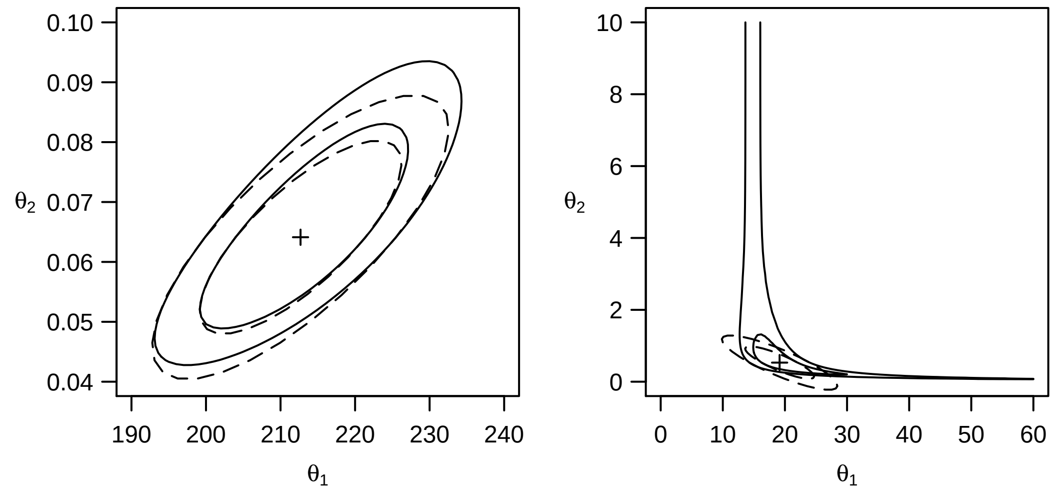

Figure 4.c: Nominal 80 and 95% likelihood contures (——) and the confidence ellipses from the asymptotic approximation (– – – –). + denotes the least squares solution. In the Puromycin example (left) the agreement is good and in the oxygen consumption example (right) it is bad.

d **F Test for Individual Parameters.** It should be checked whether an individual parameter θ_k can be equal to a certain value θ_k^*. Such a null hypothesis makes no statement about the other parameters. The model that corresponds to the null hypothesis that fits the data best is determined at a fixed $\theta_k = \theta_k^*$ through a least squares estimation of the remaining parameters. So, $S\langle\theta_1, \ldots, \theta_k^*, \ldots, \theta_p\rangle$ is minimized with respect to θ_j, $j \neq k$. We denote the minimum with $\widetilde{S}_k$ and the value θ_j that leads to it as $\widetilde{\theta}_j$. Both values depend on θ_k^*. We therefore write $\widetilde{S}_k\langle\theta_k^*\rangle$ and $\widetilde{\theta}_j\langle\theta_k^*\rangle$.

The F test statistics for the test "$\theta_k = \theta_k^*$" is

$$\widetilde{T}_k = (n-p)\,\frac{\widetilde{S}_k\langle\theta_k^*\rangle - S\left\langle\underline{\widehat{\theta}}\right\rangle}{S\left\langle\underline{\widehat{\theta}}\right\rangle}\ .$$

It has an (approximate) $F_{1,n-p}$ distribution.

We get a confidence interval from this by solving the equation $\widetilde{T}_k = q_{0.95}^{F_{1,n-p}}$ numerically for θ_k^*. It has a solution that is smaller than $\widehat{\theta}_k$ and one that is larger.

e **t Test via F Test.** In linear regression and in the previous chapter we have calculated tests and confidence intervals from a test value that follows a t-distribution (t-test for the coefficients). Is this another test?

It turns out that the test statistic of the t-test in linear regression turns into the test statistic of the F-test if we square it, and both tests are equivalent. In nonlinear regression, the F-test is not equivalent with the t-test discussed in the last chapter (3.d). However, we can transform the F-test into a t-test that is more precise than that of the last chapter:

From the test statistics of the F-tests, we drop the root and provide then with the signs of $\widehat{\theta}_k - \theta_k^*$,

$$T_k\langle\theta_k^*\rangle := \operatorname{sign}\left\langle\widehat{\theta}_k - \theta_k^*\right\rangle\,\frac{\sqrt{\widetilde{S}_k\langle\theta_k^*\rangle - S\left\langle\underline{\widehat{\theta}}\right\rangle}}{\widehat{\sigma}}\ .$$

(sign $\langle a\rangle$ denotes the sign of a, and is $\widehat{\sigma}^2 = S\left\langle\underline{\widehat{\theta}}\right\rangle/(n-p)$.) This test statistic is (approximately) t_{n-p} distributed.

In the linear regression model, T_k, is, as mentioned, equal to the test statistic of the usual t-test,

$$T_k\left\langle\theta_k^*\right\rangle = \frac{\widehat{\theta}_k - \theta_k^*}{\mathrm{se}\left\langle\widehat{\theta}_k\right\rangle}\ .$$

f **Confidence Intervals for Function Values via F-Test.** With this technique we can also determine confidence intervals for a function value at a point x_o. For this we reparameterize the original problem so that a parameter, say ϕ_1, represents the function value $h\langle x_o\rangle$ and proceed as in 4.d.

5. Profile t-Plot and Profile Traces

a **Profile t-Function and Profile t-Plot.** The graphical tools for checking the linear approximation are based on the just discussed t-test, that actually doesn't use this approximation. We consider the test statistic T_k (4.e) as a function of its arguments θ_k and call it **profile t-function** (in the last chapter the arguments were denoted with θ_k^*, now for simplicity we leave out the *). For linear regression we get, as is apparent from 4.e, a line, while for nonlinear regression the result is a monotone increasing function. The graphical comparison of $T_k\langle\theta_k\rangle$ with a line enables the so-called **profile t-plot**. Instead of θ_k, it is common to use a standardized version

$$\delta_k\left\langle\theta_k\right\rangle := \frac{\theta_k - \widehat{\theta}_k}{se\left\langle\widehat{\theta}_k\right\rangle}$$

on the horizontal axis because of the linear approximation. The comparison line is then the "diagonal", so the line with slope 1 and intercept 0.

The more strongly the profile t-function is curved, the stronger is the nonlinearity in a neighborhood of θ_k. Therefore, this representation shows how good the linear approximation is in a neighborhood of $\widehat{\theta}_k$. (The neighborhood that is statistically important is approximately determined by $|\delta_k\langle\theta_k\rangle| \leq 2.5$.) In Figure 5.a it is apparent that in the Puromycin example the nonlinearity is minimal, but in the oxygen consumption example it is large.

From the illustration we can read off the confidence intervals according to 4.e. For convenience, on the right vertical axis are marked the probabilites $P\left\langle T_k \leq t\right\rangle$ according to the t-distribution. In the oxygen consumption example, this gives a confidence interval without an upper bound!

Example b **from Membrane Separation Technology.** As 5.a shows, from the profile t-plot we can graphically read out corresponding confidence intervals that are based on the profile t-function. The R function `confint(...)` numerically calculates the desired confidence interval on the basis of the profile t-function.In Table 5.b is shown the corresponding R output from the membrane separation example. In this case, no large differences from the classical calculation method are apparent.

```
> confint(Mem.fit, level=0.95)
Waiting for profiling to be done...
            2.5%        97.5%
 T1  163.4661095  163.9623685
 T2  159.3562568  160.0953953
 T3    1.9262495    3.6406832
 T4   -0.6881818   -0.3797545
```

Table 5.b: Membrane separation technology example: R output for the confidence intervals that are based on the profile t-function.

c **Likelihood Profile Traces.** The **likelihood profile traces** are another useful tool. Here the estimated parameters $\widetilde{\theta}_j$, $j \neq k$ at fixed θ_k (see 4.d) are considered as functions $\widetilde{\theta}_j^{(k)}\langle\theta_k\rangle$ of these values.

The graphical representation of these functions would fill a whole matrix of diagrams, but without diagonals. It is worthwhile to combine the "opposite" diagrams of this matrix: Over the representation of $\widetilde{\theta}_j^{(k)}\langle\theta_k\rangle$ we superimpose $\widetilde{\theta}_k^{(j)}\langle\theta_j\rangle$ – in mirrored form, so that the axes have the same meaning for both functions.

In Figure 5.c ist shown each of these diagrams for our two examples. Additionally, are shown contours of confidence regions for $[\theta_1, \theta_2]$. We see that the profile traces cross the contours at points of contact of the horizontal and vertical tangents.

The representation shows not only the nonlinearities, but also holds useful clues for **how the parameters influence each other**. To understand this, we now consider the case of a linear regression function. The profile traces in the individual diagrams then consist of two lines, that cross at the point $[\widehat{\theta}_1, \widehat{\theta}_2]$. We standardize the parameter by using $\delta_k\langle\theta_k\rangle$ from 5.a, so we can show that the slope of the trace $\widetilde{\theta}_j^{(k)}\langle\theta_k\rangle$ is equal to the correlation coefficient c_{kj} of the estimated coefficients $\widehat{\theta}_j$ and $\widehat{\theta}_k$. The "reverse

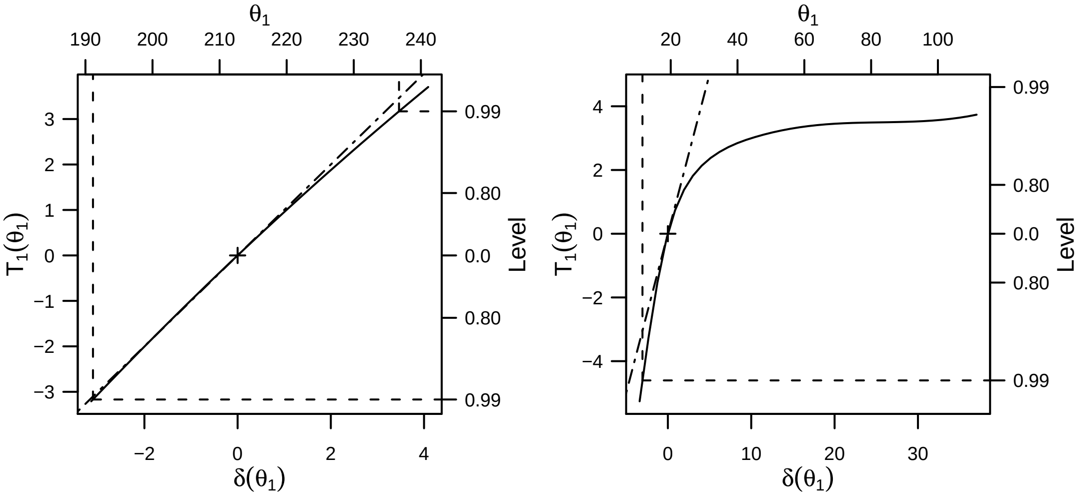

Figure 5.a: Profile t-plot for the first parameter is each of the Puromycin and oxygen consumption examples. The dashed lines show the applied linear approximation and the dotted line the construction of the 99% confidence interval with the help of $T_1\langle\theta_1\rangle$.

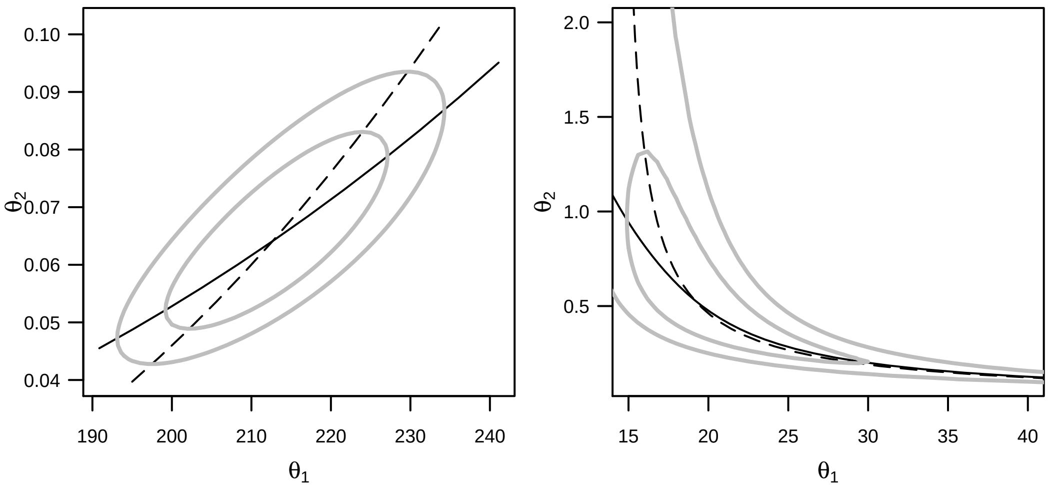

Figure 5.c: Likelihood profile traces for the Puromycin and oxygen consumption examples, with 80%- and 95% confidence regions (gray curves).

trace" $\widetilde{\theta}_k^{(j)}\langle\theta_j\rangle$ then has, compared with the horizontal axis, a slope of $1/c_{kj}$. The angle that the lines enclose is thus a monotone function of this correlation. It therefore measures the **collinearity** between the two predictor variables. If the correlation between the parameter estimations is zero, then the traces are parallel to each other.

For a nonlinear regression function, both traces are curved. The angle between them still shows how strongly the two parameters θ_j and θ_k hold together, so their estimations are correlated.

Example d **Membrane Separation Technology.** All profile t-plots and profile traces can be assembled into a triangle matrix of diagrams, as Figure 5.d shows for the membrane separation technology example.

Most profile traces are strongly curved, i.e. the regression function tends to a strong nonlinearity around the estimated parameter values. Even though the profile traces for θ_3 and θ_4 are lines, a further problem is apparent: The profile traces lie on top of each other! This means that the parameters θ_3 and θ_4 are extremely strongly collinear. Parameter θ_2 is also collinear with θ_3 and θ_4, although more weakly.

e * **Good Approximation of Two Dimensional Likelihood Contours.** The profile traces can be used to construct very accurate approximations for two dimensional projections of the likelihood contours (see Bates and Watts, 1988). Their calculation is computationally less difficult than for the corresponding exact likelihood contours.

6. Parameter Transformations

a **Parameter transformations** are primarily used to improve the linear approximation and therefore improve the convergence behavior and the **quality of the confidence interval**.

It is here expressly noted that parameter transformations, unlike transformations of the variable of interest (see 1.h) does *not* change the statistical part of the model. So, they are not helpful if the assumptions about the distribution of the random deviations

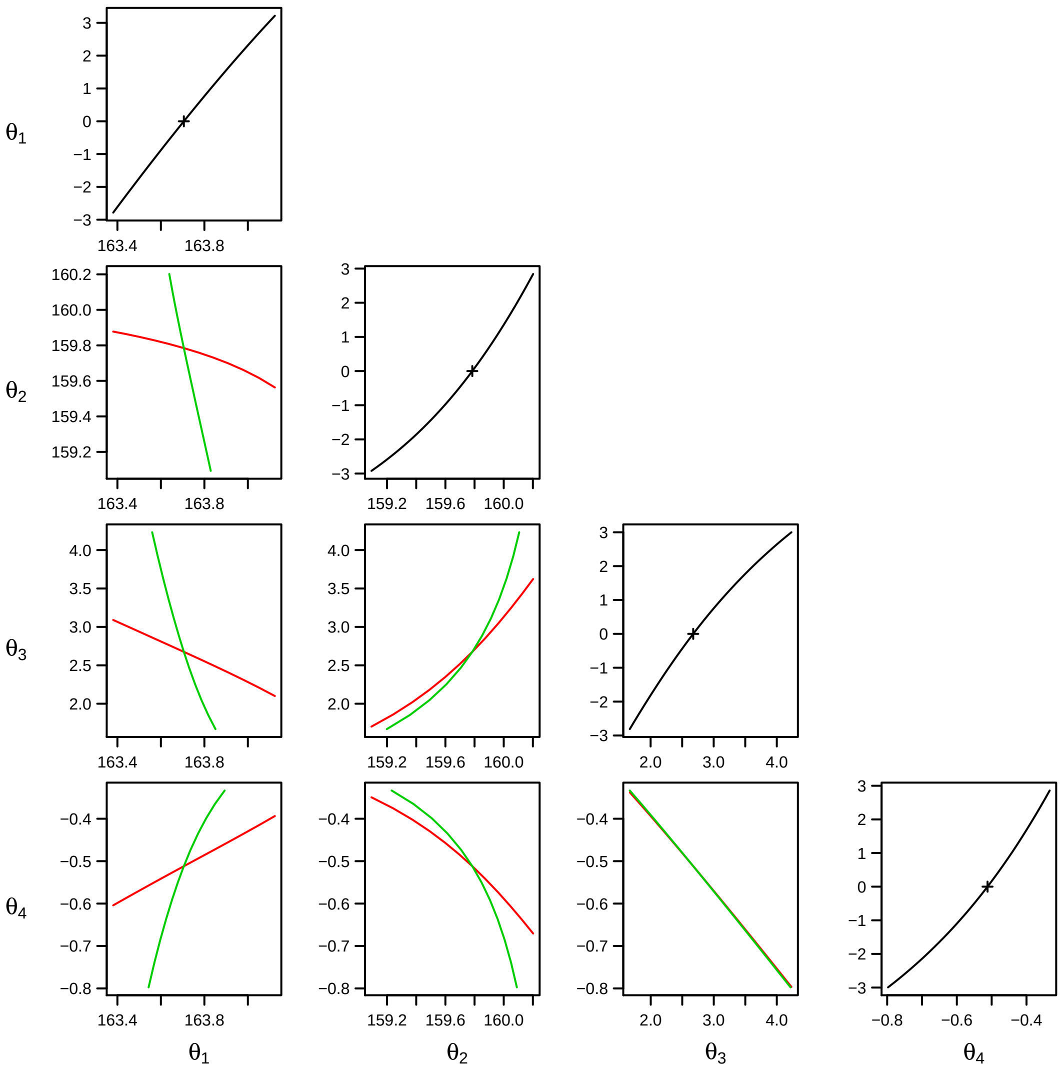

Figure 5.d: Profile t-plots and Profile Traces for the Example from Membrane Separation Technology. The + in the profile t-plot denotes the least squares solution.

are violated. It is the quality of the linear approximation and the statistical statements based on it that are changed.

Often, the transformed parameters are very difficult **to interpret** for the application. The important questions often concern individual parameters – the original parameters. Despite this, we can work with transformations: We derive more accurate confidence regions for the transformed parameters and back transform these to get results for the original variables.

b **Restricted Parameter Regions.** Often the permissible region of a parameter is restricted, e.g. because the regression function is only defined for positive values of a parameter. Usually, such a constraint is ignored to begin with and we wait to see whether and where the algorithm converges. According to experience, the parameter estimation will end up in a reasonable range if the model describes the data well and the data give enough information for determining the parameter.

Sometimes, though, problems occur in the course of the computation, especially if the parameter value that best fits the data lies near the edge of the permissible range. The simplest way to deal with such problems is via transformation of the parameter.

Beispiele:

- The parameter θ should be positive. Through a transformation $\theta \longrightarrow \phi = \ln\langle\theta\rangle$, $\theta = \exp\langle\phi\rangle$ is always positive for all possible values of $\phi \in \mathbb{R}$:
$$h\langle x, \theta\rangle \;\longrightarrow\; h\langle x, \exp\langle\phi\rangle\rangle$$
- The parameter should lie in the interval (a, b). With the log transformation $\theta = a + (b-a)/(1+\exp\langle-\phi\rangle)$, θ can, for arbitrary $\phi \in \mathbb{R}$, only take values in (a, b).
- In the model
$$h\langle x, \underline{\theta}\rangle = \theta_1 \exp\langle-\theta_2 x\rangle + \theta_3 \exp\langle-\theta_4 x\rangle$$
with $\theta_2, \theta_4 > 0$ the parameter pairs (θ_1, θ_2) and (θ_3, θ_4) are interchangeable, i.e. $h\langle x, \underline{\theta}\rangle$ does not change. This can create uncomfortable optimization problems, because, for one thing, the solution is not unique. The constraint $0 < \theta_2 < \theta_4$ that ensures the uniqueness is achieved via the transformation $\theta_2 = \exp\langle\phi_2\rangle$ und $\theta_4 = \exp\langle\phi_2\rangle(1+\exp\langle\phi_4\rangle)$. The function is now
$$h\langle x, (\theta_1, \phi_2, \theta_3, \phi_4)\rangle = \theta_1 \exp\left\langle -\exp\langle\phi_2\rangle x\right\rangle + \theta_3 \exp\left\langle -\exp\langle\phi_2\rangle(1+\exp\langle\phi_4\rangle) x\right\rangle \ .$$

c **Parameter Transformation for Collinearity.** A simultaneous variable and parameter transformation can be helpful to weaken **collinearity** in the partial derivative vectors. So, for example, the model $h\langle x, \underline{\theta}\rangle = \theta_1 \exp\langle-\theta_2 x\rangle$ has the derivatives

$$\frac{\partial h}{\partial \theta_1} = \exp\langle-\theta_2 x\rangle \ , \qquad \frac{\partial h}{\partial \theta_2} = -\theta_1 x \exp\langle-\theta_2 x\rangle \ .$$

If all x values are positive, both vectors

$$\begin{aligned} \underline{a}_1 &:= (\exp\langle-\theta_2 x_1\rangle, \ldots, \exp\langle-\theta_2 x_n\rangle)^T \\ \underline{a}_2 &:= (-\theta_1 x_1 \exp\langle-\theta_2 x_1\rangle, \ldots, -\theta_1 x_n \exp\langle-\theta_2 x_n\rangle)^T \end{aligned}$$

tend to disturbing collinearity. This collinearity can be avoided through **centering**. The model can be written as $h\left\langle x, \underline{\theta}\right\rangle = \theta_1 \exp\left\langle-\theta_2(x - x_0 + x_0)\right\rangle$ With the reparameterization $\phi_1 := \theta_1 \exp\left\langle-\theta_2 x_0\right\rangle$ and $\phi_2 := \theta_2$ we get

$$h\left\langle x, \underline{\phi}\right\rangle = \phi_1 \exp\left\langle-\phi_2(x - x_0)\right\rangle \ .$$

The derivative vectors become approximately orthogonal if for x_0 the mean value of the x_i is chosen.

	θ_1	θ_2	θ_3		θ_1	θ_2	$\widetilde{\theta}_3$
θ_2	-0.256			θ_2	-0.256		
θ_3	-0.434	0.771		$\widetilde{\theta}_3$	0.323	0.679	
θ_4	0.515	-0.708	-0.989	θ_4	0.515	-0.708	-0.312

Table 6.d: Correlation matrices for the Membrane Separation Technology for the original parameters (left) and the transformed parameters $\widetilde{\theta}_3$ (right).

Example d **Membrane Separation Technology.** In this example it is apparent from the approximate correlation matrix (Table 6.d, left half) that the parameters θ_3 and θ_4 are strongly correlated. (We have already found this in 5.d from the profile traces).

If the model is reparameterized to

$$y_i = \frac{\theta_1 + \theta_2 \, 10^{\widetilde{\theta}_3 + \theta_4 (x_i - \mathrm{med}\langle x_j \rangle)}}{1 + 10^{\widetilde{\theta}_3 + \theta_4 (x_i - \mathrm{med}\langle x_j \rangle)}} + E_i, \; i = 1 \ldots n$$

with $\widetilde{\theta}_3 = \theta_3 + \theta_4 \, \mathrm{med}\langle x_j \rangle$, an improvement is achieved (right half of Table 6.d).

e **Reparameterization.** In Chapter 5 we have presented means for graphical evaluation of the linear approximation. If the approximation is considered in adequate we would like to improve it. An appropriate reparameterization can contribute to this. So, for example, for the model

$$h\langle \underline{x}, \underline{\theta} \rangle = \frac{\theta_1 \theta_3 (x_2 - x_3)}{1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_3}$$

the reparameterization

$$h\langle \underline{x}, \underline{\phi} \rangle = \frac{x_2 - x_3}{\phi_1 + \phi_2 x_1 + \phi_3 x_2 + \phi_4 x_3}$$

is recommended by (Ratkowsky, 1985). (Also see exercises)

Example f **from Membrane Separation Technology.** The parameter transformation given in 6.d leads to a satisfactory result, as far as the correlation is concerned. We look at the likelihood contours or the profile t-plot and the profile traces, and the parametrization is still not satisfactory.

An intensive search for further improvements leads to the following transformations, for which the profile traces turn out satisfactorily (Figure 6.f):

$$\widetilde{\theta}_1 := \frac{\theta_1 + \theta_2 \, 10^{\widetilde{\theta}_3}}{10^{\widetilde{\theta}_3} + 1}, \qquad \widetilde{\theta}_2 := \log_{10}\left(\frac{\theta_1 - \theta_2}{10^{\widetilde{\theta}_3} + 1} \, 10^{\widetilde{\theta}_3} \right),$$

$$\widetilde{\theta}_3 := \theta_3 + \theta_4 \, \mathrm{med}\langle x_j \rangle \qquad \widetilde{\theta}_4 := 10^{\theta_4} .$$

The model is then

$$Y_i = \widetilde{\theta}_1 + 10^{\widetilde{\theta}_2} \, \frac{1 - \widetilde{\theta}_4^{\,(x_i - \mathrm{med}\langle x_j))\rangle}}{1 + 10^{\widetilde{\theta}_3} \, \widetilde{\theta}_4^{\,(x_i - \mathrm{med}\langle x_j \rangle)}} + E_i \, .$$

and we get the result shown in Table 6.f .

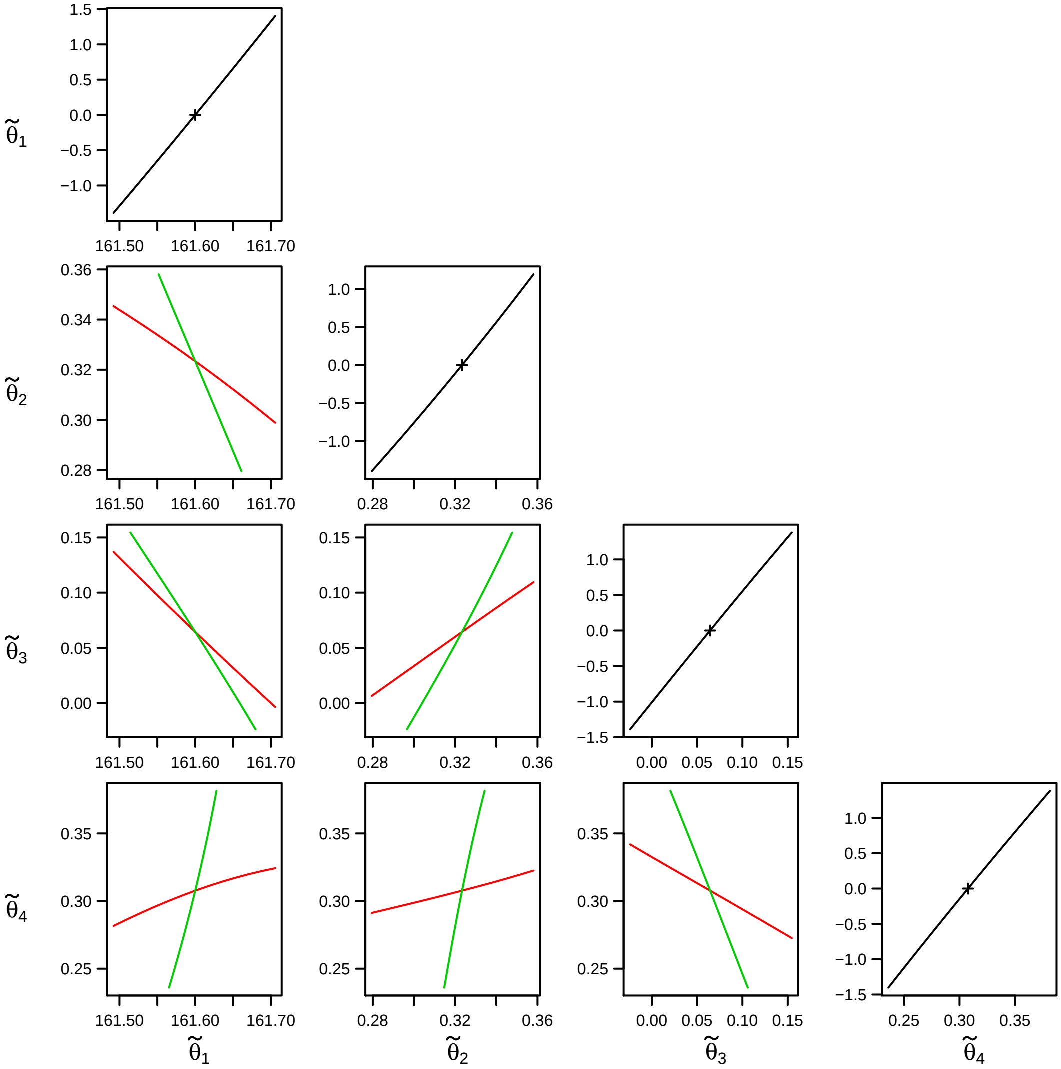

Figure 6.f: Profile t-plot and profile traces for the membrane separation technology example according to the given transformations.

g **More Successful Reparametrization.** It is apparent that a **successful reparametrization of the data set depends**, for one thing, that the nonlinearties and correlations between estimated parameters depend on the estimated value $\underline{\widehat{\theta}}$ itself. Therefore, no generally valid recipe can be given, which makes the search for appropriate reparametrizations often very tiresome.

h **Failure of Gaussian Error Propagation.** Even if a parameter transformation helps us deal with difficulties with the convergence behavior of the algorithm or the quality of the confidence intervals, the **original parameters** often have a physical interpretation. We take the simple transformation example $\theta \longrightarrow \phi = \ln\langle\theta\rangle$ from 6.b. The fitting of the model opens with an estimator $\widehat{\phi}$with estimated standard error $\widehat{\sigma}_{\widehat{\phi}}$. An obvious estimator for θ is then $\widehat{\theta} = \exp\langle\widehat{\phi}\rangle$. The standard error for $\widehat{\theta}$ can be determined with

```
Formula: delta ~ TT1 + 10^TT2 * (1 - TT4^pHR)/(1 + 10^TT3 * TT4^pHR)
Parameters:
       Estimate  Std. Error   t value  Pr(> |t|)
 TT1  161.60008     0.07389  2187.122   < 2e-16
 TT2    0.32336     0.03133    10.322  3.67e-12
 TT3    0.06437     0.05951     1.082     0.287
 TT4    0.30767     0.04981     6.177  4.51e-07

Residual standard error: 0.2931 on 35 degrees of freedom
Correlation of Parameter Estimates:
        TT1   TT2    TT3
 TT2  -0.56
 TT3  -0.77  0.64
 TT4   0.15  0.35  -0.31

Number of iterations to convergence: 5
Achieved convergence tolerance: 9.838e-06
```

Table 6.f: Membrane separation technology: R summary of the fit after the parameter transformation.

the help of the Gaussian error propagation law (see Stahel, 2000, Abschnitt 6.10):

$$\widehat{\sigma}^2_{\widehat{\theta}} \approx \left(\frac{\partial \exp\langle\phi\rangle}{\partial\phi}\bigg|_{\phi=\widehat{\phi}} \right)^2 s^2_{\widehat{\phi}} = \left(\exp\langle\widehat{\phi}\rangle\right)^2 \widehat{\sigma}^2_{\widehat{\phi}}$$

or

$$\widehat{\sigma}_{\widehat{\theta}} \approx \exp\langle\widehat{\phi}\rangle \, \widehat{\sigma}_{\widehat{\phi}} \, .$$

From this we get the approximate 95% confidence interval for θ:

$$\exp\langle\widehat{\phi}\rangle \pm \widehat{\sigma}_{\widehat{\theta}} \, q_{0.975}^{t_{n-p}} = \exp\langle\widehat{\phi}\rangle \left(1 \pm \widehat{\sigma}_{\widehat{\phi}} \, q_{0.975}^{t_{n-p}}\right) \, .$$

The Gaussian error propagation law is based on the linearization of the transformation function $g\langle\cdot\rangle$; concretely on $\exp\langle\cdot\rangle$. We have carried out the parameter transformation because the quality of the confidence intervals left a lot to be desired, then unfortunately this linearization negates what has been achieved and we are back where we started before the transformation.

The correct approach in such cases is to determine the confidence interval as presented in Chapter 4. If this is impossible for whatever reason, we can fall back on the following approximation.

i **Alternative Procedure.** Compared to 6.h, we get a better approximation of the confidence interval via the interval that consists of all values θ for which $\ln\langle\theta\rangle$ lie in the interval $\widehat{\phi} \pm \widehat{\sigma}_{\widehat{\phi}} q_{0.975}^{t_{df}}$. Generally formulated: Let g be the transformation of ϕ to $\theta = g\langle\phi\rangle$. Then

$$\left\{\theta \,:\, g^{-1}\langle\theta\rangle \in \left[\widehat{\phi} - \widehat{\sigma}_{\widehat{\phi}} q_{0.975}^{t_{df}},\, \widehat{\phi} + \widehat{\sigma}_{\widehat{\phi}} q_{0.975}^{t_{df}}\right]\right\}$$

is an approximate 95% interval for θ. If $g^{-1}\langle\cdot\rangle$ is strictly monotone increasing, this interval is identical to

$$\left[g\left\langle\widehat{\phi} - \widehat{\sigma}_{\widehat{\phi}} q_{0.975}^{t_{df}}\right\rangle,\, g\langle\widehat{\phi} + \widehat{\sigma}_{\widehat{\phi}} q_{0.975}^{t_{df}}\rangle\right].$$

This procedure also ensures that, unlike the Gaussian error propagation law, the confidence interval is entirely in the region that is predetermined for the parameter. It is thus impossible that, for example, the confidence interval in the example $\theta = \exp\langle\phi\rangle$, unlike the interval from 6.h, can contain negative values.

As already noted, the approach just discussed should only be used if the way via the F-test from Chapter 4 is not possible. The Gaussian error propagation law should not be used in this context.

7. Forecasts and Calibration

Forecasts

a Besides the question of which parameters are plausible with respect to the given data, the question of in what range future observations will lie is often of central interest. The difference in these two questions was already discussed in 3.h. In this chapter we want to go into, among other things, the answer to the second question. We assume that the parameter $\underline{\theta}$ is estimated from the data with the least squares method. (If you want, you can call this dataset the training dataset.) Was can we now, with the help of our model, say about a future observation Y_o at a given point x_o?

b **Cress Example.** The concentration of an agrochemical material in soil samples can be studied through the growth behavior of a certain type of cress (nasturtium). 6 measurements of the variable of interest Y were made on each of 7 soil samples with predetermined (or measured with the largest possible precision) concentrations x. In each case we act as though the x values have no measurement error. The variable of interest is the weight of the cress per unit area after 3 weeks. A "logit-log" model is used to describe the relationship between concentration and weight:

$$h\langle x, \underline{\theta}\rangle = \begin{cases} \theta_1 & \text{if } x = 0 \\ \frac{\theta_1}{1+\exp\langle\theta_2+\theta_3\ln\langle x\rangle\rangle} & \text{if } x > 0. \end{cases}$$

(The data and the function $h\langle\cdot\rangle$ are graphically represented in Figure 7.b) We can now ask ourselves which weight values are expected for a concentration of, e.g. $x_0 = 3$?

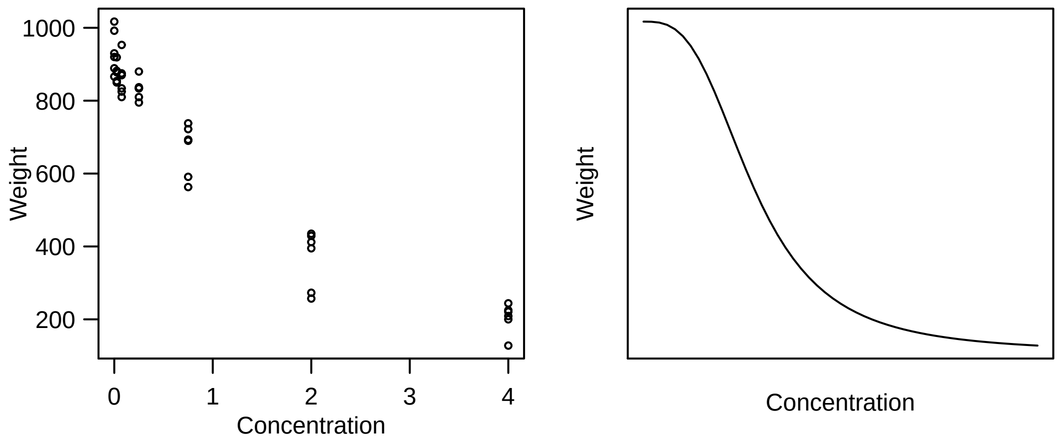

Figure 7.b: Cress Example. Left: Representation of the data. Right: A typical shape of the applied regression function.

c **Approximate Forecast Intervals.** We can estimate the expected value $E\langle Y_o\rangle = h\langle x_o, \theta\rangle$ of the variable of interest Y at the point x_o through $\widehat{\eta}_o := h\langle x_o, \widehat{\theta}\rangle$ We also want to get an interval where a future observation will lie with high probability, so we have to take into account not only the scatter of the estimate $\widehat{\eta}_o$ but also the random error E_o. Analogous to linear regression, an at least approximate $(1-\alpha/2)$ forecast interval can be given by

$$\widehat{\eta}_o \pm q^{t_{n-p}}_{1-\alpha/2} \cdot \sqrt{\widehat{\sigma}^2 + \left(\mathrm{se}\langle\widehat{\eta}_o\rangle\right)^2}$$

The calculation of $\mathrm{se}\langle\widehat{\eta}_o\rangle$ is found in 3.f.

* **Derivation.** The random variable Y_o is thus the value of the variable of interest for an observation with predictor variable value x_o. Since we do not know the true curve (actually only the parameters), we have no choice but to study the deviations of the observations from the estimated curve,

$$R_o = Y_o - h\left\langle x_o, \widehat{\underline{\theta}}\right\rangle = \left(Y_o - h\langle x_o, \underline{\theta}\rangle\right) - \left(h\left\langle x_o, \widehat{\underline{\theta}}\right\rangle - h\langle x_o, \underline{\theta}\rangle\right) .$$

Even if $\underline{\theta}$ is unknown, we know the distribution of the expressions in the big parentheses: Both are normally distributed random variables and they are independent because the first only depends on the "future" observation Y_o, the second only on the observations $Y_1, \ldots, Y_n$ that lead to the estimated curve. Both have expected value 0; the variances add up to

$$\mathrm{var}\langle R_o\rangle \approx \sigma^2 + \sigma^2 \underline{a}_o^T (A^T A)^{-1} \underline{a}_o .$$

The above described forecast interval follows by replacing the unknown values with their estimations.

d **Forecast Versus Confidence Intervals.** If the sample size n of the training dataset is very large, the estimated variance $\widehat{\sigma}_v^2$ is dominated by the error variance $\widehat{\sigma}^2$. This means that the uncertainty in the forecast is then determined primarily by the observation error. The second summand in the expression for $\widehat{\sigma}_v^2$ reflects the uncertainty that is caused by the estimation of $\underline{\theta}$.

It is therefore clear that the forecast interval is wider than the confidence interval for the expected value, since the random scatter of the observation must be taken into account. The endpoint of such an interval for x_o values in a chosen region are shown in Figure 7.i left, bound together as a band.

e * **Quality of the Approximation.** The determination of the forecast interval in 7.c is based on the same approximation as is used in Chapterber 3. The quality of the approximation can again be checked graphically as in Chapter 5.

f **Interpretation of the "Forecast Band".** The interpretation of the "forecast band", as is shown in Figure 7.i, is not totally simple: From the derivation it holds that

$$P\big\langle V_0^*\langle x_o\rangle \leq Y_o \leq V_1^*\langle x_o\rangle\big\rangle = 0.95$$

where $V_0^*\langle x_o\rangle$ is the lower and $V_1^*\langle x_o\rangle$ the upper boundaries of the prediction interval for $h\langle x_o\rangle$. However, if we want to make a prediction about more than one future observation, then the number of the observations in the forecast interval is *not* binomial distributed with $\pi = 0.95$. The events, that the individual future observations fall in the bad, are, specifically, not independent; they depend on each other over the random borders V_0 and V_1. If, for example, the estimation of $\widehat{\sigma}$ randomly turns out too small, for all future observations the band remains too narrow, and too many observations would lie outside of the band.

Calibration

g The actual goal of the experiment in the **cress example** is to estimate the concentration of the agrochemical materials from the weight of the cress. Consequently, we would like to use the regression relationship in the "wrong" direction. Problems of inference arise from this. Despite that, such a procedure is often desired to **calibrate** a measurement method or to estimate the result of a more expensive measurement method from a cheaper one. The regression curve in this relationship is often called a **calibration curve**. Another key word for finding this topic is **inverse regression**.

We would like to present here a simple method that gives a useable result if simplifying assumptions apply.

h **Procedure under Simplifying Assumptions .** We assume that the x values have no measurement error. This is achieved in the example if the concentration of the agrochemical material concentrations are very carefully determined. For several such soil samples with the most different possible concentrations we carry out several independent measurements of the value Y. This gives a training dataset, with which the unknown parameters can be estimated and the standard error of the estimations determined.

Now, if the value y_o is read for a trial to be measured, it is obvious to determine the corresponding x_o value as possible:

$$\widehat{x}_o = h^{-1}\langle y_o, \widehat{\underline{\theta}}\rangle \; .$$

Here, h^{-1} denotes the inverse function of h. This procedure is, however, only correct if $h\langle\cdot\rangle$ is monotone increasing or decreasing. However, this condition is usually fulfilled in calibration problems.

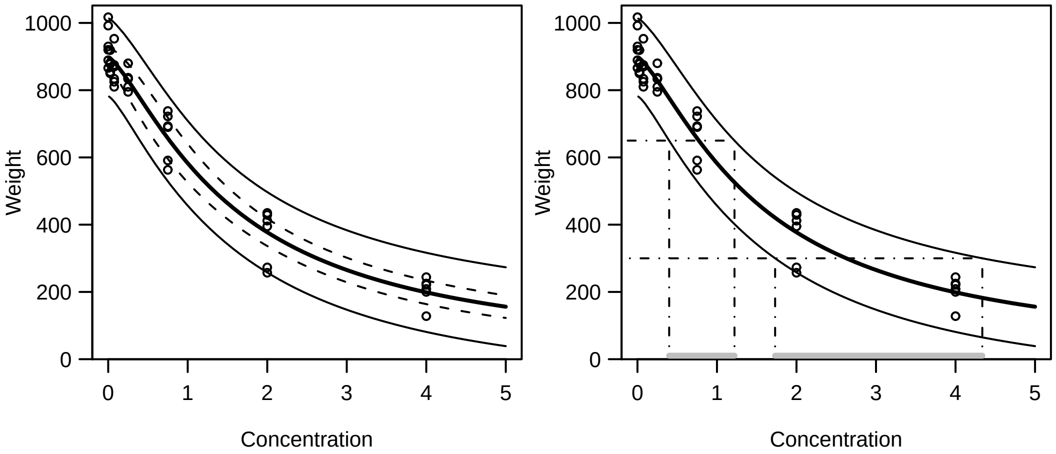

Figure 7.i: Cress example. Left: Confidence band for the estimated regression curve (dashed) and forecast band (solid). Right: Schematic representation of how a calibration interval is determined, at the points $y_0 = 650$ and $y_0 = 350$. The resulting intervals are [0.4, 1.22] and [1.73, 4.34] respectively.

i **Accuracy of the Obtained Values.** At the end of this calculation, the question arises of how accurate this value really is. The problem appears similar to the prediction problem at the start of this chapter. Unlike the prediction problem, y_o is observed and the corresponding value x_o must be estimated.

The answer can be formulated as follows: We see x_o as a parameter for which we want a confidence interval. Such an interval arises (as always) from a test. We take as null hypothesis $x_H = x_o$! As we have seen in 7.c, Y lies with probability 0.95 in the forecast interval

$$\widehat{\eta}_o \pm q_{1-\alpha/2}^{t_{n-p}} \cdot \sqrt{\widehat{\sigma}^2 + \left(\operatorname{se}\langle\widehat{\eta}_o\rangle\right)^2} .$$

Where $\widehat{\eta}_o$ was a compact notation for $h\langle x_o, \widehat{\theta}\rangle$. This interval therefore forms an acceptance interval for the value Y_o (which here plays the role of a test statistic) under the null hypothesis $x_o = x_H$. In Figure 7.i are graphically represented the prediction intervals for all possible x_o in a set interval for the **cress example**.

j **Illustration.** Figure 7.i right illustrates now the thought process for the **Cress example**: Measured values y_o are compatible with parameter values x_o in the sense of the test, if the point $[x_o, y_o]$lies between the curves shown. In the figure, we can thus determine without difficulty the set of x_o values that are compatible with the observation y_o. They form the shown interval, which can also be described as the set

$$\left\{ x \,:\, |y_0 - h\langle x, \underline{\widehat{\theta}}\rangle| \le q_{1-\alpha/2}^{t_{n-p}} \cdot \sqrt{\widehat{\sigma}^2 + \left(\operatorname{se}\langle h\langle x, \underline{\widehat{\theta}}\rangle\rangle\right)^2} \right\} .$$

This interval is now the desired confidence interval (or **calibration interval**) for x_o.

If we have m values to determine y_o, we apply the above method to $\bar{y}_o = \sum_{j=0}^{m} y_{oj}/m$ an:

$$\left\{ x \,:\, |\bar{y}_o - h\langle x, \underline{\widehat{\theta}}\rangle| \le \sqrt{\widehat{\sigma}^2 + \left(\operatorname{se}\langle h\langle x, \underline{\widehat{\theta}}\rangle\rangle\right)^2} \cdot q_{1-\alpha/2}^{t_{n-p}} \right\} .$$

k In this chapter, one of many possibilities for determining the calibration interval was presented. The diversity of methods is because their determination is based on some strongly simplified assumptions.

8. Closing Comments

a **Reason for the Difficulty in the Oxygen Consumption Example.** Why do we have so much difficulty with the oxygen consumption example? We consider Figure 1.e and remind ourselves that the parameter θ_1 represents the expected oxygen consumption for infinite incubation time, so it is clear that θ_1 is difficult to estimate, because the horizontal asymptote through the data is badly determined. If we had more observations with longer incubation times, we could avoid the difficulties with the quality of the confidence intervals of θ.

Also in nonlinear models, a good (statistical) **Experimental Design** is essential. The information content of the data is determined through the choice of the experimental conditions and no (statistical) procedure can deliver information that is not contained in the data.

b **Bootstrap.** For some time the bootstrap has also been used for determining confidence, prediction, and calibration intervals. Regarding this, see, e.g., Huet, Bouvier, Gruet and Jolivet (1996). In this book the case of inconstant variance (heteroskedastic models) is also discussed. It is also worth taking a look in the book from Carroll and Ruppert (1988). ccc

c **Correlated Error.** In this document is is always assumed that the errors E_i are independent. Like linear regression analysis, nonlinear regression can also be extended by bringing in **correlated errors** and **random effects** (Find software hints in 8.d).

d **Statistics Programs.** Today most statistics packages contain a procedure that can calculate asymptotic confidence intervals for the parameters. In principle it is then possible to calculate ”t-profiles” and profile traces, because they are also based on the fitting of nonlinear models, although with one fewer parameter.

In both implementations of the statistics language S, S-Plus and R, the function `nls` is available, that is based on the work of Bates and Watts (1988). The ”library” `nlme` contains S functions that can fit nonlinear regression models with correlated errors (`gnls`) and random effects (`nlme`). These implementations are based on the book “Mixed Effects Models in S and S-Plus” from Pinheiro and Bates (2000).

e **Literature Notes.** This document is based mainly on the book from Bates and Watts (1988). A mathematical discussion about the statistical and numerical methods in nonlinear regression can be found in Seber and Wild (1989). The book from Ratkowsky (1989) enumerates many possibly nonlinear functions $h\langle\cdot\rangle$ that primarily occur in biological applications.

A short introduction to this theme in relationship with the statistics program R can be found in Venables and Ripley (2002) and in Ritz and Streibig (2008).

A. Gauss-Newton Method

a **The Optimization Problem.** To determine the least squares estimator we must minimize the squared sum

$$S(\underline{\theta}) := \sum_{i=1}^{n} (y_i - \eta_i\langle\underline{\theta}\rangle)^2$$

. Unfortunately this can not be carried out explicitly like in linear regression. With local linear approximations of the regression function, however, the difficulties can be overcome.

b **Solution Procedure.** The procedure is set out in four steps. We consider the $(j+1)$-th iteration in the procedure.To simplify, we assume that the regression function $h\langle\rangle$ has only one unknown parameter. The solution from the previous iteration is denoted by $\theta^{(j)}$. $\theta^{(0)}$ is then the notation for the initial value.

1. **Approximation:** The regression function $h\langle x_i, \theta\rangle = \eta\langle\theta\rangle_i$ with the one dimensional parameter θ at the point $\theta^{(j)}$ is approximated by a line:

$$\eta_i\langle\theta\rangle \approx \eta_i\langle\theta^{(j)}\rangle + \left.\frac{\partial\eta_i\langle\theta\rangle}{\partial\theta}\right|_{\theta^{(j)}} (\theta - \theta^{(j)}) = \eta_i\langle\theta^{(j)}\rangle + a_i\langle\theta^{(j)}\rangle(\theta - \theta^{(j)}) \,.$$

* For a multidimensional $\underline{\theta}$ with help of a hyper plane it is approximated:

$$h\langle x_i, \underline{\theta}\rangle = \eta_i\langle\underline{\theta}\rangle \approx \eta_i\langle\underline{\theta}^{(j)}\rangle + a_i^{(1)}\langle\underline{\theta}^{(j)}\rangle(\theta_1 - \theta_1^{(j)}) + \ \ldots\ + a_i^{(p)}\langle\underline{\theta}^{(j)}\rangle(\theta_p - \theta_p^{(j)}),$$

where

$$a_i^{(k)}\langle\underline{\theta}\rangle := \frac{\partial h\langle x_i, \underline{\theta}\rangle}{\partial\theta_k}, \quad k = 1, \ldots, p.$$

With vectors and matrices the above equation can be written as

$$\underline{\eta}\langle\underline{\theta}\rangle \approx \underline{\eta}\langle\underline{\theta}^{(j)}\rangle + \boldsymbol{A}^{(j)}(\underline{\theta} - \underline{\theta}^{(j)}) \,.$$

Here the $(n \times p)$ derivative matrix $\boldsymbol{A}^{(j)}$ consists of the j-th iteration of the elements $\{a_{ik} = a_i^{(k)}\langle\underline{\theta}\rangle\}$ at the point $\underline{\theta} = \underline{\theta}^{(j)}$.

2. **Local Linear Model:** We now assume that the approximation in the 1st step holds exactly for the true model. With this we get for the residuals

$$r_i^{(j+1)} = y_i - \{\eta\langle\theta^{(j)}\rangle_i + a_i\langle\theta^{(j)}\rangle(\theta - \theta^{(j)})\} = \widetilde{y}_i^{(j+1)} - a_i\langle\theta^{(j)}\rangle\beta^{(j)}$$

with $\widetilde{y}_i^{(j+1)} := y_i - \eta\langle\theta^{(j)}\rangle$ and $\beta^{(j+1)} := \theta - \theta^{(j)}$.

* For a multidimensional $\underline{\theta}$ it holds that:

$$\underline{r}^{(j+1)} = \underline{y} - \{\underline{\eta}(\underline{\theta}^{(j)}) + \boldsymbol{A}^{(j)}(\underline{\theta} - \underline{\theta}^{(j)})\} = \widetilde{\mathbf{y}}^{(j+1)} - \boldsymbol{A}^{(j)}\underline{\beta}^{(j+1)}$$

with $\tilde{\mathbf{y}}^{(j+1)} := \mathbf{y} - \underline{\eta}(\underline{\theta}^{(j)})$ and $\underline{\beta}^{(j+1)} := \underline{\theta} - \underline{\theta}^{(j)}$.

3. **Least Square Estimation in the Locally Linear Model:** To find the best-fitting $\widehat{\beta}^{(j+1)}$ for the data, we minimize the sum of squared residuals: $\sum_{i=1}^{n}(r_i^{(j+1)})^2$. This is the usual linear least squares problem with ”‘$\widetilde{y}_i^{(j+1)} \equiv y_i$”’, ”‘$a_i\langle\theta^{(j)}\rangle \equiv x_i$”’ and ”‘$\beta^{(j+1)} \equiv \beta$”’ (The line goes through the origin). The solution to this problem, $\widehat{\beta}^{(j+1)}$, gives the best solution on the approximated line.

* To find the best-fitting $\underline{\widehat{\beta}}^{(j+1)}$ in the multidimensional case, we minimize the sum of squared residuals: $\|r^{(j+1)}\|^2$. This is the usual linear least squares problem with "' $\tilde{\mathbf{y}}^{(j+1)} \equiv \mathbf{y}$"', "' $\boldsymbol{A}^{(j)} \equiv X$"' and "' $\underline{\beta}^{(j+1)} \equiv \underline{\beta}$"'. The solution to this problem gives a $\underline{\widehat{\beta}}^{(j+1)}$, so that the point

$$\underline{\eta}\langle\underline{\theta}^{(j+1)}\rangle \quad \text{with } \underline{\theta}^{(j+1)} = \underline{\theta}^{(j)} + \underline{\widehat{\beta}}^{(j)}$$

lies nearer to $\mathbf{y}$ than $\underline{\eta}\langle\underline{\theta}^{(j)}\rangle$.

4. Iteration: Now with $\theta^{(j+1)} = \theta^{(j)} + \widehat{\beta}^{(j)}$ we return to step 1 and repeat steps 1, 2, and 3 until this procedure converges. The converged solution minimizes $S\langle\theta\rangle$ and thus corresponds to the desired estimation value $\widehat{\theta}$.

c **Further Details.** This minimization procedure that is known as the Gauss-Newton method, can be further refined. However, there are also other minimization methods available. It must be addressed in more detail what ”converged” should mean and how the convergence can be achieved. The details about these technical questions can be read in, e.g. Bates and Watts (1988). For us it is of primary importance to see that iterative procedures must be applied to solve the optimization problems in 2.a.

References

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis & Its Applications*, John Wiley & Sons.

Carroll, R. and Ruppert, D. (1988). *Transformation and Weighting in Regression*, Wiley, New York.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, John Wiley & Sons, New York.

Huet, S., Bouvier, A., Gruet, M.-A. and Jolivet, E. (1996). *Statistical Tools for Nonlinear Regression: A Practical Guide with S-Plus Examples*, Springer-Verlag, New York.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*, Statistics and Computing, Springer.

Rapold-Nydegger, I. (1994). *Untersuchungen zum Diffusionsverhalten von Anionen in carboxylierten Cellulosemembranen*, PhD thesis, ETH Zurich.

Ratkowsky, D. A. (1985). A statistically suitable general formulation for modelling catalytic chemical reactions, *Chemical Engineering Science* **40**(9): 1623–1628.

Ratkowsky, D. A. (1989). *Handbook of Nonlinear Regression Models*, Marcel Dekker, New York.

Ritz, C. and Streibig, J. C. (2008). *Nonlinear Regression with R*, Use R!, Springer Verlag.

Seber, G. and Wild, C. (1989). *Nonlinear regression*, Wiley, New York.

Stahel, W. (2000). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 3. Auflage edn, Vieweg, Braunschweig/Wiesbaden.

Venables, W. N. and Ripley, B. (2002). *Modern Applied Statistics with S*, Statistics and Computing, fourth edn, Springer-Verlag, New York.