Statistical and Numerical Methods for Chemical Engineers

Part on Statistics

Lukas Meier, ETH Zürich

Lecture Notes of W. Stahel (ETH Zürich) and A. Ruckstuhl (ZHAW)

November 2014

Contents

1 Preliminary Remarks

a Several types of problems lead to statistical models that are highly relevant for chemical engineers:

- A response variable like yield or quality of a product or the duration of a production process may be influenced by a number of variables – plausible examples are temperature, pressure, humidity, properties of the input material (educts).
  - In a first step, we need a model for describing the relations. This leads to **regression** and **analysis of variance** models. Quite often, simple or multiple **linear regression** already give good results.
  - **Optimization of production processes:** If the relations are modelled adequately, it is straightforward to search for those values of the variables that drive the response variable to an optimal value. Methods to efficiently find these optimum values are discussed under the label of **design of experiments**.
- Chemical processes develop according to clear laws ("law and order of chemical change", Swinbourne, 1971), which are typically modelled by differential equations. In these systems there are constants, like the reaction rates, which can be determined from data of suitable experiments. In the simplest cases this leads to linear regression, but usually, the methods of **non-linear regression**, possibly combined with the numerical solution of differential equations, are needed. We call this combination **system analysis**.
- As an efficient surrogate for chemical determination of concentrations of different compounds, indirect determination by spectroscopical measurements are often suitable. Methods that allow for inferring amounts or concentrations of chemical compounds from spectra belong to the field of **multivariate statistics**.

b In the very limited time available in this course we will present an introduction to these topics. We start with linear regression, a topic you should already be familiar with. The simple linear regression model is used to recall basic statistical notions. The following steps are common for statistical methods:

1. State the scientific question and characterize the data which are available or will be obtained.
2. Find a suitable probability model that corresponds to the knowledge about the processes leading to the data. Typically, a few unknown constants remain, which we call "parameters" and which we want to learn from the data. The model can (and should) be formulated *before* the data is available.
3. The field of statistics encompasses the methods that bridge the gap between models and data. Regarding parameter values, statistics answers the following questions:
   a) Which value is the **most plausible** one for a parameter? The answer is given by **estimation**. An estimator is a function that determines a parameter value from the data.
   b) Is a given value for the parameter plausible? The decision is made by using

a statistical **test**.

c) Which value*s* are plausible? The answer is given by a set of all plausible values, which is usually an interval, the so called **confidence interval**.

4. In many applications the **prediction** of measurements (observations) that are not yet available is of interest.

C Linear regression was already discussed in "Grundlagen der Mathematik II". Please have a look at your notes to (again) get familiar with the topic.

You find additional material for this part of the course on

`http://stat.ethz.ch/~meier/teaching/cheming`

2 Summary of Linear Regression

2.1 Simple Linear Regression

a Assume we have n observations (x_i, Y_i), $i = 1, \ldots, n$ and we want to model the relationship between a **response variable** Y and a **predictor variable** x.

The **simple linear regression model** is

$$Y_i = \alpha + \beta x_i + E_i, \quad i = 1, \ldots, n.$$

The x_i's are fixed numbers while the E_i's are random, called "random deviations" or "random errors". Usual assumptions are

$$E_i \sim \mathcal{N}(0, \sigma^2),\ E_i \text{ independent.}$$

The parameters of the simple linear regression model are the **coefficients** α, β and the standard deviation σ of the random error.

Figure 2.1.a illustrates the model.

b **Estimation of the coefficients** follows the principle of **least squares** and yields

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ , \qquad \widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\bar{x}\ .$$

The estimates $\widehat{\beta}$ and $\widehat{\alpha}$ fluctuate around the true (but unknown) parameters. More precisely, the estimates are normally distributed,

$$\widehat{\beta} \sim \mathcal{N}(\beta,\, \sigma^2/\mathrm{SS}_X)\ , \quad \widehat{\alpha} \sim \mathcal{N}\left(\alpha,\, \sigma^2\left(\tfrac{1}{n} + \bar{x}^2/\mathrm{SS}_X\right)\right),$$

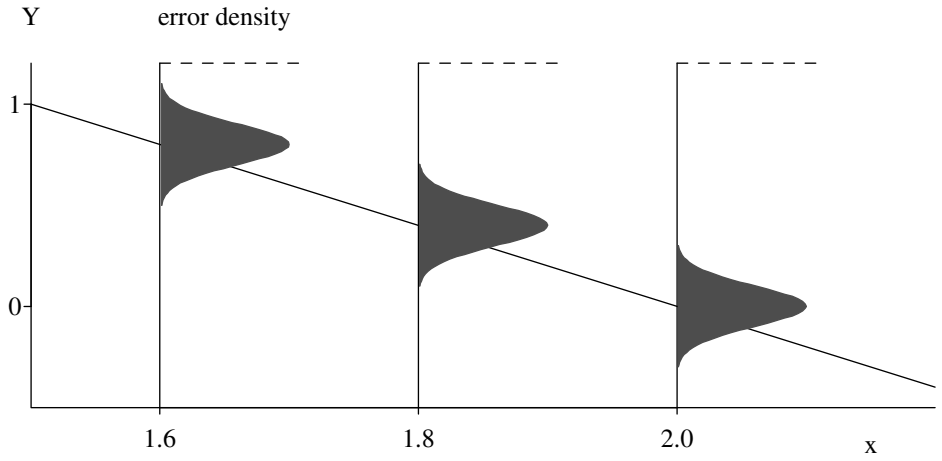

Figure 2.1.a: Display of the probability model $Y_i = 4 - 2x_i + E_i$ for 3 observations Y_1, Y_2 and Y_3 corresponding to the x values $x_1 = 1.6$, $x_2 = 1.8$ and $x_3 = 2$.

where $\mathrm{SS}_X = \sum_{i=1}^n (x_i - \bar{x})^2$.

c The deviations of the observed Y_i from the **fitted values** $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$ are called **residuals** $R_i = Y_i - \widehat{y}_i$ and are "estimators" of the random errors E_i.
They lead to an estimate of the standard deviation σ of the error,

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n R_i^2 \, .$$

d **Test** of the null hypothesis $\beta = \beta_0$: The test statistic

$$T = \frac{\widehat{\beta} - \beta_0}{\mathrm{se}(\widehat{\beta})} \, , \quad \mathrm{se}(\widehat{\beta}) = \sqrt{\widehat{\sigma}^2 / \mathrm{SS}_X}$$

has a t-distribution with $n-2$ degrees of freedom under the null-hypothesis.
This leads to the confidence interval of

$$\widehat{\beta} \pm q_{0.975}^{t_{n-2}} \, \mathrm{se}(\widehat{\beta}).$$

e The "**confidence band**" for the value of the regression function connects the end points of the confidence intervals for $\mathrm{E}(Y \mid x) = \alpha + \beta x$.
A **prediction interval** shall include a (yet unknown) value Y_0 of the response variable for a given x_0 – with a given "statistical certainty" (usually 95%). Connecting the end points for all possible x_0 produces the **"prediction band"**.

2.2 Multiple Linear Regression

a Compared to the simple linear regression model we now have **several predictors** $x^{(1)}, \ldots, x^{(m)}$.
The **multiple linear regression model** is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + E_i \\ E_i &\sim \mathcal{N}(0, \sigma^2), \; E_i \text{ independent.} \end{aligned}$$

In **matrix notation**:

$$\underline{Y} = \boldsymbol{X} \underline{\beta} + \underline{E} \, , \qquad \underline{E} \sim \mathcal{N}_n(\underline{0}, \sigma^2 \boldsymbol{I}),$$

where the response vector $\underline{Y} \in \mathbb{R}^n$, the **design matrix** $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, the parameter vector $\underline{\beta} \in \mathbb{R}^p$ and the error vector $\underline{E} \in \mathbb{R}^n$ for $p = m + 1$ (number of parameters).

$$\underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \; \boldsymbol{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(m)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(m)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(m)} \end{pmatrix}, \; \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \; \underline{E} = \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{pmatrix}.$$

Different rows of the design matrix $\boldsymbol{X}$ are different observations. The variables (predictors) can be found in the corresponding columns.

b Estimation is again based on least squares, leading to

$$\underline{\widehat{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\underline{Y},$$

i.e. we have a closed form solution.

From the distribution of the estimated coefficients,

$$\widehat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)_{jj}\right)$$

t-tests and confidence intervals for individual coefficients can be derived as in the linear regression model. The test statistic

$$T = \frac{\widehat{\beta}_j - \beta_{j,0}}{\mathrm{se}(\widehat{\beta}_j)}\,, \quad \mathrm{se}(\widehat{\beta}_j) = \sqrt{\widehat{\sigma}^2\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)_{jj}}$$

follows a t-distribution with $n-(m+1)$ parameters under the null-hypothesis H_0 : $\beta_j = \beta_{j,0}$.

The standard deviation σ is estimated by

$$\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} R_i^2.$$

c Table 2.2.c shows a typical **computer output**, annotated with the corresponding mathematical symbols.

The **multiple correlation** R is the correlation between the fitted values $\widehat{y}_i$ and the observed values Y_i. Its square measures the portion of the variance of the Y_i's that is "explained by the regression", and is therefore called **coefficient of determination**:

$$R^2 = 1 - \mathrm{SS}_E/\mathrm{SS}_Y,$$

where $SS_E = \sum_{i=1}^{n}(Y_i - \widehat{y}_i)^2$, $SS_Y = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

```
Coefficients:
                Value β̂_j      Std. Error      t value       Pr(>|t|)
(Intercept)      19.7645          2.6339       7.5039         0.0000
pH               -1.7530          0.3484      -5.0309         0.0000
lSAR             -1.2905          0.2429      -5.3128         0.0000

Residual standard error: σ̂ = 0.9108 on n − p = 120 degrees of freedom
Multiple R-Squared: R² = 0.5787

Analysis of variance
                      Df        Sum of Sq      Mean Sq        F Value    Pr(F)
Regression       m =   2   SS_R = 136.772       68.386      T = 82.43   0.0000
Residuals    n−p = 120   SS_E =  99.554   σ̂² = 0.830                   p-value
Total              122   SS_Y = 236.326
```

Table 2.2.c: Computer output for a regression example, annotated with mathematical symbols.

d

The model is called linear because it is **linear in the parameters** $\beta_0, \ldots, \beta_m$.

It could well be that some predictors are non-linear functions of other predictors (e.g., $x^{(2)} = (x^{(1)})^2$). It is still a linear model as long as the parameters appear in linear form!

e In general, it is not appropriate to replace a multiple regression model by many simple regressions (on single predictor variables).

In a multiple linear regression model, the coefficients describe how Y is changing when varying the corresponding predictor **and keeping the other predictor variables constant**. I.e., it is the effect of the predictor on the response *after* having subtracted the effect of all other predictors on Y. Hence we need to have all predictors in the model at the same time in order to estimate this effect.

f **Many applications** The model of multiple linear regression model is suitable for describing many different situations:

- **Transformations** of the predictors (and the response variable) may turn originally non-linear relations into linear ones.
- A comparison of two groups is obtained by using a binary predictor variable. Several groups need a "block of dummy variables". Thus, **nominal (or categorical) explanatory variables** can be used in the model and can be combined with continuous variables.
- The idea of different linear relations of the response with some predictors in different groups of data can be included into a single model. More generally, **interactions** between explanatory variables can be incorporated by suitable terms in the model.
- **Polynomial regression** is a special case of multiple linear (!) regression (see example above).

g The **F-Test for comparison of models** allows for testing whether several coefficients are zero. This is needed for testing whether a categorical variable has an influence on the response.

2.3 Residual Analysis

a The assumptions about the errors of the regression model can be split into

(a) their expected values are zero: $\mathrm{E}(E_i) = 0$ (or: **the regression function is correct**),

(b) they have **constant variance**, $\mathrm{Var}(E_i) = \sigma^2$,

(c) they are **normally distributed**,

(d) they are **independent** of each other.

These assumptions should be checked for

- deriving a better model based on deviations from it,
- justifying tests and confidence intervals.

Deviations are detected by inspecting graphical displays. Tests for assumptions play a less important role.

b **Fitting a regression model without examining the residuals is a risky exercise!**

c The following displays are useful:

(a) **Non-linearities**: Scatterplot of (unstandardized) residuals against fitted values **(Tukey-Anscombe plot)** and against the (original) **explanatory variables**. **Interactions**: Pseudo-threedimensional diagram of the (unstandardized) residuals against pairs of explanatory variables.

(b) **Equal scatter**: Scatterplot of (standardized) absolute residuals against fitted values **(Tukey-Anscombe plot)** and against (original) **explanatory variables**. Usually no special displays are given, but scatter is examined in the plots for (a).

(c) **Normal distribution: QQ-plot** (or histogram) of (standardized) residuals.

(d) **Independence**: (unstandardized) residuals against time or location.

(e) **Influential observations** for the fit: Scatterplot of (standardized) residuals against **leverage**.
Influential observations for individual coefficients: added-variable plot.

(f) **Collinearities**: Scatterplot matrix of explanatory variables and numerical output (of R_j^2 or VIF_j or "tolerance").

d Remedies:

- **Transformation (monotone non-linear) of the response**: if the distribution of the residuals is skewed, for non-linearities (if suitable) or unequal variances.
- **Transformation** (non-linear) of **explanatory variables**: when seeing non-linearities, high leverages (can come from skewed distribution of explanatory variables) and interactions (may disappear when variables are transformed).
- **Additional terms**: to model non-linearities and interactions.
- Linear transformations of several explanatory variables: to avoid **collinearities**.
- **Weighted regression**: if variances are unequal.
- Checking the correctness of observations: for all **outliers** in any display.
- Rejection of outliers: if robust methods are not available (see below).

More advanced methods:

- Generalized least squares: to account for correlated random errors.
- Non-linear regression: if non-linearities are observed and transformations of variables do not help or contradict a physically justified model.
- Robust regression: should always be used, suitable in the presence of outliers and/or long-tailed distributions.

Note that correlations among errors lead to wrong test results and confidence intervals which are most often too short.

3 Nonlinear Regression

3.1 Introduction

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

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

Here, h is an appropriate function that depends on the predictor variables and parameters, that we want to summarize with vectors $\underline{x} = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}]^T$ and $\underline{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^T$. We assume that the errors are all normally distributed and independent, i.e.

$$E_i \sim \mathcal{N}\left(0, \sigma^2\right), \text{ independent.}$$

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

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

where the $\widetilde{x}^{(j)}$ can be arbitrary functions of the original explanatory variables $x^{(j)}$. There, the parameters were usually denoted by β_j instead of θ_j.

c **The Nonlinear Regression Model** In nonlinear regression, we use functions h that are *not* 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 of an enzymatic reaction depends on the concentration of a substrate. As outlined in Bates and Watts (1988), an experiment was performed to examine how a treatment of the enzyme with an additional substance called Puromycin influences the reaction speed. The initial speed of the reaction is chosen as the response variable, which is measured via radioactivity (the unit of the response variable is count/min^2; the number of registrations on a Geiger counter per time period measures the quantity of the substance, and the reaction speed is proportional to the change per time unit).

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

$$h(x; \underline{\theta}) = \frac{\theta_1 x}{\theta_2 + x}\ .$$

An infinitely large substrate concentration ($x \to \infty$) leads to the "asymptotic" speed θ_1. It was hypothesized that this parameter is influenced by the addition of Puromycin. The experiment is therefore carried out once with the enzyme treated with Puromycin

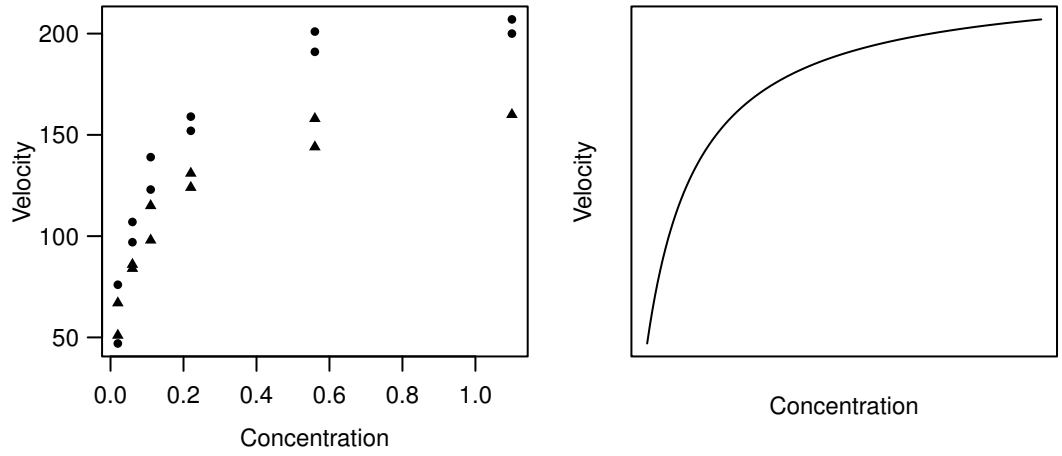

Figure 3.1.d: Puromycin. (a) Data (• treated enzyme; △ untreated enzyme) and (b) typical shape of the regression function.

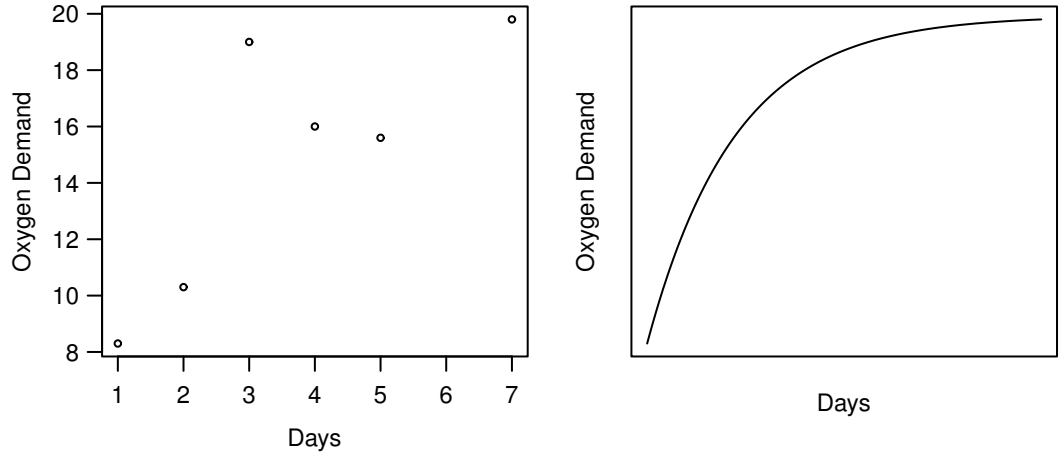

Figure 3.1.e: Biochemical Oxygen Demand. (a) Data and (b) typical shape of the regression function.

and once with the untreated enzyme. Figure 3.1.d shows the data and the shape of the regression function. In this section only the data of the treated enzyme is used.

Example e **Biochemical Oxygen Demand** To determine the biochemical oxygen demand, stream water samples were enriched with soluble organic matter, with inorganic nutrients and with dissolved oxygen, and subdivided into bottles (Marske, 1967, see Bates and Watts, 1988). Each bottle was inoculated with a mixed culture of microorganisms, sealed and put in a climate chamber with constant temperature. The bottles were periodically opened and their dissolved oxygen concentration was analyzed, from which the biochemical oxygen demand [mg/l] was calculated. The model used to connect the cumulative biochemical oxygen demand Y with the incubation time x is based on exponential decay:

$$h\left(x;\underline{\theta}\right) = \theta_1 \left(1 - e^{-\theta_2 x}\right).$$

Figure 3.1.e shows the data and the shape of the regression function.

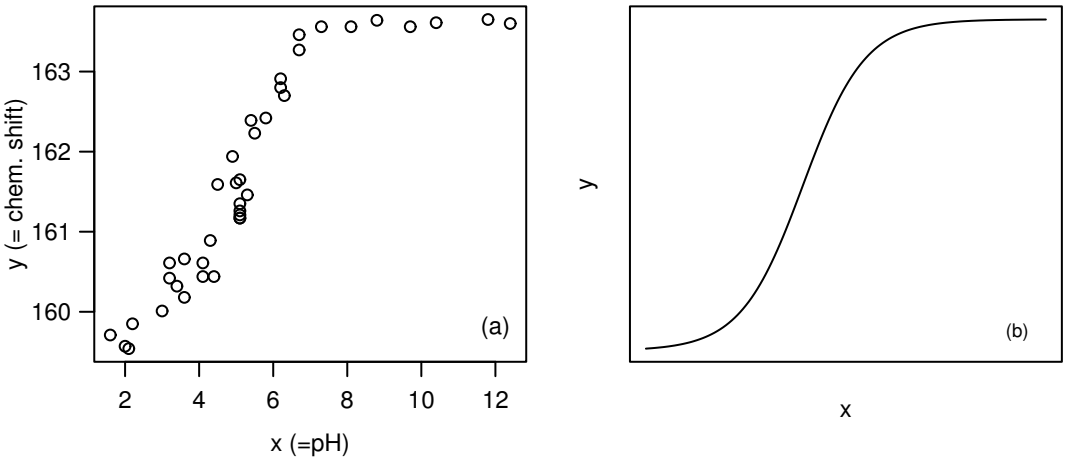

Figure 3.1.f: Membrane Separation Technology. (a) Data and (b) a typical shape of the regression function.

Example f **Membrane Separation Technology** See Rapold-Nydegger (1994). The ratio of protonated to deprotonated carboxyl groups in the pores of cellulose membranes depends 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

$$\log_{10}\left(\frac{\theta_1 - y}{y - \theta_2}\right) = \theta_3 + \theta_4\, x\ ,$$

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

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

The regression function $h(x_i, \underline{\theta})$ for a reasonably chosen $\underline{\theta}$ is shown in Figure 3.1.f next to the data.

g **A Few Further Examples of Nonlinear Regression Functions**

- Hill model (enzyme kinetics): $h(x_i, \underline{\theta}) = \theta_1 x_i^{\theta_3}/(\theta_2 + x_i^{\theta_3})$
  For $\theta_3 = 1$ this is also known as the Michaelis-Menten model (3.1.d).
- Mitscherlich function (growth analysis): $h(x_i, \underline{\theta}) = \theta_1 + \theta_2 \exp(\theta_3 x_i)$.
- From kinetics (chemistry) we get the function

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

- Cobbs-Douglas production function

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

Since useful regression functions are often derived from the theoretical background of the application of interest, a general overview of nonlinear regression functions is of very 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** by transformations of the response variable and the explanatory variables.

For example, a power function

$$h(x;\underline{\theta}) = \theta_1 x^{\theta_2}$$

can be transformed to a linear (in the parameters!) function

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

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

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

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

The last one is the Cobbs-Douglas Model from 3.1.g.

i A linear regression with the linearized regression function of the example above is based on the model

$$\ln(Y_i) = \beta_0 + \beta_1 \widetilde{x}_i + E_i \,,$$

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

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

with $\widetilde{E}_i = \exp(E_i)$. The errors $\widetilde{E}_i$, $i = 1, \ldots, n$, now have a multiplicative effect and are log-normally distributed! The assumptions about the random deviations are thus now drastically different than for a model that is based directly 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 errors can be better satisfied – in our example, if the errors actually act multiplicatively rather than additively and are log-normally rather than normally distributed. These assumptions must be checked with residual analysis.

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

k We have almost exclusively seen regression functions that only depend on one predictor variable x. This was primarily because it was possible to graphically illustrate the model. The following theory also works well for regression functions $h(\underline{x};\underline{\theta})$ that depend on several predictor variables $\underline{x} = [x^{(1)}, x^{(2)}, \ldots, x^{(m)}]$.

3.2 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 – the principle of least squares. The sum of the squared deviations

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

should be minimized. The notation that replaces $h(x_i;\underline{\theta})$ with $\eta_i(\underline{\theta})$ is reasonable because $[x_i, y_i]$ is given by the data and only the parameters $\underline{\theta}$ remain to be determined.

Unfortunately, the minimum of $S(\underline{\theta})$ and hence the estimator have *no* explicit solution (in contrast to the linear regression case). **Iterative numeric procedures** are therefore needed. We will sketch the basic ideas of the most common algorithm. It is also the basis for the easiest way to derive tests and confidence intervals.

b **Geometrical Illustration** The observed values $\underline{Y} = [Y_1, Y_2, \ldots, Y_n]^T$ define a point in n-dimensional space. The same holds true for the "model values" $\underline{\eta}(\underline{\theta}) = [\eta_1(\underline{\theta}), \eta_2(\underline{\theta}), \ldots, \eta_n(\underline{\theta})]^T$ for a given $\underline{\theta}$.

Please take note: In multivariate statistics where an observation consists of m variables $x^{(j)}$, $j = 1, 2, \ldots, m$, it's common to illustrate the observations in the m-dimensional space. Here, we consider the Y- and η-values of all n observations as points in the n-dimensional space.

Unfortunately, geometrical interpretation stops with three dimensions (and thus with three observations). Nevertheless, let us have a look at such a situation, first for simple linear regression.

c As stated above, the observed values $\underline{Y} = [Y_1, Y_2, Y_3]^T$ determine a point in three-dimensional space. For given parameters $\beta_0 = 5$ and $\beta_1 = 1$ we can calculate the model values $\eta_i\left(\underline{\beta}\right) = \beta_0 + \beta_1 x_i$ and represent the corresponding vector $\underline{\eta}\left(\underline{\beta}\right) = \beta_0\underline{1} + \beta_1\underline{x}$ as a point. We now ask: Where are all the points that can be achieved by varying the parameters? These are the possible linear combinations of the two vectors $\underline{1}$ and $\underline{x}$: they form a plane "spanned by $\underline{1}$ and $\underline{x}$". By estimating the parameters according to the principle of least squares, the squared distance between $\underline{Y}$ and $\underline{\eta}\left(\underline{\beta}\right)$ is minimized. This means that we are looking for the point on the plane that is closest 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 $\underline{\widehat{\beta}} = [\widehat{\beta}_0, \widehat{\beta}_1]^T$. An illustration can be found in Figure 3.2.c.

d Now we want to fit a nonlinear function, e.g. $h(\underline{x};\underline{\theta}) = \theta_1 \exp(1 - \theta_2 x)$, to the same three observations. We can again ask ourselves: Where are all the points $\underline{\eta}(\underline{\theta})$ that can be achieved by varying 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 is closest to $\underline{Y}$. The parameter values that correspond to this point $\underline{\widehat{\eta}}$ are then the

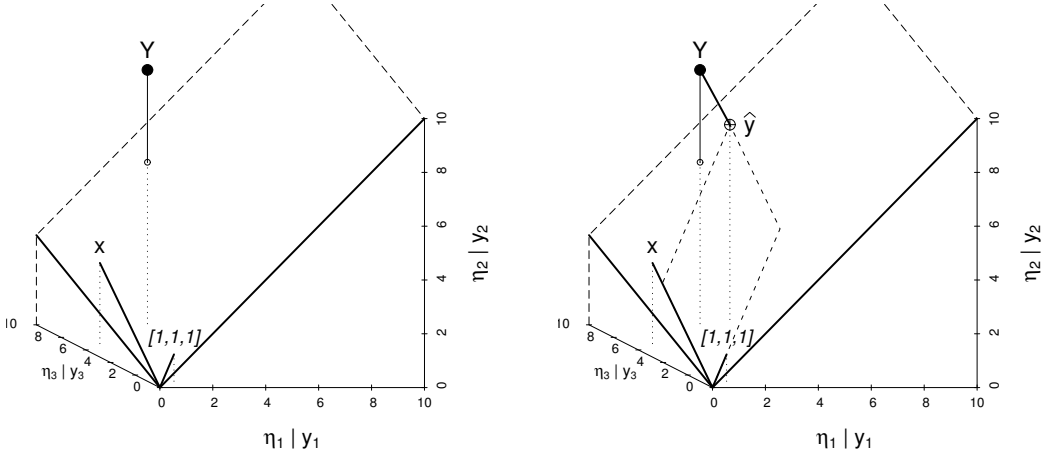

Figure 3.2.c: Illustration of simple linear regression. Values of $\underline{\eta}\left(\underline{\beta}\right) = \beta_0 + \beta_1 \underline{x}$ for varying parameters $[\beta_0, \beta_1]$ lead to a plane in three-dimensional space. The right plot also shows the point on the surface that is closest to $\underline{Y} = [Y_1, Y_2, Y_3]$. It is the fitted value $\widehat{\underline{y}}$ and determines the estimated parameters $\widehat{\underline{\beta}}$.

estimated parameter values $\widehat{\underline{\theta}} = [\widehat{\theta}_1, \widehat{\theta}_2]^T$. Figure Figure 3.2.d illustrates the nonlinear case.

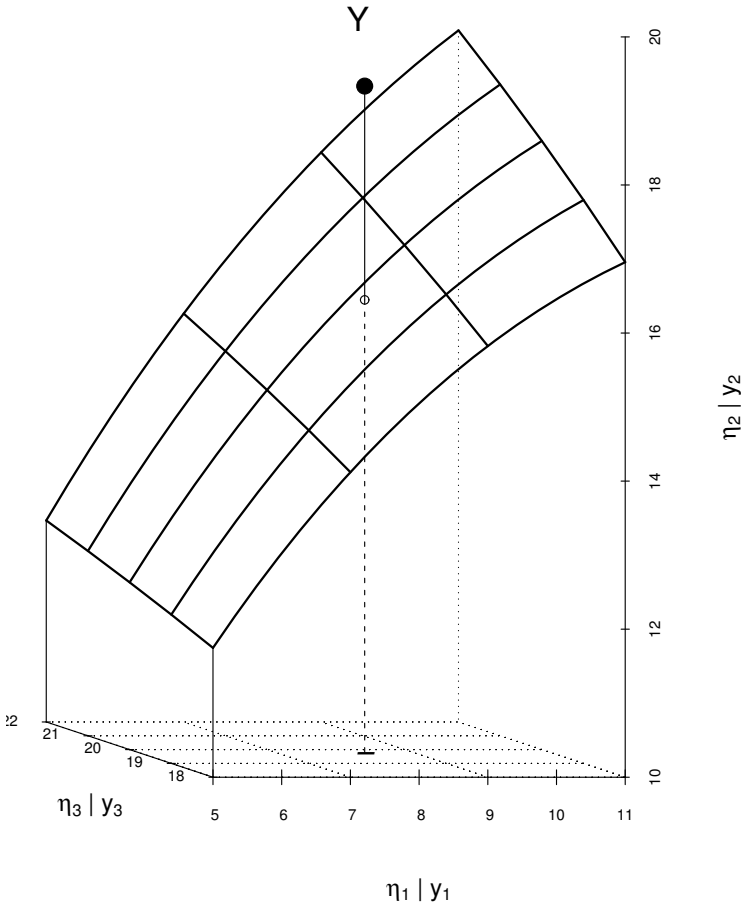

Figure 3.2.d: Geometrical illustration of nonlinear regression. The values of $\underline{\eta}\left(\underline{\theta}\right) = h\left(\underline{x}; \theta_1, \theta_2\right)$ for varying parameters $[\theta_1, \theta_2]$ lead to a two-dimensional "model surface" in three-dimensional space. The lines on the model surface correspond to constant η_1 and η_3, respectively.

e **Biochemical Oxygen Demand (cont'd)** The situation for our Biochemical Oxygen Demand example can be found in Figure 3.2.e. Basically, we can read the estimated parameters directly off the graph here: $\widehat{\theta}_1$ is a bit less than 21 and $\widehat{\theta}_2$ is a bit larger than 0.6. In fact the (exact) solution is $\underline{\widehat{\theta}} = [20.82, 0.6103]$ (note that these are the parameter estimates for the reduced data set only consisting of three observations).

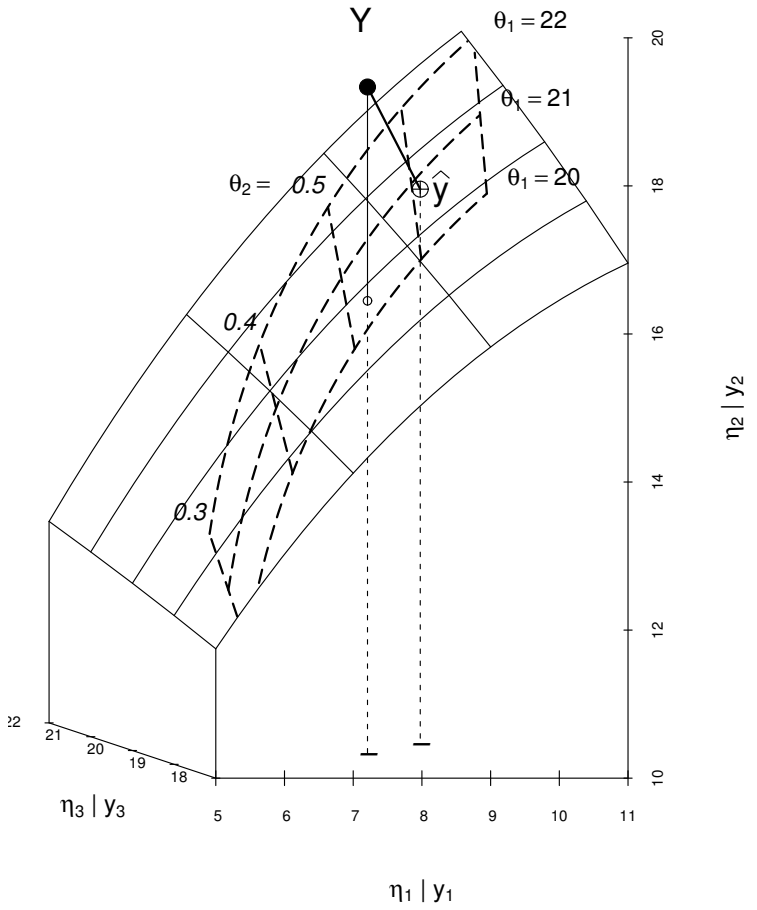

Figure 3.2.e: Biochemical Oxygen Demand: Geometrical illustration of nonlinear regression. In addition, we can see here the lines of constant θ_1 and θ_2, respectively. The vector of the estimated model values $\underline{\widehat{y}} = h\left(\underline{x}; \underline{\widehat{\theta}}\right)$ is the point on the model surface that is closest to $\underline{Y}$.

f **Approach for the Minimization Problem** The main idea of the usual algorithm for minimizing the sum of squares (see 3.2.a) is 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 $\underline{\eta}\left(\underline{\theta}^{(\ell)}\right) = h\left(\underline{x}; \underline{\theta}^{(\ell)}\right)$ (the so called tangent plane). Now, we are looking for the point on that plane that lies closest to $\underline{Y}$. This is the same as 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 $\underline{\theta}^{(\ell+1)}$ that we use as starting value for the next iteration.

g **Linear Approximation** To determine the tangent plane we need the partial derivatives

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

that can be summarized by an $n \times p$ matrix $\boldsymbol{A}$. The approximation of the model

surface $\underline{\eta}(\underline{\theta})$ by the tangent plane at a parameter value $\underline{\theta}^*$ is

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

or, in matrix notation,

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

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

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

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

h Gauss-Newton Algorithm The Gauss-Newton algorithm starts with an initial 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 hence an improved value $\underline{\theta}^{(1)} = \underline{\theta}^{(0)} + \underline{\beta}$. Again, the approximated model is calculated, and thus the "preliminary residuals" $\underline{Y} - \underline{\eta}\left(\underline{\theta}^{(1)}\right)$ and the partial derivatives $\boldsymbol{A}\left(\underline{\theta}^{(1)}\right)$ are determined, leading to $\underline{\theta}_2$. This iteration step is continued until the the correction $\underline{\beta}$ is small enough.

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

* Algorithms usually determine the derivative matrix $\boldsymbol{A}$ numerically. In more complex problems the numerical approximation can be insufficient and cause convergence problems. For such situations it is an advantage if explicit expressions for the partial derivatives can be used to determine the derivative matrix more reliably (see also Chapter 3.6).

i Initial Values An iterative procedure always requires an initial value. Good initial values help to find a solution more quickly and more reliably. Some possibilities to arrive at good initial values are now being presented.

j Initial Value from Prior Knowledge As already noted in the introduction, nonlinear models are often based on theoretical considerations of the corresponding application area. Already existing **prior knowledge** from similar experiments can be used to get an initial value. To ensure the quality of the chosen initial value, it is advisable to graphically represent the regression function $h(x;\underline{\theta})$ for various possible initial values $\underline{\theta} = \underline{\theta}^0$ together with the data (e.g., as in Figure 3.2.k, right).

k Initial Values via Linearizable Regression Functions Often – because of the distribution of the error term – one is forced to use a nonlinear regression function even though it would be linearizable. However, the linearized model can be used to get initial values.

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

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

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 3.2.k, left). This leads to the initial values

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

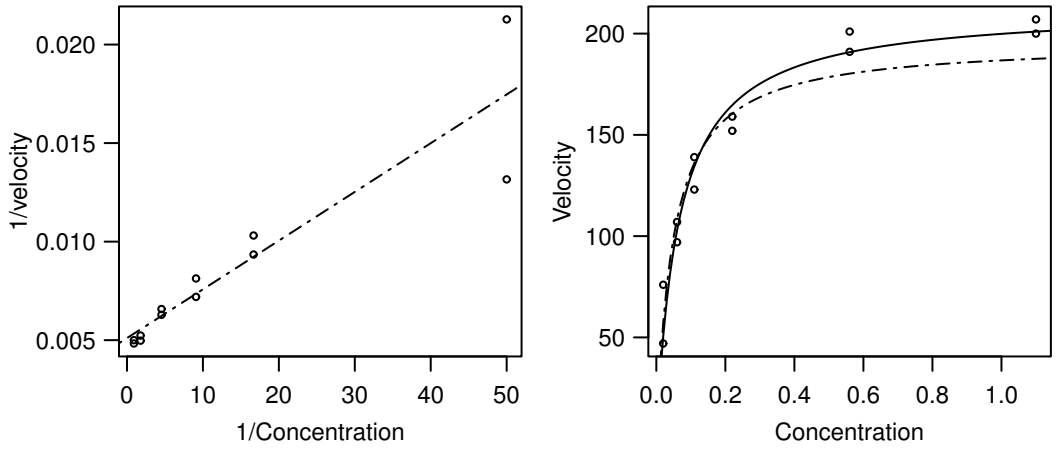

Figure 3.2.k: Puromycin. Left: Regression function in the linearized problem. Right: Regression function $h(x;\underline{\theta})$ for the initial values $\underline{\theta} = \underline{\theta}^{(0)}$ (- - - - -) and for the least squares estimation $\underline{\theta} = \widehat{\underline{\theta}}$ (———).

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

In the Puromycin Example we can derive an initial value in another way: θ_1 is the response value for $x = \infty$. Since the regression function is monotonically 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, such that y reaches half of the asymptotic value θ_1. This leads to $\theta_2^{(0)} = 0.06$.

The initial values thus result from a geometrical interpretation of the parameters and a rough estimate can be determined by "fitting by eye".

Example m **Membrane Separation Technology (cont'd)** In the Membrane Separation Technology example we let $x \to \infty$, so $h(x;\underline{\theta}) \to \theta_1$ (since $\theta_4 < 0$); for $x \to -\infty$, $h(x;\underline{\theta}) \to \theta_2$. From Figure 3.1.f (a) we see that $\theta_1 \approx 163.7$ and $\theta_2 \approx 159.5$. Once we know θ_1 and θ_2, we can linearize the regression function by

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

This is called a **conditional linearizable** function. The linear regression model 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 $h(\cdot;\underline{\theta}^{(0)})$ and $h(\cdot;\widehat{\underline{\theta}})$ are shown in Figure 3.2.m (b).

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

3.3 Approximate Tests and Confidence Intervals

a The estimator $\widehat{\underline{\theta}}$ is the value of $\underline{\theta}$ that optimally fits the data. We now ask *which parameter values $\underline{\theta}$ are compatible with the observations.* The **confidence region** is the set of all these values. For an individual parameter θ_j the confidence region is a **confidence interval**.

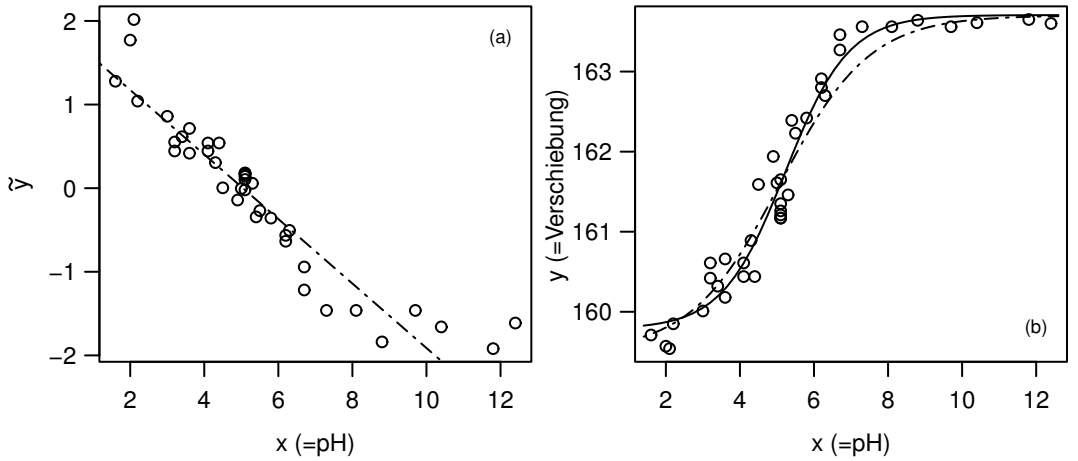

Figure 3.2.m: Membrane Separation Technology. (a) Regression line that is used for determining the initial values for θ_3 and θ_4. (b) Regression function $h(x; \underline{\theta})$ for the initial value $\underline{\theta} = \underline{\theta}^{(0)}$ (- - - - -) and for the least squares estimator $\underline{\theta} = \widehat{\underline{\theta}}$ (———).

The following results 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 multiple parameters the corresponding Chi-Square test is used and leads to 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 3.2.g

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

if the parameter vector $\underline{\theta}^*$ that is used for the linearization is close 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 $\widehat{\underline{\beta}}$ – or more generally the covariance matrix of $\widehat{\underline{\beta}}$ – then approximate the corresponding values of $\widehat{\underline{\theta}}$.

c **Asymptotic Distribution of the Least Squares Estimator** It follows that the least squares estimator $\widehat{\underline{\theta}}$ is asymptotically normally distributed

$$\widehat{\underline{\theta}} \overset{as.}{\sim} \mathcal{N}\,(\underline{\theta},\, \boldsymbol{V}\,(\underline{\theta}))\,,$$

with asymptotic covariance matrix $\boldsymbol{V}\,(\underline{\theta}) = \sigma^2(\boldsymbol{A}\,(\underline{\theta})^T\,\boldsymbol{A}\,(\underline{\theta}))^{-1}$, where $\boldsymbol{A}\,(\underline{\theta})$ is the $n \times p$ matrix of partial derivatives (see 3.2.g).

To explicitly determine the covariance matrix $V\,(\underline{\theta})$, $\boldsymbol{A}\,(\underline{\theta})$ is calculated using $\widehat{\underline{\theta}}$ instead of the unknown $\underline{\theta}$. For the error variance σ^2 we plug-in the usual estimator

$$\widehat{\boldsymbol{V}}\,(\underline{\theta}) = \hat{\sigma}^2 \left(\boldsymbol{A}\left(\widehat{\underline{\theta}}\right)^T \boldsymbol{A}\left(\widehat{\underline{\theta}}\right)\right)^{-1}$$

where

$$\hat{\sigma}^2 = \frac{S(\widehat{\underline{\theta}})}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \eta_i\left(\widehat{\underline{\theta}}\right)\right)^2.$$

Hence, the distribution of the estimated parameters is approximately determined and we can (like in linear regression) derive standard errors and confidence intervals, or confidence ellipses (or ellipsoids) if multiple variables are considered jointly.

The denominator $n-p$ in the estimator $\widehat{\sigma}^2$ was already introduced in linear regression to ensure that the estimator is unbiased. Tests and confidence intervals were not based on the normal and Chi-square distribution but on the **t- and F-distribution**. They take into account that the estimation of σ^2 causes additional random fluctuation. Even if the distributions are no longer exact, the approximations are more exact if we do this in nonlinear regression too. Asymptotically, the difference between the two approaches goes to zero.

Example d **Membrane Separation Technology (cont'd)** A computer output for the Membrane Separation Technology example can be found in Table 3.3.d. The parameter estimates are in column `Estimate`, followed by the estimated approximate standard error (`Std. Error`) and the test statistics (`t value`), that are approximately t_{n-p} distributed. The corresponding p-values can be found in column `Pr(>|t|)`. The estimated standard deviation $\widehat{\sigma}$ of the random error E_i is here labelled as "`Residual standard error`".

As in linear regression, we can now construct (approximate) confidence intervals. The 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 3.3.d: Summary of the fit of the Membrane Separation Technology example.

Example e **Puromycin (cont'd)** In order to check the influence of treating an enzyme with Puromycin a general model for the data (with and without 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, $z_i = 0$ otherwise). Table 3.3.e shows that the parameter θ_4 is not significantly different from 0 at the 5% level since the p-value of 0.167 is larger then the level (5%). However, the treatment has a clear influence that 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 97.5% 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 3.3.e: Computer output of the fit for the Puromycin example.

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

$$\widehat{\eta}_0 \pm q_{1-\alpha/2}^{t_{n-p}} \cdot \operatorname{se}(\widehat{\eta}_0)$$

where

$$\operatorname{se}(\widehat{\eta}_0) = \widehat{\sigma} \sqrt{\underline{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \underline{x}_0}.$$

Using asymptotic approximations, we can specify confidence intervals for the function values $h(\underline{x}_0; \underline{\theta})$ for nonlinear h. If the function $\eta_0\left(\widehat{\underline{\theta}}\right) := h\left(x_0, \widehat{\underline{\theta}}\right)$ is linearly approximated at $\underline{\theta}$ we get

$$\eta_0(\widehat{\underline{\theta}}) \approx \eta_0(\underline{\theta}) + \underline{a}_0^T \left(\widehat{\underline{\theta}} - \underline{\theta}\right) \text{ where } \underline{a}_0 = \frac{\partial h(x_0, \underline{\theta})}{\partial \underline{\theta}}.$$

If $\underline{x}_0$ is equal to an observed $\underline{x}_i$, $\underline{a}_0$ equals the corresponding row of the matrix $\boldsymbol{A}$ from 3.2.g. The $(1-\alpha)$ confidence interval for the function value $\eta_0(\underline{\theta}) := h(\underline{x}_0, \underline{\theta})$ is then approximately

$$\eta_0\left(\widehat{\underline{\theta}}\right) \pm q_{1-\alpha/2}^{t_{n-p}} \cdot \operatorname{se}\left(\eta_0\left(\widehat{\underline{\theta}}\right)\right),$$

where

$$\operatorname{se}\left(\eta_0\left(\widehat{\underline{\theta}}\right)\right) = \widehat{\sigma} \sqrt{\widehat{\underline{a}}_0^T \left(\boldsymbol{A}\left(\widehat{\underline{\theta}}\right)^T \boldsymbol{A}\left(\widehat{\underline{\theta}}\right)\right)^{-1} \widehat{\underline{a}}_0}.$$

Again, the unknown parameter values are replaced by the corresponding estimates.

g **Confidence Band** The expression for the $(1-\alpha)$ confidence interval for $\eta_0(\underline{\theta}) := h(x_0, \underline{\theta})$ also holds for arbitrary x_0. As in linear regression, it is illustrative to represent the limits of these intervals as a "confidence band" that is a function of x_0. See Figure 3.3.g for the confidence bands for the examples "Puromycin" and "Biochemical Oxygen Demand".

Confidence bands for linear and nonlinear regression functions behave differently: For linear functions the confidence band has minimal width at the center of gravity of the predictor variables and gets wider the further away one moves from the center (see Figure 3.3.g, left). In the nonlinear case, the bands can have arbitrary shape. Because the functions in the "Puromycin" and "Biochemical Oxygen Demand" examples 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 3.3.g, right).

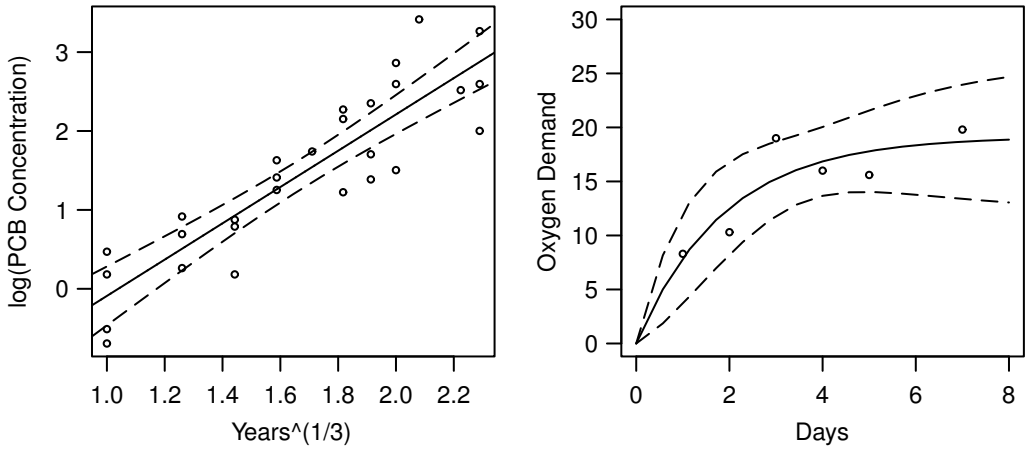

Figure 3.3.g: Left: Confidence band for an estimated line for a linear problem. Right: Confidence band for the estimated curve $h(x, \underline{\theta})$ in the oxygen demand example.

h **Prediction Interval** The confidence band gives us an idea of the **function values** $h(x)$ (the expected values of Y for a given x). However, it does not answer the question where **future observations** Y_0 for given $\underline{x}_0$ will lie. This is often more interesting than the question of the function value itself; for example, we would like to know where the measured value of oxygen demand will lie for an incubation time of 6 days.

Such a statement is a prediction about a **random variable** and should be distinguished from a confidence interval, which says something about a **parameter**, which is a fixed (but unknown) number. Hence, we call the region **prediction interval** or **prognosis interval**. More about this in Chapter 3.7.

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

- there is no one-to-one relationship between parameters and predictor variables. Usually, the number of parameters is different than the number of predictors.
- there are seldom problems where we need to clarify whether an explanatory variable is necessary or not – the model is derived from the underlying theory (e.g., "enzyme kinetics").

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

3.4 More Precise Tests and Confidence Intervals

a The quality of the approximate confidence region that we have seen so far strongly depends on the quality of the linear approximation. Also, the convergence properties of the optimization algorithm are influenced by the quality of the linear approximation. With a somewhat larger computational effort we can check the linearity graphically and – at the same time – we can derive more precise confidence intervals.

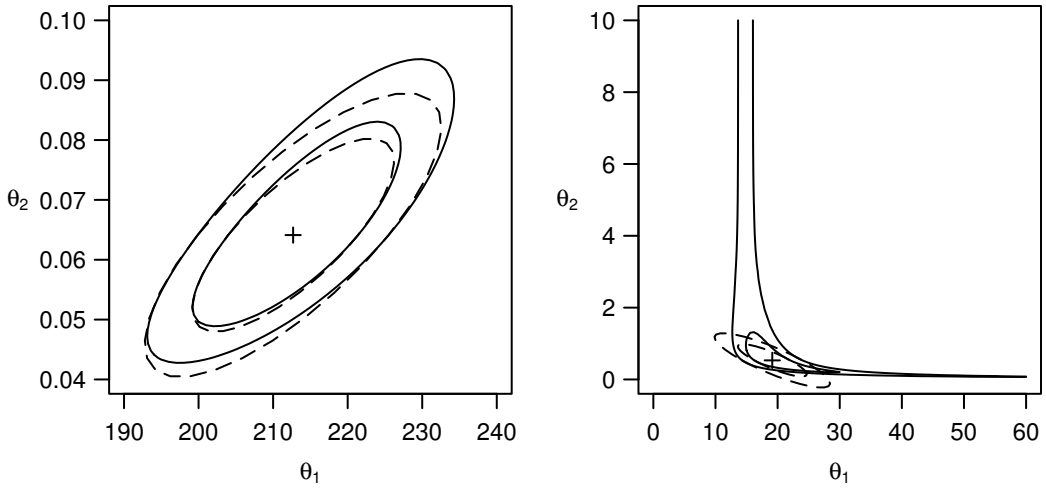

Figure 3.4.c: Nominal 80 and 95% likelihood contours (——) and the confidence ellipses from the asymptotic approximation (– – – –). + denotes the least squares solution. In the Puromycin example (left) the agreement is good and in the oxygen demand example (right) it is bad.

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 $S(\underline{\theta}^*)$ that arises under the null hypothesis with the sum of squares $S(\widehat{\underline{\theta}})$ (for $n \to \infty$ the F- test is the same as the so-called likelihood-ratio test, and the sum of squares is, up to a constant, equal to the negative log-likelihood).

Let us first consider a null-hypothesis $\underline{\theta} = \underline{\theta}^*$ for the whole parameter vector. The test statistic is

$$T = \frac{n-p}{p} \frac{S(\underline{\theta}^*) - S(\widehat{\underline{\theta}})}{S(\widehat{\underline{\theta}})} \overset{(as.)}{\sim} F_{p,n-p}.$$

Searching for all null-hypotheses that are not rejected leads us to the confidence region

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

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-test is *not* based on the linear approximation in 3.2.g and hence is (much) more exact.

c **Confidence Regions for p=2** For $p = 2$, we can find the confidence regions by calculating $S(\underline{\theta})$ on a grid of $\underline{\theta}$ values and determine the borders of the region through interpolation, as is common for contour plots. Figure 3.4.c illustrates both the confidence region based on the linear approximation and based on the F-test for the example "Puromycin" (left) and for "Biochemical Oxygen Demand" (right).

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

d **F-Test for Individual Parameters** Now we focus on the the question whether an individual parameter θ_k is equal to a certain value θ_k^*. Such a null hypothesis makes *no* statement about the remaining parameters. The model that fits the data best for a fixed $\theta_k = \theta_k^*$ is given by the least squares solution of the remaining parameters. So, $S(\theta_1, \ldots, \theta_k^*, \ldots, \theta_p)$ is minimized with respect to θ_j, $j \neq k$. We denote the minimum by $\widetilde{S}_k$ and the minimizer θ_j by $\widetilde{\theta}_j$. Both values depend on θ_k^*. We therefore write $\widetilde{S}_k(\theta_k^*)$ and $\widetilde{\theta}_j(\theta_k^*)$.

The test statistic for the F-test (with null hypothesis $H_0 : \theta_k = \theta_k^*$) is given by

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

It follows (approximately) an $F_{1,n-p}$ distribution.

We can now construct a confidence interval by (numerically) solving the equation $\widetilde{T}_k = q_{0.95}^{F_{1,n-p}}$ for θ_k^*. It has a solution that is less 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. Hence, both tests are equivalent. In nonlinear regression, the F-test is not equivalent to the t-test discussed in the last chapter (3.3.d). However, we can transform the F-test to a t-test that is more accurate than the one of the last chapter (that was based on the linear approximation):

From the test statistic of the F-test, we take the square-root and add the sign of $\widehat{\theta}_k - \theta_k^*$,

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

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

In the linear regression model, T_k is – as already pointed out – equal to the test statistic of the usual t-test,

$$T_k(\theta_k^*) = \frac{\widehat{\theta}_k - \theta_k^*}{\operatorname{se}\left(\widehat{\theta}_k\right)}.$$

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_0. For this we reparameterize the original problem so that a parameter, say ϕ_1, represents the function value $h(x_0)$ and proceed as in 3.4.d.

3.5 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 (3.4.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 can be

seen from 3.4.e, a straight line, while for nonlinear regression the result is a monotone increasing function. The graphical comparison of $T_k(\theta_k)$ with a straight line is the so-called **profile t-plot**. Instead of θ_k, it is common to use a standardized version

$$\delta_k(\theta_k) := \frac{\theta_k - \widehat{\theta}_k}{\mathrm{se}\left(\widehat{\theta}_k\right)}$$

on the horizontal axis because it is used in the linear approximation. The comparison line is then the "diagonal", i.e. the line with slope 1 and intercept 0.

The more the profile t-function is curved, the stronger 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 $|\delta_k(\theta_k)| \leq 2.5$). In Figure 3.5.a it is evident that in the Puromycin example the nonlinearity is minimal, while in the Biochemical Oxygen Demand example it is large.

In Figure 3.5.a we can also read off the confidence intervals according to 3.4.e. For convenience, the probabilites $P(T_k \leq t)$ of the corresponding t-distributions are marked on the right vertical axis. For the Biochemical Oxygen Demand example this results in a confidence interval without upper bound!

b **Likelihood Profile Traces** The **likelihood profile traces** are another useful graphical tool. Here the estimated parameters $\widetilde{\theta}_j$, $j \neq k$ for fixed θ_k (see 3.4.d) are considered as functions $\widetilde{\theta}_j^{(k)}(\theta_k)$.

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 $\widetilde{\theta}_j^{(k)}(\theta_k)$ we superimpose $\widetilde{\theta}_k^{(j)}(\theta_j)$ in mirrored form so that the axes have the same meaning for both functions.

Figure 3.5.b shows one of these diagrams for both our two examples. Additionally, contours of confidence regions for $[\theta_1, \theta_2]$ are plotted. It can be seen that that the profile traces intersect the contours at points where they have horizontal or vertical tangents.

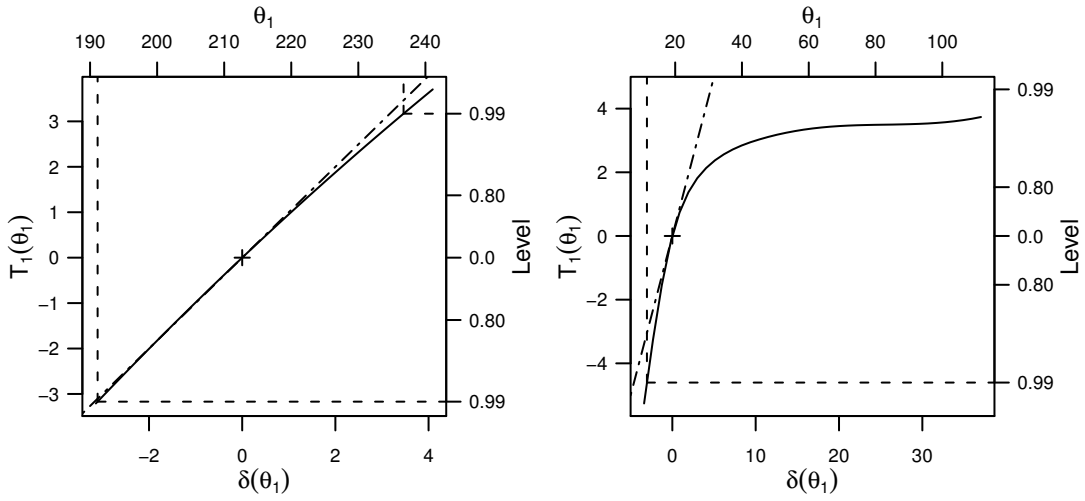

Figure 3.5.a: Profile t-plot for the first parameter for both the Puromycin (left) and the Biochemical Oxygen Demand example (right). The dashed lines show the applied linear approximation and the dotted line the construction of the 99% confidence interval with the help of $T_1(\theta_1)$.

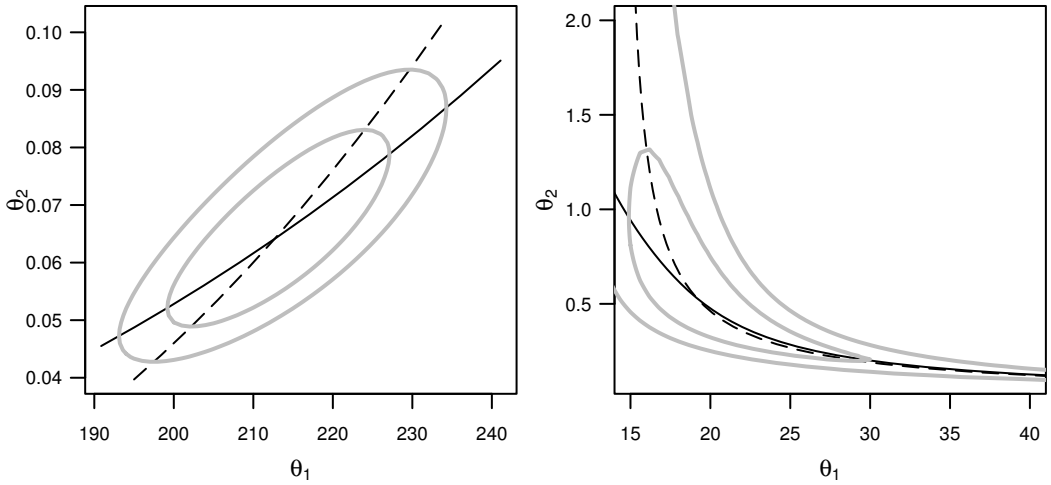

Figure 3.5.b: Likelihood profile traces for the Puromycin and Oxygen Demand examples, with 80%- and 95% confidence regions (gray curves).

The representation does not only show the nonlinearities, but is also useful for the understanding of **how the parameters influence each other**. To understand this, we go back to the case of a linear regression function. The profile traces in the individual diagrams then consist of two lines, that intersect at the point $[\widehat{\theta}_1, \widehat{\theta}_2]$. If we standardize the parameter by using $\delta_k(\theta_k)$ from 3.5.a, one can show that the slope of the trace $\widetilde{\theta}_j^{(k)}(\theta_k)$ is equal to the correlation coefficient c_{kj} of the estimated coefficients $\widehat{\theta}_j$ and $\widehat{\theta}_k$. The "reverse line" $\widetilde{\theta}_k^{(j)}(\theta_j)$ then has, compared with the horizontal axis, a slope of $1/c_{kj}$. The angle between the lines is thus a monotone function of the correlation. It therefore measures the **collinearity** between the two predictor variables. If the correlation between the parameter estimates is zero, then the traces are orthogonal 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 interplay, and hence how their estimators are correlated.

Example c **Membrane Separation Technology (cont'd)** All profile t-plots and profile traces can be put in a triangular matrix, as can be seen in Figure 3.5.c. Most profile traces are strongly curved, meaning that 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 straight 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 strongly collinear. Parameter θ_2 is also collinear with θ_3 and θ_4, although more weakly.

d * **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 demanding than for the corresponding exact likelihood contours.

3.6 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**.

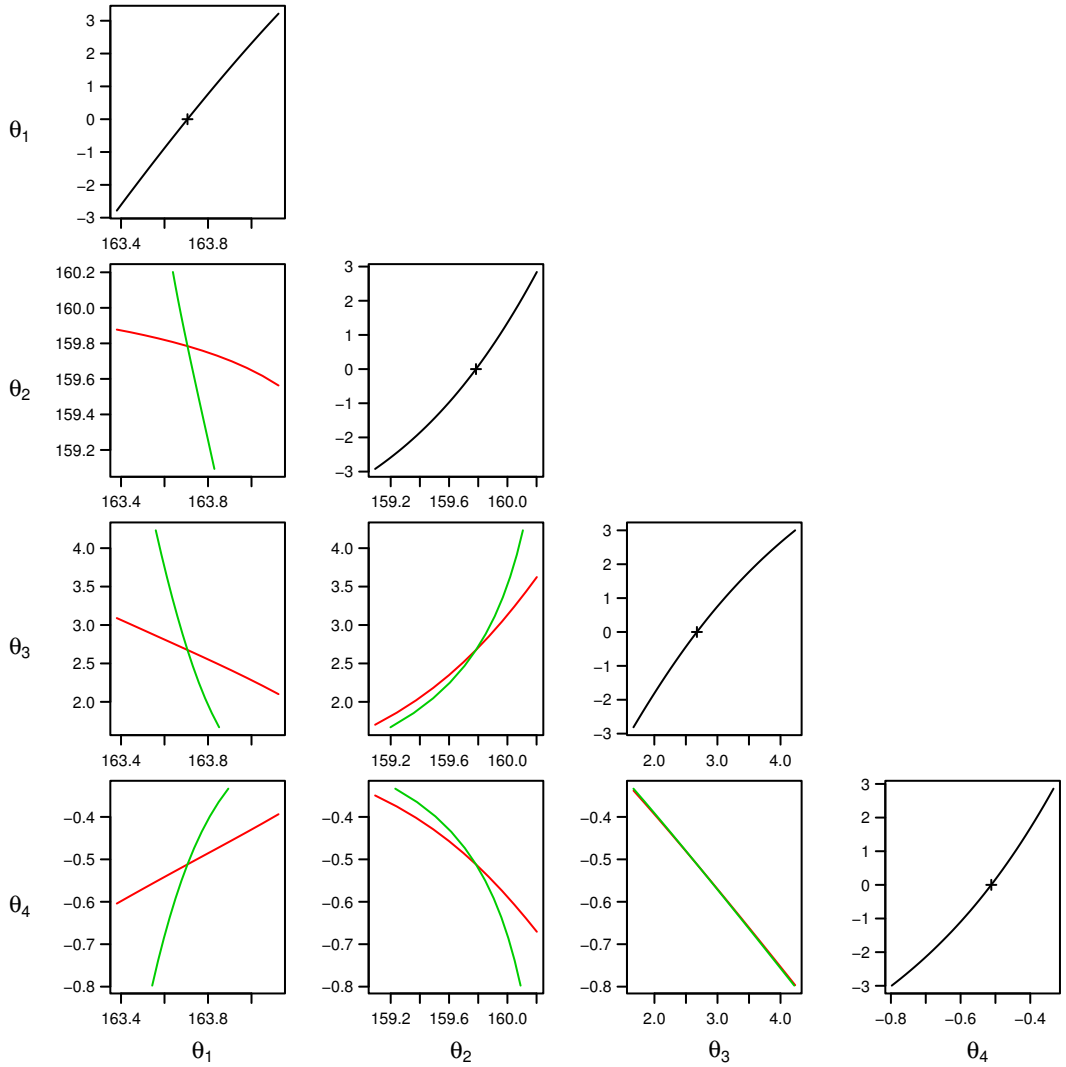

Figure 3.5.c: Profile t-plots and Profile Traces for the Example "Membrane Separation Technology". The + in the profile t-plot denotes the least squares solution.

We point out that parameter transformations, unlike transformations of the response variable (see 3.1.h), do *not* change the statistical part of the model. Hence, they are *not* helpful if the assumptions about the distribution of the random error are violated. It is the quality of the linear approximation and the statistical statements based on it that are being changed!

Sometimes the transformed parameters are very difficult **to interpret**. The important questions often concern individual parameters – the original parameters. Nevertheless, we can work with transformations: We derive more accurate confidence regions for the transformed parameters and can transform them (the confidence regions) back to get results for the original parameters.

b **Restricted Parameter Regions** Often, the admissible 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, parameter estimation will end up in a reasonable range if the model describes the data well and the data contain 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 border of the admissible region. The simplest way to deal with such problems is via transformation of the parameter.

Examples

- The parameter θ should be positive. Through a transformation $\theta \longrightarrow \phi = \ln(\theta)$, $\theta = \exp(\phi)$ is always positive for all possible values of $\phi \in \mathbb{R}$:

$$h(x, \theta) \longrightarrow h(x, \exp(\phi)).$$

- The parameter should lie in the interval (a, b). With the log transformation $\theta = a + (b - a)/(1 + \exp(-\phi))$, θ can (for arbitrary $\phi \in \mathbb{R}$) only take values in (a, b).
- In the model

$$h(x, \underline{\theta}) = \theta_1 \exp(-\theta_2 x) + \theta_3 \exp(-\theta_4 x)$$

with $\theta_2, \theta_4 > 0$ the parameter pairs (θ_1, θ_2) and (θ_3, θ_4) are interchangeable, i.e. $h(x, \underline{\theta})$ does not change. This can create uncomfortable optimization problems, because the solution is not unique. The constraint $0 < \theta_2 < \theta_4$ that ensures the uniqueness is achieved via the transformation $\theta_2 = \exp(\phi_2)$ und $\theta_4 = \exp(\phi_2)(1 + \exp(\phi_4))$. The function is now

$$h(x, (\theta_1, \phi_2, \theta_3, \phi_4)) = \theta_1 \exp\left(-\exp(\phi_2)x\right) + \theta_3 \exp\left(-\exp(\phi_2)(1 + \exp(\phi_4))x\right).$$

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

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

If all x values are positive, both vectors

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

tend to disturbing collinearity. This collinearity can be avoided if we use **centering**. The model can be written as $h(x; \underline{\theta}) = \theta_1 \exp(-\theta_2(x - x_0 + x_0))$ With the re-parameterization $\phi_1 := \theta_1 \exp(-\theta_2 x_0)$ and $\phi_2 := \theta_2$ we get

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

The derivative vectors are approximately orthogonal if we chose the mean value of the x_i for x_0.

Example d Membrane Separation Technology (cont'd) In this example it is apparent from the approximate correlation matrix (Table 3.6.d, left half) that the parameters θ_3 and θ_4 are strongly correlated (we have already observed this in 3.5.c using the profile traces).

If the model is re-parameterized to

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

with $\widetilde{\theta}_3 = \theta_3 + \theta_4 \, \mathrm{med}(x_j)$, an improvement is achieved (right half of Table 3.6.d).

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

Example e Membrane Separation Technology (cont'd) The parameter transformation in 3.6.d leads to a satisfactory result, as far as correlation is concerned. If we look at the likelihood contours or the profile t-plot and the profile traces, the parameterization is still not satisfactory.

An intensive search for further improvements leads to the following transformations that turn out to have satisfactory profile traces (see Figure 3.6.e):

$$\widetilde{\theta}_1 := \frac{\theta_1 + \theta_2\, 10^{\widetilde{\theta}_3}}{10^{\widetilde{\theta}_3} + 1}\,, \qquad \widetilde{\theta}_2 := \log_{10}\left(\frac{\theta_1 - \theta_2}{10^{\widetilde{\theta}_3} + 1}\, 10^{\widetilde{\theta}_3}\right)\,,$$
$$\widetilde{\theta}_3 := \theta_3 + \theta_4 \operatorname{med}(x_j) \qquad \widetilde{\theta}_4 := 10^{\theta_4}\,.$$

The model is now

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

and we get the result shown in Table 3.6.e

```
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 3.6.e: Membrane Separation Technology: Summary of the fit after parameter transformation.

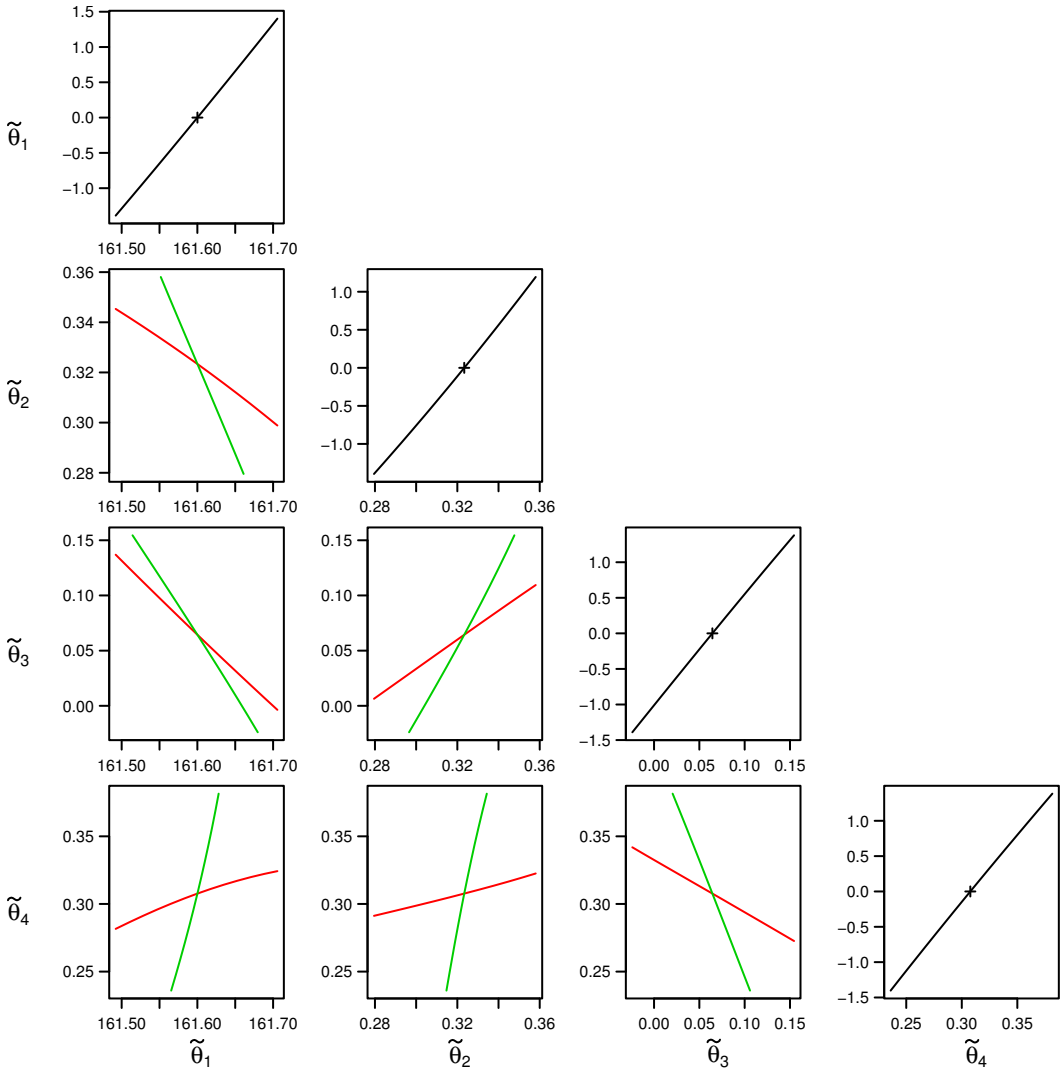

Figure 3.6.e: Profile t-plot and profile traces for the Membrane Separation Technology example according to the given transformations.

f **More Successful Reparametrization** It turned out that a **successful reparametrization is very data set specific**. A reason is that nonlinearities and correlations between estimated parameters depend on the (estimated) parameter vector itself. Therefore, no generally valid recipe can be given. This makes the search for appropriate reparametrizations often very difficult.

g **Confidence Intervals on the Original Scale (Alternative Approach)** Even though parameter transformations help us in situations where we have problems with convergence of the algorithm or the quality of confidence intervals, the original parameters often remain the quantity of interest (e.g., because they have a nice physical interpretation). Consider the transformation $\theta \longrightarrow \phi = \ln(\theta)$. Fitting the model results in an estimator $\widehat{\phi}$ and an estimated standard error $\widehat{\sigma}_{\widehat{\phi}}$. Now we can construct a confidence interval for θ. We have to search all θ for which $\ln(\theta)$ lies 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 $\theta = g(\phi)$. Then

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

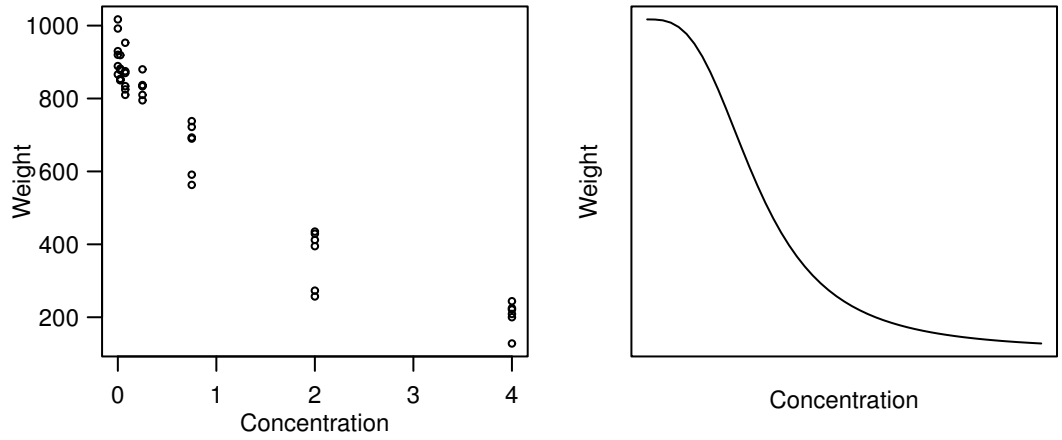

Figure 3.7.b: Cress Example. Left: Representation of the data. Right: A typical shape of the applied regression function.

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

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

However, this approach should only be used if the way via the F-test from Chapter 3.4 is not possible.

3.7 Forecasts and Calibration

Forecasts

a Besides the question of the set of plausible parameters (with respect to the given data, which we also call training data set), the question of the range of future observations is often of central interest. The difference between these two questions was already discussed in 3.3.h. In this chapter we want to answer the second question. We assume that the parameter $\underline{\theta}$ is estimated using the least squares method. What can we now say about a future observation Y_0 at a given point x_0?

Example b **Cress** 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 response variable Y were made on each of 7 soil samples with predetermined (or measured with the largest possible precision) concentrations x. Hence, we assume that 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:

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

The data and the function $h(\cdot)$ are illustrated in Figure 3.7.b. We can now ask ourselves which weight values will we see at a concentration of e.g. $x_0 = 3$?

c **Approximate Forecast Intervals** We can estimate the expected value $E(Y_0) = h(x_0, \theta)$ of the variable of interest Y at the point x_0 by $\widehat{\eta}_0 := h(x_0, \widehat{\theta})$. We also want to get an interval where a future observation will lie with high probability. So, we do not only have to take into account the randomness of the estimate $\widehat{\eta}_0$, but also the random error E_0. Analogous to linear regression, an at least approximate $(1 - \alpha)$ forecast interval is given by

$$\widehat{\eta}_0 \pm q_{1-\alpha/2}^{t_{n-p}} \cdot \sqrt{\widehat{\sigma}^2 + \left(\operatorname{se}(\widehat{\eta}_0)\right)^2}.$$

The calculation of $\operatorname{se}(\widehat{\eta}_0)$ can be found in 3.3.f.

* **Derivation** The random variable Y_0 is the value of interest for an observation with predictor variable value x_0. 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,

$$R_0 = Y_0 - h\left(x_0, \widehat{\underline{\theta}}\right) = \left(Y_0 - h\left(x_0, \underline{\theta}\right)\right) - \left(h\left(x_0, \widehat{\underline{\theta}}\right) - h\left(x_0, \underline{\theta}\right)\right).$$

Even if $\underline{\theta}$ is unknown, we know the distribution of the expressions in parentheses: Both are normally distributed random variables and they are independent because the first only depends on the "future" observation Y_0, the second only on the observations $Y_1, \ldots, Y_n$ that led to the estimated curve. Both have expected value 0; the variances add up to

$$\operatorname{Var}(R_0) \approx \sigma^2 + \sigma^2 \underline{a}_0^T (A^T A)^{-1} \underline{a}_0.$$

The described forecast interval follows by replacing the unknown values by their corresponding estimates.

d **Forecast Versus Confidence Intervals** If the sample size n of the training data set is very large, the estimated variance is dominated by the error variance $\widehat{\sigma}^2$. This means that the uncertainty in the forecast is then primarily caused by the random error. The second term in the expression for the variance 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 error of the observation must also be taken into account. The endpoints of such intervals are shown in Figure 3.7.i (left).

e * **Quality of the Approximation** The derivation of the forecast interval in 3.7.c is based on the same approximation as in Chapter 3.3. The quality of the approximation can again be checked graphically.

f **Interpretation of the "Forecast Band"** The interpretation of the "forecast band" (as shown in Figure 3.7.i), is not straightforward. From our derivation it holds that

$$P\left(V_0^*(x_0) \leq Y_0 \leq V_1^*(x_0)\right) = 0.95,$$

where $V_0^*(x_0)$ is the lower and $V_1^*(x_0)$ the upper bound of the prediction interval for $h(x_0)$. 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* binomially distributed with $\pi = 0.95$. The events that the individual future observations fall in the band are not independent; they depend on each other through the random borders V_0 and V_1. If, for example, the estimation of $\widehat{\sigma}$ randomly turns out to be too small, the band is too narrow for *all* future observations, and too many observations would lie outside the band.

Calibration

g The actual goal of the experiment in the **cress example** is to estimate the concentration of the agrochemical material from the weight of the cress. This means that we would like to use the regression relationship in the "wrong" direction. This will cause problems with statistical inference. Such a procedure is often desired to **calibrate** a measurement method or to predict 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 keyword for finding this topic is **inverse regression**.

Here, we would like to present a simple method that gives useable results if simplifying assumptions hold.

h **Procedure under Simplifying Assumptions** We assume that the predictor values x have no measurement error. In our example this holds true if the concentrations of the agrochemical material are determined very carefully. For several soil samples with many different possible concentrations we carry out several independent measurements of the response value Y. This results in a training data set that is used to estimate the unknown parameters and the corresponding parameter errors.

Now, for a given value y_0 it is obvious to determine the corresponding x_0 value by simply inverting the regression function:

$$\widehat{x}_0 = h^{-1}(y_0, \underline{\widehat{\theta}}).$$

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

i **Accuracy of the Obtained Values** Of course we now face the question about the accuracy of $\widehat{x}_0$. The problem seems to be similar to the prediction problem. However, here we observe y_0 and the corresponding value x_0 has to be estimated.

The answer can be formulated as follows: We treat x_0 as a *parameter* for which we want a confidence interval. Such an interval can be constructed (as always) from a test. We take as null hypothesis $x = x_0$. As we have seen in 3.7.c, Y lies with probability 0.95 in the forecast interval

$$\widehat{\eta}_0 \pm q_{1-\alpha/2}^{t_{n-p}} \cdot \sqrt{\widehat{\sigma}^2 + \left(\operatorname{se}(\widehat{\eta}_0)\right)^2},$$

where $\widehat{\eta}_0$ was a compact notation for $h(x_0, \widehat{\theta})$. Therefore, this interval is an acceptance interval for the value Y_0 (which here plays the role of a test statistic) under the null hypothesis $x = x_0$. Figure 3.7.i illustrates all prediction intervals for all possible values of x_0 for the given interval in the Cress example.

j **Illustration** Figure 3.7.i (right) illustrates the approach for the Cress example: Measured values y_0 are compatible with parameter values x_0 in the sense of the test, if the point $[x_0, y_0]$ lies in the (prediction interval) band. Hence, we can thus determine the set of values of x_0 that are compatible with a given observation y_0. They form the dashed interval, which can also be described as the set

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

This interval is now the desired confidence interval (or **calibration interval**) for x_0.

If we have m values to determine y_0, we apply the above method to $\bar{y}_0 = \sum_{j=0}^{m} y_{0j}/m$ and get

$$\left\{ x \,:\, |\bar{y}_0 - h(x, \underline{\widehat{\theta}})| \leq \sqrt{\widehat{\sigma}^2 + \left(\operatorname{se}(h(x, \underline{\widehat{\theta}}))\right)^2} \cdot q_{1-\alpha/2}^{t_{n-p}} \right\}.$$

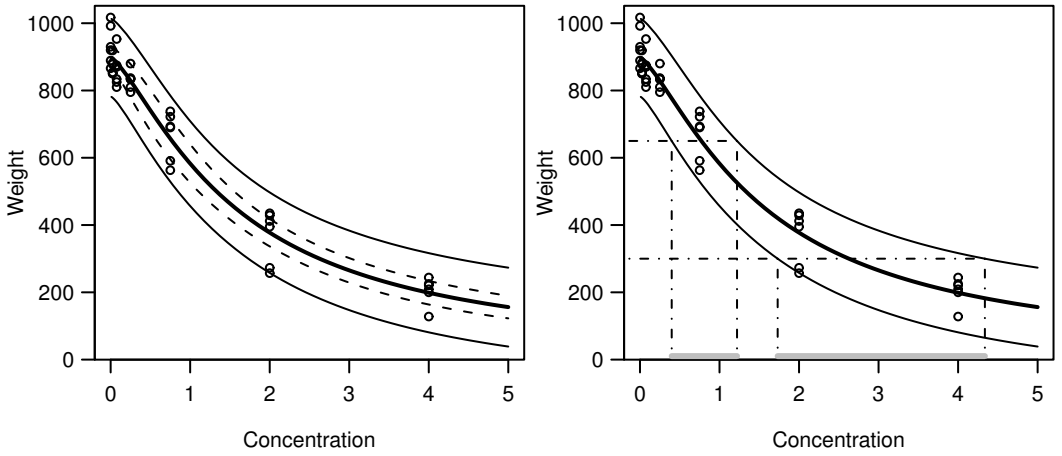

Figure 3.7.i: Cress example. Left: Confidence band for the estimated regression curve (dashed) and forecast band (solid). Right: Schematic representation of how a calibration interval is determined, at the points $y_0 = 650$ and $y_0 = 350$. The resulting intervals are [0.4, 1.22] and [1.73, 4.34], respectively.

k In this chapter, only one of many possibilities for determining a calibration interval was presented.

3.8 Closing Comments

a **Reason for the Difficulty in the Biochemical Oxygen Demand Example** Why did we have so many problems with the Biochemical Oxygen Demand example? Let us have a look at Figure 3.1.e and remind ourselves that the parameter θ_1 represents the expected oxygen demand for infinite incubation time, so it is clear that it is difficult to estimate θ_1, because the horizontal asymptote is badly determined by the given data. 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. See, e.g. Huet, Bouvier, Gruet and Jolivet (1996) where also the case of non-constant variance (heteroscedastic models) is discussed. It is also worth taking a look at the book of Carroll and Ruppert (1988).

c **Correlated Errors** Here we always assumed that the errors E_i are independent. Like in linear regression analysis, nonlinear regression models can also be extended to handle **correlated errors** and **random effects**.

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 (on a reduced set of parameters).

e Literature Notes This chapter is mainly based on the book of 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 of Ratkowsky (1989) contains many nonlinear functions $h(\cdot)$ that are primarily used in biological applications.