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,

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

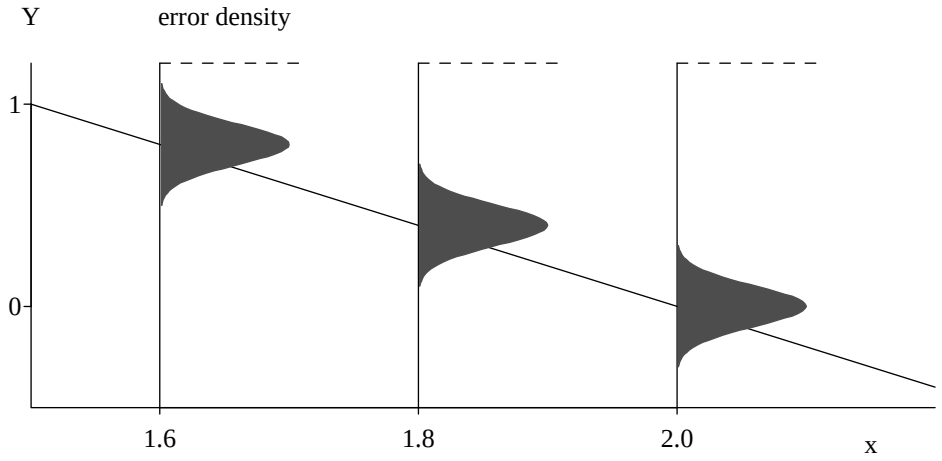

Figure 2.1.a: Display of the probability model $Y_i = 4 - 2x_i + E_i$ for 3 observations Y_1, Y_2 and Y_3 corresponding to the x values $x_1 = 1.6$, $x_2 = 1.8$ and $x_3 = 2$.

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

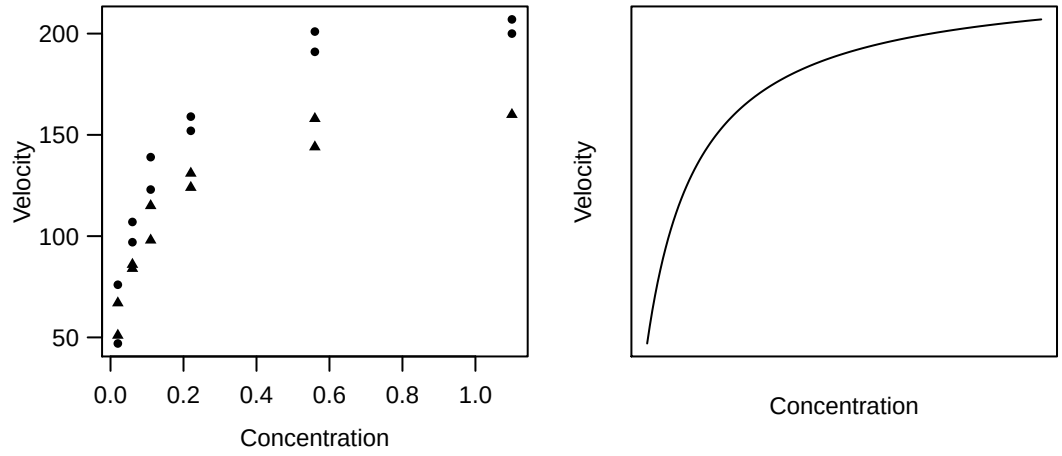

Figure 3.1.d: Puromycin. (a) Data (• treated enzyme; △ untreated enzyme) and (b) typical shape of the regression function.

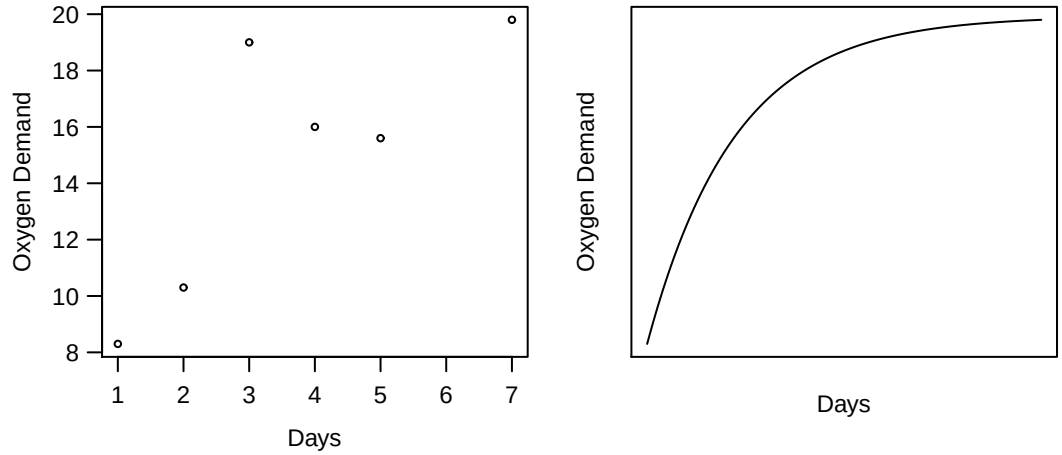

Figure 3.1.e: Biochemical Oxygen Demand. (a) Data and (b) typical shape of the regression function.

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.

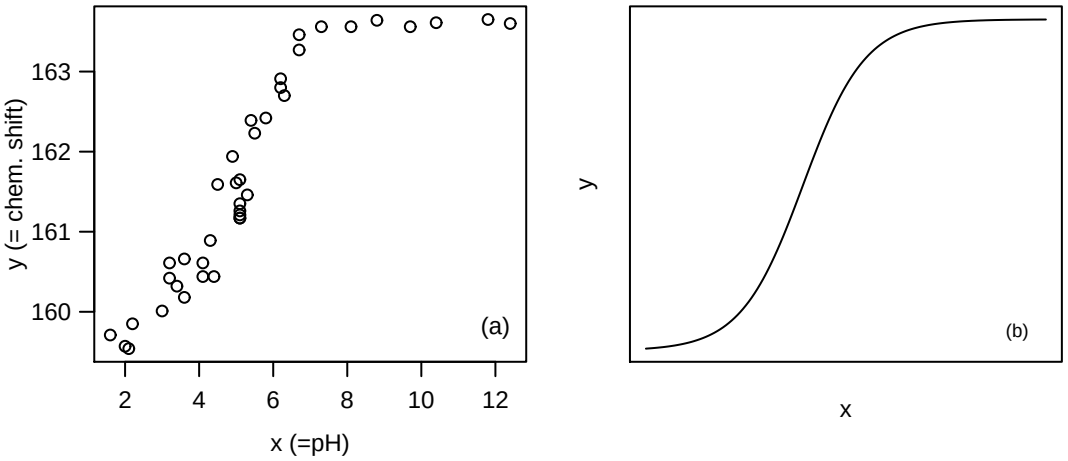

Figure 3.1.f: Membrane Separation Technology. (a) Data and (b) a typical shape of the regression function.

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):

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

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

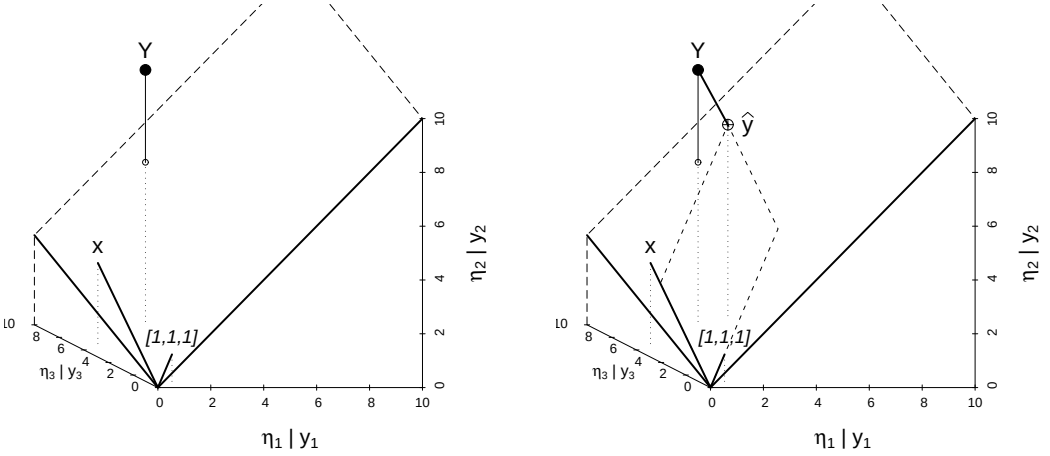

Figure 3.2.c: Illustration of simple linear regression. Values of $\underline{\eta}(\underline{\beta}) = \beta_0 + \beta_1 \underline{x}$ for varying parameters $[\beta_0, \beta_1]$ lead to a plane in three-dimensional space. The right plot also shows the point on the surface that is closest to $\underline{Y}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

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

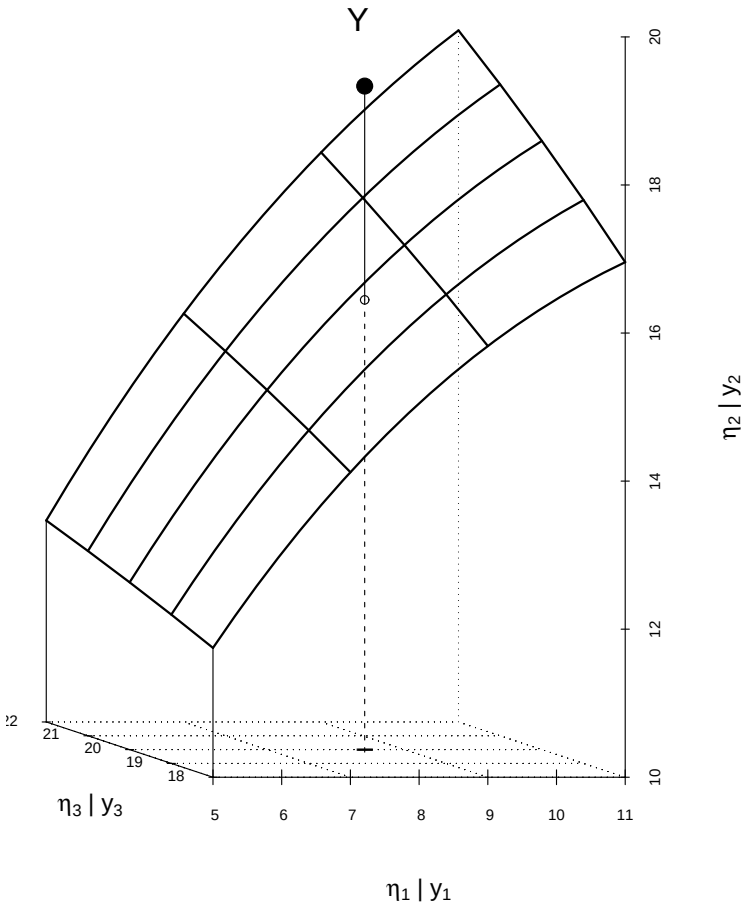

Figure 3.2.d: Geometrical illustration of nonlinear regression. The values of $\underline{\eta}(\underline{\theta}) = h(\underline{x}; \theta_1, \theta_2)$ for varying parameters $[\theta_1, \theta_2]$ lead to a two-dimensional "model surface" in three-dimensional space. The lines on the model surface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

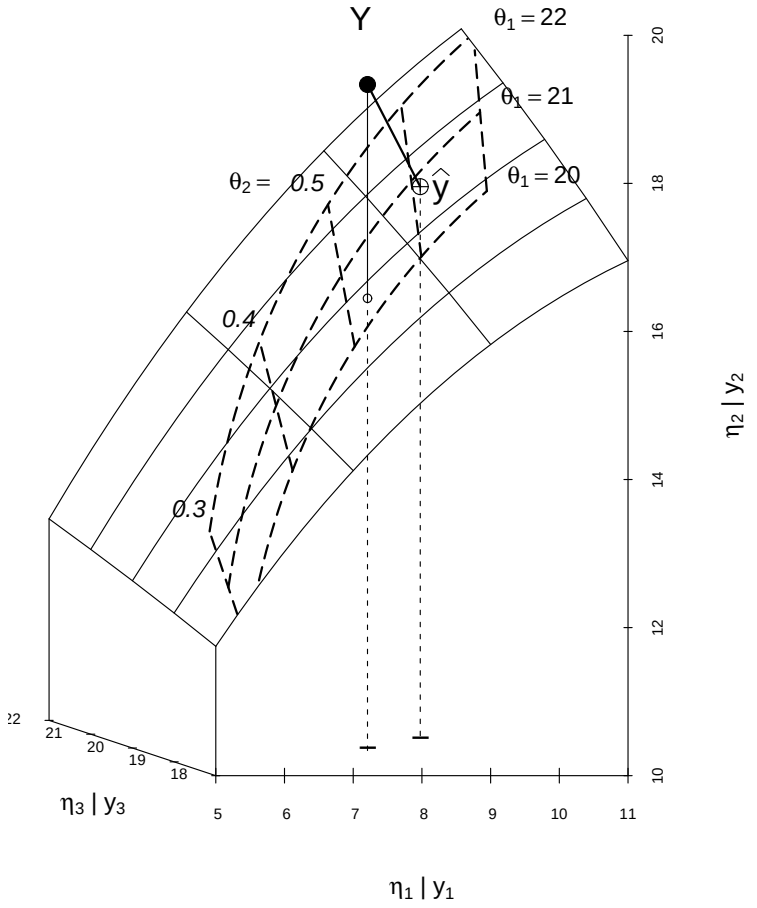

Figure 3.2.e: Biochemical Oxygen Demand: Geometrical illustration of nonlinear regression. In addition, we can see here the lines of constant θ_1 and θ_2, respectively. The vector of the estimated model values $\underline{\widehat{y}} = h\left(\underline{x}; \underline{\widehat{\theta}}\right)$ is the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

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

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

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

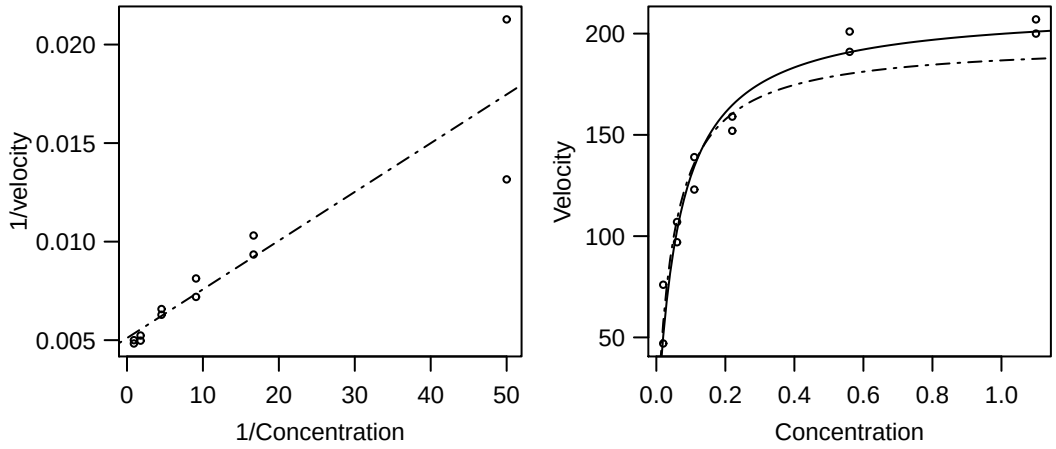

Figure 3.2.k: Puromycin. Left: Regression function in the linearized problem. Right: Regression function $h(x;\underline{\theta})$ for the initial values $\underline{\theta} = \underline{\theta}^{(0)}$ (- - - - -) and for the least squares estimation $\underline{\theta} = \widehat{\underline{\theta}}$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

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**.

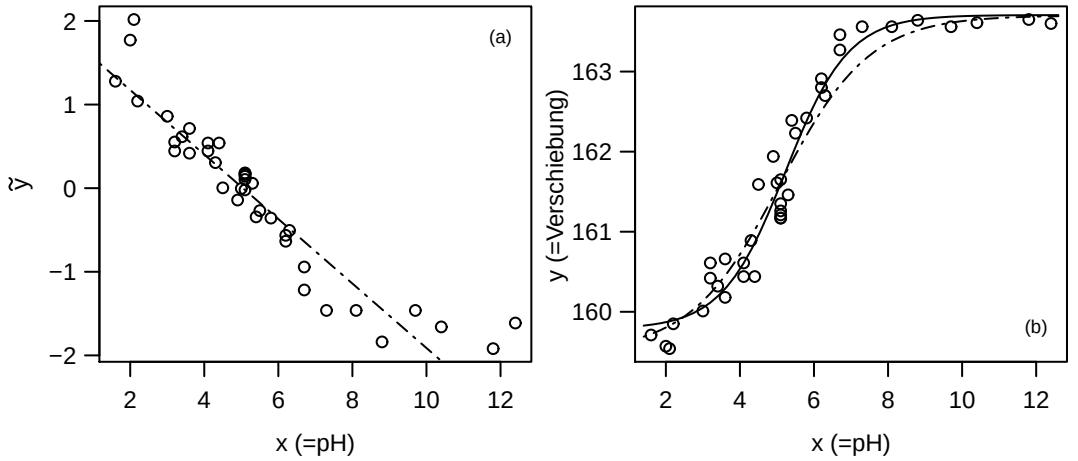

Figure 3.2.m: Membrane Separation Technology. (a) Regression line that is used for determining the initial values for θ_3 and θ_4. (b) Regression function $h(x; \underline{\theta})$ for the initial value $\underline{\theta} = \underline{\theta}^{(0)}$ (- - - - -) and for the least squares estimator $\underline{\theta}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

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 $\eta_0\left(\underline{\theta}\right) := h\left(x_0, \underline{\theta}\right)$ is linearly approximated at $\underline{\theta}$ we get

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

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

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

where

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

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).

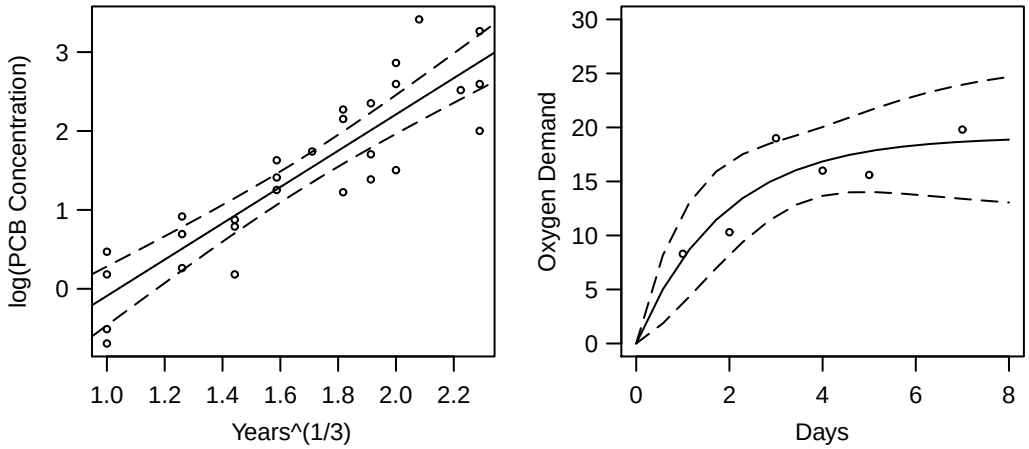

Figure 3.3.g: Left: Confidence band for an estimated line for a linear problem. Right: Confidence band for the estimated curve $h(x, \underline{\theta})$ in the oxygen demand example.

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.

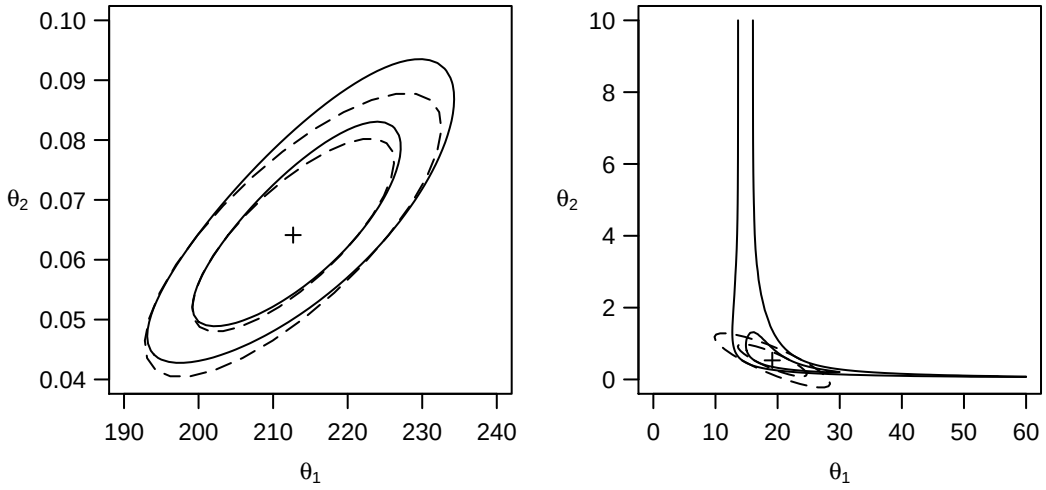

Figure 3.4.c: Nominal 80 and 95% likelihood contours (——) and the confidence ellipses from the asymptotic approximation (– – – –). + denotes the least squares solution. In the Puromycin example (left) the agreement is good and in the oxygen demand example (right) it is bad.

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

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

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

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

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.

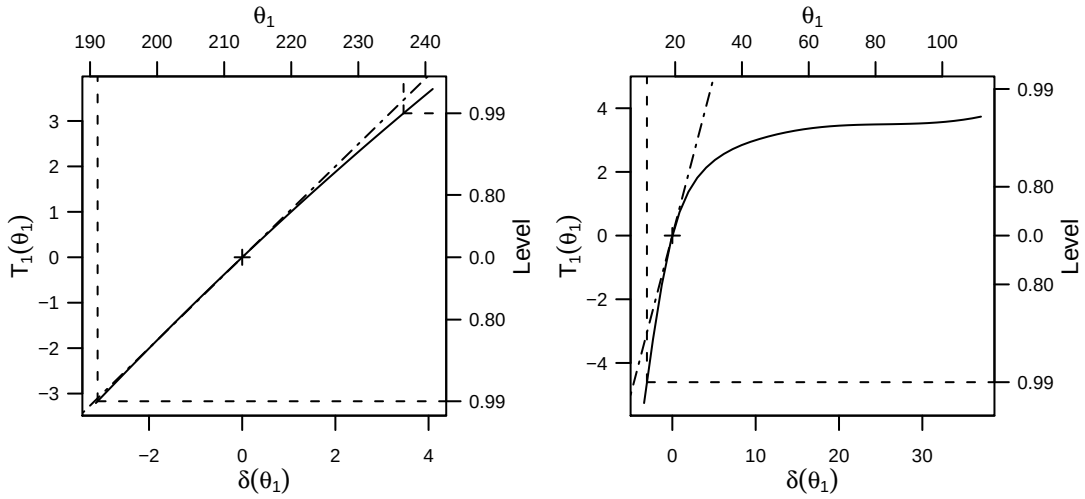

Figure 3.5.a: Profile t-plot for the first parameter for both the Puromycin (left) and the Biochemical Oxygen Demand example (right). The dashed lines show the applied linear approximation and the dotted line the construction of the 99% confidence interval with the help of $T_1(\theta_1)$.

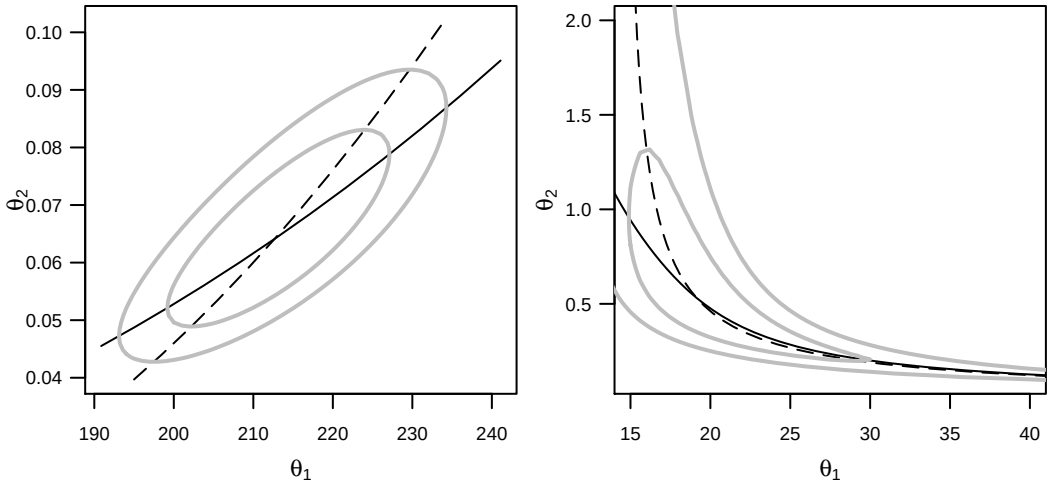

Figure 3.5.b: Likelihood profile traces for the Puromycin and Oxygen Demand examples, with 80%- and 95% confidence regions (gray curves).

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**.

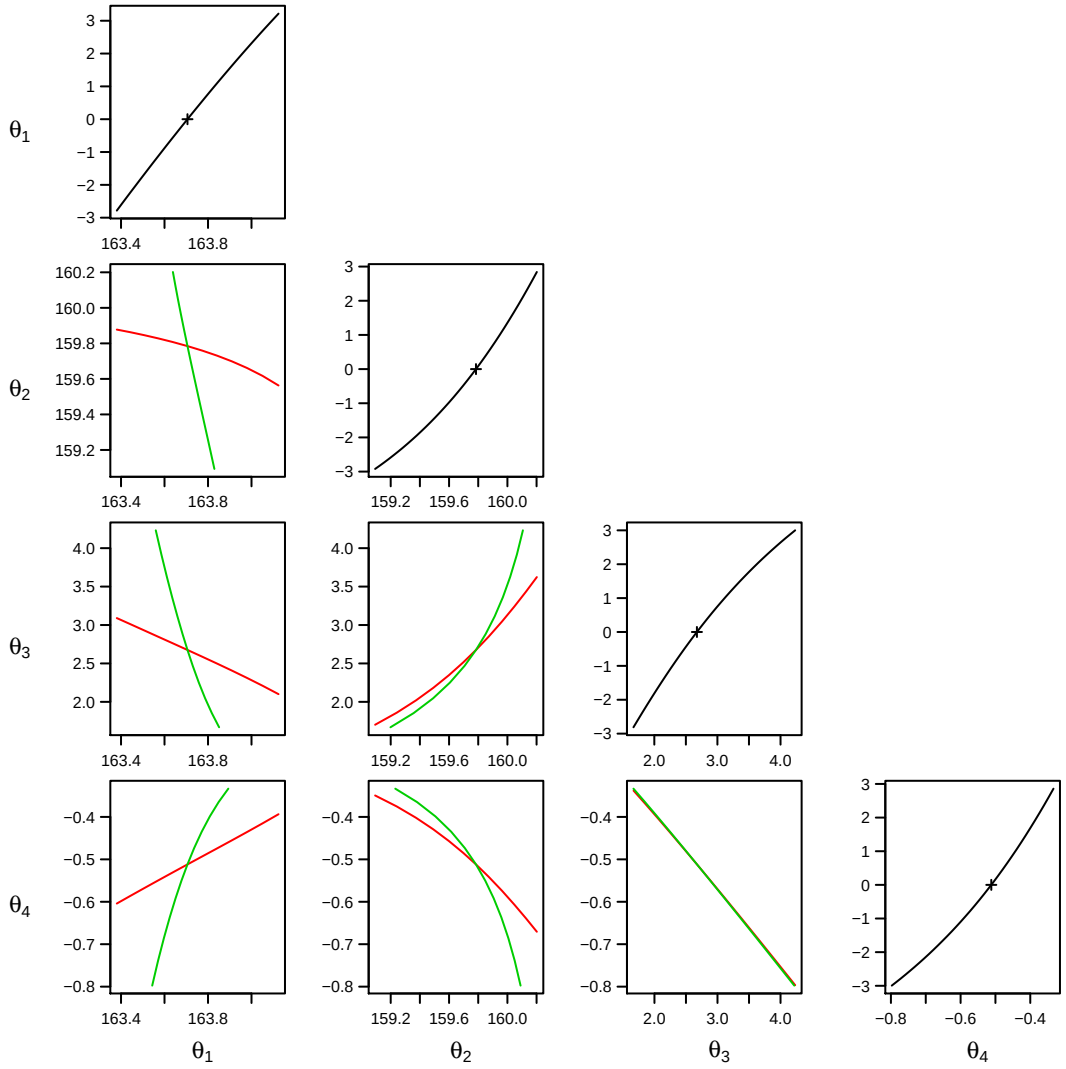

Figure 3.5.c: Profile t-plots and Profile Traces for the Example "Membrane Separation Technology". The + in the profile t-plot denotes the least squares solution.

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.

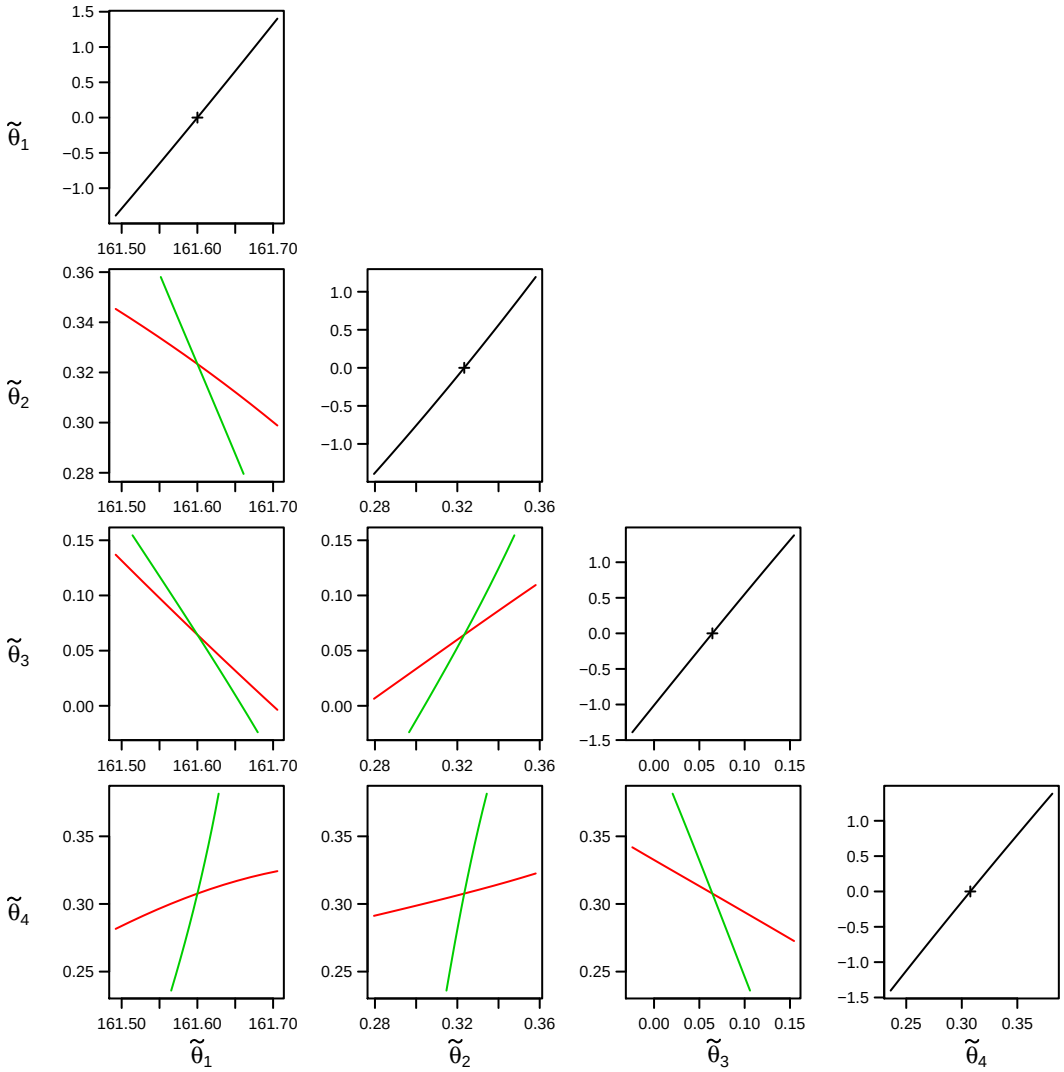

Figure 3.6.e: Profile t-plot and profile traces for the Membrane Separation Technology example according to the given transformations.

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

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

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,

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

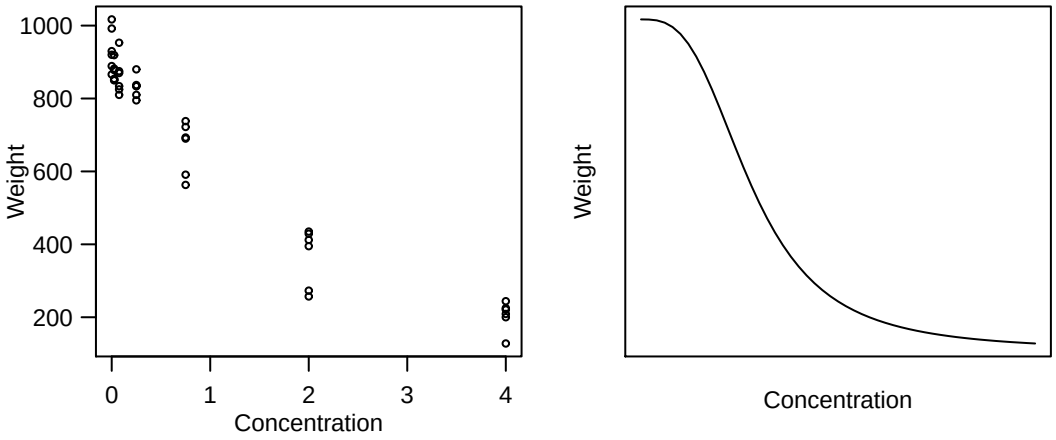

Figure 3.7.b: Cress Example. Left: Representation of the data. Right: A typical shape of the applied regression function.

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

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

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:

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

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

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

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

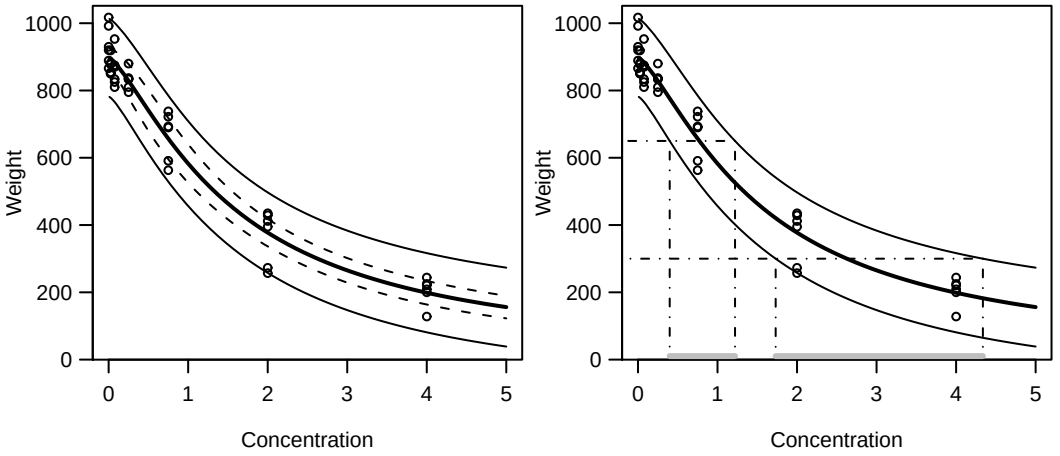

Figure 3.7.i: Cress example. Left: Confidence band for the estimated regression curve (dashed) and forecast band (solid). Right: Schematic representation of how a calibration interval is determined, at the points $y_0 = 650$ and $y_0 = 350$. The resulting intervals are $[0.4, 1.22]$ and $[1.73, 4.34]$,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

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

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

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.

4 Analysis of Variance and Design of Experiments

Preliminary Remark Analysis of variance (ANOVA) and design of experiments are both topics that are usually covered in separate lectures of about 30 hours. Here, we can only give a very brief overview. However, for many of you it may be worthwhile to study these topics in more detail later.

Analysis of variance addresses models where the response variable Y is a function of **categorical predictor variables** (so called **factors**). We have already seen how such predictors can be applied in a linear regression model. This means that analysis of variance can be viewed as a special case of regression modeling. However, it is worthwhile to study this special case separately. Analysis of variance and linear regression can be summarized under the term **linear model**.

Regarding design of experiments we only cover one topic, the **optimization of a response variable**. If time permits, we will also discuss some more general aspects.

4.1 Multiple Groups, One-Way ANOVA

a We observe g groups of values

$$Y_{hi} = \mu_h + E_{hi} \qquad i = 1, 2, \ldots, n_h; \qquad h = 1, 2, \ldots, g,$$

where $E_{hi} \sim \mathcal{N}(0, \sigma^2)$, independent.

The question of interest is whether there is a difference between the μ_h's.

b **Null hypothesis** $H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$.

Alternative $H_A : \mu_h \neq \mu_k$ for at least one pair (h, k).

Test statistic
Based on the average of each group $\overline{Y}_{h.} = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$ we get the "mean squared error between the different groups"

$$\text{MSG} = \frac{1}{g-1} \sum_{h=1}^{g} n_h (\overline{Y}_{h.} - \overline{Y}_{..})^2.$$

This can be compared to the "mean squared error within the groups"

$$\text{MSE} = \frac{1}{n-g} \sum_{h,i} (Y_{hi} - \overline{Y}_{h.})^2,$$

leading to the test statistics of the **F-test**:

$$T = \frac{MSG}{MSE},$$

which follows an F-distribution with $g-1$ and $n-g$ degrees of freedom under H_0.

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
Treatment	4	520.69	130.173	1.508	0.208
Error	77	6645.15	86.301		
Total	81	7165.84			

Table 4.1.b: Example of an ANOVA table.

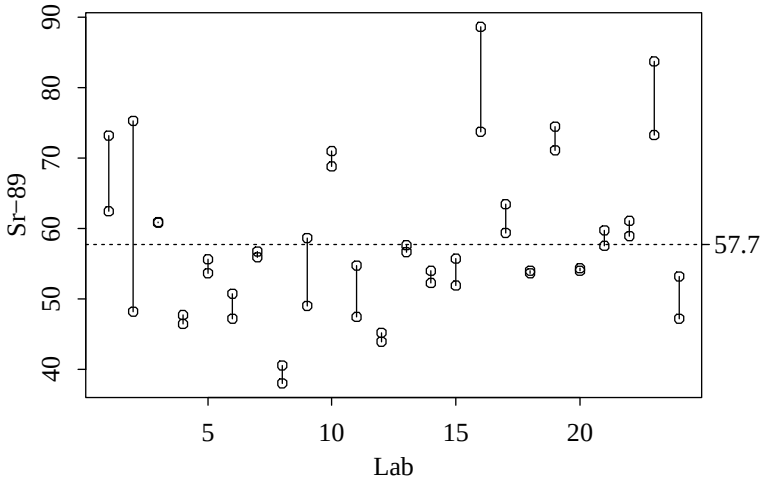

Figure 4.2.a: Sr-89-values for the 24 laboratories

c **Non-parametric tests** (if errors are not normally distributed):
"Kruskal-Wallis-Test", based on the ranks of the data.
For $g = 2$ groups: "Wilcoxon-Mann-Whitney-Test", also called "U-Test".

4.2 Random Effects, Ring Trials

Example **a** **Strontium in Milk** Figure 4.2.a illustrates the results of a ring trial (an inter-laboratory comparison) to determine the concentration of the radioactive isotope **Sr-89 in milk** (the question was of great interest after the Chernobyl accident). In 24 laboratories in Germany two runs to determine this quantity in artificially contaminated milk were performed. For this special situation the "true value" is known: it is 57.7 Bq/l. Source: G. Haase, D. Tait und A. Wiechen: "Ergebnisse der Ringanalyse zur Sr-89/Sr-90-Bestimmung in Milch im Jahr 1991". Kieler Milchwirtschaftliche Forschungsberichte 43, 1991, S. 53-62).

Figure 4.2.a shows that the two measurements of the same laboratory are in general much more similar than measurements between different laboratories.

b **Model**: $Y_{hi} = \mu + A_h + E_{hi}$. A_h **random**, $A_h \sim \mathcal{N}(0, \sigma_A^2)$.

Special quantities can tell us now how far two measurements can be from each other such that it is still safe to assume that the difference is only random.

- Comparisons **within** laboratory: **"Repeatability"** $2\sqrt{2} \cdot \widehat{\sigma}_E$,
- Comparisons **between** laboratories: **"Comparability"** $2\sqrt{2(\widehat{\sigma}_E^2 + \widehat{\sigma}_A^2)}$

4.3 Two and More Factors

Example a **Fisher's Potato Crop Data** Sir Ronald A. **Fisher** who established ANOVA (and many other things), used to work in the agricultural research center in Rothamstead, England. In an experiment to increase the yield of potatoes, the influence of **two treatment factors**, the addition of ammonium- and potassium-sulphate (each having 4 levels: 1, 2, 3, 4), was studied. Figure 4.3.a illustrates the data. Source: T. Eden and R. A. Fisher, Studies in Crop Variation. VI. Experiments on the Response of the Potato to Potash and Nitrogen, J. Agricultural Science, 19, 201-213, 1929; available through Bennett, 1971.

		Potassium-sulphate K			
		1	2	3	4
Ammonium-sulphate N	1	404	308	356	439
	2	318	434	402	422
	3	456	544	484	504
	4	468	534	500	562

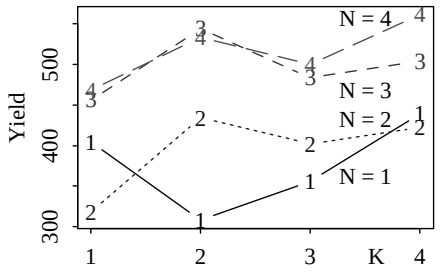

Figure 4.3.a: Fisher's Potato Crop Data.

b Model:

$$Y_{h,k} = \mu + \alpha_h + \beta_k + E_{hk} \,, \qquad \sum_h \alpha_h = 0 \quad \text{und} \quad \sum_k \beta_k = 0$$

c Estimates:

$$\widehat{\mu} = \overline{Y}_{..} \,, \qquad \widehat{\alpha}_h = \overline{Y}_{h.} - \overline{Y}_{..} \,, \qquad \widehat{\beta}_k = \overline{Y}_{.k} - \overline{Y}_{..} \,.$$

d **Tests**. Null-hypotheses: No influence of factor A (B). F-Tests. See table.

	DF	SS	MS	F	Pr(F)
N	3	59793	19931	10.84	0.0024
K	3	10579	3526	1.92	0.1973
Resid.	9	16552	1839		
Total	15	86924			

Table 4.3.d: ANOVA table for Fisher's potato crop data.

e **Interaction Effects** Model 4.3.b assumes that the effect of factor B is given by β_k, independent of the value of factor A. Or in other words, the model postulates that the effects of the two factors are additive. In general, so called **interaction effects** can occur. E.g., for fertilizers, further increasing one fertilizer is of little effect if another substance (fertilizer) is missing.

The general model for two factors with interaction effect can be written as

$$Y_{h,k} = \mu_{h,k} + E_{h,k} = \mu + \alpha_h + \beta_k + \gamma_{hk} + E_{h,k} \,.$$

Side constraints for the the interaction effect γ_{hk} are needed in order to obtain an identifiable model: $\sum_h \gamma_{hk} = 0$ for all k and $\sum_k \gamma_{hk} = 0$ for all h.

However, parameters can only be estimated if there are two or more observations for each combination of (h, k) (replicates).

f It's not difficult to extend model 4.3.b for **more than two factors**. The general model then also contains "interactions of higher order".

g For product development it's often necessary to check the effect of several (many) factors. In order to avoid too many experiments, it's often useful to restrict each factor to two levels and to avoid replicates. Such a series of experiments for k factors is called 2^k**-design** and will be discussed in more detail in the next section.

4.4 Response Surface Methods

Example a **Antibody Production** Large amounts of antibodies are obtained in biotechnological processes: Host animals (e.g. mice) are injected with modified cells that can produce the corresponding antibody. After a certain time these cells start to produce antibodies that can be collected in excreted fluid for further processing.

The cells can only produce antibodies if the immune system of the host animal is being weakened at the same time. This can be done with 4 factors. Moreover, it is believed that the amount of injected cells and their development stage has an influence on antibody production.

As there are no theoretical models for such complex biological processes, the relevant process factors have to be determined by an experiment. Such an experiment needs many mice, is time-intensive and usually costs a lot of money. Using a clever design, we can find out the important process factors with the lowest possible effort. That's where **statistical design of experiments** comes into play.

Two relevant process factors were identified in this study: the dose of Co^{60} gamma rays and the number of days between radiation and the injection of a pure oil. Now, the question is to find the levels for these two factors such that an **optimal amount** of antibodies is being produced by the modified cells.

b We have already seen a model which models a response variable Y that depends on two factors. It was

$$Y_{h,k} = \mu_{h,k} + E_{h,k} = \mu + \alpha_h + \beta_k + \gamma_{hk} + E_{h,k}\ , \qquad h,\, k = 1,\, 2.$$

If the two factors are based on continuous variables $x^{(1)}$, $x^{(2)}$, as is the case here with radiation dose and the number of days between radiation and injection, we have the corresponding general model

$$Y_i = h\left(x_i^{(1)}, x_i^{(2)}\right) + E_i,$$

(analogous for more than two factors). The function $h\left(x^{(1)}, x^{(2)}\right)$, which depends on $x^{(1)}$ and $x^{(2)}$, is the so-called **response surface**. Usually a quadratic polynomial (see below) in the variables $x^{(1)}$ and $x^{(2)}$ is used for h (sometimes the function h is available from theory). Once we have h, we can find the optimal setting $[x_0^{(1)}, x_0^{(2)}]$ of the process factors. Usually, h must be estimated from data.

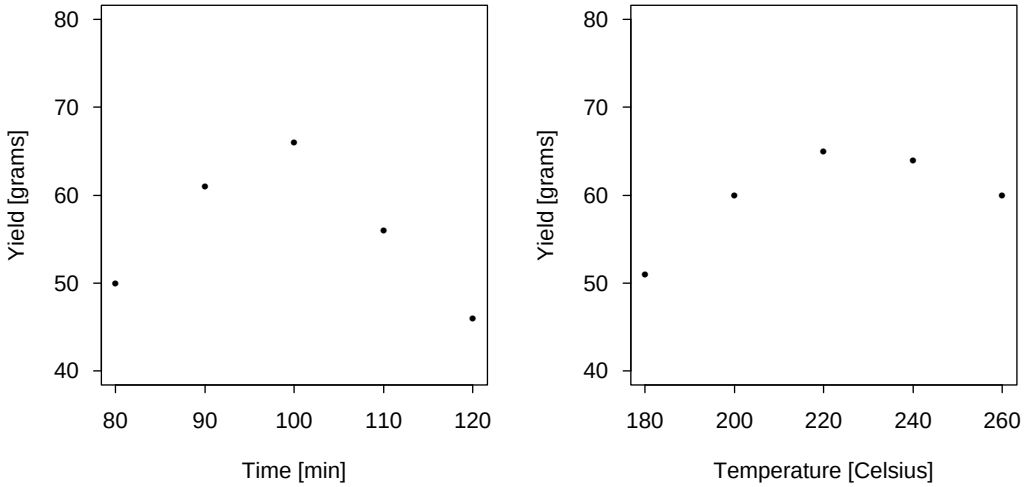

Figure 4.4.c: Varying the variables one by one. Left: Yield vs. reaction time, reaction temperature held constant at $220°C$. Right: Yield vs. reaction temperature, reaction time held constant at 100 minutes.

c A naive approach to find the optimum would be to optimize the variables **one by one**. The weakness of such an approach is now being illustrated with an artificial example.

Example d **Reaction Analysis** A chemist wants to maximize the yield of a chemical reaction by varying reaction time and reaction temperature. First, he performs an experiment where he uses a constant reaction temperature of $T = 220°C$ and reaction times 80, 90, 100, 110, and 120 minutes. Results are illustrated in Figure 4.4.d. According to this data, the maximum is attained with a reaction time of about 100 minutes.

In a second stage, reaction time is held constant at its optimal value of $t = 100$ minutes. Reaction temperature is varied at 180, 200, 220, 240, and $260°C$. Now, the conclusion is that maximal yield is attained with a reaction temperature of about $220°C$. This is not too far away from the value that was used in the first stage. Hence, the final conclusion is that the maximal yield of about 65 grams is attained using a reaction time of about 100 minutes and a reaction temperature of about $220°C$.

e To see that this conclusion is wrong, we have to make use of a two-dimensional view. Let us put time on the x- and temperature on the y-axis. Yield is illustrated by the corresponding contours (Figure 4.4.e). In this example, maximal yield of about 70 grams is attained with a reaction time of about 85 minutes and a reaction temperature of about $270°C$.

The approach of "varying the variables one by one" is misleading because it tacitly assumes that the maximal value of one variable is independent of the other ones. This assumption is usually not fulfilled.

f Even though the original setting of the process variables was "far away" from the optimal value, an appropriate sequence of well chosen experimental set-ups leads to the optimum. For that purpose, we start with a so called **first-order design**, a 2^k-design with additional measurements in the center. Experience from earlier experiments should guide us in selecting appropriate levels for the factors.

Example g **Reaction Analysis (cont'd)** From earlier experiments we know that a reaction temperature of $140°C$ and a reaction time of 60 minutes gives good results. Now we want to vary reaction time by 10 minutes and reaction temperature by $20°C$. The corre-

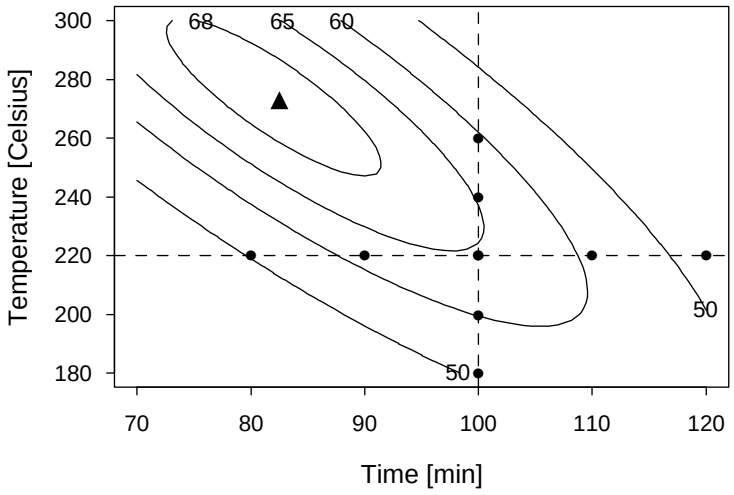

Figure 4.4.e: A hypothetical response surface, illustrated by contours in the diagram reaction temperature vs. reaction time.

sponding first-order design and the corresponding measurement results can be found in Table 4.4.g. Usually, coded variables are used in literature. They can also be found in Table 4.4.g.

	Variable in original units		Variable in coded units		Yield
Run	Temperature [$°C$]	Time [min]	Temperature	Time	Y [grams]
1	120	50	–1	–1	52
2	160	50	+1	–1	62
3	120	70	–1	+1	60
4	160	70	+1	+1	70
5	140	60	0	0	63
6	140	60	0	0	65

Table 4.4.g: First-order design and measurement results for the example "Reaction Analysis". The single experiments (runs) were performed in random order : 5, 4, 2, 6, 1, 7, 3.

h Because the response surface h (see 4.4.b) is unknown, we approximate it with the simplest possible surface, a plane. Hence, we have the model

$$Y_i = \theta_0 + \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + E_i,$$

which has to be fitted to the data. We have already seen how the parameter estimates can be obtained.

The fitted plane, the so called **first-order response surface**, is given by

$$\widehat{y} = \widehat{\theta}_0 + \widehat{\theta}_1 x^{(1)} + \widehat{\theta}_2 x^{(2)}.$$

Of course, this is only an approximation of the real response surface.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	62.000	0.882	70.30	6.3e-06 ***
xt1	5.000	1.080	4.63	0.019 *
xt2	4.000	1.080	3.70	0.034 *

Table 4.4.i: Estimated coefficients for the coded variables in the example "Reaction Analysis".

Example i **Reaction Analysis (cont'd)** The parameters with respect to the coded variables can be found in Table 4.4.i.

j On the first-order response surface we can find those points $[x^{(1)},\ x^{(2)}]^T$ which have a constant yield $\widehat{y} = \widehat{y}_0$. From equation

$$\widehat{y}_0 = \widehat{\theta}_0 + \widehat{\theta}_1 x^{(1)} + \widehat{\theta}_2 x^{(2)}$$

we find the straight line

$$x^{(2)} = \frac{\widehat{y}_0 - \widehat{\theta}_0 - \widehat{\theta}_1 x^{(1)}}{\widehat{\theta}_2}$$

with slope $b = -\widehat{\theta}_1/\widehat{\theta}_2$ and intercept $a = (\widehat{y}_0 - \widehat{\theta}_0)/\widehat{\theta}_2$. Orthogonal to this straight line is the **direction of steepest ascent (descent)**. This straight line has slope $\widehat{\theta}_2/\widehat{\theta}_1$. The two-dimensional vector $[\widehat{\theta}_1, \widehat{\theta}_2]^T$ is called estimated **gradient**; this direction is the fastest way to get large values of $\widehat{y}$.

Of course we also get large values when following any direction that is "close" to the gradient.

k Observations that are in the center of the 2^k-design have no influence on the estimates of the parameters $\theta_1, \ldots, \theta_k$ and hence no influence on the estimated gradient, either. This can be seen from the normal equations.

But why should we do experiments in the center?

- It's possible to estimate the measurement error without using the assumption that the plane is a good approximation of the true response surface if several observations are available in the center.
- Possible curvature of the true response surface can be detected. If there is no curvature and if the plane is a "good" approximation of the true response surface in the range of the experimental set-up, the average of the observations in the center, $\overline{Y}_c$, and the average of the observations of the 2^k-design, $\overline{Y}_f$, are estimates of the mean of Y for the set-up in the center. Hence, they should be "more or less equal". If the difference is obviously different from zero it's a hint that there is curvature.

A statistical test for curvature is as follows: The empirical variance s^2 that was estimated from the n_c observations in the center can be used to determine the standard deviation of the difference. The variance of $\overline{Y}_c$ can be estimated by s^2/n_c and the variance of $\overline{Y}_f$ by $s^2/2^k$. Because $\overline{Y}_c$ and $\overline{Y}_f$ are independent, the variance of the difference $\overline{Y}_c - \overline{Y}_f$ is estimated by $s^2(1/n_c + 1/2^k)$. Now we can perform a t-test or we can construct the corresponding confidence interval: If the interval

$$\overline{Y}_c - \overline{Y}_f \pm q_{0.975}^{t_{n_c-1}} \cdot \sqrt{s^2(1/n_c + 1/2^k)}$$

does not cover zero, the difference is statistically different from zero.

If we face relevant curvature of the true response surface (in the range of the experimental set-up), a linear approximation is not appropriate. Hence, we may also have problems determining the direction of steepest ascent. Usually we will use a second-order response surface (see below) for such situations.

l If the measurements of the center observations are all performed in a row, we face the danger that the observed variation (measured by s^2) is not really the variation between "independent" observations. Usually we will get significant curvature, even for cases where the response surface is a plane.

In general, it's **important to randomize the different experimental set-ups** – even though this usually needs much more effort because we always have to arrange a new setting for each new run.

m If it's plausible to use a linear response surface, we will search for an optimum along the direction of steepest ascent. Along

$$\begin{bmatrix} x_0^{(1)} \\ x_0^{(2)} \end{bmatrix} + k \begin{bmatrix} c^{(1)} \\ c^{(2)} \end{bmatrix}$$

we will perform additional experiments for $k = 1, 2, \ldots$ until yield starts to decrease. $[x_0^{(1)},\ x_0^{(2)}]^T$ is the point in the center of our experimental design and $[c^{(1)},\ c^{(2)}]^T$ is the direction of steepest ascent.

Example n **Reaction Analysis (cont'd)** The first-order response surface is

$$\widehat{y} = 62 + 5\widetilde{x}^{(1)} + 4\widetilde{x}^{(2)} = 3 + 0.25 \cdot x^{(1)} + 0.4 \cdot x^{(2)},$$

where $[\widetilde{x}^{(1)},\ \widetilde{x}^{(2)}]^T$ are the coded x-values (taking values ± 1) (see Table 4.4.i). Note that the gradient for the coded and the non-coded (original) x-values lead to different directions of steepest ascent. An other ascent direction can be identified by observing that individually increasing temperature by $4°C$ or time by 2.5 minutes both leads to a yield increase of 1 gram.

Further experiments are now performed along

$$\begin{bmatrix} 140 \\ 60 \end{bmatrix} + k \cdot \begin{bmatrix} 25 \\ 10 \end{bmatrix},$$

which corresponds to the steepest ascent direction with respect to the coded x-values, (see Table 4.4.n).

	Temperature [$°C$]	Time [min]	Y
1	165	70	72
2	190	80	77
3	215	90	79
4	240	100	76
5	265	110	70

Table 4.4.n: Experimental design and measurement results for experiments along the steepest ascent direction for the example "Reaction Analysis".

Based on the results in Table 4.4.n (plot the profile of yield vs. runs), the optimum should be in the neighborhood of a reaction temperature of $215°C$ and a reaction time of 90 minutes.

o Once the optimum along the gradient is identified, a further first-order design experiment can be performed (around the optimum) to get a new gradient. However, in general the hope is that we are already close to the optimal solution and we will continue as illustrated in the next section.

4.5 Second-Order Response Surfaces

a Once we are close to the optimal solution, the estimated plane will be nearly parallel to the $(x^{(1)}, x^{(2)})$-plane. Hence, $\widehat{\theta}_1$ and $\widehat{\theta}_2$ will be nearly zero. We expect that the optimal solution is a (flat) peak in the range of our experimental set-up and hence we expect the difference $\overline{Y}_c - \overline{Y}_f$ to be significantly different from zero. Such a peak can be modelled by a second-order polynomial:

$$Y_i = \theta_0 + \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + \theta_{11}(x_i^{(1)})^2 + \theta_{22}(x_i^{(2)})^2 + \theta_{12} x_i^{(1)}\, x_i^{(2)} + E_i$$

b However, the 2^k-design does not contain enough data to estimate the parameters of the second-order polynomial. The reason is that we need at least 3 levels for each factor. There are now several ways of expanding our original design. The more levels we have for each factor, the better we can estimate the curvature. So called **rotatable central composite designs** (also called second-order central composite designs) are very famous. As can be seen from the graphical representation in Figure 4.5.b we can get such a design by extending our original first-order design. In total we have 9 different experimental set-ups if we have two predictor variables. All points (except the center point) have the **same distance from the center** $(0, 0)$. Five levels are used for each factor. If we use replicates at $(0, 0)$ we get a more precise estimate of the quadratic part of the model.

Example **c** **Reaction Analysis (cont'd)** A rotatable central composite design was applied. The results can be found in Table 4.5.c.

The parameter estimates lead to the estimated **second-order response surface**:

$$\begin{aligned}\widehat{y} &= \widehat{\theta}_0 + \widehat{\theta}_1 x^{(1)} + \widehat{\theta}_2 x^{(2)} + \widehat{\theta}_{11}(x^{(1)})^2 + \widehat{\theta}_{22}(x^{(2)})^2 + \widehat{\theta}_{12} x^{(1)}\, x^{(2)} \\ &= -278 + 2.0 \cdot x^{(1)} + 3.2 \cdot x^{(2)} + 0.0060 \cdot (x^{(1)})^2 + 0.026 \cdot (x^{(2)})^2 + 0.006 \cdot x^{(1)}\, x^{(2)}.\end{aligned}$$

d Depending on the parameters, a second-order response surface can take different shapes. The most important ones are those that have a maximum (minimum) or a saddle (rather rare). A schematic contour plot of these two types is illustrated in Figure 4.5.d.

Surfaces with a maximum (minimum) don't need further explanations: Once we leave the optimum in any direction, yield Y is decreasing (increasing). For a saddle, it depends on the direction whether yield Y increases or decreases. Hence, the surface is like a horse saddle.

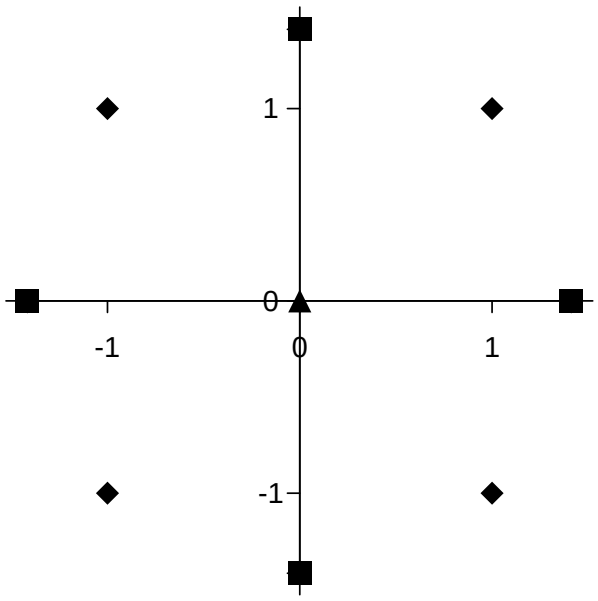

Figure 4.5.b: Rotatable central composite design. It consists of a 2^2-design (◆) with additional experiments in the center and along the axes (■).

Run	Variables in original units Temperature [$^\circ C$]	Time [min]	Variables in coded units Temperature	Time	Yield Y [grams]
1	195	80	-1	-1	78
2	235	80	$+1$	-1	76
3	195	100	-1	$+1$	72
4	235	100	$+1$	$+1$	75
5	187	90	$-\sqrt{2}$	0	74
6	243	90	$+\sqrt{2}$	0	76
7	215	76	0	$-\sqrt{2}$	77
8	215	104	0	$+\sqrt{2}$	72
9	215	90	0	0	80

Table 4.5.c: Rotatable central composite design and measurement results for the example "Reaction Analysis".

e It's possible to find an analytical solution for the critical point by calculating partial derivatives. In the critical point they have to be zero:

$$\begin{aligned} \frac{\partial \widehat{Y}}{\partial x^{(1)}} &= \widehat{\theta}_1 + 2\,\widehat{\theta}_{11}\,x_0^{(1)} + \widehat{\theta}_{12}\,x_0^{(2)} = 0 \\ \frac{\partial \widehat{Y}}{\partial x^{(2)}} &= \widehat{\theta}_2 + 2\,\widehat{\theta}_{22}\,x_0^{(2)} + \widehat{\theta}_{12}\,x_0^{(1)} = 0\,. \end{aligned}$$

Solving this linear equation system for $x_0^{(1)}$ and $x_0^{(2)}$ leads us to

$$x_0^{(1)} = \frac{\widehat{\theta}_{12}\,\widehat{\theta}_2 - 2\,\widehat{\theta}_{22}\,\widehat{\theta}_1}{4\,\widehat{\theta}_{11}\,\widehat{\theta}_{22} - \widehat{\theta}_{12}^2} \qquad x_0^{(2)} = \frac{\widehat{\theta}_{12}\,\widehat{\theta}_1 - 2\,\widehat{\theta}_{11}\,\widehat{\theta}_2}{4\,\widehat{\theta}_{11}\,\widehat{\theta}_{22} - \widehat{\theta}_{12}^2}\,.$$

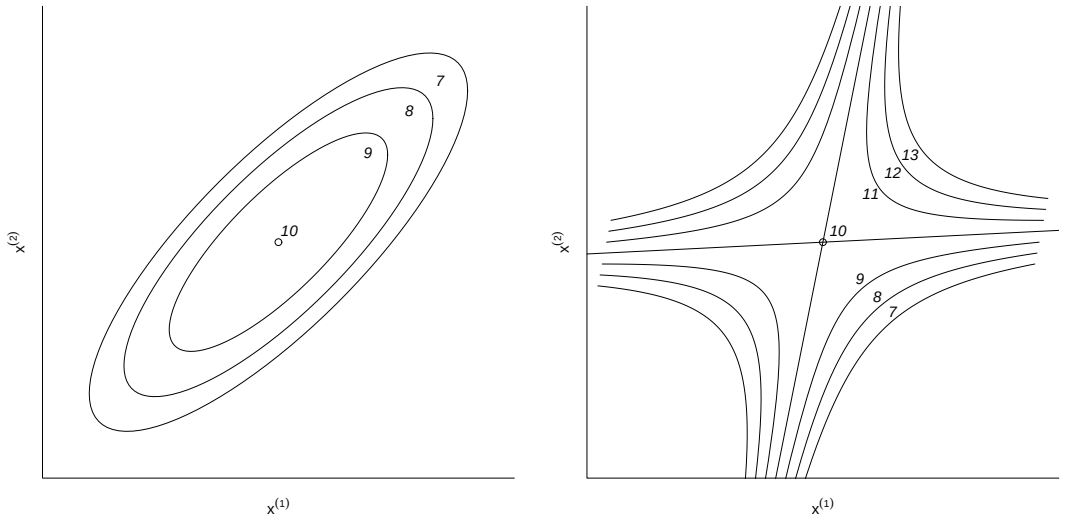

Figure 4.5.d: Contour plots of second-order response surfaces with a maximum (left) and a saddle (right).

Example f **Reaction Analysis (cont'd)** We can directly read the critical values off the contour plot in Figure 4.5.f: ($220°C$, 85 minutes).

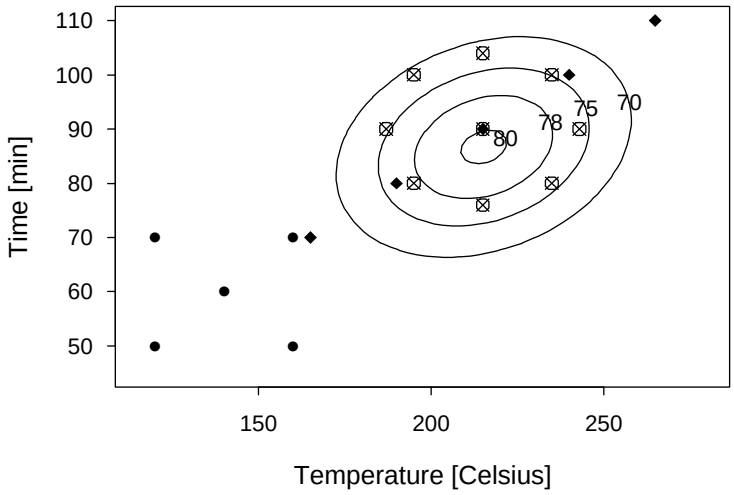

Figure 4.5.f: Estimated response surface and experimental designs for the example "Reaction Analysis". The first-order design is marked with ●, the experiments along the steepest ascent with ◆ and those of the rotatable central composite design with ⊠.

Example g **Antibody production (cont'd)** Let's come back to our original example. A radioactive dose of 200 rads and a time of 14 days is used as starting point. Around this point we use a first-order design, see Table 4.5.g.

Based on the measurements in the center we can calculate the standard deviation of the error term: $\widehat{\sigma} = 53.9$. Now we check whether there is significant curvature. The confidence interval for the difference can be calculated as outlined in 4.4.k:

$$\begin{aligned} \overline{Y}_c - \overline{Y}_f \pm q_{0.975}^{t_{n_c-1}} \cdot \sqrt{s^2(1/n_c + 1/2^k)} &= 589 - 335 \pm 4.30 \cdot 41.1 \\ &= [77,\ 431]. \end{aligned}$$

As this interval does not cover 0, the difference is statistically different from zero on the

Run	Variables in original units RadDos [rads]	Time [days]	Variables in coded units RadDos	Time	Yield Y
1	100	7	-1	-1	207
2	100	21	-1	$+1$	257
3	300	7	$+1$	-1	306
4	300	21	$+1$	$+1$	570
5	200	14	0	0	630
6	200	14	0	0	528
7	200	14	0	0	609

Table 4.5.g: First-order design and measurement results for the example "Antibody Production".

5% level. As yield decreases at the border of our experimental set-up, we conjecture that the optimal value must lie somewhere within our set-up range.

Hence, we expand our design to a rotatable central composite design by doing additional measurements (see Table 4.5.g).

Run	Variables in original units RadDos [rads]	Time [days]	Variables in coded units RadDos	Time	Yield Y
8	200	4	0	$-\sqrt{2}$	315
9	200	24	0	$+\sqrt{2}$	154
10	59	14	$-\sqrt{2}$	0	100
11	341	14	$+\sqrt{2}$	0	513

Table 4.5.g: Rotatable central composite design and measurement results for the example "Antibody Production".

The estimated response surface is

$$\begin{aligned}\hat{Y} &= -608.4 + 5.237 \cdot \text{RadDos} + 77.0 \cdot \text{Time} \\ &\quad -0.0127 \cdot \text{RadDos}^2 - 3.243 \cdot \text{Time}^2 + 0.0764 \cdot \text{RadDos} \cdot \text{Time}.\end{aligned}$$

We can identify the optimal conditions in the contour plot (Figure 4.5.g):
$\text{RadDos}_{opt} \approx 250$ rads and $\text{time}_{opt} \approx 15$ days.

h **Summary** To find the optimal setting of our variables (leading to maximal yield) we have to iteratively do experiments using special designs.

- If we are still "far away" from the optimum, we use first-order designs and we estimate the corresponding first-order response surface (a plane).
- On the estimated response surface we determine an ascent direction. Along that direction we do additional experiments until the response variable decreases again ("extrapolation").
- Further first-order experiments may be performed (hence we repeat the last two steps).
- As soon as we are close to the optimum, we perform (e.g.) a rotatable central composite design (or we expand our first-order design) to estimate the second-

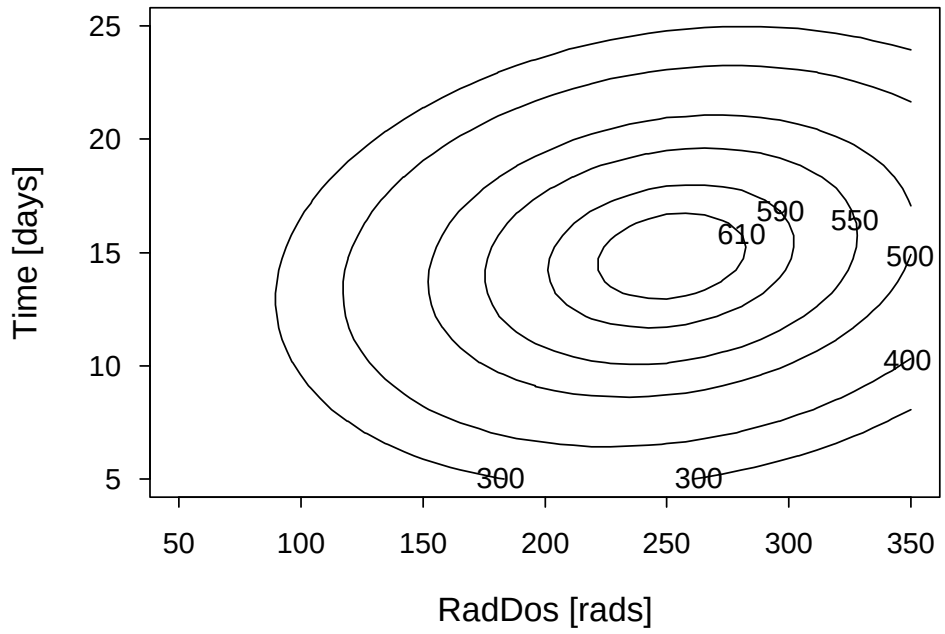

Figure 4.5.g: Estimated second-order response surface for the example "antibody production".

order response surface. The optimum on that response surface can be determined either analytically or graphically.

4.6 Experimental Designs, Robust Designs

a Here we discussed two types of experimental designs in more detail. Of course there are books full of other designs that are useful for various scopes. The subject is called **design of experiments**.

It may be worthwhile to note that for situations where little is known about the influence of the predictors on the response, so called "**screening designs**" can be very useful. They allow a (rough) analysis of k factors with less than 2^k experiments.

b The idea of "**robust product design**" is that products should have constant quality even if production conditions vary. To reach this goal we do not only have to optimize the expected quality (or yield or other response variables) but also the variability. There are special designs for that purpose, e.g. the **Taguchi designs**.

4.7 Further Reading

- **ANOVA and Design of Experiments** Short overviews of simple ANOVA models can be found in Hartung, Elpelt and Klösener (2002, Chap. XI) and Sachs (2004, Chap. 7). Linder and Berchtold (1982) give a more detailed introduction.
- Applied books about ANOVA and design of experiments are the famous book of Box, Hunter and Hunter (1978) and the book of Daniel (1976).
- A special book that uses unusual ways to introduce known (and unknown) methods with focus an explorative analysis is Hoaglin, Mosteller and Tukey (1991).
- A classical mathematically oriented book about ANOVA is Scheffe (1959).
- **Design of Experiments** Federer (1972, 1991) is an introduction to statistics

where design of experiments often takes center stage. More details can be found in the already mentioned book of Box et al. (1978) but also in Mead (1988).

- A systematic overview of experimental design can be found in Petersen (1985).
- A few books discuss topics about practical application of statistics that can't be dealt with mathematics. Recommendations are Boen and Zahn (1982) and Chatfield (1996).
- **Response Surfaces** An introduction to response surfaces is available in Box et al. (1978, Chapter 15) or in Hogg and Ledolter (1992).

  Box and Draper (1987) and Myers and Montgomery (1995) cover more details.

5 Multivariate Analysis of Spectra

5.1 Introduction

a "Spectrum" means here: We measure the "intensity" for certain "wave lengths". Such a function characterizes a chemical mixture (or as a special case a pure substance).

There are many spectra in chemistry. For some of them, pure substances have a spectrum that consists of a single "peak". As long as the peaks are not overlapping, we can identify the different components of a mixture and their proportions.

b **NIR-Spectra** (near infrared): The NIR-Spectra of pure substances is "any" function with some more or less characteristic peaks. Hence, it's rather difficult to identify the type and the quantity of the different components based on the spectrum of a chemical mixture. On the other side, these spectra are very cheap: No extra processing is needed, they can be measured on-line.

Example **c** **Quality Control via NIR-Spectra** We have data of reflections of NIR-waves on 52 granulate samples with wave length 1100, 1102, 1104, ..., 2500 nm. Figure 5.1.c shows the spectra in "centered" form; for each wave length j the median value $\text{med}_i(X_i^{(j)})$ was subtracted from the $X_i^{(j)}$'s.

Wl.	1800	1810	...	2500
a	0.003097	0.017238	...	−0.02950
b	0.002797	0.016994		−0.03095
c	0.002212	0.015757		−0.03095
	...			
Z	0.001165	0.014237	...	−0.03110

Table 5.1.c: Data for the example "NIR-spectra" (for wavelengths larger than 1800nm).

Questions

- There are outliers. Are there other "structures"?
- The amount of an active ingredient was determined with a chemical analysis. Can we estimate it sufficiently accurate with the spectrum?

d In other applications we measure spectra to follow a reaction on-line. It is used for

- estimating the order of a reaction and to determine potential intermediate products and reaction constants,
- determining the end of a process,
- monitoring a process.

We can also automatically monitor slow processes of all kinds. For example stock-keeping: Are there any (unwanted) aging effects?

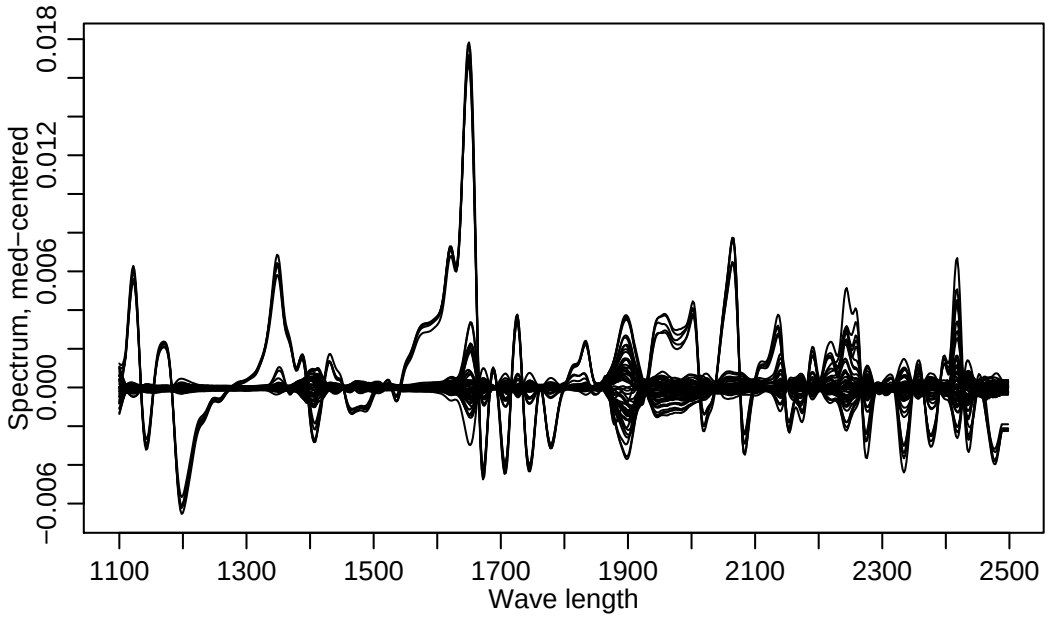

Figure 5.1.c: NIR-Spectra of granulate samples, centered at the median curve.

e For each observation (sample) we have many variables (a whole spectrum).

Questions

- Is it reasonable to plot the different samples on a plane or is it possible to catch most information and see structure from just a few dimensions (instead of using all variables)?
- Can we identify dimensions (with technical interpretation) in the high-dimensional space that contain most of the information?
- Is it possible to identify and to quantify the different components of a chemical mixture based on its spectrum?
- For a regression analysis, 70 variables (or 700 at a higher resolution) are too much if we only have 52 observations. How should we reduce dimensionality?

5.2 Multivariate Statistics: Basics

a **Notation** The vector $\underline{X}_i = [X_i^{(1)}, X_i^{(2)}, ..., X_i^{(m)}]^T$ denotes the ith spectrum. It's a point in m-dimensional space. Hence, for each observation we measure m different quantities.

Remark In statistics and probability theory vectors are usually column vectors. Row vectors are denoted by the symbol ^{T} (transposed vector).

This is inconvenient in statistics because the **data matrix**

$$\boldsymbol{X} = [X_i^{(j)}] \,,$$

that consists of n observations of m variables is built up the other way round: The ith row contains the values for the ith observation. For most applications this is a useful table (see e.g. the design matrix of a linear regression model). Here, it's often the other way round: In a table of spectra, a column often contains a single spectrum (i.e., it's one observation of a spectrum).

b **Definitions** We define the following quantities for an m-dimensional random vector $\underline{X} = [X^{(1)}, X^{(2)}, ..., X^{(m)}]^T \in \mathbb{R}^m$.

- **Expectation** $\underline{\mu} \in \mathbb{R}^m$
  $\underline{\mu} = (\mu_1, \ldots, \mu_m)^T$, where $\mu_k = \mathrm{E}[X^{(k)}]$, $k = 1, \ldots, m$.
  In other words: a vector that consists of the (univariate) expectations.
  We write $\underline{\mu}_X$ in situations where we also have other random variables.
- **Covariance Matrix** $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$
  $\boldsymbol{\Sigma}$ is an $m \times m$ matrix with elements

$$\boldsymbol{\Sigma}_{jk} = \mathrm{Cov}(X^{(j)}, X^{(k)}) = \mathrm{E}\left[(X^{(j)} - \mu_j)(X^{(k)} - \mu_k)\right].$$

  We also use the notation $\mathrm{Var}(\underline{X})$ or $\mathrm{Cov}(\underline{X})$.
  Note that
  - $\boldsymbol{\Sigma}_{jj} = \mathrm{Cov}(X^{(j)}, X^{(j)}) = \mathrm{Var}(X^{(j)})$.
    This means that the diagonal elements of the matrix are the variances.
  - $\mathrm{Corr}(X^{(j)}, X^{(k)}) = \dfrac{\boldsymbol{\Sigma}_{jk}}{\sqrt{\boldsymbol{\Sigma}_{jj}\boldsymbol{\Sigma}_{kk}}}$.

  Again, sometimes we write $\boldsymbol{\Sigma}_X$ if we want to point out that this is the covariance matrix that corresponds to $\underline{X}$.

c **Linear Transformations**

- For a simple (one-dimensional) random variable: $Y = a + bX$, where $a, b \in \mathbb{R}$.
  Expectation: $\mathrm{E}[Y] = a + b\mu_X$.
  Variance: $\mathrm{Var}(Y) = b^2\sigma_X^2$.
- For random vectors: $\underline{Y} = \underline{a} + \boldsymbol{B}\underline{X}$, where $\underline{a} \in \mathbb{R}^m$, $b \in \mathbb{R}^{m \times m}$.
  Expectation: $\mathrm{E}[\underline{Y}] = \underline{a} + \boldsymbol{B}\underline{\mu}_X$.
  Covariance: $\mathrm{Cov}(\underline{Y}) = \boldsymbol{B}\boldsymbol{\Sigma}_X\boldsymbol{B}^T$.

d **Remark** The multivariate normal distribution $\underline{X} \sim \mathcal{N}(\underline{\mu}, \boldsymbol{\Sigma})$ is fully characterized by the mean $\underline{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. It is the most common distribution in multivariate statistics. See e.g. Chapter 15.3 in Stahel (2000).

Figure 5.2.d illustrates two two-dimensional normal distributions with the "contours" of their densities. The mean vector is responsible for the location of the distribution and the covariance matrix for the shape of the contours.

e **Estimators**

$$\begin{aligned}
\widehat{\underline{\mu}} &= \left[\overline{X}^{(1)}, \overline{X}^{(2)}, \ldots, \overline{X}^{(m)}\right]^T = \text{ vector of means} \\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{n-1}\sum_{i=1}^{n}(\underline{X}_i - \widehat{\underline{\mu}})(\underline{X}_i - \widehat{\underline{\mu}})^T \\
&= \text{matrix of the empirical variances and covariances.}
\end{aligned}$$

This means that

$$\widehat{\boldsymbol{\Sigma}}_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^{(j)} - \overline{X}^{(j)})(X_i^{(k)} - \overline{X}^{(k)}).$$

The covariance matrix plays a crucial role in multivariate models that are based on the **normal distribution** or that want to model **linear relationships**.

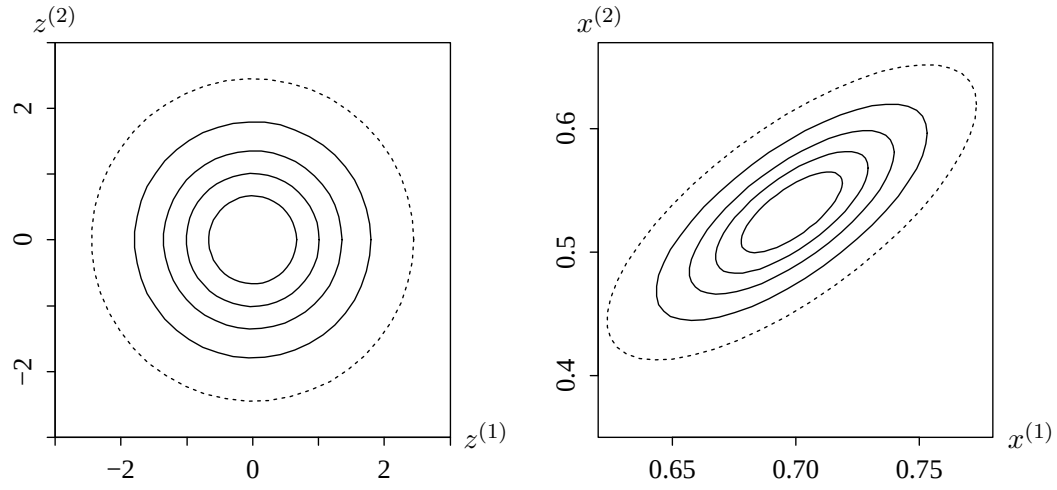

Figure 5.2.d: Contours of the probability densities for a standard normal (left) and a general (right) multivariate normal distribution.

5.3 Principal Component Analysis (PCA)

a Our goal is **dimensionality reduction**. We are looking for a few dimensions in the m-dimensional space that can explain "most of the variation in the data".

We define variation in the data as the sum of the individual m variances

$$\sum_{j=1}^{m} \widehat{\mathrm{Var}}(X^{(j)}) = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (\widetilde{X}_i^{(j)})^2,$$

where $\widetilde{X}_i^{(j)}$ are the centered observations: $\widetilde{X}_i^{(j)} = X_i^{(j)} - \overline{X}^{(j)}$.

We want to find a new "coordinate system" with certain properties. This will lead to

- new basis vectors $\underline{b}_k$ ($\|\underline{b}_k\| = 1$), the so called **principal components**. The individual components of these basis vectors are called **loadings**.
- new coordinates $Z_i^{(k)} = \underline{\widetilde{X}}_i^T \underline{b}_k$, the so called **scores** (projections of the data on the directions above).

What properties should the new coordinate system have?

- The first basis vector $\underline{b}_1$ should be chosen such that $\mathrm{Var}(Z^{(1)})$ is maximal.
- The second basis vector $\underline{b}_2$ should be orthogonal to the first one ($\underline{b}_2^T \underline{b}_1 = 0$) such that $\mathrm{Var}(Z^{(2)})$ is maximized.
- And so on...

Figure 5.3.a illustrates the idea using a two-dimensional distribution.

To summarize, we are performing a **transformation to new variables**

$$\underline{Z}_i = \widehat{\boldsymbol{B}}^T (\underline{X}_i - \widehat{\underline{\mu}}),$$

where the transformation matrix $\widehat{\boldsymbol{B}}$ is orthogonal.

It can be shown that $\widehat{\boldsymbol{B}}$ is the matrix of (standardized) eigenvectors and $\widehat{\lambda}_k$ are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_X$.

Remember that $\widehat{\boldsymbol{\Sigma}}_X$ is a symmetric matrix and therefore we can decompose it into

$$\widehat{\boldsymbol{\Sigma}}_X = \widehat{\boldsymbol{B}}\, \widehat{\boldsymbol{D}}\, \widehat{\boldsymbol{B}}^T,$$

where $\widehat{\boldsymbol{B}}$ is the matrix with the eigenvectors in the different columns and $\widehat{\boldsymbol{D}}$ is the diagonal matrix with the eigenvalues on the diagonal (this is a fact from linear algebra).

Therefore we have

$$\widehat{\mathrm{Var}}(\underline{Z}) = \widehat{\boldsymbol{B}}^T \widehat{\boldsymbol{\Sigma}}_X \widehat{\boldsymbol{B}} = \widehat{\boldsymbol{D}} = \begin{bmatrix} \widehat{\lambda}_1 & 0 & \dots & 0 \\ 0 & \widehat{\lambda}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \widehat{\lambda}_m \end{bmatrix}$$

$$\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_m \geq 0.$$

Hence, the individual components of $\underline{Z}$ are **uncorrelated** and the first component of $\underline{Z}$ has largest variance. By construction it holds that $\widehat{\lambda}_1 = \widehat{\mathrm{Var}}(Z^{(1)})$. It is the maximal variance of a projection:

$$\widehat{\lambda}_1 = \max_{\underline{b}:\|\underline{b}\|=1} (\widehat{\mathrm{Var}}(\boldsymbol{X}\underline{b})).$$

Accordingly for $\widehat{\lambda}_m$: It's the smallest variance.

Because the $\widehat{\lambda}_k$ are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_X$, we know from linear algebra that

$$\sum_{k=1}^{m} \widehat{\lambda}_k = \sum_{k=1}^{m} \widehat{\boldsymbol{\Sigma}}_{kk} = \sum_{k=1}^{m} \widehat{\mathrm{Var}}(X^{(j)}).$$

Hence

$$\frac{\sum_{j=1}^{k} \widehat{\lambda}_j}{\sum_{j=1}^{m} \widehat{\lambda}_j}$$

is the **proportion of the total variance** that is explained by the first k principal components.

Of course we can always go back to the original data using the new variables by doing a simple back-transformation

$$\underline{X}_i - \widehat{\underline{\mu}} = \left(\widehat{\boldsymbol{B}}^T\right)^{-1} \underline{Z}_i = \widehat{\boldsymbol{B}}\underline{Z}_i = \sum_{k=1}^{m} Z_i^{(k)} \underline{b}^{(k)}.$$

b **Graphical Representation** By reducing dimensionality it gets easier to visualize the data. For that reason we only consider the first two (or three) components and forget about the other ones. Figure 5.3.b (i) illustrates the first two components for the "NIR-spectra" example (for technical reasons we only consider wave lengths larger than 1800 nm). We can see 5 outliers – they were already visible in the spectra. Figure 5.3.b (ii) shows the first three components of a principal component analysis without the outliers.

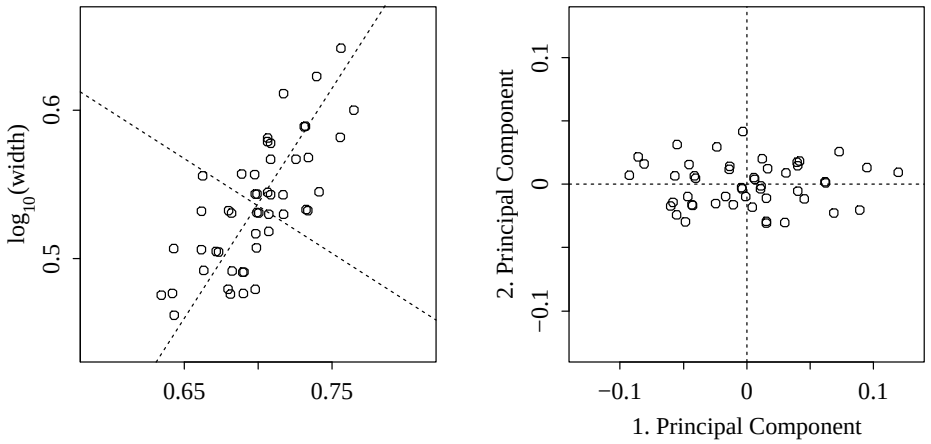

Figure 5.3.a: Principal component rotation.

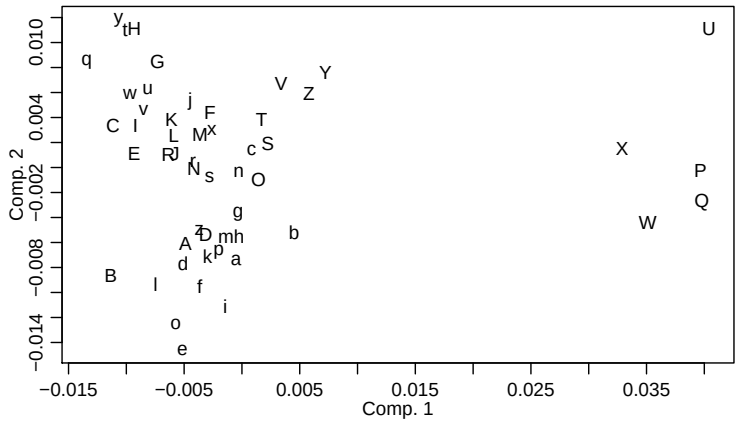

Figure 5.3.b: (i) Scatterplot of the first two principal components for the example "NIR-spectra".

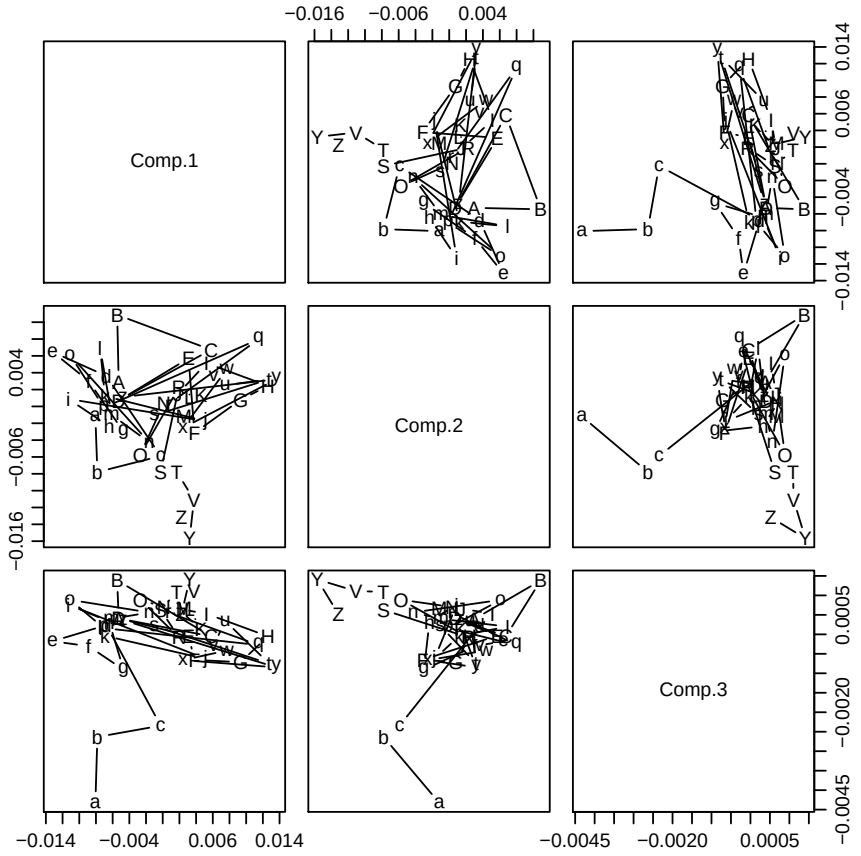

Figure 5.3.b: (ii) Scatterplot matrix of the first three principal components for the example "NIR-spectra" without the 5 outliers.

c PCA is suitable for many multivariate data sets. If we are analyzing spectra we have the special case that the variables (the intensities of different wavelengths) have a special ordering. Hence, we can plot each observation as a "function". We can also illustrate the principal component directions (the loadings) $\underline{b}_k$ as spectra!

d **Scaling Issues** If the variables are measured in different units, they should be standardized to (empirical) variance 1 (otherwise comparing variances doesn't make sense). This leads to a PCA (= eigenanalysis) of the correlation- instead of the covariance matrix.

For spectra this is *not* useful because wavelengths with very variable intensities contain the most important information. If we would standardize the variables in that setup, we would down-weight these variables compared to the unstandardized data set.

e **Choosing the number p of components**: $(p < m)$

- 2 (maybe 3) for illustrational purposes.
- Plot the explained variance (eigenvalues) in decreasing order and look for a breakpoint (**"scree plot"**: plot $\widehat{\lambda}_k$ vs. k), see Figure 5.3.e.
- "Explain 95% of the variance": The sum of the eigenvalues $\sum_{j=1}^{p} \widehat{\lambda}_j$ should be 95% of the total sum $\sum_{j=1}^{m} \widehat{\lambda}_j$.

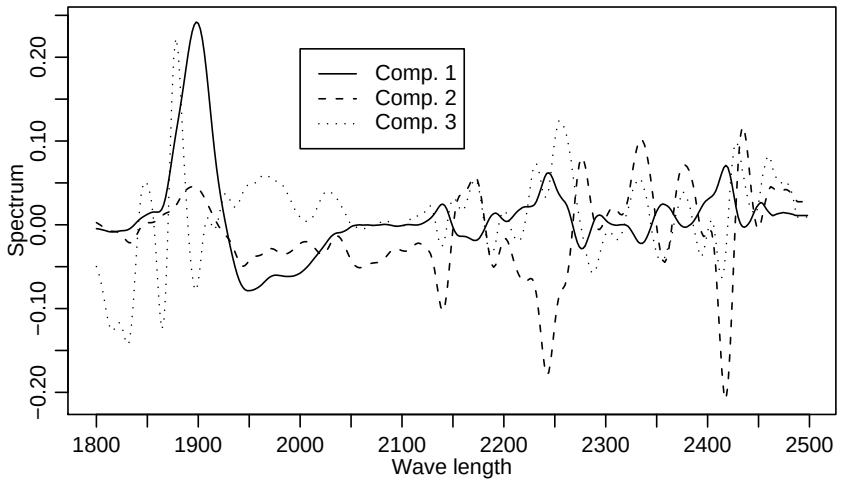

Figure 5.3.c: Spectra of "loadings" of the first three principal components for the example "NIR-spectra".

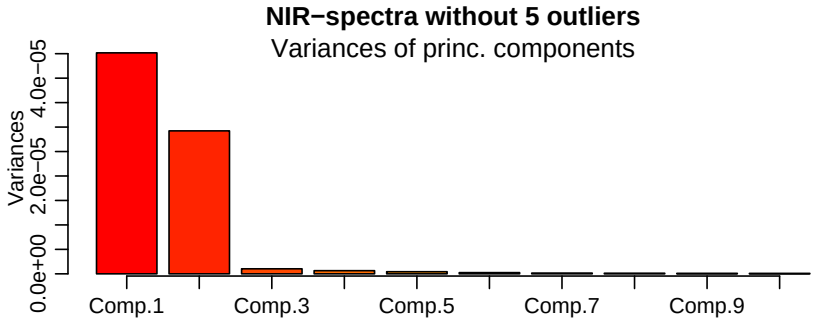

Figure 5.3.e: Variances of the principal components (scree plot) for the example "NIR-spectra".

But: "Variance" $\sum_{j=1}^{m} \lambda_j = \sum_{j=1}^{m} \mathrm{Var}(X^{(j)})$ is the sum of all variances. There could be (many) noise variables among them!

Restriction to the first p principal components: In the transformation formula (5.3.a) we simply ignore the last $m - p$ terms:

$$\underline{X}_i - \widehat{\underline{\mu}} = \widehat{\underline{X}}_i + \widehat{\underline{E}}_i \ , \quad \widehat{\underline{X}}_i = \sum_{k=1}^{p} Z_i^{(k)} \underline{b}^{(k)} \ , \quad \widehat{\underline{E}}_i = \sum_{k=p+1}^{m} Z_i^{(k)} \underline{b}^{(k)}.$$

This can be interpreted in the following two ways.

- In Linear Algebra terminology:
  The "data matrix" of the $\widehat{\underline{X}}_i$ is the best approximation of the data matrix of the $\underline{X}_i - \widehat{\underline{\mu}}$ if we restrict ourselves to matrices with rank p (in the sense of the so-called Frobenius norm of matrices: $\| \boldsymbol{E} \|^2 = \sum_{ij} E_{ij}^2$).
- In statistical terminology:
  We were looking for p variables $Z^{(k)} = \sum_{j=1}^{m} B_{kj} X^{(j)}$, $k = 1, \ldots p$, such that the differences $\underline{E}_i = \underline{X}_i - \widehat{\underline{X}}_i$ of $\widehat{\underline{X}}_i = \sum_{k=1}^{p} Z_i^{(k)} \underline{b}^{(k)}$ show minimial variance (in the sum): $\sum_{j=1}^{m} \widehat{\mathrm{Var}}(E^{(j)}) = \sum_{k=p+1}^{m} \lambda_k$ is minimal (there will be no better choice than the variables $Z^{(k)}$).

5.4 Linear Mixing Models, Factor Analysis

a **Model for Spectra** Let $\underline{c}_k$ be the spectrum of the chemical component k and consider a mixture of the components with coefficients $\underline{s} = [s^{(k)}]$. For the ith mixture we have the coefficients $\underline{s}_i$. According to Lambert-Beer the spectrum of the ith mixture is

$$\underline{X}_i = \sum_k \underline{c}^{(k)} \underline{s}_i^{(k)} + \underline{E}_i = \boldsymbol{C}\underline{s}_i + \underline{E}_i$$

where $\underline{E}_i$ are measurement errors. $\boldsymbol{C}$ is the matrix of spectra $\underline{c}_k$ (in the different columns).

This looks very similar to 5.3.e. The differences are

- $\boldsymbol{C}$ not orthogonal
- $\underline{X}_i$ instead of $\underline{X}_i - \underline{\widehat{\mu}}$, not centered
- $\underline{E}_i$ random vector (measurement error)
- $s_i^{(k)} \geq 0$ or $C_{jk} \geq 0$, $X_i^{(j)} \geq 0$ if we use the original spectra.

b This model can be used for many applications where there are m measurements that are linear superimpositions of $p < m$ components.

Examples are:

- Chemical elements in rocks that consist of several bed-rocks.
- Trace elements in spring water that ran through different soil layers.

c If the source profiles (spectra) $\underline{c}_k$ are known, the "contributions" $s_i^{(k)}$ can be estimated for each observation i separately using linear regression.

However, it's more interesting if both the source profiles and their contributions have to be estimated from data. This can be achieved using a combination of statistical methods, professional expertise and application specific properties.

5.5 Regression with Many Predictors

a In the introductory example about NIR-spectra we discussed the question whether we can "predict" the amount of an active ingredient based on a spectrum.

Hence, we have a response variable Y and several predictors $[x^{(1)}, ..., x^{(m)}]$. If we set up a linear regression model we face the problem that there are many more predictors than observations. Hence, it's not possible to fit a "full model" (it would lead to a perfect fit).

A possible remedy is to use **"stepwise"** regression: We start with just one predictor and add the most significant predictor in the next step (until some stopping criterion is met).

Example: Granulate Samples.
Y = yield. $n = 44$ (without "outliers"). Table 5.5.a shows a computer output. For comparison: Simple correlation between L2450 and yield: $r = -0.57$, $R^2 = 0.32$.

b Better known are the following methods to handle the problem of having too many predictors

1. **Principal Component-Regression**,
2. Ridge Regression
3. New methods like Lasso, Elastic Net, ...

```
                  Value  Std. Error  t value  Pr(>|t|)  Signif
(Intercept)    75.30372     0.07175  1049.52     0.000  ***
L2450        -395.43390    76.70623    -5.16     0.000  ***
L2010        -465.28939   142.44458    -3.27     0.002  **
L1990         585.20468   128.49676     4.55     0.000  ***
L2360         875.33702   160.04160     5.47     0.000  ***
L2400         532.91971   117.74430     4.53     0.000  ***
L2480        -301.44225    77.70208    -3.88     0.000  ***
L2130        -501.39852    88.17596    -5.69     0.000  ***

Residual standard error: 0.2268 on 36 degrees of freedom
Multiple R-Squared: 0.7212
```

Table 5.5.a: Computer output for a regression model after variable selection with stepwise forward.

c **Principal Component-Regression** PCA of the predictors leads to new variables $[Z^{(1)}, \ldots, Z^{(p)}]$. The principal components are usually selected without examining the relationship with the response Y.
Variant of Brown Brown (1993): Select them according to simple correlation with Y!

d **Ridge Regression** An easy way to ensure that the matrix $\boldsymbol{X}^T\boldsymbol{X}$ (that needs to be invertible for least squares) is non-singular is to add a diagonal matrix $\lambda\boldsymbol{I}$, leading to

$$\underline{\widehat{\beta}}_\lambda = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\underline{Y}.$$

Bibliography

Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression analysis and its applications*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Bennett, J. H. (ed.) (1971). *Collected Papers of R. A. Fischer; 1912-24*, Vol. I, The University of Adelaide.

Boen, J. R. and Zahn, D. A. (1982). *The Human Side of Statistical Consulting*, Wadsworth Inc. Belmont.

Bortz, J. (2005). *Statistik für Sozialwissenschaftler*, 6 edn, Springer, Berlin.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experimenters*, Wiley, N. Y.

Brown, P. J. (1993). *Measurement, Regression, and Calibration*, Clarendon Press, Oxford, UK.

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

Chatfield, C. (1996). *The Analysis of Time Series; An Introduction*, Texts in Statistical Science, 5 edn, Chapman and Hall, London, NY.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*, Wiley Series in Probability & Mathematical Statistics, Wiley, New York.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, 2 edn, Wiley, N. Y. 1st ed. 1971.

Federer, W. T. (1972, 1991). *Statistics and Society: Data Collection and Interpretation*, Statistics: Textbooks and Monographs, Vol.117, 2 edn, Marcel Dekker, N.Y.

Harman, H. H. (1960, 1976). *Modern Factor Analysis*, 3 edn, University of Chicago Press, Chicago.

Hartung, J., Elpelt, B. and Klösener, K. (2002). *Statistik. Lehr- und Handbuch der angewandten Statistik*, 13 edn, Oldenbourg, München.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (1991). *Fundamentals of Exploratory Analysis of Variance*, Wiley, N. Y.

Hogg, R. V. and Ledolter, J. (1992). *Applied Statistics for Engineers and Physical Scientists*, 2 edn, Maxwell Macmillan International Editions.

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.

Lawley, D. N. and Maxwell, A. E. (1963, 1967). *Factor Analysis as a Statistical Method*, Butterworths Mathematical Texts, Butterworths, London.

Linder, A. and Berchtold, W. (1982). *Statistische Methoden II: Varianzanalyse und Regressionsrechnung*, Birkhäuser, Basel.

Mead, R. (1988). *The design of experiments*, Cambridge University Press, Cambridge.

Myers, R. H. and Montgomery, D. C. (1995). *Response Surface Methodology; Process and Product Optimization Using Designed Experiments*, Wiley Series in Probability and Statistics, Wiley, NY.

Petersen, R. G. (1985). *Design and Analysis of Experiments*, Statistics Textbooks and Monographs, Marcel Dekker, N.Y.

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

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

Renner, R. M. (1993). The resolution of a compositional data set into mixtures of fixed source compositions, *Applied Statistics — Journal of the Royal Statistical Society C* **42**: 615–631.

Sachs, L. (2004). *Angewandte Statistik*, 11 edn, Springer, Berlin.

Scheffe, H. (1959). *The Analysis of Variance*, Wiley, N. Y.

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

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

Swinbourne, E. S. (1971). *Analysis of Kinetic Data*, Nelson, London.