Linear Regression 18

Linear regression is one of the most important methods of data analysis. It serves the determination of model parameters, model fitting, assessing the importance of influencing factors, and prediction, in all areas of human, natural and economic sciences. Computer scientists who work closely with people from these areas will definitely come across regression models.

The aim of this chapter is a first introduction into the subject. We deduce the coefficients of the regression models using the method of least squares to minimise the errors. We will only employ methods of descriptive data analysis. We do not touch upon the more advanced probabilistic approaches which are topics of statistics. For that, as well as for nonlinear regression, we refer to the specialised literature.

We start with simple (or univariate) linear regression—a model with a single input and a single output quantity—and explain the basic ideas of analysis of variance for model evaluation. Then we turn to multiple (or multivariate) linear regression with several input quantities. The chapter closes with a descriptive approach to determine the influence of the individual coefficients.

18.1 Simple Linear Regression

A first glance at the basic idea of linear regression was already given in Sect. 8.3. In extension to this, we will now allow more general models, in particular regression lines with nonzero intercept.

Consider pairs of data $(x_1, y_1), \dots, (x_n, y_n)$, obtained as observations or measurements. Geometrically they form a scatter plot in the plane. The values x_i and y_i may appear repeatedly in this list of data. In particular, for a given x_i there can be data points with different values $y_{i1}, \dots, y_{ip}$. The general task of *linear regression*

M. Oberguggenberger and A. Ostermann, *Analysis for Computer Scientists*,
Undergraduate Topics in Computer Science,
https://doi.org/10.1007/978-3-319-91155-7_18

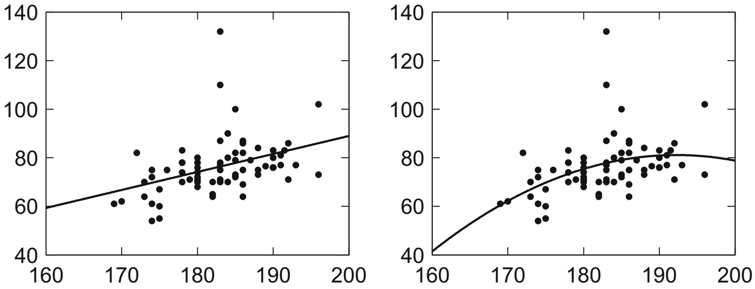

Fig. 18.1 Scatter plot height/weight, line of best fit, best parabola

is to fit the graph of a function

$$y = \beta_0\varphi_0(x) + \beta_1\varphi_1(x) + \cdots + \beta_m\varphi_m(x)$$

to the n data points $(x_1, y_1), \ldots, (x_n, y_n)$. Here the shape functions $\varphi_j(x)$ are given and the (unknown) coefficients β_j are to be determined such that the sum of squares of the errors is minimal (*method of least squares*):

$$\sum_{i=1}^{n} \big(y_i - \beta_0\varphi_0(x_i) - \beta_1\varphi_1(x_i) - \cdots - \beta_m\varphi_m(x_i)\big)^2 \to \min$$

The regression is called *linear* because the function y depends linearly on the unknown coefficients β_j. The choice of the shape functions ensues either from a possible theoretical model or empirically, where different possibilities are subjected to statistical tests. The choice is made, for example, according to the proportion of data variability which is explained by the regression—more about that in Sect. 18.4. The standard question of (simple or univariate) linear regression is to fit a *linear model*

$$y = \beta_0 + \beta_1 x$$

to the data, i.e., to find the *line of best fit* or *regression line* through the scatter plot.

Example 18.1 A sample of $n = 70$ computer science students at the University of Innsbruck in 2002 yielded the data depicted in Fig. 18.1. Here x denotes the height [cm] and y the weight [kg] of the students. The left picture in Fig. 18.1 shows the regression line $y = \beta_0 + \beta_1 x$, the right one a fitted quadratic parabola of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2.$$

Note the difference to Fig. 8.8 where the *line of best fit through the origin* was used; i.e., the intercept β_0 was set to zero in the linear model.

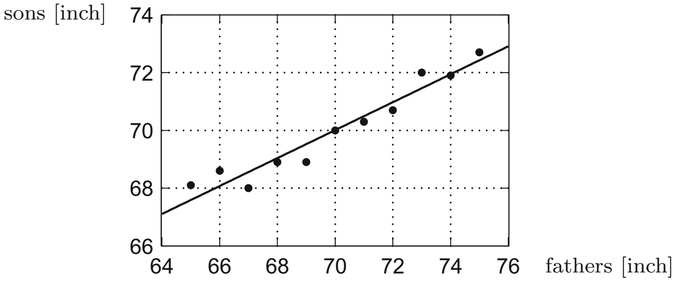

Fig. 18.2 Scatter plot height of fathers/height of the sons, regression line

A variant of the standard problem is obtained by considering the linear model

$$\eta = \beta_0 + \beta_1 \xi$$

for the transformed variables

$$\xi = \varphi(x), \quad \eta = \psi(y).$$

Formally this problem is identical to the standard problem of linear regression, however, with transformed data

$$(\xi_i, \eta_i) = \bigl(\varphi(x_i), \psi(y_i)\bigr).$$

A typical example is given by the *loglinear regression* with $\xi = \log x$, $\eta = \log y$

$$\log y = \beta_0 + \beta_1 \log x,$$

which in the original variables amounts to the *exponential model*

$$y = \mathrm{e}^{\beta_0} x^{\beta_1}.$$

If the variable x itself has several components which enter linearly in the model, then one speaks of *multiple linear regression*. We will deal with it in Sect. 18.3.

The notion of *regression* was introduced by Galton[1] who observed, while investigating the height of sons/fathers, a tendency of *regressing* to the average size. The data taken from [15] clearly show this effect, see Fig. 18.2. The method of least squares goes back to Gauss.

After these introductory remarks about the general concept of linear regression, we turn to *simple linear regression*. We start with setting up the model. The postulated relationship between x and y is linear

$$y = \beta_0 + \beta_1 x$$

[1] F. Galton, 1822–1911.

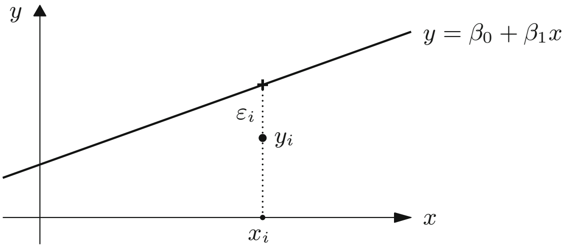

Fig. 18.3 Linear model and error ε_i

with unknown coefficients β_0 and β_1. In general, the given data will not exactly lie on a straight line but deviate by ε_i, i.e.,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

as represented in Fig. 18.3.

From the given data we want to obtain estimated values $\widehat{\beta_0}, \widehat{\beta_1}$ for β_0, β_1. This is achieved through minimising the sum of squares of the errors

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

so that $\widehat{\beta_0}, \widehat{\beta_1}$ solve the minimisation problem

$$L(\widehat{\beta_0}, \widehat{\beta_1}) = \min \Big(L(\beta_0, \beta_1) \; ; \; \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R} \Big).$$

We obtain $\widehat{\beta_0}$ and $\widehat{\beta_1}$ by setting the partial derivatives of L with respect to β_0 and β_1 to zero:

$$\begin{aligned} \frac{\partial L}{\partial \beta_0}\big(\widehat{\beta_0}, \widehat{\beta_1}\big) &= -2 \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i) = 0, \\ \frac{\partial L}{\partial \beta_1}\big(\widehat{\beta_0}, \widehat{\beta_1}\big) &= -2 \sum_{i=1}^{n} x_i (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i) = 0. \end{aligned}$$

This leads to a linear system of equations for $\widehat{\beta_0}, \widehat{\beta_1}$, the so-called *normal equations*

$$\begin{aligned} n\,\widehat{\beta_0} \;+ \big(\textstyle\sum x_i\big)\widehat{\beta_1} &= \textstyle\sum y_i, \\ \big(\textstyle\sum x_i\big)\widehat{\beta_0} \;+ \big(\textstyle\sum x_i^2\big)\widehat{\beta_1} &= \textstyle\sum x_i y_i. \end{aligned}$$

Proposition 18.2 *Assume that at least two x-values in the data set (x_i, y_i), $i = 1, \dots, n$ are different. Then the normal equations have a unique solution*

$$\widehat{\beta_0} = \big(\tfrac{1}{n} \textstyle\sum y_i\big) - \big(\tfrac{1}{n} \textstyle\sum x_i\big)\widehat{\beta_1}, \qquad \widehat{\beta_1} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

which minimises the sum of squares $L(\beta_0, \beta_1)$ of the errors.

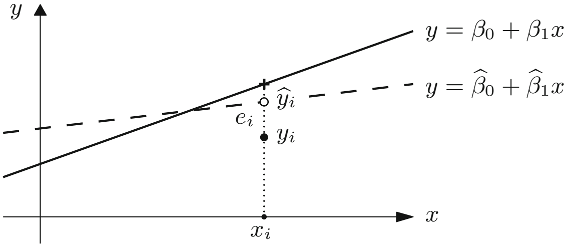
y
$y = \beta_0 + \beta_1 x$
$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$
$\widehat{y}_i$
e_i
y_i
x
x_i

Fig. 18.4 Linear model, prediction, residual

Proof With the notations $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{1} = (1, \dots, 1)$ the determinant of the normal equations is $n \sum x_i^2 - (\sum x_i)^2 = \|\mathbf{x}\|^2 \|\mathbf{1}\|^2 - \langle \mathbf{x}, \mathbf{1} \rangle^2$. For vectors of length $n = 2$ and $n = 3$ we know that $\langle \mathbf{x}, \mathbf{1} \rangle = \|\mathbf{x}\| \|\mathbf{1}\| \cdot \cos \measuredangle(\mathbf{x}, \mathbf{1})$, see Appendix A.4, and thus $\|\mathbf{x}\| \|\mathbf{1}\| \geq |\langle \mathbf{x}, \mathbf{1} \rangle|$. This relation, however, is valid in any dimension n (see for instance [2, Chap. VI, Theorem 1.1]), and equality can only occur if $\mathbf{x}$ is parallel to $\mathbf{1}$, so all components x_i are equal. As this possibility was excluded, the determinant of the normal equations is greater than zero and the solution formula is obtained by a simple calculation.

In order to show that this solution minimises $L(\beta_0, \beta_1)$, we compute the Hessian matrix

$$H_L = \begin{bmatrix} \frac{\partial^2 L}{\partial \beta_0^2} & \frac{\partial^2 L}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 L}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 L}{\partial \beta_1^2} \end{bmatrix} = 2 \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = 2 \begin{bmatrix} \|\mathbf{1}\|^2 & \langle \mathbf{x}, \mathbf{1} \rangle \\ \langle \mathbf{x}, \mathbf{1} \rangle & \|\mathbf{x}\|^2 \end{bmatrix}.$$

The entry $\partial^2 L / \partial \beta_0^2 = 2n$ and $\det H_L = 4\left(\|\mathbf{x}\|^2 \|\mathbf{1}\|^2 - \langle \mathbf{x}, \mathbf{1} \rangle^2\right)$ are both positive. According to Proposition 15.28, L has an isolated local minimum at the point $(\widehat{\beta}_0, \widehat{\beta}_1)$. Due to the uniqueness of the solution, this is the only minimum of L. □

The assumption that there are at least two different x_i-values in the data set is not a restriction since otherwise the regression problem is not meaningful. The result of the regression is the *predicted regression line*

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

The *values predicted by the model* are then

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \quad i = 1, \dots, n.$$

Their deviations from the data values y_i are called *residuals*

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i, \quad i = 1, \dots, n.$$

The meaning of these quantities can be seen in Fig. 18.4.

With the above specifications, the *deterministic regression model* is completed. In the *statistical regression model* the errors ε_i are interpreted as random variables with mean zero. Under further probabilistic assumptions the model is made accessible to statistical tests and diagnostic procedures. As mentioned in the introduction, we will not pursue this path here but remain in the framework of descriptive data analysis.

In order to obtain a more lucid representation, we will reformulate the normal equations. For this we introduce the following vectors and matrices:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

By this, the relations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

can be written simply as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Further

$$\mathbf{X}^{\mathsf{T}}\mathbf{X} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix},$$

$$\mathbf{X}^{\mathsf{T}}\mathbf{y} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{bmatrix} \begin{bmatrix} y_i \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix},$$

so that the normal equations take the form

$$\mathbf{X}^{\mathsf{T}}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^{\mathsf{T}}\mathbf{y}$$

with solution

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathsf{T}}\mathbf{y}.$$

The predicted values and residuals are

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}, \quad \mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}.$$

Example 18.3 (Continuation of Example 18.1) The data for x = height and y = weight can be found in the M-file `mat08_3.m`; the matrix $\mathbf{X}$ is generated in MATLAB by

```
X = [ones(size(x)), x];
```

the regression coefficients are obtained by

```
beta = inv(X'*X)*X'*y;
```

The command `beta = X\y` permits a more stable calculation in MATLAB. In our case the result is

$$\begin{aligned}\widehat{\beta}_0 &= -85.02,\\ \widehat{\beta}_1 &= 0.8787.\end{aligned}$$

This gives the regression line depicted in Fig. 18.1.

18.2 Rudiments of the Analysis of Variance

First indications for the quality of fit of the linear model can be obtained from the *analysis of variance* (ANOVA), which also forms the basis for more advanced statistical test procedures.

The arithmetic mean of the y-values $y_1, \dots, y_n$ is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

The deviation of the measured value y_i from the mean value $\bar{y}$ is $y_i - \bar{y}$. The *total sum of squares* or *total variability* of the data is

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

The total variability is split into two components in the following way:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

The validity of this relationship will be proven in Proposition 18.4 below. It is interpreted as follows: $\widehat{y}_i - \bar{y}$ is the deviation of the predicted value from the mean value, and

$$SS_R = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$$

the *regression sum of squares*. This is interpreted as the part of the data variability accounted for by the model. On the other hand $e_i = y_i - \widehat{y}_i$ are the residuals, and

$$SS_E = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

is the *error sum of squares* which is interpreted as the part of the variability that remains unexplained by the linear model. These notions are best explained by considering the two extremal cases.

(a) The data values y_i themselves already lie on a straight line. Then all $\widehat{y}_i = y_i$ and thus $S_{yy} = SS_R$, $SS_E = 0$, and the regression model describes the data record exactly.

(b) The data values are in no linear relation. Then the line of best fit is the horizontal line through the mean value (see Exercise 13 of Chap. 8), so $\widehat{y}_i = \bar{y}$ for all i and hence $S_{yy} = SS_E$, $SS_R = 0$. This means that the regression model does not offer any indication for a linear relation between the values.

The basis of these considerations is the validity of the following formula.

Proposition 18.4 (Partitioning of total variability) $S_{yy} = SS_R + SS_E$.

Proof In the following we use matrix and vector notation. In particular, we employ the formulas

$$\mathbf{a}^\mathsf{T}\mathbf{b} = \mathbf{b}^\mathsf{T}\mathbf{a} = \sum a_i b_i, \quad \mathbf{1}^\mathsf{T}\mathbf{a} = \mathbf{a}^\mathsf{T}\mathbf{1} = \sum a_i = n\bar{a}, \quad \mathbf{a}^\mathsf{T}\mathbf{a} = \sum a_i^2$$

for vectors $\mathbf{a}$, $\mathbf{b}$, and the matrix identity $(\mathbf{AB})^\mathsf{T} = \mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$. We have

$$\begin{aligned} S_{yy} &= (\mathbf{y} - \bar{y}\mathbf{1})^\mathsf{T}(\mathbf{y} - \bar{y}\mathbf{1}) = \mathbf{y}^\mathsf{T}\mathbf{y} - \bar{y}(\mathbf{1}^\mathsf{T}\mathbf{y}) - (\mathbf{y}^\mathsf{T}\mathbf{1})\bar{y} + n\bar{y}^2 \\ &= \mathbf{y}^\mathsf{T}\mathbf{y} - n\bar{y}^2 - n\bar{y}^2 + n\bar{y}^2 = \mathbf{y}^\mathsf{T}\mathbf{y} - n\bar{y}^2, \\ SS_E &= \mathbf{e}^\mathsf{T}\mathbf{e} = (\mathbf{y} - \widehat{\mathbf{y}})^\mathsf{T}(\mathbf{y} - \widehat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol\beta}) \\ &= \mathbf{y}^\mathsf{T}\mathbf{y} - \widehat{\boldsymbol\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} - \mathbf{y}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol\beta} + \widehat{\boldsymbol\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol\beta} = \mathbf{y}^\mathsf{T}\mathbf{y} - \widehat{\boldsymbol\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y}. \end{aligned}$$

For the last equality we have used the normal equations $\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol\beta} = \mathbf{X}^\mathsf{T}\mathbf{y}$ and the transposition formula $\widehat{\boldsymbol\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} = (\mathbf{y}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol\beta})^\mathsf{T} = \mathbf{y}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol\beta}$. The relation $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol\beta}$ implies in particular $\mathbf{X}^\mathsf{T}\widehat{\mathbf{y}} = \mathbf{X}^\mathsf{T}\mathbf{y}$. Since the first line of $\mathbf{X}^\mathsf{T}$ consists of ones only, it follows that $\mathbf{1}^\mathsf{T}\widehat{\mathbf{y}} = \mathbf{1}^\mathsf{T}\mathbf{y}$ and thus

$$\begin{aligned} SS_R &= (\widehat{\mathbf{y}} - \bar{y}\mathbf{1})^\mathsf{T}(\widehat{\mathbf{y}} - \bar{y}\mathbf{1}) = \widehat{\mathbf{y}}^\mathsf{T}\widehat{\mathbf{y}} - \bar{y}(\mathbf{1}^\mathsf{T}\widehat{\mathbf{y}}) - (\widehat{\mathbf{y}}^\mathsf{T}\mathbf{1})\bar{y} + n\bar{y}^2 \\ &= \widehat{\mathbf{y}}^\mathsf{T}\widehat{\mathbf{y}} - n\bar{y}^2 - n\bar{y}^2 + n\bar{y}^2 = \widehat{\boldsymbol\beta}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol\beta}) - n\bar{y}^2 = \widehat{\boldsymbol\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} - n\bar{y}^2. \end{aligned}$$

Summation of the obtained expressions for SS_E and SS_R results in the sought after formula. □

The partitioning of total variability

$$S_{yy} = SS_R + SS_E$$

and its above interpretation suggests using the quantity

$$R^2 = \frac{SS_R}{S_{yy}}$$

for the assessment of the goodness of fit. The quantity R^2 is called *coefficient of determination* and measures the fraction of variability explained by the regression. In the limiting case of an exact fit, where the regression line passes through all data points, we have $SS_E = 0$ and thus $R^2 = 1$. A small value of R^2 indicates that the linear model does not fit the data.

Remark 18.5 An essential point in the proof of Proposition 18.4 was the property of $\mathbf{X}^\mathsf{T}$ that its first line was composed of ones only. This is a consequence of the fact that β_0 was a model parameter. In the regression where a straight line through the origin is used (see Sect. 8.3) this is not the case. For a regression which does not have β_0 as a parameter the variance partition is not valid and the coefficient of determination is meaningless.

Example 18.6 We continue the investigation of the relation between height and weight from Example 18.1. Using the MATLAB program `mat18_1.m` and entering the data from `mat08_3.m` results in

$$S_{yy} = 9584.9, \quad SS_E = 8094.4, \quad SS_R = 1490.5$$

and

$$R^2 = 0.1555, \quad R = 0.3943.$$

The low value of R^2 is a clear indication that height and weight are not in a linear relation.

Example 18.7 In Sect. 9.1 the fractal dimension $d = d(A)$ of a bounded subset A of $\mathbb{R}^2$ was defined by the limit

$$d = d(A) = -\lim_{\varepsilon \to 0^+} \log N(A, \varepsilon)/\log \varepsilon,$$

where $N(A, \varepsilon)$ denoted the smallest number of squares of side length ε needed to cover A. For the experimental determination of the dimension of a fractal set A, one rasters the plane with different mesh sizes ε and determines the number $N = N(A, \varepsilon)$ of boxes that have a non-empty intersection with the fractal. As explained in Sect. 9.1, one uses the approximation

$$N(A, \varepsilon) \;\approx\; C \cdot \varepsilon^{-d}.$$

Applying logarithms results in

$$\log N(A, \varepsilon) \;\approx\; \log C + d \log \frac{1}{\varepsilon},$$

which is a linear model

$$y \approx \beta_0 + \beta_1 x$$

for the quantities $x = \log 1/\varepsilon$, $y = \log N(A, \varepsilon)$. The regression coefficient $\widehat{\beta_1}$ can be used as an estimate for the fractal dimension d.

In Exercise 1 of Sect. 9.6 this procedure was applied to the coastline of Great Britain. Assume that the following values were obtained:

$1/\varepsilon$	4	8	12	16	24	32
$N(A, \varepsilon)$	16	48	90	120	192	283

A linear regression through the logarithms $x = \log 1/\varepsilon$, $y = \log N(A, \varepsilon)$ yields the coefficients

$$\widehat{\beta_0} = 0.9849, \quad d \approx \widehat{\beta_1} = 1.3616$$

with the coefficient of determination

$$R^2 = 0.9930.$$

This is very good fit, which is also confirmed by Fig. 18.5. The given data thus indicate that the fractal dimension of the coastline of Great Britain is $d = 1.36$.

A word of caution is in order. Data analysis can only supply indications, but never a proof that a model is correct. Even if we choose among a number of wrong models the one with the largest R^2, this model will not become correct. A healthy amount of skepticism with respect to purely empirically inferred relations is advisable; models should always be critically questioned. Scientific progress arises from the interplay between the invention of models and their experimental validation through data.

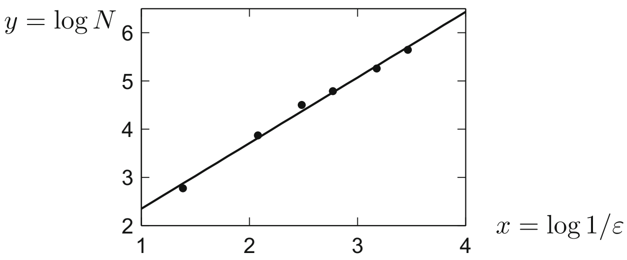

Fig. 18.5 Fractal dimension of the coastline of Great Britain

18.3 Multiple Linear Regression

In multiple (multivariate) linear regression the variable y does not just depend on one regressor variable x, but on several variables, for instance $x_1, x_2, \ldots, x_k$. We emphasise that the notation with respect to Sect. 18.1 is changed; there x_i denoted the ith data value, and now x_i refers to the ith regressor variable. The measurements of the ith regressor variable are now denoted with two indices, namely $x_{i1}, x_{i2}, \ldots, x_{in}$. In total, there are $k \times n$ data values. We again look for a linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

with the yet unknown coefficients $\beta_0, \beta_1, \ldots, \beta_k$.

Example 18.8 A vending machine company wants to analyse the delivery time, i.e., the time span y which a driver needs to refill a machine. The most important parameters are the number x_1 of refilled product units and the distance x_2 walked by the driver. The results of an observation of 25 services are given in the M-file `mat18_3.m`. The data values are taken from [19]. The observations $(x_{11}, x_{21}), (x_{12}, x_{22}), (x_{13}, x_{23}), \ldots, (x_{1,25}, x_{2,25})$ with the corresponding service times $y_1, y_2, y_3, \ldots, y_{25}$ yield a scatter plot in space to which a plane of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ should be fitted (Fig. 18.6; use the M-file `mat18_4.m` for visualisation).

Remark 18.9 A special case of the general multiple linear model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ is simple linear regression with several nonlinear form functions (as mentioned in Sect. 18.1), i.e.,

$$y = \beta_0 + \beta_1 \varphi_1(x) + \beta_2 \varphi_2(x) + \cdots + \beta_k \varphi_k(x),$$

where $x_1 = \varphi_1(x), x_2 = \varphi_2(x), \cdots, x_k = \varphi_k(x)$ are considered as regressor variables. In particular one can allow polynomial models

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

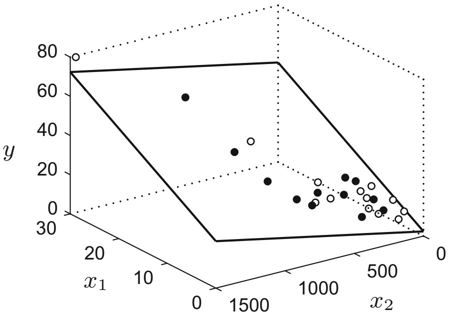

Fig. 18.6 Multiple linear regression through a scatter plot in space

or still more general interactions between several variables, for instance

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

All these cases are treated in the same way as the standard problem of multiple linear regression, after renaming the variables.

The data values for the individual regressor variables are schematically represented as follows:

Variable	y	x_1	x_2	$\dots$	x_k
Observation 1	y_1	x_{11}	x_{21}	$\dots$	x_{k1}
Observation 2	y_2	x_{12}	x_{22}	$\dots$	x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
Observation n	y_n	x_{1n}	x_{2n}	$\dots$	x_{kn}

Each value y_i is to be approximated by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, \dots, n$$

with the errors ε_i. The estimated coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_k$ are again obtained as the solution of the minimisation problem

$$L(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 \to \min$$

Using vector and matrix notation

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

the linear model can again be written for short as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The coefficients of best fit are obtained as in Sect. 18.1 by the formula

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

with the predicted values and the residuals

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}, \quad \mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}.$$

The partitioning of total variability

$$S_{yy} = SS_R + SS_E$$

is still valid; the *multiple coefficient of determination*

$$R^2 = SS_R/S_{yy}$$

is an indicator of the goodness of fit of the model.

Example 18.10 We continue the analysis of the delivery times from Example 18.8. Using the MATLAB program `mat18_2.m` and entering the data from `mat18_3.m` results in

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 2.3412 \\ 1.6159 \\ 0.0144 \end{bmatrix}.$$

We obtain the model

$$\widehat{y} = 2.3412 + 1.6159\,x_1 + 0.0144\,x_2$$

with the multiple coefficient of determination of

$$R^2 = 0.9596$$

and the partitioning of total variability

$$S_{yy} = 5784.5, \quad SS_R = 5550.8, \quad SS_E = 233.7$$

In this example merely $(1 - R^2) \cdot 100\% \approx 4\%$ of the variability of the data is not explained by the regression, a very satisfactory goodness of fit.

18.4 Model Fitting and Variable Selection

A recurring problem is to decide which variables should be included in the model. Would the inclusion of $x_3 = x_2^2$ and $x_4 = x_1x_2$, i.e., the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_1 x_2,$$

lead to better results, and can, e.g., the term $\beta_2 x_2$ be eliminated subsequently? It is not desirable to have too many variables in the model. If there are as many variables as data points, then one can fit the regression exactly through the data and the model would loose its predictive power. A criterion will definitely be to reach a value of R^2 which is as large as possible. Another aim is to eliminate variables that do not contribute essentially to the total variability. An algorithmic procedure for identifying these variables is the sequential partitioning of total variability.

Sequential partitioning of total variability. We include variables stepwise in the model, thus consider the increasing sequence of models with corresponding SS_R:

$$\begin{array}{rl} y = \beta_0 & SS_R(\beta_0), \\ y = \beta_0 + \beta_1 x_1 & SS_R(\beta_0, \beta_1), \\ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 & SS_R(\beta_0, \beta_1, \beta_2), \\ \vdots & \quad \vdots \\ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k & SS_R(\beta_0, \beta_1, \ldots, \beta_k) = SS_R. \end{array}$$

Note that $SS_R(\beta_0) = 0$, since in the initial model $\beta_0 = \bar{y}$. The additional explanatory power of the variable x_1 is measured by

$$SS_R(\beta_1|\beta_0) = SS_R(\beta_0, \beta_1) - 0,$$

the power of variable x_2 (if x_1 is already in the model) by

$$SS_R(\beta_2|\beta_0, \beta_1) = SS_R(\beta_0, \beta_1, \beta_2) - SS_R(\beta_0, \beta_1),$$

the power of variable x_k (if $x_1, x_2, \ldots, x_{k-1}$ are in the model) by

$$SS_R(\beta_k|\beta_0, \beta_1, \ldots, \beta_{k-1}) = SS_R(\beta_0, \beta_1, \ldots, \beta_k) - SS_R(\beta_0, \beta_1, \ldots, \beta_{k-1}).$$

Obviously,

$$\begin{aligned} SS_R(\beta_1|\beta_0) + SS_R(\beta_2|\beta_0, \beta_1) + SS_R(\beta_3|\beta_0, \beta_1, \beta_2) + \cdots \\ + SS_R(\beta_k|\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1}) = SS_R. \end{aligned}$$

This shows that one can interpret the *sequential, partial coefficient of determination*

$$\frac{SS_R(\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1})}{S_{yy}}$$

as explanatory power of the variables x_j, under the condition that the variables $x_1, x_2, \ldots, x_{j-1}$ are already included in the model. This partial coefficient of determination depends on the order of the added variables. This dependency can be eliminated by averaging over all possible sequences of variables.

Average explanatory power of individual coefficients. One first computes all possible sequential, partial coefficients of determination which can be obtained by adding the variable x_j to all possible combinations of the already included variables. Summing up these coefficients and dividing the result by the total number of possibilities, one obtains a measure for the contribution of the variable x_j to the explanatory power of the model.

Average over orderings was proposed by [16]; further details and advanced considerations can be found, for instance, in [8,10]. The concept does not use probabilistically motivated indicators. Instead it is based on the data and on combinatorics, thus belongs to descriptive data analysis. Such descriptive methods, in contrast to the commonly used statistical hypothesis testing, do not require additional assumptions which may be difficult to justify.

Example 18.11 We compute the explanatory power of the coefficients in the delivery time problem of Example 18.8. First we fit the two univariate models

$$y = \beta_0 + \beta_1 x_1, \quad y = \beta_0 + \beta_2 x_2$$

and from that obtain

$$SS_R(\beta_0, \beta_1) = 5382.4, \quad SS_R(\beta_0, \beta_2) = 4599.1,$$

with the regression coefficients $\widehat{\beta_0} = 3.3208$, $\widehat{\beta_1} = 2.1762$ in the first and $\widehat{\beta_0} = 4.9612$, $\widehat{\beta_2} = 0.0426$ in the second case. With the already computed values of the bivariate model

$$SS_R(\beta_0, \beta_1, \beta_2) = SS_R = 5550.8, \quad S_{yy} = 5784.5$$

from Example 18.10 we obtain the two sequences

$$\begin{aligned} SS_R(\beta_1|\beta_0) &= 5382.4 \approx 93.05\% \text{ of } S_{yy} \\ SS_R(\beta_2|\beta_0, \beta_1) &= 168.4 \approx 2.91\% \text{ of } S_{yy} \end{aligned}$$

and

$$\begin{aligned} SS_R(\beta_2|\beta_0) &= 4599.1 \approx 79.51\% \text{ of } S_{yy} \\ SS_R(\beta_1|\beta_0, \beta_2) &= 951.7 \approx 16.45\% \text{ of } S_{yy}. \end{aligned}$$

The average explanatory power of the variable x_1 (or of the coefficient β_1) is

$$\frac{1}{2}\Big(93.05 + 16.45\Big)\% = 54.75\%,$$

the one of the variable x_2 is

$$\frac{1}{2}\Big(2.91 + 79.51\Big)\% = 41.21\%;$$

the remaining 4.04% stay unexplained. The result is represented in Fig. 18.7.

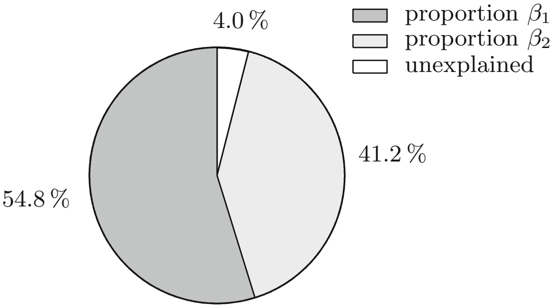

Fig. 18.7 Average explanatory powers of the individual variables

Numerical calculation of the average explanatory powers. In the case of more than two independent variables one has to take care that all possible sequences (represented by permutations of the variables) are considered. This will be exemplarily shown with three variables x_1, x_2, x_3. In the left column of the table there are the $3! = 6$ permutations of $\{1, 2, 3\}$, the other columns list the sequentially obtained values of SS_R.

1 2 3	$SS_R(\beta_1\|\beta_0)$	$SS_R(\beta_2\|\beta_0, \beta_1)$	$SS_R(\beta_3\|\beta_0, \beta_1, \beta_2)$
1 3 2	$SS_R(\beta_1\|\beta_0)$	$SS_R(\beta_3\|\beta_0, \beta_1)$	$SS_R(\beta_2\|\beta_0, \beta_1, \beta_3)$
2 1 3	$SS_R(\beta_2\|\beta_0)$	$SS_R(\beta_1\|\beta_0, \beta_2)$	$SS_R(\beta_3\|\beta_0, \beta_2, \beta_1)$
2 3 1	$SS_R(\beta_2\|\beta_0)$	$SS_R(\beta_3\|\beta_0, \beta_2)$	$SS_R(\beta_1\|\beta_0, \beta_2, \beta_3)$
3 1 2	$SS_R(\beta_3\|\beta_0)$	$SS_R(\beta_1\|\beta_0, \beta_3)$	$SS_R(\beta_2\|\beta_0, \beta_3, \beta_1)$
3 2 1	$SS_R(\beta_3\|\beta_0)$	$SS_R(\beta_2\|\beta_0, \beta_3)$	$SS_R(\beta_1\|\beta_0, \beta_3, \beta_2)$

Obviously the sum of each row is always equal to SS_R, so that the sum of all entries is equal to $6 \cdot SS_R$. Note that amongst the 18 SS_R-values there are actually only 12 different ones.

The average explanatory power of the variable x_1 is defined by M_1/S_{yy}, where

$$M_1 = \frac{1}{6}\Big(SS_R(\beta_1|\beta_0) + SS_R(\beta_1|\beta_0) + SS_R(\beta_1|\beta_0, \beta_2) + SS_R(\beta_1|\beta_0, \beta_3) + SS_R(\beta_1|\beta_0, \beta_2, \beta_3) + SS_R(\beta_1|\beta_0, \beta_3, \beta_2)\Big)$$

and analogously for the other variables. As remarked above, we have

$$M_1 + M_2 + M_3 = SS_R,$$

and thus the total partitioning adds up to one

$$\frac{M_1}{S_{yy}} + \frac{M_2}{S_{yy}} + \frac{M_3}{S_{yy}} + \frac{SS_E}{S_{yy}} = 1.$$

For a more detailed analysis of the underlying combinatorics, for the necessary modifications in the case of collinearity of the data (linear dependence of the columns of the matrix $\mathbf{X}$) and for a discussion of the significance of the average explanatory power, we refer to the literature quoted above. The algorithm is implemented in the applet *Linear regression*.

Experiment 18.12 Open the applet *Linear regression* and load data set number 9. It contains experimental data quantifying the influence of different aggregates on a mixture of concrete. The meaning of the output variables x_1 through x_4 and the input variables x_5 through x_{13} is explained in the online description of the applet. Experiment with different selections of the variables of the model. An interesting initial model is obtained, for example, by choosing $x_6, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}$ as independent and x_1 as dependent variable; then remove variables with low explanatory power and draw a pie chart.

18.5 Exercises

1. The total consumption of electric energy in Austria 1970–2015 is given in Table 18.1 (from [26, Table 22.13]). The task is to carry out a linear regression of the form $y = \beta_0 + \beta_1 x$ through the data.

(a) Write down the matrix $\mathbf{X}$ explicitly and compute the coefficients $\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_0, \widehat{\beta}_1]^\mathsf{T}$ using the MATLAB command `beta = X\y`.

(b) Check the goodness of fit by computing R^2. Plot a scatter diagram with the fitted straight line. Compute the forecast $\widehat{y}$ for 2020.

Table 18.1 Electric energy consumption in Austria, year = x_i, consumption = y_i [GWh]

x_i	1970	1980	1990	2000	2005	2010	2013	2014	2015
y_i	23.908	37.473	48.529	58.512	66.083	68.931	69.934	68.918	69.747

2. A sample of $n = 44$ civil engineering students at the University of Innsbruck in the year 1998 gave the values for x = height [cm] and y = weight [kg], listed in the M-file `mat18_ex2.m`. Compute the regression line $y = \beta_0 + \beta_1 x$, plot the scatter diagram and calculate the coefficient of determination R^2.

3. Solve Exercise 1 using `Excel`.

4. Solve Exercise 1 using the statistics package SPSS.
Hint. Enter the data in the worksheet *Data View*; the names of the variables and their properties can be defined in the worksheet *Variable View*. Go to *Analyze* → *Regression* → *Linear*.

5. The stock of buildings in Austria 1869–2011 is given in the M-file `mat18_ex5.m` (data from [26, Table 12.01]). Compute the regression line $y = \beta_0 + \beta_1 x$ and the regression parabola $y = \alpha_0 + \alpha_1 (x - 1860)^2$ through the data and test which model fits better, using the coefficient of determination R^2.

6. The monthly share index for four breweries from November 1999 to November 2000 is given in the M-file `mat18_ex6.m` (November 1999 = 100%, from the Austrian magazine profil 46/2000). Fit a univariate linear model $y = \beta_0 + \beta_1 x$

to each of the four data sets ($x \dots$ date, $y \dots$ share index), plot the results in four equally scaled windows, evaluate the results by computing R^2 and check whether the caption provided by profil is justified by the data. For the calculation you may use the MATLAB program `mat18_1.m`.
Hint. A solution is suggested in the M-file `mat18_exsol6.m`.

7. Continuation of Exercise 5, stock of buildings in Austria. Fit the model

$$y = \beta_0 + \beta_1 x + \beta_2 (x - 1860)^2$$

and compute $SS_R = SS_R(\beta_0, \beta_1, \beta_2)$ and S_{yy}. Further, analyse the increase of explanatory power through adding the respective missing variable in the models of Exercise 5, i.e., compute $SS_R(\beta_2|\beta_0, \beta_1)$ and $SS_R(\beta_1|\beta_0, \beta_2)$ as well as the average explanatory power of the individual coefficients. Compare with the result for data set number 5 in the applet *Linear regression*.

8. The M-file `mat18_ex8.m` contains the mileage per gallon y of 30 cars depending on the engine displacement x_1, the horsepower x_2, the overall length x_3 and the weight x_4 of the vehicle (from: Motor Trend 1975, according to [19]). Fit the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

and estimate the explanatory power of the individual coefficients through a simple sequential analysis

$$SS_R(\beta_1|\beta_0), \quad SS_R(\beta_2|\beta_0, \beta_1), \quad SS_R(\beta_3|\beta_0, \beta_1, \beta_2), \quad SS_R(\beta_4|\beta_0, \beta_1, \beta_2, \beta_3).$$

Compare your result with the average explanatory power of the coefficients for data set number 2 in the applet *Linear regression*.
Hint. A suggested solution is given in the M-file `mat18_exsol8.m`.

9. Check the results of Exercises 2 and 6 using the applet *Linear regression* (data sets 1 and 4); likewise for the Examples 18.1 and 18.8 with the data sets 8 and 3. In particular, investigate in data set 8 whether height, weight and the risk of breaking a leg are in any linear relation.

10. Continuation of Exercise 14 from Sect. 8.4. A more accurate linear approximation to the relation between shear strength τ and normal stress σ is delivered by Coulomb's model $\tau = c + k\sigma$ where $k = \tan\varphi$ and c [kPa] is interpreted as cohesion. Recompute the regression model of Exercise 14 in Sect. 8.4 with nonzero intercept. Check that the resulting cohesion is indeed small as compared to the applied stresses, and compare the resulting friction angles.

11. (Change point analysis) The consumer prize data from Example 8.21 suggest that there might be a change in the slope of the regression line around the year 2013, see also Fig. 8.9. Given data $(x_1, y_1), \dots, (x_n, y_n)$ with ordered data points $x_1 < x_2 < \dots < x_n$, phenomena of this type can be modelled by a piecewise linear regression

$$y = \begin{cases} \alpha_0 + \alpha_1 x, & x \le x_*, \\ \beta_0 + \beta_1 x, & x \ge x_*. \end{cases}$$

If the slopes α_1 and α_2 are different, x_* is called a *change point*. A change point can be detected by fitting models

$$y_i = \begin{cases} \alpha_0 + \alpha_1 x_i, & i = 1, \dots, m, \\ \beta_0 + \beta_1 x_i, & i = m+1, \dots, n \end{cases}$$

and varying the index m between 2 and $n-1$ until a two-line model with the smallest total residual sum of squares $SS_R(\alpha_0, \alpha_1) + SS_R(\beta_0, \beta_1)$ is found. The change point x_* is the point of intersection of the two predicted lines. (If the overall one-line model has the smallest SS_R, there is no change point.)
Find out whether there is a change point in the data of Example 8.21. If so, locate it and use the two-line model to predict the consumer price index for 2017.

12. Atmospheric CO_2 concentration has been recorded at Mauna Loa, Hawai, since 1958. The yearly averages (1959–2008) in ppm can be found in the MATLAB program `mat18_ex12.m`; the data are from [14].

(a) Fit an exponential model $y = \alpha_0 \,\mathrm{e}^{\alpha_1 x}$ to the data and compare the prediction with the actual data (2017: 406.53 ppm).
Hint. Taking logarithms leads to the linear model $z = \beta_0 + \beta_1 x$ with $z = \log y$, $\beta_0 = \log \alpha_0$, $\beta_1 = \alpha_1$. Estimate the coefficients $\widehat{\beta}_0$, $\widehat{\beta}_1$ and compute $\widehat{\alpha}_0$, $\widehat{\alpha}_1$ as well as the prediction for y.

(b) Fit a square exponential model $y = \alpha_0 \,\mathrm{e}^{\alpha_1 x + \alpha_2 x^2}$ to the data and check whether this yields a better fit and prediction.