Statistical Regression Models

Linear Regression

Werner Stahel
Seminar fÃijr Statistik, ETH ZÃijrich

January 2018

Course given at AIMS Rwanda, Jan 8 - 27, 2018

1 Introduction to statistical regression

1.1 Examples for linear regression

a In science, engineering and daily life a frequent type of questions asks how a certain variable of interest depends on other variables. Statistical regression treats this basic question and therefore is (apart from gaphical descriptions) the most widely applied statistical methodology. In this section we introduce typical types of questions by examples for "ordinary" linear regression before giving an overview of more general regression models in the next section.

b ▹ **Example nitrogen dioxide.** Nitrogen dioxyde (NO_2) is an air pollutant produced mainly by car engines. It is a irritant gas that harms the mucous membranes (surfaces of nose and lung). Its concentration is monitored by continuous measuring stations, together with other pollutants and the three weather characteristics temperature, global radiation, and precipitation. Data for Swiss stations can be found in
`www.bafu.admin.ch/bafu/de/home/themen/luft/zustand/daten/datenabfrage-nabel.html`
In order to plan remedies reducing the impact, it is important to know the impact of the weather conditions on the NO_2 concentration. Figure 1.1.b shows its relation to temperature with a symbol indicating rain. We restrict attention to hourly means, 3 to 4 p.m., of the year 2016, for the station in Zurich to obtain a suitable size of the dataset for this introduction and to avoid technical issues for modeling it later. (For the figure, we have further reduced the data zu Thursdays.) ◃

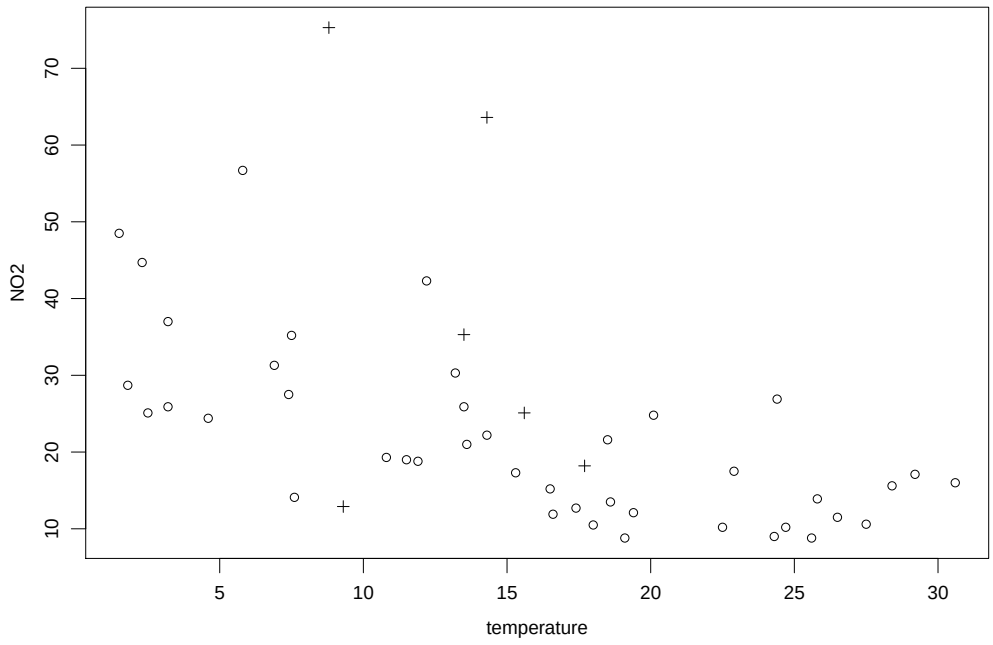

Figure 1.1.*b*: NO_2 *concentration and temperature. Means for* 3-4 *pm on Thursdays, station Zurich. Observations with precipitation* > 0 *are marked by* $+$.

c ▷ **Example blasting.** When building a road tunnel under a town, blasting is needed. The tremor shaking the buildings shall not exceed a given security threshold. Therefore, blasting near houses must be done with care, which causes extra costs. This asks for a rule that specifies the load of the blasting permitted in any given situation.

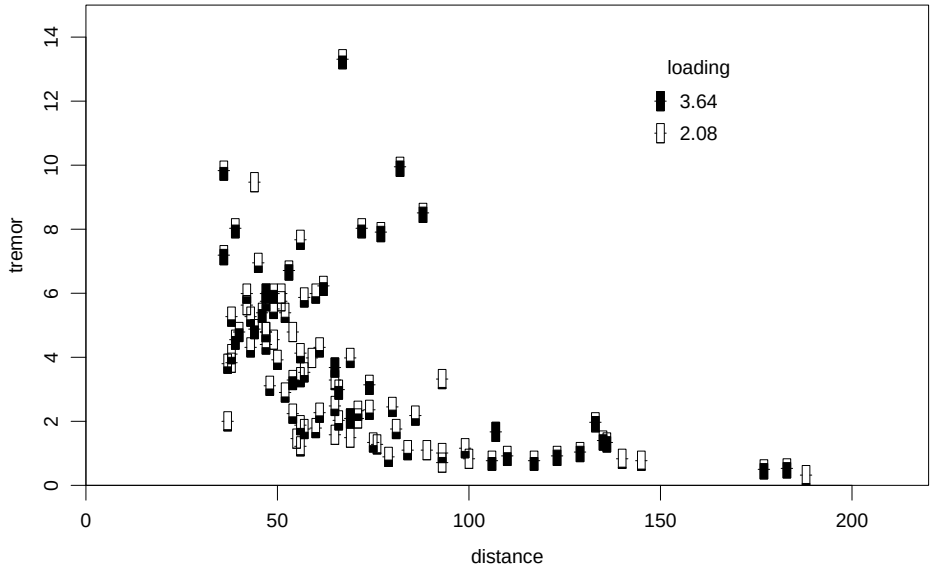

Figure 1.1.*c*: *Tremor as a function of distance for diferent loads*

Tremor depends on load, distance between the locations of blasting and measurement, type of soil material between these points, location of blasting within the tunnel's profile, and possibly other characteristics.

Figure 1.1.c shows how tremor depends on distance for different loads. (The dataset originates from digging a tunnel under the Swiss town Schaffhausen and is a coutesy of Basler and Hoffmann, engineers, Zurich.)

If tremor was a precise, known function of these properties and could one measure them exactly, it would be possible to calculate the load that leads to the tolerable tremor. In reality, one id confined to using a formula which holds only approximately. The deviations will be introduced as random variables, which leads to a probability model. ◁

d ▷ **Example birthrates.** How do birthrates depend on socio-ecomonic characteristics? A historic dataset has been collected, desribing such variables for the 182 district in Switzerland back in 1870, when this was a rather poor agricultural country. The characteristics include

- the variable of interest, an index of **fertility**, which is a more suitable index than raw birthrates, since it takes the percentages of women in their birthprone years into account;
- main language (there are four language regions);
- religious affiliation: Percent catholic – the complement being protestant back in these times;
- the percentage of locally born inhabitants, an index for rural regions;
- the percentages of employees in the main economic sectors;
- and some more.

Source: `https://opr.princeton.edu/archive/pefp/switz.aspx`
Figure 1.1.d shows a scatterplot of `fertility` against the proportion employed in the agricultural sector. ◁

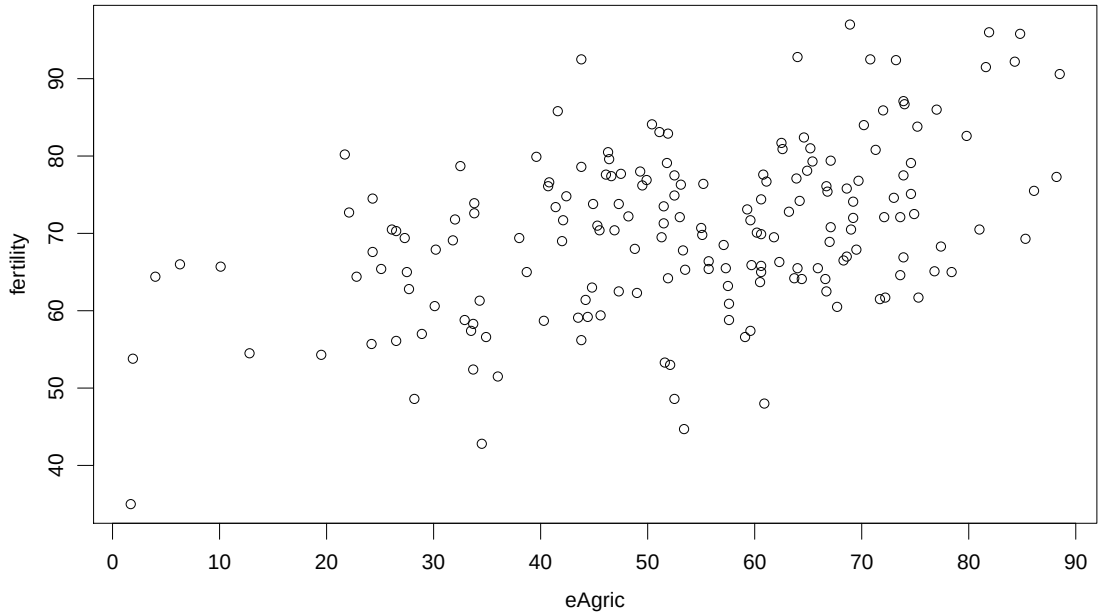

Figure 1.1.*d*: *Fertility and proportion working in agriculture in the example birthrates*

e Let us start by introducing mathematical symbols and technical terms!

The **target variable** y – the NO_2 concentration, tremor, or fertility – depends on the **input variables** or **explanatory variables** varianle!explanatory – temperature and precipitation; load, distance, situation, soil, ...; or socioeconomic characteristics of districts, respectively – through a function h

Remarks on the jargon. The expression "explanatory variables" is suitable if these are causal determinants of the target variable. A regression model can however serve to infer the values of a causal variable from the effect. Then, the roles are interchanged, and the cause is treated as the target variable in the model. Therefore, we prefer the name "input variables", which is more neutral.

There are the names **independent variables** for the input variables, and **dependent variable** for y. They are misguiding since there is no implication of any type of stochastic dependence. Very often, the independent variables are correlated!

f **Random deviations.** Ideally,[1]

$$y_i = h\left\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)} \right\rangle$$

should hold for each **observation** i (each hour, blasting, or district, respectively).

Unfortunately, no such formula exists. In practice, there are **deviations** for all (reasonable) functions h. Among other reasons, they stem from variables that have not been measured, but nevertheless have an effect on the target variable. and in addition, the type of soil cannot be described adequately.

▷ In the **nitrogen oxyde example,** these are further weather variables and certainly the varying

[1]The unusual brackets $\langle ... \rangle$ enclose arguments of a function. This enables a clear distinction between a function $f\langle a+b\rangle$ with argument $a+b$ and a product $f(a+b) = f \cdot (a+b)$.

traffic on nearby streets. The effect of **blasting** depends on the soil that cannot sufficiently be characterized. Finally, even if society could be desribed adequately by socio-economic variables, there are random effects that influence the actual number of births in the third example. ◁

g Statistical regression postulates that there is a formula that is "approximately" correct – up to deviations called "random". We write

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

and call E_i **random errors** or **random deviations**. Ideas about the size of these deviations are described by a **probability distribution**. Usually, the **normal distribution** is employed. – Function h is called the **regression function**.

▷ In the **nitrogen dioxyde example** one may be interested if for a given temperature, the limit for the pollutant is respected. It is not possible to say this with certainty: Even if the value of the regression function is below it, the limit may still be exceeded because of the random deviation E . Therefore, we can only give a probability for keeping the pollution below the threshold.

Analogously the goal in the **blasting example** is to keep the tremor below the threshold by adjusting the charge if necessary. Based on the model, it will be possible to derive a rule for choosing the acceptable charge in a given situation in spite of the uncertainty. However, it must be accepted that the tremor exceeds the given threshold with a certain probability, even if the charge has been chosen such that the regression function yields an undercritical value. If this probability should be small, the blasting must be applied with a cautiously chosen charge. The statistical regression gives a quantitative relation between the charge and the probability of an exceeding tremor for a given distance. ◁

These two examples will guide us through the following sections. You therefore need to wait patiently for the solutions of these problems.

2 Simple linear regression

2.1 The model

a ▷ **Example Nitrogen Oxyde** (1.1.b). Let us first consider the relation between the NO_2 concentration and temperature on days without precipitation. In the scatter diagram Figure 2.1.a, the vertical axis is shown with logarithmic scale. The log of the pollutant's concentration appears to depend linearly on the temperature. In other words, the points in the diagram scatter around a straight line. ◁

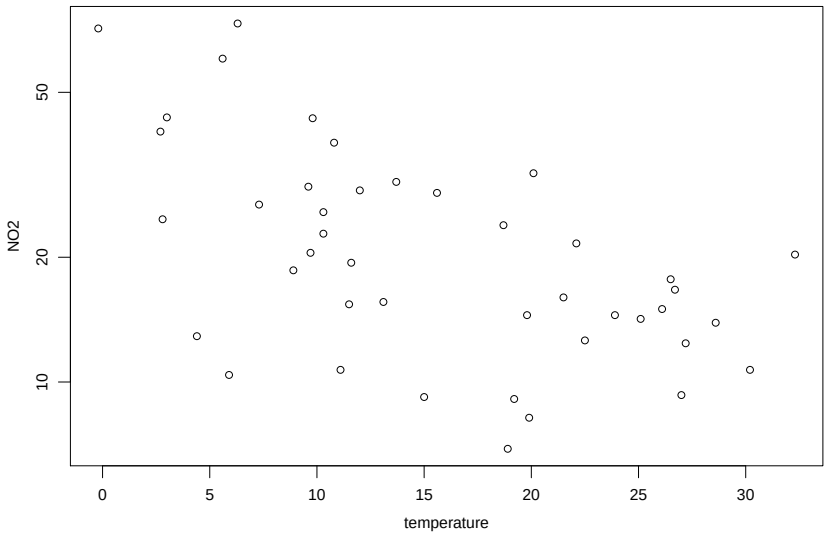

Figure 2.1.*a*: NO_2 *concentration against temperature for observations without precipitation. The vertical axis is shown on logarithmic scale.*

b ▷ **Example Blasting.** For readers with a more technical motivation, here is the Example Blasting (1.1.c). Let us first examine how tremor depends on distance for a given charge. In the scatterplot in Figure 2.1.b both axes use a logarithmic scale. It can be seen that the points scatter around a straight line, which means that the logarithmically transformed tremor depends approximately linearly on the log transformed distance. ◁

c A **straight line** is clearly the simplest function to express a dependence. All points $[x, y]$ on a straight line fulfill

$$y = \alpha + \beta x$$

with suitable numbers α and β. The first one, α, is the **intercept** and β measures the **slope**. Since β appears as a factor with the input variable, it is also called **(regression) coefficient** of X. It measures how the target variable changes when the input variable is augmented by one unit.

If $\alpha = 0$, the line passes through the origin of the coordinates.

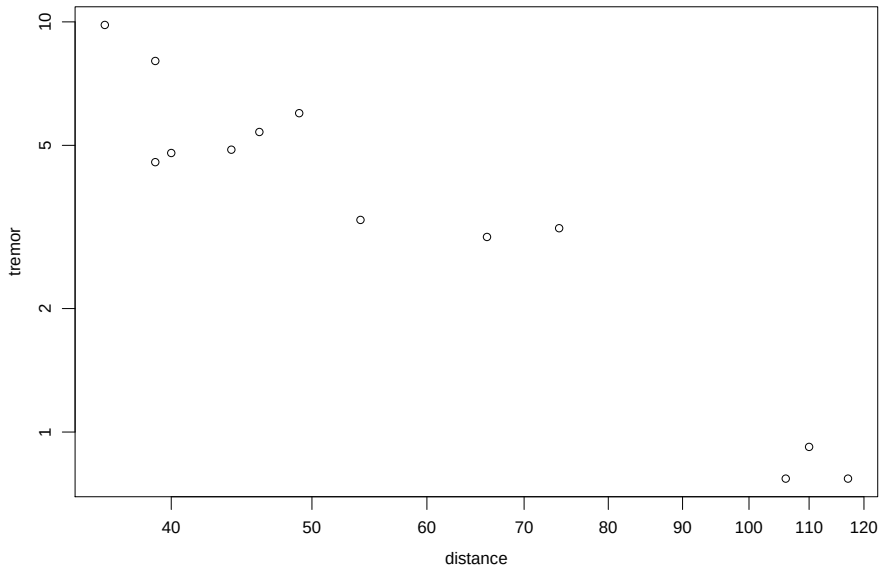

Figure 2.1.*b*: *Distance and tremor for blastings with a charge or* 3.12. *The axes are on logarithmic scale.*

d **Transformation.** In our examples, the *logarithmically transformed* data show the desired simple relation expressed by the straight line. The question arises if a **transformation** of original data is permitted or is an illegitimate **manipulation**. Here we adhere to the following statement:

Data do not ask for justice. It is our goal to detect relationships and structures and hopefully to understand them. To this end, we build models that combine deterministic, well interpretable relationships with random deviations. It will be important to examine how well the models correspond to the data. Whether the models are formulated in terms of original or transformed data is therefore no question of scientific integrity but at most one of simple **interpretability**..

In our examples, there will rarely be scientists who oppose the use of logarithmic axes. This corresponds to models and calculations based on logarithmically transformes data.

There often is a desire to describe how the target variable in the untransformed version depends on the input variable. For the log transformation, this is simple: If the input variable is increased by one unity, we expect $Y = \log_{10}\left\langle \widetilde{Y} \right\rangle$ to change by a value β. The original target $\widetilde{Y}$ then changes by a *factor* 10^{β}, and this amounts to a proportional change. The deviations E_i may also be back transformed. Then, they will be log-normally distributed instead of normally. This may be somewhat more difficult to apply, but it often fits the data better.

e **Which logarithm?** We use the logarithm to base 10 here. Why not taking the more common natural logarithm?

For the models, this does not make a relevant difference. The two versions only differ by a factor $\log\langle 10\rangle \approx 2.3$, since $\log_{10}\langle y\rangle = \log\langle y\rangle / \log\langle 10\rangle$, where log without subscript denotes the natural logarithm. Thus, the target variable is simply expressed with a

different "unit of measurement", and this leads to a corresponding multiplication of the coefficients α and β.

So, what is the advantage of log-10? The number in front of the decimal point shows directly the order of magnitude, and if one can memorize a few original values for the first decimal digit of the logarithm, the original number can be read off the log-10 value in quite good approximation

* It suffices to remember that $\log_{10}\langle 2\rangle \approx 0.3$. It then follow that $\log_{10}\langle 4\rangle = \log_{10}\langle 2^2\rangle \approx 0.6$, $\log_{10}\langle 8\rangle \approx 0.9$ and $\log_{10}\langle\sqrt{2}\rangle \approx 0.15$. Furthermore, $\log_{10}\langle 5 = 10/2\rangle = \log_{10}\langle 10\rangle - \log_{10}\langle 2\rangle \approx 0.7$, $\log_{10}\langle 2.5 = 10/4\rangle \approx 0.4$ and $\log_{10}\langle 1.25\rangle \approx 0.1$ etc. This leads to evaluating $\log_{10}\langle y\rangle = 2.1$ as $y = 100 \cdot 1.25 = 125$.

f **Theoretical arguments.** In many examples there are theories from the scientific context that lead to linear relations.

▷ In the **blasting example,** theory suggests that tremor is proportional to charge and inversely proportional to squared distance,

$$\begin{aligned} \text{tremor} &\approx \text{const} \cdot \text{charge}/(\text{distance})^2 \qquad \text{or} \\ \log_{10}(\text{tremor}) &\approx \log_{10}(\text{const}) + \log_{10}(\text{charge}) - 2 \cdot \log_{10}(\text{distance}) \,. \end{aligned}$$

Thus, the log transformed data follow a linear relationship. Since charge is fixed in the present considerations, the points [lg(distance), lg(tremor)] should ideally lie on a straight line. – According to this physical model, the slope of the straight line is known – a rare case! We shall assume an approximately linear relationship between the log transformed data, but refrain from fixing the slope of the straight line. ◁

g **Why do we need a statistical model?** The natural next step is to draw a straight line fitting the points as well as possible. This is a task for summarizing description, i.e., for descriptive statistics. The best known rule to determine the "best fitting" line is called "Least Squares". It is introduced below. The result is shown in Figure 2.2.a.

If data is taken as "the truth", then this may be declared "the correct" line. However, it is clear that the data could have resulted somewhat differently, due to random variations. But different data would have lead to a (slightly) different fitting straight line. Hence, the line is itself random, unprecise. How should we describe the randomness or imprecision?

The answer is given by inferential statistics, which is based on probability theory. We therefore need a model that describes what data could have occured as well as the dataset shown in Figure 2.1.b. To this end, we temporarily forget the given data and construct a probability model that characterizes the given situation.

h **Model.** Let us first ask what the value Y_i of the target variable will be if the input variable has a given value x_i – in the **nitrogen oxyde example,** what will the value of the logarithmized concentration be for a temperature of $15\,^0$ C, or in the **blasting example,** how large will the logarithmized tremor be for a logarithmized distance of $x_i = \log_{10}\langle 50\rangle$? According to the preceding considerations, this is the value of the linear function, $\alpha + \beta x_i$, up to a deviation E_i which we now regard as a random variable,

$$Y_i = \alpha + \beta x_i + E_i \ .$$

We assume that the deviations E_i , $i = 1, \ldots, n$, follow a certain distribution – the same one for all i – and that they are stochastically independent (and therefore uncorrelated) Thus, they form a random sample. It turns out the assuming a *normal* distribution leads to mathematically simple results. The normal distribution shall have expectation 0 and variance σ^2 . We denote this as $E_i \sim \mathcal{N}\langle 0, \sigma^2\rangle$.

i **Parameters.** The model is well-defined only after specifying the numbers α , β and σ . This is the usual situation on probability: In a first step, the model is fixed only up to one or a few constants. These are called **parameters** of the distribution. Thus, the "normal distribution" is really not a distribution but a "distribution family". Only after fixing the expected value and the variance (or standard deviation), this becomes *one* distribution.

In many areas of application, the word "parameter" is used for an observed property or characteristic – and this is called "variable" in Statisics. We rely on your comprehension for this confusing terminology.

j **Graphical representation.** A model emerges in our brains. Let us go further to fix even the parameters. Figure 2.1.j displays the model of simple linear regression with the parameters $\alpha = 4$, $\beta = -2$ and $\sigma = 0.1$. The probabilities for the possible Y values are shown by probability density curves.

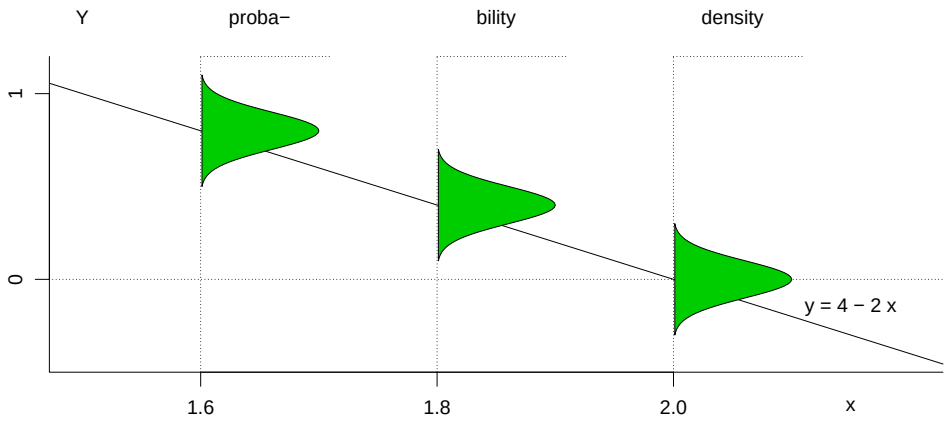

Figure 2.1.*j*: *Visualization of the regression model* $Y_i = 4 - 2x_i + E_i$ *for three observations* Y_1 , Y_2 *and* Y_3 *corresponding to the* x *values* $x_1 = 1.6$, $x_2 = 1.8$ *and* $x_3 = 2$

k **Simulation.** For a second kind of visualization we draw **random numbers** according to the model for the random deviations E_i. This then yields values Y_i that we can vizualise in combination with the given x_i values. They are called **simulated** values. Three standard normal random numbers, which are multiplied by $\sigma = 0.1$ form a possible result for the three deviations E_1, E_2 and E_3. A random number generator produced the four tripels

$$\begin{aligned} -0.419,\ -1.536,\ -0.671\ ; \qquad & 0.253,\ -0.587,\ -0.065\ ; \\ 1.287,\ 1.623,\ -1.442\ ; \qquad & -0.417,\ 1.427,\ 0.897\ . \end{aligned}$$

Multiplying by 0.1 and adding them to $4-2x_i$ with $x_1 = 1.6$, $x_2 = 1.8$ and $x_3 = 2$ yields the values Y_1, Y_2 and Y_3. Figure 2.1.k shows these simulated results.

Figure 2.1.*k*: *Four simulated sets of three measurements according to the model* $Y_i = 4 - 2x_i + E_i$. *The dashed lines mark the "true" relationship* $y = 4 - 2x$, *which is know in this situation.*

2.2 Parameter estimation

a ▷ Let us return to concrete data. Figure 2.2.a displays the **blasting example** data with a straight line fitting them. ◁

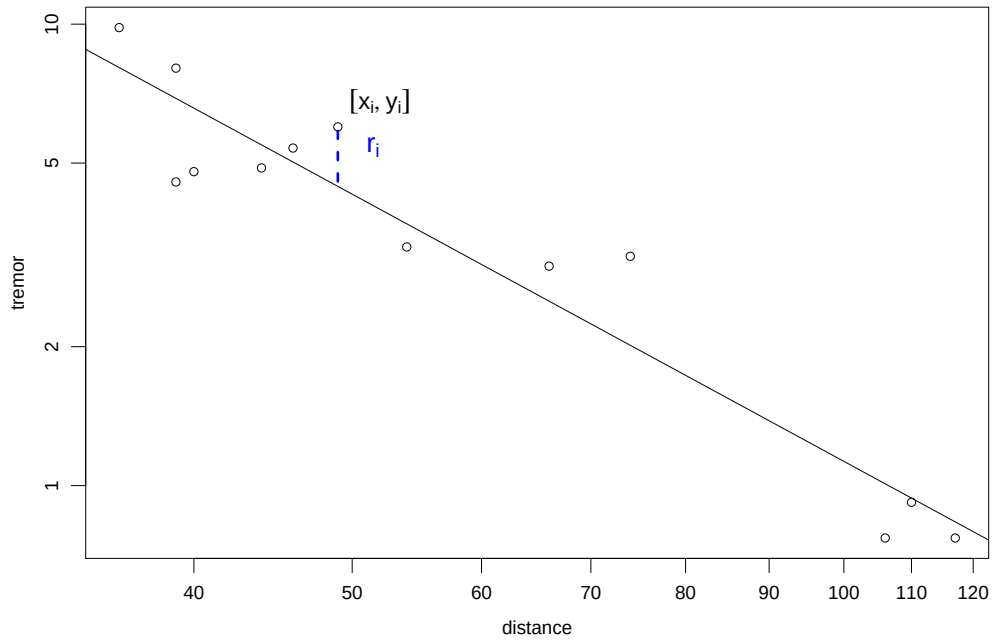

Figure 2.2.*a*: *Estimated straight line for the blasting example*

b **Estimation.** For relating a best fitting model to data in a general case, a rule determining the parameters is needed. The functions producing best fitting parameter values for given data are called **estimators** or, somewhat sloppy, **estimates**.

There are some general rules which determine such functions. The best known one for our problem is the Least Squares principle.

c **Least Squares.** This rule asks that the sum of the squared deviations

$$\sum_{i=1}^{n} r_i^2 , \qquad r_i = y_i - (\alpha + \beta x_i)$$

shall be minimized over α and β. If the random errors E_i have a normal distribution, then this criterion is equivalent to the principle of maximum likelihood.

The estimator then results in the function

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

Further details are described in 2.A.

There is an additional parameter in our model, the variance σ^2 of the random errors E_i. This parameter also needs to be estimated from the data. Since it is not needed for finding the best fitting straight line, we postpone this topic (2.2.m).

d* **Orthogonal regression.** To fit a straight line to a cloud of points, the most natural way might be to minimize the sum of (squared) distances from the points to the line – perpendicular to the line, not in the vertical direction as discussed above. This method is called orthogonal regression. In a different context than the model in 2.1.h, this is indeed the optimal way to estimate the parameters of the line, see 6.1.k.

e **Estimates are random.** An estimator is a function that produces, on the basis of n observations, *one* number as a result. When we assume a probability model, the n observations turn into n random variable, and the result will itself be a random variable. **Thus, estimators are random variables.** Commonly, they are denoted by writing a hat symbol above the symbol for the parameter, which leads to $\widehat{\alpha}$, $\widehat{\beta}$.

Random variable scatter. This is seen in Figure 2.2.e, where the best fitting straight lines are shown for the triples of points from Figure 2.1.k. The estimated line and thus the estimated parameters scatter around the "true" line and the "true" parameter values, respectively.

Since estimators are random variables, we want to study their **distribution.** This will result from studying the probability model. To this end, we need to dismiss the concrete data of the example once more. Let us assume that we know the model including the parameter values. Let us imagine the estimated values for the slope to be obtained by a poor researcher who does not know the true value, and let us find out how probable the possible values are.

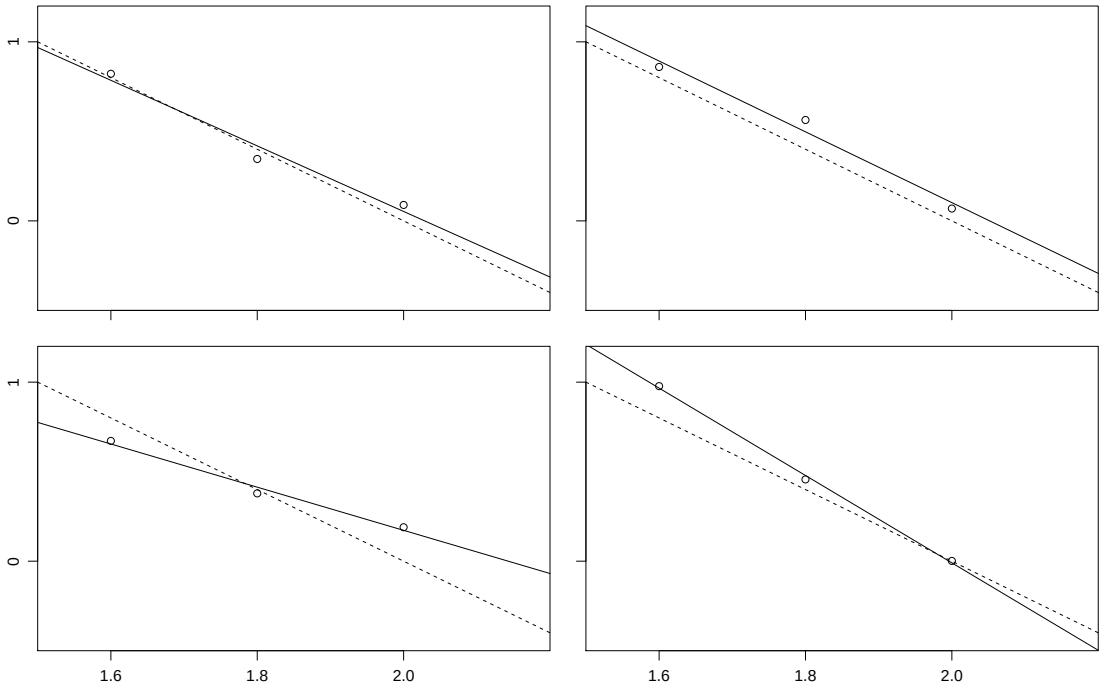

Figure 2.2.*e*: *Four simulated results with the corresponding estimated straight lines* (*solid lines*)

f **Simulated distribution of estimates.** This distribution can be determined by probability calculations. It amy be more easy to understand this step by considering **model experiments** realized by simulating values of the random variables as we did in Figure 2.2.e. We then estimate the parameters for the simulated datasets. When this is done m times, we get m values for the estimators $\widehat{\alpha}$ and $\widehat{\beta}$. Figure 2.2.f shows a histogram of $m = 1000$ values for the estimated slope $\widehat{\beta}$.

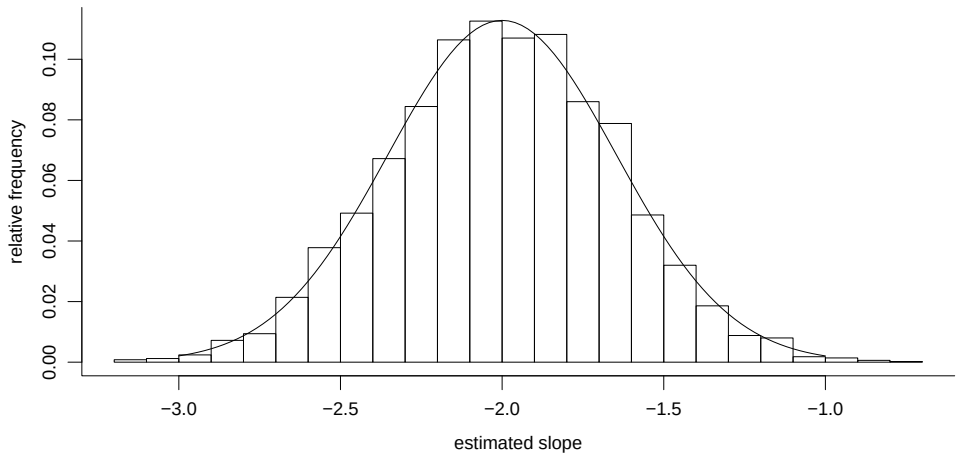

Figure 2.2.*f*: *Simulated and theoretical distribution of the estimator* $\widehat{\beta}$ *of the slope*

g **Theoretical distribution.** As said above, the distribution of the estimators can be determined from the assumptions about the random errors E_i by probability calculation. We have assumed that they are independent and have a normal distribution. Probability calculation then shows that the estimates $\widehat{\alpha}$ and $\widehat{\beta}$ also are normally distributed, i.e.,

$$\widehat{\beta} \sim \mathcal{N}\left\langle \beta, \sigma^{(\beta)2} \right\rangle \quad \text{und} \quad \widehat{\alpha} \sim \mathcal{N}\left\langle \alpha, \sigma^{(\alpha)2} \right\rangle \ ,$$

where $\sigma^{(\beta)}$, $\sigma^{(\alpha)}$ and the "sum of squares" $\mathrm{SSQ}^{(X)}$ of the x values are defined by

$$\sigma^{(\beta)2} = \sigma^2/\mathrm{SSQ}^{(X)} \qquad \sigma^{(\alpha)2} = \sigma^2\left(\tfrac{1}{n} + \bar{x}^2/\mathrm{SSQ}^{(X)}\right)$$
$$\mathrm{SSQ}^{(X)} \ = \ \sum_{i=1}^{n}(x_i - \bar{x})^2 \ .$$

h* **Properties of the estimator.** The Least Squares method is clearly the best know, but not the only way to estimate the parameters. You could simply connect the point with the smalles x value with the one with the largest x, and you would get a straight line that does not need any calculations and would most often be quite acceptable. Nevertheless, it would be difficult to find anybody who would recommend this rule in practice. Why not? We can only answer this question rationally if we examine the distribution of the potential estimators.

i* **Bias.** The results shown above imply that the expected value of the estimator $\widehat{\beta}$ of the slope equals its "true" value, and the same holds for the intersect α. This property is generally called **unbiasedness**. It is certainly a desirable feature: If the estimator necessarily scatters (since it is random), it should at least scatter around the value it is designed to estimate.

(If this fails for some estimator, there is a non-zero **bias**, defined as the difference between the expectation of the estimator $\widehat{\theta}$ and the given parameter θ.)

j* **Statistical efficiency.** As said, each estimator scatters. It is of course desirable that it scatter as little as possible. A suitable measure for the degree of scattering ist the **variance of the estimator**, which turned out to be $\mathrm{var}\langle\widehat{\beta}\rangle = \sigma^2/\mathrm{SSQ}^{(X)}$ for $\widehat{\beta}$. (If an estimator $\widehat{\theta}$ is biased, the **Mean Squared Error (MSE)** $\mathrm{MSE} = \mathcal{E}\langle(\widehat{\theta} - \theta)^2\rangle$ is an appropriate measure.)

The larger the variance (or the MSE), the worse the estimator. In order to compare two estimators, the reciprocal ratio of the variances is commonly used and called the **relative efficiency** of the estimators. The (absolute) efficiency of an estimator is its relative efficienc in comparison with the best estimator, i.e. with the estimator with minimal variance among all unbiased estimators. It can be shown that under the assumptions used above, the Least Squares Estimator is the best in this sense.

k* **Robustness.** Why such a multitude of concepts? If the optimal estimators are so easy to describe and calculate, let us just forget about alternatives! Agreed, we will do this for a long while. Later, we will remember that all this nice theory relies on the normal distribution of the random errors. If this assumption is violated, the estimators introduced here are no longer optimal; so called **robust estimators** are then usually better. For the time being, we summarize:

l The **Least Squares Estimators** $\widehat{\alpha}$ and $\widehat{\beta}$ are

- unbiased and normally distributed with the variances indicated above and
- the best estimators

if the random errors follow the same normal distribution $\mathcal{N}\langle 0, \sigma^2\rangle$ and are independent.

m **Estimation of the variance.** Up to now, we have only discussed the estimation of the two parameters that determine the straight line. Let us now turn to the parameter $\sigma^2 = \mathrm{var}\langle E_i \rangle$ that determines the **variance of the random errors**. The random errors E_i cannot be observed directly, nor can they be derived from the differences $E_i = Y_i - (\alpha + \beta x_i)$ since α and β are unknown – otherwise one could calculate their empirical variance. At least we have, as an approximation to the E_i, the so called **residuals**

$$R_i = Y_i - (\widehat{\alpha} + \widehat{\beta} x_i) ,$$

the differences between the observe Y_i and the "**fitted values**" $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$ .Their empirical value is $\frac{1}{n-1} \sum_{i=1}^{n} (R_i - \bar{R})^2$. The denominator $n-1$ in the definition of the empirical variance was introduced to make it unbiased in the case of a simple random sample. It is not difficult to show that in the case of a simple linear regression, a denominator of $n-2$ is needed for this purpose. Since we always have $\bar{R} = 0$,

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

is the commonly used, unbiased estimator of σ^2.

n* **Distribution of $\widehat{\sigma}^2$.** A multiple of the estimated variance, $(n-2)\widehat{\sigma}^2/\sigma^2$, has a Chi-Squared distribution with $n-2$ degrees of freedom. We refrain here from deriving this result.

2.3 Tests and confidence intervals

a After finding the parameters that best fit the data we may ask if the data are compatible with a (partly) predetermined value of the parameter(s) – in the example, if the slope of the straight line may be -2 as determined by theory (cf. 2.1.f).

Even though the estimated slope is $\widehat{\beta} = -1.92$, this might still be the case since it deviates randomly from the supposedly true value $\beta = -2$. Thus, we cannot conclude stringently that the data contradicts the theory. We therefore ask if the estimated value differs only by chance from the hypothesized value or we are lead to *reject the model with* $\beta_0 = -2$ *as incompatible with the data*. This question is answered by a **statistical test**.

More generally, one may ask which values of the parameter appear compatible with the dat. This question leads to **confidence intervals**.

Here we outline the procedures to answer these questions.

b **Test.** The statistical test examines the null hypothesis

$$H_0 : \quad \beta = \beta_0 = -2; .$$

The complete null hypotheses is: the observations follow the model of a simple linear regression with $\beta = -2$ and any α and σ.

As **alternative hypothesis** H_A we entertain the idea that $\beta \neq \beta_0$, but all other assumptions (distribution of random error, independence) are still valid. Thus, the alternative $\beta \neq \beta_0$

comprises all parameter values except for β_0, that is, values on both sides of β_0 are included – a **two-sided alternative**.

In some applications, only alternatives on one side are of interest, e.g., if deviations to the other side cannot occur. Then, one concentrates on a **one-sided alternative** – in our example $\beta > \beta_0$ (or $\beta < \beta_0$). The null hypothesis to be examined then includes, besides the limiting case, also the other side – here $\beta \leq \beta_0$ (or $\beta \geq \beta_0$).

An suitable **test statistic** is (as usual) a standardized form of the difference between estimate and hypothesized value,

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

The item $\mathrm{se}^{(\beta)}$ corrsponds to $\sigma^{(\beta)}$ of 2.2.g. Since the parameter σ in that formula is usually unknown, it is replaced by $\widehat{\sigma}$. $\mathrm{se}^{(\beta)}$ (sometimes also $\sigma^{(\beta)}$) is called **standard error**.

The test statistic T follows, if the null hypothesis is valid, a so called t distribution with $n-2$ degrees of freedom, which therefore defined the critical value that discriminates rejection and acceptance of the null hypothesis.

The whole resulting procedure is called "**t test**" for the coefficient β.

c **P value.** The p value provides a standardized measure for the "usualness" of the value of the test statistic or for the degree of agreement between the data and the null hypothesis. Its calculation is based on the cumulative distribution function $F^{(T)}$ of the test statistic that applies if the null hypothesis is valid. Figure 2.3.c illustrates the calculation for the case of a two-sided test. (To avoid complications, $\widehat{\beta}$ is used as the test statistic. This would make sense if σ was known.)

The p value equals to the area under the density curve for the range of test statistic values that is at least as "extreme" as the one observed. It therefore equals the probability of obtaining such values, assuming that the null hypothesis is true. If it is small enough, we say that "the data deviates significantly from the null hypothesis", or, if $\beta_0 = 0$ is tested, that the influence of the input variable on the target varialble is "statistically significant". "Small enough" is determined by the usual *convention* as smaller than 0.05.

d **Level.** The threshold of $0.05 = 5\%$ is called the level of the test. It equals the probability of a "error of the first kind", which consists of rejecting the null hypotheses in case it is true. If you do not know this concept, an explanation is called for: Probabilities only exist under the assumption of an assumed model for the observations. Assuming the model with the null hypothesis, the probability that the test result "significant deviation" comes out is calculated – and this result is then wrong. This happens if the p value is below 5%. By construction of the p value, the probability of this result is 5% itself. The same can be ascerted for other values of the level. Thus, the p value allows to assess the significance of the deviation between the data and the null hypothesis for any level without more calculations.

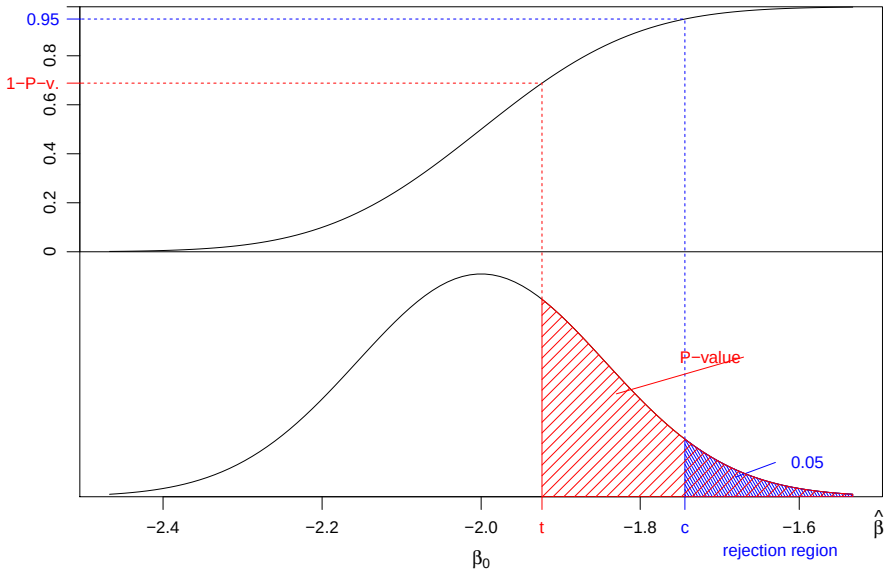

Figure 2.3.*c*: *Illustration of the p value and the rejection region for a two-sided test. The upper curve displays the cumulative distribution function, the lower, the density of the distribution of the test statistic.*

e ▷ **Outputs** for the **nitrogen oxyd and blasting examples** are given in Tabelle 2.3.e (i) and (ii). They have been generated by the function `lm` of the software `R`, see appendix 2.A.c of the current chapter. The estimates for α and β constitute the column "Estimate". When the temperature in the first example increases by 10^o C, then $\log_{10}\langle \mathtt{NO}_2 \rangle$ will decrease by 0.1549. Thus, the $\mathtt{NO}_2$ concentration will be drop by a factor of $10^{-0.1549} = 0.70$ or by 30 percent.

For a test of the null hypothesis $\beta = 0$ (and also for $\alpha = 0$), the value of the test statistic $T = T^{(\beta)}$ (and the corresponding test statistic for $T^{(\alpha)}$) as well as the respective p value(s) are listed. The test statistics have a t distribution under the null hypothesis. Therefore, slope and intercept are tested using a **t test**. ◁

Parameter	Estimate	Standard Error	T Value	(*P-Wert*) Prob. Level
(Intercept)	$\widehat{\alpha} = 1.54492$	$\mathrm{se}^{(\alpha)} = 0.06639$	$T^{(\alpha)} = 23.3$	< 2e-16
temp	$\widehat{\beta} = -0.01611$	$\mathrm{se}^{(\beta)} = 0.00374$	$T^{(\beta)} = -4.3$	1e-04

Residual standard error: $= \widehat{\sigma} = 0.211$ on $n - 2 = 41$ degrees of freedom
R-squared = 0.503 = r^2_{XY}
F-statistic: 40.6 on 1 and 40 DF, p-value: 1.47e-07

Table 2.3.*e* (*i*): *Reduced Output for the nitrogen oxyde example, enriched by the appropriate mathematical symbols*

The third line from below contains the estimate $\widehat{\sigma}$ of the of the random deviations' standard deviation (2.2.m). (The name *Residual standard error* is clearly mistaken, however, since the variance of the *residuals* is (somewhat) smaller than σ^2 as we have just seen, and a *standard error* usually refers to the standard deviation of an *estimator*.)

Parameter	Estimate	Standard Error	T Value	(P-Wert)	Prob. Level
(Intercept)	$\widehat{\alpha} = 3.8996$	$\mathrm{se}^{(\alpha)} = 0.3156$	$T^{(\alpha)} = 12.36$		0
log10(dist)	$\widehat{\beta} = -1.9235$	$\mathrm{se}^{(\beta)} = 0.1783$	$T^{(\beta)} = -10.79$		0

Residual standard error: $= \widehat{\sigma} = 0.1145$ on $n - 2 = 11$ degrees of freedom

Table 2.3.*e* (*ii*): *Reduced Computer-Output for the blasting example with mathematical symbols*

f ▷ In the **blasting example,** theory suggests a slope of $\beta = \beta_0 = -2$ (2.1.f). For this null hypothesis we obtain $T = (\widehat{\beta} - \beta_0)/\mathrm{se}^{(\beta)} = (-1.92 - (-2))/0.1783 = 0.429$. The critical value c for the t distribution with 11 degrees of freedom is, according to a table, 2.201. Hence, the deviation is far from significant. This can also be concluded by having the program calculate the p value, which turns out as 0.676 and is way higher than 0.05. ◁

g Let us now turn to the question which parameter values are plausible in the light of the data.

The confidence interval collects all the parameter values that are not rejected by a given statistical test. A confidence interval thus corresponds to a certain test. For the slope of a simple linear regression we find the interval

$$\widehat{\beta} - q\ \mathrm{se}^{(\beta)} \le \beta \le \widehat{\beta} + q\ \mathrm{se}^{(\beta)}$$

where $q = q_{0.975}^{t_{n-2}}$ is the 0.975 quantile of the mentioned t distribution. We write

$$\widehat{\beta} \pm q\ \mathrm{se}^{(\beta)}\ , \quad \mathrm{se}^{(\beta)} = \widehat{\sigma} \Big/ \sqrt{\mathrm{SSQ}^{(X)}}\ .$$

h ▷ The necessary ingredients for a confidence interval for β are contained in the computer output (Table 2.3.e): We get in the **nitrogen oxyde example** $-0.01769 \pm 2.02 \cdot 0.00278 = -1.9235 \pm 0.3924$, that is, the interval from -0.0233 to -0.0121 (Good programs deliver the confidence interval directly, see 3.1.l)

For the **blasting example,** we get $-1.9235 \pm 2.201 \cdot 0.1783 = -1.9235 \pm 0.3924$, that is, the interval from -2.32 to -1.53. The hypothesized value of -2 lies well within this interval, which again makes it clear that a model with slope -2 fits well with the data. ◁

i At this point, we have answered all **three fundamental questions** of inferential parametric statistics:

1. Which value is the **most plausible** for the parameter? The answer is given by an **estimator**.
2. Is a given parameter value plausible in the light of the data? This decision is made by a **test**.
3. Which values of the parameter(s) appear plausible? The answer is given by *set*, which usually consist of an interval – the confidence interval. confidence interval

2.4 Confidence and prediction band

a In the **blasting example** an important question is: How large will the tremor be if we perform a blasting at 50 m distance from the measurement location? We first ask for the expected value of the tremor for this distance. In a general setting one asks for the **value of the regression function** $h\langle x_0\rangle$ for a given x-value x_0. Can one obtain a **confidence interval** for it?

According to the model, $h\langle x_0\rangle = \alpha + \beta x_0$. First, we are interested in testing the hypothesis $h\langle x_0\rangle = \eta_0$ ("eta")Up to now, a hypothesis has always referred to a given value of a *parameter* of the model. The basic scheme of constructing a test can however be applied to any number calculated from the parameters, like $\eta = \alpha + \beta x$.

b **Test for η_0.** A suitable choice for a test statistic the hypothesis just mentioned is, as usual, the corresponding estimator

$$\widehat{\eta} = \widehat{\alpha} + \widehat{\beta} x_0 \ .$$

It is not difficult to calculate the expectation and variance of $\widehat{\eta}$.

* One gets $\mathcal{E}\langle\widehat{\eta}\rangle = \mathcal{E}\langle\widehat{\alpha}\rangle + \mathcal{E}\left\langle\widehat{\beta}\right\rangle x_0 = \alpha + \beta x_0 = \eta_0$. For finding the variance, we note that $\widehat{\eta} = \widehat{\gamma} + \widehat{\beta}(x_0 - \bar{x})$ with $\widehat{\gamma} = \widehat{\alpha} + \widehat{\beta}\bar{x} = \overline{Y}$ and obtain, since $\operatorname{cov}\left\langle\overline{Y}, \widehat{\beta}\right\rangle = 0$,

$$\operatorname{var}\langle\widehat{\eta}\rangle = \operatorname{var}\langle\widehat{\gamma}\rangle + \operatorname{var}\left\langle\widehat{\beta}\right\rangle (x_0 - \bar{x})^2 = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{\mathrm{SSQ}^{(X)}} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SSQ}^{(X)}} \right) \ .$$

In the common case of unknown σ^2, the suitable test statistic is

$$T = \frac{\widehat{\eta} - \eta_0}{\mathrm{se}^{(\eta)}} \ , \qquad \mathrm{se}^{(\eta)} = \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SSQ}^{(X)}}} \ ,$$

which follows a t distribution with $n-2$ degrees of freedom under the null hypothesis.

The confidence interval for $\eta = h\langle x_0\rangle$ therefore results in

$$(\widehat{\alpha} + \widehat{\beta} x_0) \pm q \ \mathrm{se}^{(\eta)} \ ,$$

where $q = q_{0.975}^{t_{n-2}}$ again marks the 0.975 quantile of the t distribution with $n-2$ degrees of freedom.

c **Confidence band.** This expression for the confidence interval is valid for any x_0. It is natural to regard the limits of the interval as functions of x_0 and to draw these (Figure 2.4.c, inner curves). This forms a "band" that is narrowest for $x_0 = \bar{x}$ and slowly widens to both sides. The estimated straight line $\widehat{\alpha} + \widehat{\beta}x$ forms the middle of the band. This figure allows for reading off the **confidence interval for the value of the regression function** $h\langle x_0\rangle$.

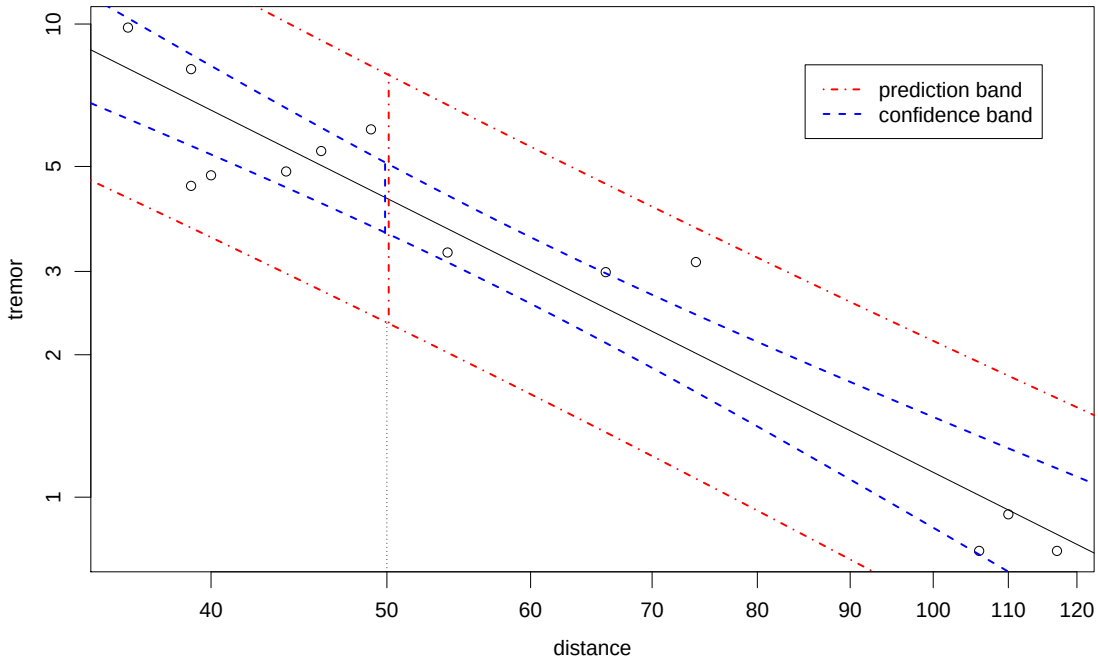

Figure 2.4.*c*: *Confidence band for the value of the regression function* $h\langle x\rangle$ *and prediction band for a further observation in the blasting example*

d **Prediction band.** The confidence band indicates where the *ideal function values* $h\langle x\rangle$ – the expected value of Y for given x must be. This does not answer the question where a *future observation* should lie. But this question is often of more interest than the question for the ideal value: In the example, we want to know how large the value of the tremor that we will obtain from a new blasting event a a distance of $x = 50$ m will be – since it should not trespass the threshold.
This asks for a statement about a *random variable* and has a different nature than the confidence interval, which concerns a *parameter* – a fixed but unknown *number*. Correspondingly, we call the region that we desire now differently and name it **prediction interval**. Connecting endpoints leads to the **prediction band**

Clearly, this interval (and the band) must be wider that the confidence interval (and band) for the expected value, since the random deviation of the future observation comes into play. The band is shown in Figure 2.4.c along with the confidence band.

e* **Derivation of the prediction band.** Let Y_0 denote the (random) value of the response variable for a desired value x_0 of the input variable. Since we do not know the true line, we need to consider the deviation of the new observation from the estimated line,

$$R_0 = Y_0 - (\widehat{\alpha} + \widehat{\beta}x_0) = \big(Y_0 - (\alpha + \beta x_0)\big) - \big((\widehat{\alpha} + \widehat{\beta}x_0) - (\alpha + \beta x_0)\big) \; .$$

Even though α and β are unknown, we know the distributions of the two expressions in the large parentheses: Both of them are normally distributed variables, and they are independent, since the first one only depends on the "future" observation Y_0, whereas the second only depends on the observations $Y_1, \ldots, Y_n$ that were used to estimate the line. Both have expectation 0. Their variances add to

$$\operatorname{var}\langle R_0 \rangle = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SSQ}^{(X)}} \right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SSQ}^{(X)}} \right) \; .$$

This leads to the predition interval

$$\widehat{\alpha} \; + \; \widehat{\beta}x_0 \pm q\widehat{\sigma} \sqrt{1 + \tfrac{1}{n} + (x_0 - \bar{x})^2/\mathrm{SSQ}^{(X)}} = \widehat{\alpha} \; + \; \widehat{\beta}x_0 \pm q\sqrt{\widehat{\sigma}^2 + (\mathrm{se}^{(\eta)})^2} \; ,$$

where again $q = q_{0.975}^{t_{n-2}}$. (The second expression also applies to multiple regression.)

f The **interpretation of the prediction band** is tricky: The derivation shows that

$$P\left\langle \, V_0^*\langle x_0 \rangle \le Y_0 \le V_1^*\langle x_0 \rangle \, \right\rangle = 0.95$$

where $V_0^*\langle x_0 \rangle$ is the lower and $V_1^*\langle x_0 \rangle$ is the upper limit of the interval. However, if we want to obtain a statement about more than one future observation, then we must notice that the number of observations in the band is *not* binomially distributed with $\pi = 0.95$. This is because the events of the individual observations to fall within the band are not independent since they all are related to the same random limits V_0^* and V_1^*. Assume, for example, that the estimate $\widehat{\sigma}$ was quite small due to "bad luck" for the data from which it was obtained. Then, this will be applied for all future observations, and too many of them will fall outside the band.

In order to make sure that at least 95% of all future observations will be contained in a band, such a band must be wider than the prediction band. More details can be found under the keyword **tolerance interval**.

2.A Least Squares

a **Maximum Likelihood.** A good reason for the requirement of Least Squares is given by the principle of maximum likelihood. We have assumed that $E_i \sim \mathcal{N}(0, \sigma^2)$. It follows that the probability density for observation Y_i, if $[\alpha, \beta]$ are the true coefficients, is

$$f\langle y_i \rangle = c \cdot \exp\left\langle - \frac{\big(y_i - (\alpha + \beta x_i)\big)^2}{2\sigma^2} \right\rangle = c \cdot \exp\left\langle - \frac{r_i\langle \alpha, \beta \rangle^2}{2\sigma^2} \right\rangle \; .$$

Here, $r_i\langle \alpha, \beta \rangle = y_i - (\alpha + \beta x_i)$, analogously to 2.2.m, and c is a constant the we do not need to specify. The joint density for all observations is the product of all these expressions, $i = 1, 2, \ldots, n$.

The principle of maximum likelihood means to determine the coefficientss such that this density is maximal.

Calculations are simplified if logarithms are used. This is possible since the parameter values maximizing the density also do this for its logarithm. This leads to

$$\sum_{i=1}^{n} \left(\log\langle c \rangle - r_i\langle \alpha, \beta \rangle^2 / (2\sigma^2) \right) = n \log\langle c \rangle - \frac{1}{2\sigma^2} \sum_{i=1}^{n} r_i^2 \langle \alpha, \beta \rangle \ .$$

Since $n \log\langle c \rangle$ and σ^2 do not depend on either α or β, they can be dropped from these expressions for getting the result. This means that we maximize $-\sum_i r_i^2 \langle \alpha, \beta \rangle$ which is the same as determining the coefficients by the Least Squares.

b **Least Squares.** Therefore, we have to minimize $\sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$ as a function of α and β. Thus, we take derivatives

$$\begin{aligned} \frac{\partial}{\partial \alpha} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 &= \sum_{i=1}^{n} 2\,(y_i - (\alpha + \beta x_i))\,(-1) \\ \frac{\partial}{\partial \beta} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 &= \sum_{i=1}^{n} 2\,(y_i - (\alpha + \beta x_i))\,(-x_i) \end{aligned}$$

and set them zero. This results in

$$\begin{aligned} n\widehat{\alpha} &= \sum_{i=1}^{n} y_i - \widehat{\beta} \sum_{i=1}^{n} x_i \\ \widehat{\beta} \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i - \widehat{\alpha} \sum_{i=1}^{n} x_i \ , \end{aligned}$$

which can be reexpressed as

$$\begin{aligned} \widehat{\alpha} &= \bar{y} - \widehat{\beta}\bar{x} \\ \widehat{\beta} \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i + \widehat{\beta}\bar{x} \sum_{i=1}^{n} x_i \\ \widehat{\beta} \sum_{i=1}^{n} x_i (x_i - \bar{x}) &= \sum_{i=1}^{n} (y_i - \bar{y}) x_i \\ \widehat{\beta} &= \frac{\sum_{i=1}^{n} (y_i - \bar{y}) x_i}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} \ . \end{aligned}$$

The expression for $\widehat{\beta}$ can again be modified: Since $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n} (y_i - \bar{y}) = 0$, we may subtract $\sum_{i=1}^{n} (y_i - \bar{y})\,\bar{x} = 0$ from the numerator and $\sum_{i=1}^{n} (x_i - \bar{x})\,\bar{x} = 0$ from the denominator. Then, we obtain the usual expression

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

for the estimated slope. This completes the derivation of the estimators of α and β.

c **Distribution of the estimated coefficients.** The next point of interest is the distribution of the estimated coefficients $\widehat{\alpha}$ and $\widehat{\beta}$. Here, we defer its derivation to the more general case of multiple linear regression, 3.B.i.

2.R R Functions

a At the end of each chapter we present an appendix containing the corresponding useful R functions.

b **Function `lm`.** The fundamental R function for fitting linear regression models is `lm`. It generates an object of class `lm`, which is essential for the generic functions `print`, `summary` and `plot` – and some more – to let them select the appropriate methods.

```
> r.lm <- lm(log10(NO2) ~ temp, data = d.pollZH16)
```

c **Model formulas.** The first argument of `lm` is a "model formula." Such formulas contain names of variables and, if needed, as in the example, names of functions. They always contain the symbol ~ that relates the target variable on its left to the x variable on its right. The variables must appear either in the `data.frame` specified by the argument `data=` (see below) or be available as separate objects.

Model formulas will be discussed in the section for the next chapter, 3.R.a, in a more general context.

d **Argument `data`.** The variables mentioned in the model formula will primarily be fetched from the `data.frame` specified in the argument `data`. If they are not found there, they will be searched for in the "global environment", where all your objects are stored (or even in attached packages).

R allows for making variables of a `data.frame` available by using the function `attach`. Then, the argument `data` does not need to be specified. This feature is not recommended (since modifications of variable values are not effective in the expected way).

e **Missing values.** The simplest way to deal with data sets with missing values consists of dropping all observations containing at least a missing value in any one of the variables. This is done by default. Other ways of dealing with missing values can be specified in the argument `na.action`. If there are many missing values, this may lead to only few observations remaining effective for analysis. Methods that help in such cases are quite elaborate.

f **Argument `subset`.** This argument allows for restricting the dataset in `data` to a subset of observations. The specification may use variable names of the `data.frame`, like
`lm(log10(ersch) ~ log10(dist), data = d.spreng, subset = dist<=150)`.

g **Function `summary`.** The generic function `summary` generally extracts the "useful" information from an object. Applied to the result of a call of `lm`, the ouput shown in 2.3.e is obtained.

h **Function `predict`.** Predicted values for given values of the input variable are produced by `predict`. If requested, confidence or prediction intervals are also calculated. If the predictions for the input variable values are desired, the function `fitted` is sufficient.

For prediction for input values that are not contained in the dataset used for fitting the model, these values must be packed into a `data.frame` with the respective variable, even though there is only one input variable, like

```
> r.pred <- predict(t.r, newdata = data.frame(x=seq(5,15,0.1)),
                    interval = "prediction")
```

3 Multiple Linear Regression

3.1 Model and Statistics

a **Simple versus mulitple regression.** The dependence of a target variable on a single input variable can be displayed in a scatterplot. This describes the essential relationship readily. The whole machinery of simple regression then serves just for capturing the precision of the estimated line and of predictions. In borderline cases it is needed in addition to decide if the influence of X on Y is "significant."

When we now turn to the relationship of a target varianble Y and several input variables $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$, graphical displays will not be enough. The model, however, readily expands to

$$Y_i = h\left\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)} \right\rangle + E_i \ .$$

The random deviations E_i follow the same assumptions as before. For h, the simplest form is again the linear equation,

$$h\left\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)} \right\rangle = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} \ ,$$

and this defines the model of **multiple linear regression**. The parameters are $\beta_0, \beta_1, \ldots, \beta_m$ and the variance σ^2 of the random deviations E_i. The "intercept" (for the Y axis) is called β_0 instead of α as in simple regression for easing notation later on. The coefficients $\beta_1, \beta_2, \ldots, \beta_m$ measure the "slopes in the directions of the $x^{(j)}$ axes."

b ▷ **Example nitrogen oxyde.** Apart from $\mathtt{NO_2}$ and the temperature, precipitation was measured, and it is plausible to suppose that the target variable depends on it as well. The observation with non-zero precipitation are marked by a + in figure 1.1.b. They might relate to somewhat higher $\mathtt{NO_2}$ values that the rest. The multiple linear regression model with $m = 2$ input variables reads

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + E_i \ .$$

Once again, a linear relationship is more plausible for the logarithmized values of the target variable. We therefore use $Y = \log_{10}\langle \mathtt{NO_2} \rangle$, temperature (`temp`) and precipitation (`prec`) and write

$$\mathtt{log10(NO2)}_i = \beta_0 + \beta_1 \mathtt{temp}_i + \beta_2 \mathtt{prec}_i + E_i \ . \qquad \triangleleft$$

c ▷ Likewise, in the **blasting example**, it did not only happen at different distances, but also with various charges, see figure 1.1.c. The corresponding linear regression model is the same as in the previous example. Here again, a linear relationship is plausible for logarithmized tremor values, and there is even a physical theoretical reason for it. Thus, we use $Y = \log_{10}\langle \text{ErschÃijtterung} \rangle$, $X^{(1)} = \log_{10}\langle \text{Distanz} \rangle$ and $X^{(2)} = \log_{10}\langle \text{Ladung} \rangle$, and write

$$\mathtt{log10(ersch)}_i = \beta_0 + \beta_1 \mathtt{log10(dist)}_i + \beta_2 \mathtt{log10(ladung)}_i + E_i \ . \qquad \triangleleft$$

d **Estimation** of the coeffizients β_j is commonly based by the principle of **Least Squares** as in simple linear regression. The distribution of the estimates is not difficult to determine if based on linear algebra, as shown in the appendices 3.A.l and 3.B.i. Tests and confidence intervals are the obtained in the usual way.

Estimation of the variance σ^2 is based again on the sum of squares of the residuals. Requiring unbiasedness (2.2.m) leads to dividing it by $n - p$, where p is the number of estimated coefficients β_j, which is $m + 1$ if, as usual, the intercept β_0 appears in the model, and $p = m$ otherwise. This number, p, is also called the (number of) **degrees of freedom**. Thus,

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

Following this short account of the theory, we can now consider the interpretation of results.

e ▷ **Computer results** (of function `lm` in `R`, see appendix 3.3.o) for **examples nitrogen oxyde and blasting** are displayed in Tabelle 3.1.e (i) und (ii). They comprise the coefficients in column "Value," the estimated standard deviation of the random deviations, and the results needed for test, to which we return shortly. ◁

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.53326    0.06100   25.14  < 2e-16 ***
temp        -0.01549    0.00351   -4.41  6.4e-05 ***
prec         0.02810    0.08249    0.34     0.73
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Residual standard error: 0.205 on 45 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.307,Adjusted R-squared:  0.277
F-statistic: 9.99 on 2 and 45 DF,  p-value: 0.000257
```

Table 3.1.*e* (*i*): *Computer output in the nitrogen oxyde example*

f **Test for the model.** Before we jump to spotting the p values, we should ask **which questions** should be asked.

In the examples we might ask – if it was not so obvious – if temperature and precipitation, or distance and charge, respectively, jointly have any influence on the target variable. More generally: Is there an influence of the **collection of all input variables** on the target? The null hypothesis pretends that "all β_j (except for β_0) are $= 0$." The corresponding test results are found in the last rows of the tables 3.1.e. (We come back to the next to last row shortly.) They are based on a test statistic that shows an F distribution. Therefore, the test is called the **F test** for the entire regression model.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.832      0.223   12.71   <2e-16 ***
log10(dist)     -1.511      0.111  -13.59   <2e-16 ***
log10(ladung)    0.808      0.304    2.66    0.011 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.153 on 45 degrees of freedom
Multiple R-squared:  0.805,Adjusted R-squared:  0.796
F-statistic: 92.8 on 2 and 45 DF,  p-value: <2e-16
```

Table 3.1.*e* (*ii*): *Computer output in the blasting example*

g The number labelled "Multiple R-Squared" is the square of the "**multiple correlation**," the correlation between the observations Y_i and the **fitted values**

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i^{(1)} + \widehat{\beta}_2 x_i^{(2)} + \ldots + \widehat{\beta}_m x_i^{(m)} .$$

It can be shown that the Least Squares estimates of the coefficients do not only minimize the Sum or Squares of the residuals, but also maximize the correlation between the fitted values and the observations of the target variable, the maximum being the multiple correlation.

The scatterplot in Figure 3.1.g displays this correlation.

Figure 3.1.*g*: *Scatterplot of the observed and fitted values in the example nitrogen oxyde*

The squared multiple correlation is often called **coefficient of determination**, since it expresses the proportion of scatter that is "explained" by the regression model,

$$R^2 = \mathrm{SSQ}^{(R)}/\mathrm{SSQ}^{(Y)} = 1 - \mathrm{SSQ}^{(E)}/\mathrm{SSQ}^{(Y)} .$$

* The value of R^2 may be conceived as the estimate of a corresponding parameter ρ^2, like $\widehat{\sigma}$ estimates σ^2. ρ is the correlation between the target variable and the "true model values" $\beta_0+\beta_1 x_i^{(1)}+\beta_2 x_i^{(2)}+\ldots+\beta_m x_i^{(m)}$. (In exact terms, this in not a correlation since the x values are fixed, nonrandom quantities.) This estimate has a bias depending on the number n of observations and the degrees of freedom p. The number given as "Adjusted R-squared" is a corrected value that is approximately unbiased for ρ^2.

h The question regarding an **effect of a single input variable** $X^{(j)}$ needs a clear specification. The t and p values in the row of tables 3.1.e (i) or (ii) corresponding to $X^{(j)}$ examines whether this variable could be dropped from the model, that is, if the null hypothesis $\beta_j = 0$ is compatible with the data.

The last column in the table shows the usual symbols for the **significance**: Three asterisks

*** for highly significant (p value balow 0.1%), tow asterisks for p values between 0.1% and 1%, one asterisk for barely significant results (1% to 5%), a dot for not quite significant cases (p between 5% and 10%) and blank for higher p values. These symbols simplify finding the significant effects in large tables.

▷ In the **example of of allkaline soils**, Tabelle 3.1.h shows that the second type of alkalinity (basicity), $X^{(2)}$, !!!erfasst a part of the variability of Y that is not explained by the pH value $X^{(1)}$ (3.1.g). – In contrast, in the **Example nitrogen oxyde**, precipitation does not show a statistically significant influence on the response (Tabelle 3.1.e (i)). ◁

The question how strongly $X^{(2)}$ alone, without the competition of von $X^{(1)}$, is related to Y is answered by a simple regression. It is not reflected in the output of the multiple regression fit.

i **Confidence intervals.** The output allows to calculate a confidence interval for each coefficient β_j. It has the usual form $\widehat{\beta}_j \pm q \ \mathrm{se}^{(\beta_j)}$, where $\widehat{\beta}_j$ and $\mathrm{se}^{(\beta_j)}$ aqpear in the columns "Value" and "Std. Error" in table 3.1.e, whereas $q = q_{0.975}^{t_{n-2}}$ can be found by the quantile function of the t distribution (`qt` in `R`).

Some programs produce the confidence intervals directly, see 3.1.l.

j ▷ In the **example nitrogen oxyde**, the interval for the coefficient of temperature is (see 3.1.e (i)) $-0.01832 \pm 2.01 \cdot 0.00331 = -0.01832 \pm 0.00665 = [-0.0250, -0.0117]$.

In the **example blasting** the interval for the coefficient of `log10(dist)` becomes $-1.5107 \pm 2.014 \cdot 0.1111 = -1.5107 \pm 0.2237 = [-1.7345, -1.2869]$ (see 3.1.e (ii)). Now, the value -2 suggested by theory is no longer covered. This value corresponds to free expansion of energy in 3-dimensional space, since in this case the energy is inversely proportional to the surface of the sphere and therefore to the squared radius. If energy is reflected by some soil layers, a less rapid decrease with distance is plausible. ◁

k Let us introduce a new **measure of significance**, which may replace the column "t value" in the table, makes calculation of confidence intervals easy and has an intuitive connotation.

The t values in the common tables are not really needed for the tests of $\beta_j = 0$, since the p values are also given. Nevertheless, they measure significance in a different way than those: If one of them is considerably larger than 2 or smaller than -2, then the effect is more significant accordingly, since the $97.5\ \%$ quantile of a t distribution with a reasonable number of degrees of freedom is about 2. For clearly significant effects, this is a more natural quantitative measure than the p value, which in these cases is simply "very small".

Let us turn this comparison with "approximately 2" into a precise one. The proposed **measure of significance** shall be the "**t ratio**"

$$\widetilde{T}_j = \frac{\widehat{\beta}_j}{\mathrm{se}^{(\beta_j)} \cdot q_{0.975}^{(t_k)}} = T / q_{0.975}^{(t_k)} \ .$$

The amount of significance is no longer given by a comparison with "approximately 2", but with exactly 1. When $\widetilde{T}_j$ is greater than 1 in absolute value, the coefficient is significant.

The t ratio $\widetilde{T}_j$ measures how far within or outside the confidence interval the value 0

is found, as a proportion of the half-width of the interval. If the value is 0.8, then 0 is inside the interval, by 20% of its half length. If $\widetilde{T}_j = 1.2$, then) is outside, by the same percentage.

l ▷ **Example nitrogen oxyde.** Tabelle 3.1.l displays this measure, labeled " `signif` ", and additionally a column containing the degrees of freedom (`df`) that in the present context is always 1, as well as the confidence intervals. Column " `stcoef` " will be explained shortly, whereas " `R2.x` " waits a bit longer 5.3.f. The table shows the results of the R function `regr` to be introduced in 3.3.o. ◁

```
Terms:
               coef df   ciLow   ciHigh  stcoef  R2.x signif p.value p.symb
(Intercept)  1.5333  1  1.4104  1.65612
temp        -0.0155  1 -0.0226 -0.00841 -0.5491 0.008 -2.189   0.000    ***
prec         0.0281  1 -0.1380  0.19425  0.0424 0.008  0.169   0.735
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

St.dev.error:  0.205   on 45 degrees of freedom
Multiple R^2:  0.307    Adjusted R-squared: 0.277
F-statistic:   9.99   on 2 and 45 d.f.,  p.value: 0.000257
```

Table 3.1.*l*: *Result of the R function* `regr` *for the nitrogen oxyde example*

m* The t ratio $\widetilde{T}_j$ allows for a rather easy calculation of the confidence interval in case the latter is not given in the table: Half the width of the interval is $\widehat{\beta}_j/\widetilde{T}_j$ and thus, the full interval is

$$\widehat{\beta}_j \cdot (1 \pm 1/\widetilde{T}_j) \,.$$

For the coefficient of Ttemp we get $-0.0155(1 \pm 1/2.189) = -0.0155 \pm 0.00708$, without the need to determine the quantile of the appropriate t distribution. It space in the table is to be economized, the columns for the confidence interval may therefore be suppressed.

n* **Standardized coefficient.** Another useful indication for each x variable is the *standardized* regression coefficient. ("stcoef" in the table)

$$\widehat{\beta}^*_j = \widehat{\beta}_j \cdot \mathrm{sd}\left\langle X^{(j)} \right\rangle / \mathrm{sd}\langle Y \rangle \,.$$

(`sd` means standard deviation.) It equals the coefficient that is obtained if all x variables are standardized to have 0 mean and standard deviation 1 and the model is fitted with these modified variables. In a simple regression, the slope in this standardization is the correlation coefficient. In the case of multiple regression, the standardized coefficients also measure the strengths of effects of the individual x variables on the response variable, independently of units of measurements or the variability of the variables. These coefficients are expressed on a common scale and made comparable between variables with different measurement units. The standardized coefficients provide therefore a **measure for the relative importance** of the input variables for the response.

Furthermore, a standardized coefficient has a direct interpretation as a slope: When $X^{(j)}$ is increased by one standard deviation $\mathrm{sd}\langle X^{(j)} \rangle$, the estimated value of the response is changed by $\widehat{\beta}^*_j$ standard deviations $\mathrm{sd}\langle Y \rangle$.

3.2 The Flexibility of the Model

a The input variables $X^{(1)}$ and $X^{(2)}$ are continuous variables in both examples, just as the target variable. There is no need for this.

> **In the multiple regression model, there are no assumptions abour the x variables.** They do not have to be of a certain data type, and even less they need to follow any specific distribution. They are not even characterized as *random* variables.

b* In the **example nitrogen oxyde** the weather variable are indeed as random as the pollution measures are. For our analyses, we can and will act nevertheless as if they, e.g., the temparature values, were fixed numbers. Formally, we say that we are interested in the distribution of the target variable, conditional on the values of the x variables.

c

> An input variable may be **binary**, that is, it may be restricted to the values 0 and 1 – or on "no" and "yes" or any other pair of values. If it is the only X variable, then the model reduces to $Y_i = \beta_0 + E_i$ for $x_i = 0$ and $Y_i = \beta_0 + \beta_1 + E_i$ for $x_i = 1$. This regression model is equivalent to the model of two independent samples with potentially different location parameter – a very common and one of the most simple statistical problems.

This is seen as follows: !!!XXX Oft werden bei zwei Stichproben die Beobachtungen mit zwei Indices versehen: Y_{ki} ist die i te Beobachtung der k ten Gruppe ($k = 1$ oder 2) and $Y_{ki} \sim \mathcal{N}\langle \mu_k, \sigma^2 \rangle$. Es sei nun $x_{ki} = 0$, falls $k = 1$ ist, and $x_{ki} = 1$ for $k = 2$. Dann ist $Y_{ki} \sim \mathcal{N}\langle \beta_0 + \beta_1 x_{ki}, \sigma^2 \rangle$, mit $\beta_0 = \mu_1$ and $\beta_1 = \mu_2 - \mu_1$. Wenn man die Beobachtungen wieder mit einem einzigen Index durchnummeriert, ergibt sich das Regressionsmodell mit der binÃd́ren x-Variablen.

d ▷ **nitrogen oxyde example.** Nitrogen oxydes are generated mostly by fuel combustion in traffic. Therefore, we expect the pollution to be less severe dureing the weekend than on working days.

We first compare the values of Thursdays with those of Sundays, setting $X^{(2)} = 0$ for Thursday and $X^{(2)} = 1$ for Sunday. The first input variable shall again be the temperature. The model (3.1.b) then describes two straight lines, one for the group with $x_i^{(2)} = 0$, given by $y = \beta_0 + \beta_1 x$, the other, for group $x_i^{(2)} = 1$, resulting in $y = (\beta_0 + \beta_2) + \beta_1 x$. In both groups, the slope is the β_1. Thus, the these are **two parallel lines**.

Tabelle 3.2.d documents the clear difference between Sunday and Thursday. The effect of Sunday corresponds to the effect of more than 10 degrees of temperature increase. ◁

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.59353	0.04730	33.69	< 2e-16
temp	-0.01950	0.00246	-7.92	5.6e-12
prec	0.02530	0.05968	0.42	0.67
sunday	-0.22353	0.04105	-5.45	4.4e-07

Table 3.2.*d*: *Regression results including the input variable* `sunday` *in the example nitrogen oxyde* (*shortened*)

e ▷ In the **example nitrogen oxyde** there are more than Thursdays and Sundays. If we postulate that traffic will be the same for all working days, then three types of day remain: working, Saturday, and Sunday.

In the **example blasting**, there were four locations that have been labeled arbitrarily as 1, 2, 3, or 4. It is not sensible to include the variable `location` as another input variable $X^{(j)}$ into the model like we have done for `charge`, since a *linear* dependence of the tremor from the location number is not particularly plausible. ◁

An input variable with **nominal** or **categorical values** is often called a **factor** in connection with regression. In order to integrate it into a model, a binary **indicator variable** is generated for each one of its values as follows

$$x_i^{(j)} = \begin{cases} 1 & \text{falls die } i \text{ te Beobachtung zur } j \text{ ten Gruppe gehÃűrt,} \\ 0 & \text{sonst.} \end{cases} .$$

A model for several groups j of observations with different expected values μ_j (but equal distribution apart from that) can be written as

$$Y_i = \mu_1 x_i^{(1)} + \mu_2 x_i^{(2)} + \cdots + E_i$$

with independent, equally distributed E_i. Letting $\mu_j = \beta_j$ converts this into the multiple regression model without intercept β_0.
A binary variable expressing attribution to a group is called **dummy variable**. A categorical input variable – a factor – generates a **"block of indicator** or **dummy variables"**.

f ▷ In the examples, this block is added to the two other input variables (and the numbers j of $X^{(j)}$ need adaptation) The model with the factor `daytype` in the example nitrogen oxyde may be written, including the indicator variables `Work, Sat` and `Sun` (and γ_j instead of μ_j) as

$$\texttt{log10(NO2)}_i = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{prec}_i + \gamma_1 \texttt{working}_i + \gamma_2 \texttt{Sat}_i + \gamma_3 \texttt{sunday}_i + E_i$$

and for the blasting example with input factor `location`:

$$\begin{aligned} \texttt{log10(ersch)}_i = {} & \beta_0 + \beta_1 \texttt{log10(dist)}_i + \beta_2 \texttt{log10(ladung)}_i \\ & + \gamma_1 \texttt{location1}_i + \gamma_2 \texttt{location2}_i + \gamma_3 \texttt{location3}_i + \gamma_4 \texttt{location4}_i + E_i \end{aligned}$$ ◁

g **One dummy variable must go.** A technical point: Since we have included the intercept again in this last model "by mistake", the coefficients can **no longer be uniquely identified.** (cf. 3.B.h). This is because the model values $h\left\langle x_i^{(1)}, \ldots x_i^{(m)} \right\rangle$ do not change if we add a constant to each γ_k and subtracts it again from β_0. A combination of coefficients obtained in this way thus certainly fits any data as well as the original combination. In such situations, one calls the parameters **not identifyable**.

In order to get back to identifyzable coefficients, one can set, e.g., $\gamma_1 = 0$ or, put differently, exclude the dummy variable `working` or `location1`, respectively, from the model. For the simplest model, see 3.2.e, we get

$$Y_i = \beta_0 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \cdots + E_i .$$

Writing $\beta_0 = \mu_1$ and $\ell \geq 2$ $\beta_\ell = \mu_\ell - \mu_1$, eth model is seen to be equivalent to the former one. In the new form, β_ℓ ($\ell > 1$) measures the difference of the effects between the ℓ th and the first levels of the factor on the target variable.

Why should this be better than avoiding the intercept β_0? If two or more factors appear in a model, this solution gets into trouble since we cannot drop the intercept twice. Furthermore, the question whether a factor has an influnce on the target variable will be easier to tackle with the new form, see 3.2.m.

h ▷ The numerical results for the two examples are shown in Tables 3.2.h (i) and (ii). The t and P values belonging to the "dummy" variables `Sat` (`daytypeSat`) and `Sun` (`daytypeSun`), or `location2` to `location4`, respectively, have restricted merit. With our chioce of $\gamma_1 = 0$, they show the significance of the differences between `Sat` or `Sun` to `work`, or between locations 2 to 4 and location 1. ◁

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.56846    0.02191   71.57  < 2e-16 ***
temp        -0.01847    0.00122  -15.17  < 2e-16 ***
prec         0.00461    0.01143    0.40   0.6871
daytypeSat  -0.08256    0.02927   -2.82   0.0051 **
daytypeSun  -0.21109    0.02945   -7.17  5.1e-12 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Residual standard error: 0.184 on 328 degrees of freedom
Multiple R-squared:  0.466,Adjusted R-squared:  0.459
F-statistic: 71.5 on 4 and 328 DF,  p-value: <2e-16
```

Table 3.2.*h* (*i*): *Output for the nitrogen oxyde example with the factor* `daytype`

Coefficients:

	Value	Std. Error	t value	Pr(> |t|)	Signif
(Intercept)	2.51044	0.28215	8.90	0.000	***
log10(dist)	-1.33779	0.14073	-9.51	0.000	***
log10(charge)	0.69179	0.29666	2.33	0.025	*
location2	0.16430	0.07494	2.19	0.034	*
location3	0.02170	0.06366	0.34	0.735	
location4	0.11080	0.07477	1.48	0.146	

Residual standard error: 0.1468 on 42 degrees of freedom
Multiple R-Squared: 0.8322
F-statistic: 41.66 on 5 and 42 degrees of freedom. p-value: 3.22e-15

Table 3.2.*h* (*ii*): *Results for example blasting with the factor* `location`

i ▷ In order to visualize this idea, we suppress `prec` in the **example nitrogen oxyde** by restricting our attention to the observations with Tprec=0. Figure 3.2.i shows the observations and the fitted model. **For each of working days, Saturdays and Sundays**, a **straight line** results, and since there is only one slope for variable Ttemp, the fitted lines must be **parallel**.

Figure 3.2.*i*: *Observations and estimated straight line in the example nitrogen oxyde*

◁

j **Side conditions.** Another possibility to get unique coefficients consists of settin a side condition: **The coefficients γ_j are required to average out to** $= 0$, or equivalently, the sum must be 0. There are two versions: Either the simple average over the γ_j or the weighted version with weights equal to the numbers n_ℓ of observations in the groups ℓ shall be zero. In both versions, all the γ_j change by a constant γ_0 as compared with the earlier method to achieve uniqueness (dropping one dummy variable), and the intercept β_0 changes by $-\gamma_0$ to compensate for this.

This requirement has the advantage that the intercept β_0 can be interpreted in a more "neutral" way: It will be the average of the intercepts in Figure 3.2.i. The estimated coefficients γ_j and the respective confidence intervals now measure the difference to this neutral point, and tests for them examine if the target variable differs for group j from the average of the other groups.

The result for an example will be shown below.

k **Model formulas.** There is a very useful simplified **notation** for denoting models of the kind discussed here, called "model formulas". In the examples they read as

$$\begin{array}{rcl} \texttt{log10(NO2)} & \sim & \texttt{temp} + \texttt{prec} + \texttt{daytype} \qquad \text{resp.} \\ \texttt{log10(ersch)} & \sim & \texttt{log10(dist)} + \texttt{log10(ladung)} + \texttt{location}\,. \end{array}$$

The notation drops indices of observation, coefficients, and the random deviation term as compared to the mathematical formula for a model. The plus sign gets a different role than in usual mathematical formulas: It no longer connects numbers, but input variable, in their original form or endowed with transformations.

The language of model formulas is suitable for writing input to statistical programs. It must be known by the program that the variable `daytype` or `location`, respectively, is a categorical variable, that is, a factor. Then, the program creates the necessary dummy variables automatically. Thus, `daytype` or `location` is a **term** in the model formula that encompasses a whole group of intimately related X variables regarding interpretation.

* In some statistical program packages, the model specification does not allow for transformations. In this case, transformed variables have to be generated first, like `lNO2 = log10(NO2)`. The model then reads `lNO2 ~ temp + prec + daytype`.

l **Concepts.** The "X variables" now appear in diverse forms, which we want to distiguish by different notions: An **input variable** is one that is supposedly related to the target variable. Therefore, it should be included in the model in suitable form. This may be in a transformed version or, if it is a categorical variable, as a whole group of dummy variables.

The X variables appearing in the linear model are also called **regressors**. A **term** in the model formula can be a single regressor or a group of related regressors considered as a unity. Besides the (indicator) variables belonging to a factor there will also be interactions to be introduced soon (3.2.r).

m **Influence of a factor.** Does the `daytype` or the `location` have an influence at all on the target varianble? "No influence" means that the coeficients of all respective regressors are zero, that is, $\beta_\ell = 0$ for all $\ell \geq 2$ in 3.2.g (and $\gamma_1 = 0$ *and* $\gamma_2 = 0$ *and* ... in 3.2.j. Let us describe the usual test for this hypothis in more general form.

n **F-test for comparing models**. The question to be examined shall be if the p^* coefficients β_{j_1}, β_{j_2}, ..., $\beta_{j_{p^*}}$ in a linear regression model are all $= 0$.

- Null hypothesis: $\beta_{j_1} = 0$ *and* $\beta_{j_2} = 0$ *and* ... *and* $\beta_{j_{p^*}} = 0$
- Test statistic:
$$T = \frac{(\mathrm{SSQ}^{(E)*} - \mathrm{SSQ}^{(E)})/p^*}{\mathrm{SSQ}^{(E)}/(n-p)} \; ;$$
$\mathrm{SSQ}^{(E)*}$ is the sum of squares of the random deviation in the "small" model to be obtained from a regression with the remaining $m - p^*$ regressors, and p is the number of coefficients in the "large" model ($= m + 1$ if the model contains the intercept term, $= m$ otherwise).
- Distribution of T under the null hypolthesis:
$T \sim \mathcal{F}_{p^*, n-p}$, F-Verteilung with p^* and $n - p$ degrees of freedom.

This test is called the F-test for comparing models. However, it can only compare a smaller model with a larger one if all the regressors of the smaller also appear in the larger one, that is, if the models are hierarchical. Note that the F test discussed above for the whole model (3.1.f) is a special case, in which the small model only consists of the intercept β_0.

o **Output for testing a factor.** We wanted to test for the influence of a categorical variable. Suitable programs produce the respective test automatically. They provide a table for the influence of each term in the model (Tabelle 3.2.o). Each row in the table gives the result of the test, if the mentioned term has a significant influence on the target variable, or, in other words, if the model without it fits the data significantly worse than the full model. (In R, the function providing this information is called `drop1`, see 3.R.g.)

For the first two input variables, the table gives an equivalent answer to the preceding one (3.2.h). The "F Value" equals the square of the "t value" there. The two tests are equivalent.

The last row compares the full model with the model without the input variable `daytype`. It shows that the influence of this term is highly significant (`Pr(>F)` is much smaller than 0.05).

```
Model:
log10(NO2) ~ temp + prec + daytype
        Df Sum of Sq  RSS   AIC F value  Pr(>F)
<none>               11.2 -1121
temp     1      7.84 19.0  -946  230.24 < 2e-16 ***
prec     1      0.01 11.2 -1123    0.16    0.69
daytype  2      1.84 13.0 -1074   27.05 1.3e-11 ***
```

Table 3.2.*o*: *Tests for the effects of the individual terms in the nitrogen oxyde example. The rows relate to models*: `<none>` *contains two informations about the full model. The next row shows if* `temp` *may be dropped from the model, and the third row does the same for* `perc` *. The last row the results of the F-test examining if* `daytype` *can be dropped.*

p **Table for mixed terms.** If continuous and categorical input variables appear in a model, the relevant information is usually found in different tables: Table 3.1.e is examined to see coefficients of the continuous regressors and the corresponding P-values for the test of $\beta_j = 0$, and in the preceding table (3.2.o), that may need a separate request, the P-values for factors are checked. The result of the function `regr` shows both informations in one table (Tabelle 3.2.p).

```
Terms:
               coef df   ciLow  ciHigh  R2.x signif p.value p.symb
(Intercept)  1.52677  1  1.4854  1.5681
temp        -0.01847  1 -0.0209 -0.0161 0.000 -7.713   0.000    ***
prec         0.00461  1 -0.0179  0.0271 0.006  0.205   0.687
daytype           NA  2      NA      NA 0.003  2.991   0.000    ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

St.dev.error:  0.184   on 328 degrees of freedom
Multiple R^2:  0.466    Adjusted R-squared: 0.459
F-statistic:   71.5   on 4 and 328 d.f.,  p.value: 1.71e-43

Effects of factor levels:
$daytype
  work        Sat         Sun
  0.0417 ***  -0.0409 .   -0.1694 ***
```

Table 3.2.*p*: *Results of* `regr` *for the nitrogen oxyde example*

q **Coefficients for factors.** In the usual presentations of results (3.2.h), the coefficients of the dummy variables of factors are shown in the same table as those for continuous regressors, except for the one dropped for the reason discussed above. Their interpretation, however, depends on the "Coding" (`contrasts`) of the factor into dummy variables, see 3.2.j. For these coefficients, the table contains t- and P-values anyway. Cautious interpretation is

often possible: If "treatment contrasts" are used (the default of the R-function `lm`), then each coefficient measures the difference of effects between the level shown on the row of the table and the one that has been dropped (3.2.g), and the significance of this contrast is tested.

The presentation of `regr` shows the estimated effects for the levels of a factor at the end of the output. The default for the coding requests a weighted average of 0 for the effects (`contrasts="contr.wsum"`). Then, the significance asterisks (*, **) correspond to testing the difference of the effects of the individual level and their mean.

r **Interaction.** In the model 3.2.f, the influence of `location` or `daytype` is included in the simple form of an additive constant. In the nitrogen oxyde example, this means that on weekends, the NO_2 concentrations will be lower, but their increase with temperature remains the same, that is, the straigt lines in 3.2.d must be **parallel**. Similarly, in the example blasting, the change of location is only allowed to lead to an constant change in logarithmic tremor, while the dependence on distance and loading must remain constant.

Of course, it may be that the relationship between NO_2 and temperature differs on weekends by being attenuated. In the other example, tremor might decrease less when the distance increases for one location as compared to another one.

A straightforward generalization of the model allows for different coeffients of the continuous regressor(s) in each group given by the factor. This is called an interaction between the two terms.

The special case for a factor with two levels leads to the following question.

s **Are two straight lines equal?** Do they differ in their intercept, their slope or both? To answer these questions, we consider the model

$$Y_i = \alpha + \beta\, x_i + \Delta\alpha\, g_i + \Delta\beta\, x_i g_i + E_i$$

where g_i is the group indicator: $g_i = 0$ for observations i belonging to one line, and $g_i = 1$ for the other group.

For the group with $g_i = 0$, the straigt line $\alpha + \beta x_i$ results, whereas for $g_i = 1$, we obtain $(\alpha + \Delta\alpha) + (\beta + \Delta\beta)x_i$. Both lines are parallel – have the same slope – when $\Delta\beta = 0$. If in addition, $\Delta\alpha = 0$, then they coincide. (The case of equal intercept but different slope is rarly of interest.)

The model looks somewhat different from the multiple regression model discussed before. However, we only need to set $x_i^{(1)} = x_i$, $x_i^{(2)} = g_i$ and $x_i^{(3)} = x_i g_i$ and to denote the coefficients α, β, $\Delta\alpha$, $\Delta\beta$ as β_0, β_1, β_2, β_3 to bring it to the known form.

The null hypothesis $\Delta\beta = 0$ is tested in the usual output table. The test for "$\Delta\alpha = 0$ and $\Delta\beta = 0$" is another case of the F-test for comparison of two models (**??**).

t **Polynomial regression.** We have already seen several ways to obtain regressors from the original input variables – in the last model, $X^{(3)}$ was the product of two input variables. We may also choose $X^{(2)} = (X^{(1)})^2$. This leads to **quadratic regression**,

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i \ .$$

Figure 3.2.t shows the results of fitting this model in the **example of alkaline soils** (observations with pH > 8.5 have been deleted).
In the same way, higher exponents can be introduced, which leads to **polynomial regression**.

Figure 3.2.*t*: *Quadratic regression in the example of allkaline soils*

* Since all smooth functions can be approximated by a polynomial, this last type of regression is often used when "no" assumptions about the form of dependence between the target variable and an input variable should be taken. However, other models with labels **smoothing** or **nonparametric regression** are clearly more suitable for this situation.

u **Linear regression?** Now, the concepts become confused: We have treated a model with quadratic regression function under our heading of linear regression! **– The label *linear* in the concept of multiple linear regression does *not* refer to the relationship between Y and the $X^{(j)}$, but to the property that the *coefficients* occur linearly in the formula!!**

This means that in the regression function $\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)}$, there is only one β_j between two +'s, and it occurs as a factor in front of any function of the input variables. The β s do not appear as an exponent or in the denominator of a ratio or ... .
A counter example is given by a simple saturation curve, $y = \theta_1 \left(1 - \exp\langle -\theta_2 x \rangle\right)$, and there are numerous other formulas coming from substantive knowledge, prominently in chemistry. Some of them can be linearized by handsome transformations. Such tricks are discussed in discussions about **nonlinear regression.**

v **Optimization.** The quadratic function has ein maximum or a minimum. It is the simplest function that can used for modelling a quality criterion that has an optimal value and gets worse on both sides of them. It therefore can be used when an optimal setting of an input variable is desired.

The model can be expanded. For two input variables, it reads

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \beta_{11} x_i^{(1)2} + \beta_{22} x_i^{(2)2} + \beta_{12} x_i^{(1)} x_i^{(2)} + E_i \ .$$

This is again a linear regression model, with two input variables and five regressors.

* Optimal values for the input variables are obtained by setting partial derivatives to zero. This results in

$$\begin{bmatrix} \widehat{x}_o^{(1)} \\ \widehat{x}_o^{(2)} \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_{11} & \widehat{\beta}_{12} \\ \widehat{\beta}_{12} & \widehat{\beta}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix}$$

w This section has demonstrated that the model of multiple linear regression describes many types of relations between input and target variables, using suitable regressors:

- **Transformations** of input and target variables can be used to turn non-linear relationships into linear ones.
- Two groups are compared by using a **binary regressor.** More groups lead to a **block of dummy variables,** thereby incorporating **categorical** input variables or **factors** into the model.
- The idea of two **different straight lines** for two groups of data can be expressed in a single model by introducing an **interaction** term.
- **Polynomial regression** is also a special case of linear (!) regression.

3.3 Multiple regression is much more than several simple ones

a **Several simple regressions.** Multiple regression was introduced to capture the influence of several (or many) input variables on a target variable. An appealing simple approach to the same goal amounts to fit a simple regression to each of the input variables in turn. This also leads to an estimated coefficient and a confidence interval for each input variable. Multiple regrssion delivers the results in a single table. Is this the essential advantage?

The title of this section states that the difference between several simple and one multiple model is much more fundamental. This will be substantiate in the following.

b ▷ **Modified blasting example.** In order to demonstrate the difference of the two approaches, the dataset of the blasting example was restricted to locations 3 and 6[2] and distances smaller than 100 m.

Tabelle 3.3.b shows the numerical resultats of the simple regressions of the logarithm of tremo to the log of distance.

and, for comparison, the multiple model with input variables log(distance), log(loading) and location.

The simple regression produces a very strange result for the coefficient of the log distance, with a confidence interval of $[-0.635, 0.679]$ The multiple model gives the interval $[-1.40053, 0.0553]$, which is compatible with the results we had based on the dataset of locations 1 to 4 (3.2.h).

The estimated slopes for different models are shown in Figure 3.3.b together with the data. Simple regressions have been fitted for each of the two locations as well as for both of them together. The two other, parallel straight lines represent the result of the multiple regression, indicating the fitted values for an average loading. (The interaction between `log10(distance)` and `location`, which corresponds to different slopes in the two locations, turned out not to be significant.) ◁

Figure 3.3.*b*: *Data of the restricted blasting example* (*locationns* 3 *and* 6) *with estimated regression lines*: *The lines represent simple regressions for the two locations separately and for both of them together, as well as the two parallel lines resulting from the multiple model, for an average value of* `loading`.

[2]location 6 was not contained in the data used before.

```
(i)
regr(formula = log10(ersch) ~ log10(dist), data = dd)
Terms:
             coef df  ciLow ciHigh R2.x signif p.value p.symb
(Intercept) 0.617  1 -0.536  1.771
log10(dist) 0.022  1 -0.635  0.679    0 0.0335   0.946

St.dev.error:  0.22   on 33 degrees of freedom
Multiple R^2:  0.000141    Adjusted R-squared: -0.0302
F-statistic:   0.00466   on 1 and 33 d.f.,  p.value: 0.946
-------------------------------------------------------------------------
(ii)
regr(log10(ersch) ~ log10(dist) + log10(ladung) + st6, data = dd)
Terms:
                coef df    ciLow ciHigh  R2.x signif p.value p.symb
(Intercept)    1.045  1 -0.13304 2.2232    NA     NA      NA     NA
log10(dist)   -0.673  1 -1.40053 0.0553 0.429 -0.924   0.069      .
log10(ladung)  1.579  1  0.67744 2.4805 0.056  1.751   0.001     **
st6            0.183  1  0.00682 0.3595 0.403  1.039   0.042      *

St.dev.error:  0.183   on 31 degrees of freedom
Multiple R^2:  0.346    Adjusted R-squared: 0.282
F-statistic:   5.46   on 3 and 31 d.f.,  p.value: 0.00395
```

Table 3.3.*b*: *Results for the* (*i*) *simple regression of log tremor onto log distance and* (*ii*) *multiple regression with log distance, log loading and location.*

c ▷ Using **artificial examples**, such effects can be expressed even more clearly. Four possible cases with a continuous regressor $X^{(1)}$ and a binary grouping variable $X^{(2)}$ are shown in Figure 3.3.c. The dashed lines indicate the model from which the observations have been drawn. In all cases, they are parallel and correspond to different choices of a slope β_1 and a vertical discrepancy of β_2. The two different symbols for points represent the group membership. The solid line is the result of a simple regression of Y on $X^{(1)}$, whereas the slim rectangle in the right margin displays the difference between the two group means, which corresponds to the simple regression of Y on $X^{(2)}$. Thus, the solid line and the rectangle show the result of fitting the two simple regressions on the regressors $X^{(1)}$ and $X^{(2)}$.

The results of fitting the multiple regression are not shown. They estimate the true model (which is known since the data have been simulated) quite precisely.

The four cases demonstrate the difficulty of interpretation of simple regression drastically:

Figure 3.3.*c*: *Simple and multiple regression for a continuous variable* (*horizontal axis*) *and a binary grouping variable* (*plotting symbols*)

(A) Both variables have a positive effect, $\beta_1 > 0$, $\beta_2 > 0$. The estimated slope and the difference between group means are estimated considerably too large, but at least have the correct sign.

(B) No effect of the continuous input variable $X^{(1)}$. The estimated line gets the clear positive

slope from the group difference.

(C) Opposite effects, $\beta_1 < 0$, $\beta_2 > 0$. The estimated line shows a positive effect of $X^{(1)}$, whereas in the true model, it is negative!

(D) The effects have been chosen to annihilate each other. The wrong conclusion will be that neither variable has an influence of the target variable, even though both have a considerable influence. ◁

d **Coefficient in the multiple model.** Focussing on the multiple regression, we can understand the origin of the contradictory results: The coefficient β_1 indicates by how much the expected value of the target variable will change when the continuous input variable $X^{(1)}$ (the log distance) is increased by 1 unit – and **all other variable remain constant.**

▷ In the **blasting example** this means that the same loading will be chosen and we remain in the same location. We therefore estimate the effect of increasing `log(distance)` keeping loading and location the constant. ◁

If we now consider the simple regression of the target variable on $X^{(1)}$, the meaning of β_1 will change. The second selected location was captured at larger distances than the first one, but nevertheless showed similar tremor values. Partly, this was due to higher loadings employed – because no caution was needed there. If the distance $X^{(1)}$ is increase by 1, there is the tendency that loadings will be higher in the dataset, and that the second location will apply, which in this case also leads to higher values of the target variable. This is why the tremor values do not decrease with the distance when loading and location are ignored. Such an distortion of effects is called "**confounding**": The effect of distance is confounded by the effects of loading and location.

e **Indirect effects.** If two continuous input variables $X^{(1)}$ and $X^{(2)}$ are positivly correlated, an increase of $X^{(1)}$ by 1 will on average lead to an increase of $X^{(2)}$, and this may have an additional effect on the target variable (except if $\beta_2 = 0$).[3] There is an analogous effect if a continuous $X^{(1)}$ has different means for the different levels of a grouping variable appearing as the secondary input term in the regression model.

These consideration show that, in a more general sense, the **meaning of a regression coefficient** depends decisively on the set of other terms included in the model. Note that we are talking about the model. This problem does not have anything to do with estimation of the parameters.

xxxx

[3]* The effect, expressed by the coefficent β_2 in the multiple model and the "regression coefficient" in a separate simple regression of the "target variable" $X^{(2)}$ on $X^{(1)}$, $\beta_{21} = \text{cov}\left\langle X^{(1)}, X^{(2)}\right\rangle / \text{var}\left\langle X^{(1)}\right\rangle$, is $\beta_2 \cdot \beta_{21}$.

f **Cause and effect.** Searching for cause-effect relationships is fundamental for all science. As is well-known, such relationships cannot be proven by merely scrutinizing statistical correlations. Notwithstanding, an important application of regression consists of searching for hints to such relations. The following kinds of condlusions are commonly drawn.

g **Confirmation of a postulated effect.** First type of argument: If a coefficient differs **significantly** from 0 and a causal influence of the target variable on the input variable can be excluded by reasoning (the tremor cannot have an influence on the distance(!)), then this result is interpreted as a onfirmation of a postulated effect.

h **Phony correlation.** Often however, a correlation between an input variable and the target variable results since both of them have a common cause given by a third variable Z.

This is quite common when **time series** are examined. In the last century, the number of births has diminished in developed countries, and so have the storchs that allegedly bring the babies. Time is not the reason for the parallel trend in both variables, but there are common reasons for the development of both of them, which we could desribe as general welfare, and Time can represent these developments partly.

Such situations may also be called **indirect relations**, indirect correlations or phony correlations.

i **Put potential causes into the model!** If the variable Z appears in the model as an input variable (in the appropriate form), then indirect effects via Z are excluded. Ideally the model should therefore include **all potential causes** for the target variable's variation. Then a significant coefficient of a regressor $X^{(j)}$ is a strong indication for a causal relationship with the target variable.

j **Experiments!** An optimal basis for such interpretation is given by planned experiments, in which it is possible to make sure that only the input variable(s) under study are varied, and if there is more than one, they are set such that they are unrelated ("orthogonal") among them.

k **No influence(?)** Second type of argument: A **non-significant** coefficient is often seen as a proof that the corresponding input variable has no influence on the target variable. This is a **wrong conclusion** for several reasons:

- As is the case for all statistical tests, failing significance does not prove the null hypothesis.
- The effects discussed above regarding variables that are not in the model may lead to an annihilation of true causal influences, see the artificial example 3.3.c, case C!).
- The influence of an input variable may be non-linear. The introduction of a squared input variable may show such a relationship.

l Thus, the clearest answer to the question of a **causal effect** of an input on the target variable results from

- a suitable **planned experiment** varying just the focussed input variable(s),

... or, if this is not feasible,

- taking all possible other causes as variables into the model,
- checking the linearity of the relationships (see 4.3.h, 4.2.g),
- calculating a *confidence interval* for the coefficient(s) (instead of p-values), since this still indicates the remaining potential of an influence if the effect is not significant.

m **Orthogonal regressors.** Indirect effects, which have been found guilty for wrong interpretations, cannot occur if the regressors are unrelated, at least not linearly related, labeled as "orthogonal". We could use the word *uncorrelated* if the regressors were random variables. Orthogonal thus means: if we calculate the empirical correlation between the regressors anyhow, we get (exactly) zero. The problems of interpretation with correlated regressors will be discussed in detail in 5.2.d.

We should therefore strive to achieve orthogonality, at least approximately. This is easiest in planned experiments and constitutes a primary principle in their planning.

n If all regressors are orthogonal, then the *estimators* of the coefficients in the multiple regression model are equal to those of the several simple regressions. Even in this case, a multiple model is desirable because the standard deviation of the random deviations is reduced, and this shows a better precision – **shorter confidence intervals** – for the estimated coefficients.

o Summarizing, a multiple regression model gives clearly more valuable results than the corresponding bunch of simple regressions – in the case of correlated regressors even **much more!**

3.R R Functions

a **Model formulas** are the means to communicate the models of regression and analysis of variance to the system. They also do this for models in multivariate statistics. Their identifier is the symbol ~. They form a class of R-objects, called `formula`. Fitting functions for regression and analysis of variance ask for a formula object as their first argument.

In these instances, the target variable is mentioned to the left of the ~ symbol, and input variables or terms formed by them, to the right. The simplest case of such a formula in our context is

```
y ~ x1 + x2
```

The + symbol gets a new meaning here: The variable `x1` and `x2` are not added, but they are seen as the two regressors of the model. In mathematical language, this stands for β_1 `x1` + β_2 `x2`. A random deviations term $+E$ is tacitly added. Likewise, as **intercept** term β_0 is assumed unless suppressed by adding `-1` to the formula, like `y ~ -1 + x1 + x2`. Thus, the formula `y ~ x1 + x2` expresses the model

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + E_i \ .$$

b **Transformatations.** as mentioned in 2.R.c, transformations can be included in the formula directly,

```
log10(ersch) ~ log10(dist) + log10(ladung)
```

c **Factors** or categorical variables may be included in the formula (see 3.2.k). The R-function converts this into a suitable set of regressors. Usually, such variables are stored in the `data.frame` as factors, which will be recognized by the function. If a numerical variable, for example one with values 1, 2, 3, 4, should be interpreted in this way, the **function `factor`** must be used. For example, if `location` in `d.blasting` was not declared as a factor beforehand, we could ask for it by writing

```
log10(ersch) ~ log10(dist) + log10(ladung) + factor(St)
```

to obtain the correct model.

In 3.2.g, **side conditions** have been discussed to make the model's parameters unique in teh case of factors. Three versions have been discussed. The first one was to drop the first dummy variable that corresponds to the factor. This is the default version for the standard fitting funcitons, expressed by `contrasts="treatment"`. A different side condition has lead to "sum" contrasts (3.2.j). They are expressed by `contrasts="sum"`. More details are a subject of analysis of variance.

d **Interactions** between variables (3.2.r) may also be expressed in `formula` easily by adding `x1:x2` ,

```
log10(ersch) ~ log10(dist) + St + log10(dist):St
```

Interactions should only be included between variables for which the simple terms, called main effects, are also contained in the model. Therefore there is a short notation. `x1*x2` is equivalent to `x1+x2+x1:x2` . The preceding model can therefore be expressed as

```
log10(ersch) ~ log10(dist) * St
```

.

e **Mathematics of `formula`s and the function `I`.** This shows that not the symbol `+` only, but also `*` and `:` get new meaning in `formula` expressions, as they symbolize interactions. The same happens to `^` , and in analysis of variance models, even `/` is used for an abbreviation of a common model structure. But in some occasions, `*` should be interpreted in the usual mathematical sense. Then, the function `I()` is required. In the following specification, `^` and `*` are interpreted in the usual way:

```
y ~ x1 + x2 + I(x1^2) + I(x1*(x2-4))
```

f **functions `lm, summary`.** The functions `lm` and `summary` generate the same results as in simple regression (2.R.g), with additional rows in the table of coefficients corresponding to the additional terms in the model.

g **Function `drop1`.** If an input variable in the model formula is a factor, the tests for individual coefficients associated with the factor are of limited value since their meaning depends on the particular choice of the `contrasts` , see 3.2.n. The appropriate test, which checks if the factor as a whole has any influence on the target (the F-test) is performed by the function `drop1` ,

```
> drop1(r.lm, test="F")
```

The function is primarily used to calculate a criterion AIC that we will use later for model selection (5.2.c). It the argument `test` is not specified, no test results are produced.

h **Package `regr0`.** Some features of the functions described above appear little user friendly to the author of this text. He finds it anoying that the results of `lm` are rather incomplete and only minimal information is obtained when they are printed – and the user is asked to call `summary` before the useful results are shown. If the results should be re-used for any further steps, the user needs to know which results are available from the object produced by `lm` and which ones are available from the result of `summary` . Further functions are needed, as just mentioned, for getting confidence intervals for coefficients or tests for the influence of factors and for obtaining other useful information.

The author has therefore written a more comprehensive function `regr` that generates a new class of objects encompassing "all" useful information. If the generic functions `print` or `plot` are applied to them, methods are used that give a more comprehensive output and more diagnostic plots with additional features. (The new class "inherits methods"

from `lm` if no specific methods were deemed necessary.) The function is also useful for many more regression models, for different types of target variables, to be discussed in further chapters. For these models, different functions need to be called in `R`, and `regr` will call them internally as needed and produce information and output in a unified manner.

i **Function `regr`** (package `regr0`). The function `regr` asks for the same arguments as `lm` (and some more, if other models should be fitted). It generates an object of class `regr`.

```
> r.regr <- regr(log10(ersch) ~ log10(dist)+log10(ladung)+stelle,
 data=d.spreng)
```

The important results appear if one types

```
> r.regr
```

The main result is a table containing a row for each term in the model (component `termtable` of the `regr` object). It contains the test for "no influence" for each term. For terms with one degree of freedom (single regressor corresponding to a continuous or binary input variable), the table contains the confidence interval. Standardized coefficients can be requested in the output by setting the corresponding " `userOption` " (or calling `print` with the respective argument). There also is a column `signif` that is explained in 3.1.k.

For terms with more than 1 degree of freedom, the `termtable` contains the result of the F-test mentioned above. The respective estimated coefficients appear after the `termtable` in the output. Care is needed for their correct interpretation of the significance asterisks (*, **, ...) since they depend on the used `contrasts`.

j **Results of `regr`** shown when printed:

- Call by which the object has been generated;
- "main table" `termtable` consisting of the columns
  - `coef`: estimated coefficients $\widehat{\beta}_j$ for terms with a single degree of freedom,
  - `df`: number of degrees of freedom,
  - `ciLow, ciHigh`: confidence interval (for `df` =1),
  - if requested, `stcoef`: standardized coefficient $\widehat{\beta}_j^* = \widehat{\beta}_j \cdot \mathrm{sd}\langle X^{(j)} \rangle / \mathrm{sd}\langle Y \rangle$,
  - `R2x`: The measure R_j^2 of collinearity (5.3.f),
  - `signif`: for terms with a single degree of freedom this is the t-ratio $= T/q_{0.975}^{(t_k)}$, the ratio of the classical t-test statistic and its critical value. The null hypothesis $\beta_j = 0$ is rejected if the absolute value of the t-ratio is > 1.
    For factors and other terms with more than 1 degree of freedom, this column equals a monotone transformation of the F-test statistic that is also compared to 1 for deciding about significance, see 3.2.s.

  - `p value` : The p-value for the test used.

- The output then shows the estimated standard deviation of the random deviations (with a suitable description!), the measure of determination R^2 and the test for the full model.

- If there are factors or other terms with more than 1 degree of freedom, then the corresponding estimated coefficients are shown. (They are stored in the object with several useful columns, like confidence intervals and p-values, to be interpreted with great care, as the component `allcoef`).

- If `print` is called with argument `correlation=TRUE` , then the correlation matrix of the estimated coefficients is shown, see `?summary.lm` .

k **Functions `residuals, fitted.`** The residuals and the fitted values each form a component of the resulting object of `lm` or `regr` . They can therefore be addressed as `rr$residuals` or `rr$fitted.values` , respectively. Nevertheless, it is safer to get them by the "extractor functions" `residuals` and `fitted` , since these functions also work for other models and "know" how to handle missing values. Thus, the residuals should be obtained by typing `residuals(rr)` , and analogously for fitted values.

Caution! If the dataset contains missing values (`NA` s), the residual vector `rr$residuals` is shorter than the columns in the dataset. The statement `residuals(rr)` can return a vector of full length, with `NA` s inserted in the appropriate positions to have it match the observations in the dataset. This only occurs if the argument `na.action=na.replace` was used in the call to `lm` . In `regr` , this is done by default.

4 Residual analysis

4.1 The problem

a **Model assumptions.** The methods for estimation and tests introduced in the earlier chapters rely on assumptions: We required $E_i \sim \mathcal{N}\langle 0, \sigma^2 \rangle$, independently. This can be split up into:

(a) The expected value of E_i is $\mathcal{E}\langle E_i \rangle = 0$,

(b) they all have the same variance $\operatorname{var}\langle E_i \rangle = \sigma^2$,

(c) they are normally distributed,

(d) they are independent.

The regression function must follow a certain formula for which only a few parameters β_j are left to be specified. In the sense explained in (3.2.u), linearity in the β_j s is assumed. If the formula does not have the form that would be "really adequate" for the data, then assumption (a) is violated.

b **Model improvement.** It is often essential that the assumptions are checked for "compatibility" with the data. On one hand, the tests and confidence intervals only guarantee the properties (test and confidence level) that we expect if the assumptions hold On the other hand, we are interested in exploiting any discrepancies to improve the model. This may lead to

- transformations of variables,
- additional terms, e.g., interactions,
- using differents weights for the observations,
- using more general models and statistical methods.

Welcoming chances of improving the model forms the fundamental attitude of **exploratory data analysis**. Here, we do not ask for mathematically based propositions, optimality of statistical methods, or significance, but about methods for creative development of models that agree well with the data.

c **Symptoms, syndrom, diagnose.** (cf. 4.2.h).

4.2 Residuals and fitted values

a **Target variable against fitted values.** In a simple regression, the assumptions – except for independence (d) – can be judged well by just eyeballing a scatterplot of the target variable against the input variable. For multiple regression, a similar plot of the fitted values $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i^{(1)} + \widehat{\beta}_2 x_i^{(2)} + \ldots + \widehat{\beta}_m x_i^{(m)}$ are used on the horizontal axis, as was already done in 3.1.g. Figure 4.2.a shows the respective plot for the blasting example. What can this display tell us about assumptions?

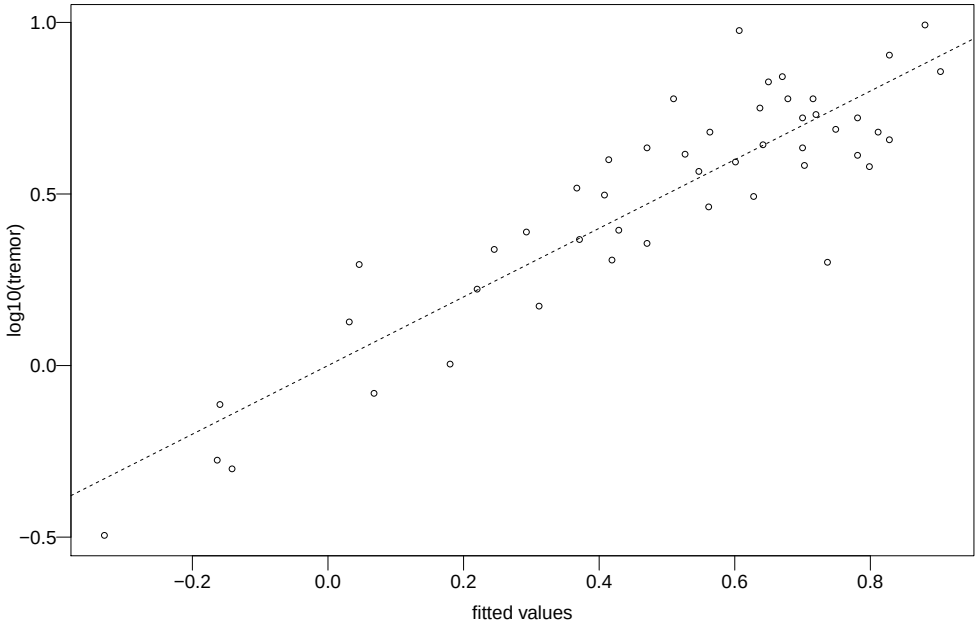

Figure 4.2.*a*: *Scatterplot of the observed and fitted values in the blasting example*

b **(a) Regression function:** ▷ The straight line fits the points in that example quite well. If one looks more carefully, the points to the right of the middle ($\widehat{y}_i$ between 0.4 and 0.7) have a tendency to occur slightly higher than the line, whereas those to the left and to the right appear somewhat more frequently below it.

A slightly bended line would fit the data somewhat better. This is a hint to a deviation of the true expected values of the target variable Y given the values of the input variables, depending on their combination into the fitted values. This in turn can be seen as a violation of $\mathcal{E}\langle E_i\rangle \neq 0$. ◁

c **(b) Equal variances.** This assumption is called **heteroscedasticity**. If it is violated, one speaks of **heteroscedasticity**.

▷ The scattering of the points around the straight line is approximately constant when going from left to right – up to one or two points that one could call "**outliers**", one at $\widehat{y}_i \approx 0.73$, deviating to the low side, and one at $\widehat{y}_i \approx 0.6$, somewhat too high. These extreme points arguably point more to a violation of the assumption of a normal distribution, (c), than of equal variances (b). ◁

A typical deviation from the assumption of equal variances leads to an increase of the width of the scattering for increasing fiited values and thus to a funnel shaped dilatation – or the other way round, which is more rare (cf. 4.4.b). If the random deviations have different variances according to any pattern that is not linked to the value of the regression function, this will not be seen in this display.

d **(c) Distribution of the random deviations.** The deviations of the points from the straight line are the **residuals** $R_i = Y_i - \widehat{y}_i$. They scatter rather **symmmetrically** around the line. The two "outliers" have already been noticed. They hint to a "long-tailed" distribution. We come back to judging the distribution below (4.3.a).

e **Assessment.** The deviations from assumptions that we have noted here are clearly to be tolerated. This is the assessment of this author, which is certainly a rather non-scientific statement! In what sense can they be tolerated? That cannot be made precise. Here are a few considerations:

- If the assumptions hold exactly, there still occur appearent deviations, just like in testing, where in 5% of the cases in which the null hypothesis is exactly fulfilled, the test shows significance anyway. Based on experience, one may be able to develop a good judgement about how large such **random discrepancies** may be. We will soon discuss a method to develop a clearer picture about possible random discrepancies.
- Even if discrepancies become (statistically) significant in some sense, the methods introduced in the previous chapter may still produce adequate results. A judgement on when this is the case must rely on knowledge and experience about the **effects of violations** on results like the distribution of estimators, on P-values of tests and on confidence intervals.
- The importance of precise statements of statistical methods depends on the **scientific questions** of a study. If precise estimation of the effect of an input variable on the target variable in a well determined model is intended, the assumptions are more relevant than if one wanted to roughly select the important variables in a number of potential input variables, or to find a model for predictions in the sense of **??**.

Let us return to more concrete matters! We intend to go into more depth with assessing the assumptions, with more precise diagnostic tools.

f **Tukey-Anscombe plot.** The considerations attached to the plot of the target variable against the fitted values (3.1.g) can be made more precise if we modify the diagram:

> Instead of using the observed values Y_i as vertical coordinates we use the **residuals** R_i. This helps to see the discrepancies better, even more so if the points in the diagram 3.1.g only scatter little around the straight line, that is, if the multiple correlation (or its square, R^2) is high and the residuals therefore get small as compared to the scatter in the Y values. The plot constructed in this way is called after the two authors who porpagated them as indispensible ingredient of an analysis of the residuals, Tukey-Anscombe plot (Figure 4.2.f). In this plot, the points should scatter uniformly around the zero line. Note that the spacings in horizontal direction do not matter here.

We now want to go through the assumptions again, using this new display.

g **(a) Regression function.** A curve in 3.1.g leads to a corresponding curve, "laid flat", in 4.2.f. Even though our eyes are quite good at detecting such patterns, it often helps to draw a curve corresponding to the points in a defined sense, a "smoother", into the plot.

Assumption (a) was $\mathcal{E}\langle E_i\rangle = 0$. If we now collect some observations with similar $\widehat{y}_i$, that is, if we form a vertical "band" in Figure 4.2.f, the average of the respective R_i should be approximately 0. Let us select such a band with preselected with h and mark the average of the residuals in its midpoint as the vertical coordinate (Figure 4.2.g). If we now move the midpoint of the band along the horizontal axis, we obtain the **running mean**.

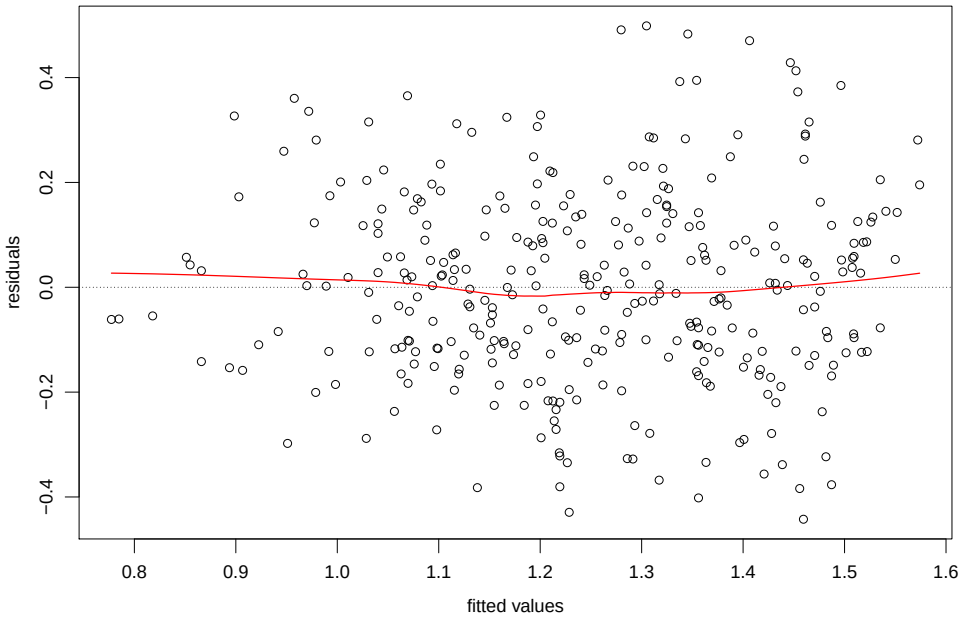

Figure 4.2.*f*: *Tukey-Anscombe plot for the nitrogen oxyde example, with a smooth line*

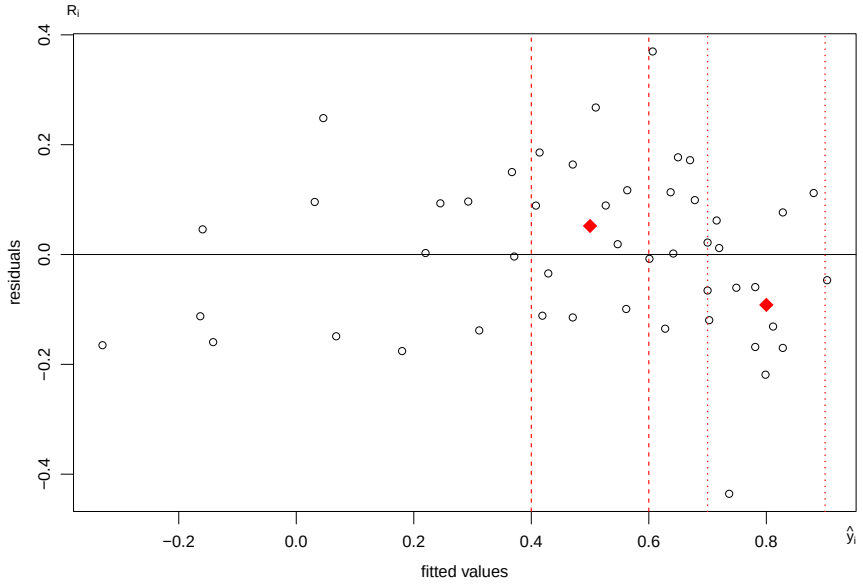

Figure 4.2.*g*: *Determination of the running mean*: *The figure shows the means for two vertical bands.*

This short description should only illustrate the basic idea of smoothing, using a very simple rule for it. The method is easy to beat in several aspects of performance and should therefore not be used. For adequate smoothing methods, consult the literature about "nonparametric regression".

h **Diagnostics should be specific.** If there are outliers, the smoother should not be overly affected by them. It should show specifically the appropriateness of the regression functions as well as possible.

In real applications, we should always guard against a **violation of more than one assumption**. Thus, diagnostics are especially useful if they specifically show a certain deficiency regardless of other flaws.

Methods that only show a limited reaction to a voilation of a certain assumption are called **robust**, cf. 4.5.a. The running mean is strongly influenced by outliers and is therefore not robust in this sense. Thus, a robust method called "loess" is preferred.

i **The curve is random.** The smooth line in Figure 4.2.f clearly displays the deviation from linearity noted in 3.1.g just by eyeballing (4.2.b). Can such a **curvature be caused by the randomness** of the data? Or is there a genuine deviation that could disappear if we improve the model?

We might think of a formal test that checks the corresponding null hypothesis. Here, we rather introduce an informal method, which could be useful more generally, too (cf. 2.2.e).

Step (1): Generate observations corresponding to the model precisely, using random numbers. More precisely: Draw n random numbers E_i^* with a standard normal distribution and obtain $Y_i^* = \widehat{y}_i + \widehat{\sigma} E_i^*$.

Step (2): Using the dataset consisting of the given input variables and the generated Y_i^* for the target variable, fit the model, calculate the smoother for the Tukey-Anscombe plot and add it in the plot for the actual dataset.

Step (rep): Repeat these steps n_{rep} times.

These curves reflect the *random* fluctuations of such curves.

j **Scattering of the curves.** Alluding to the idea of a test on the 5% = 1/20 level, Davies (1995) suggested to augment the plot with its curve for the original dataset by $n_{rep} = 19$ simulated curves. An "informal graphical test" then consists of displaying the 20 curves (including the one for the data) in the same way and asking a third person to select the "most extreme" one. If the curve coming from the actual data is selected, then the deviation should be judged as significant. Since we do not count on getting third persons, we draw the 19 curves in pale color and the one for the actual data, darker, in Figure 4.2.j.

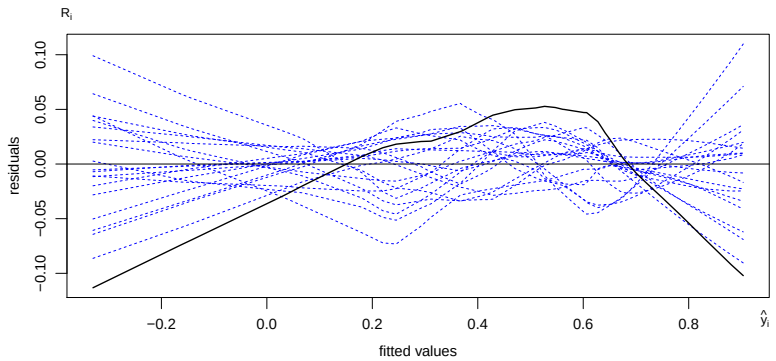

Figure 4.2.*j*: *The smoother for the actual residuals Tukey-Anscombe plot* (—) *and* 19 *simulated curves* (- - -)

▷ In Figure 4.2.j the simulated curves clearly show a wider scattering at both ends of the horizontal axis than in the middle. This is very plausible. Nevertheless, the curve corresponding to the actual residuals appears as the most bended curve. ◁

k **Function `regr`.** The idea of adding simulated smooth lines is implemented in the `plot` method that is called for graphical displays of results of the function `regr` , see Figure 4.2.k. The two additional dark curves will be explained shortly. (4.2.m). The extreme points are marked by their code, which eases their identification. Since they should not overly affect the detail to be seen for the "ordinary" points, the points appearing outside the framed range are shown much too near to those, using a highly nonlinear transformation. They therefore may be much more extreme than they appear to be on the plot. For example, the point marked as `Jp` displays a residuum of 0.499. (In the actual case, displaying vertical coordinates without any distortion would have been as good as what we see. However, in the presence of more extreme outliers, this version is clearly more helpful.)

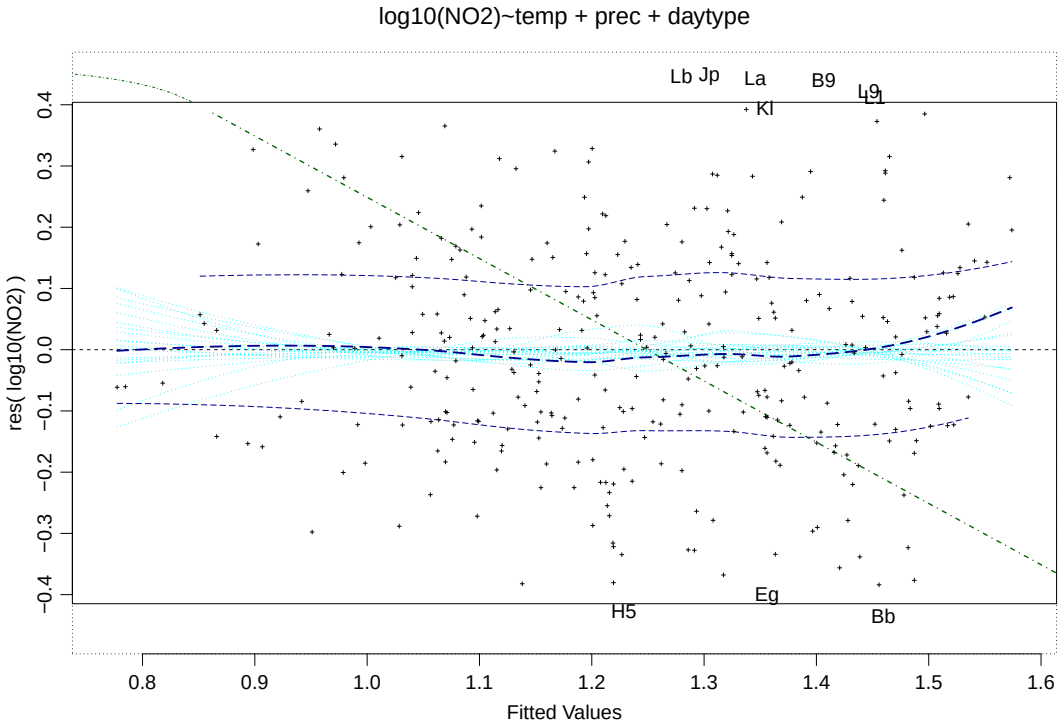

Figure 4.2.*k*: *Enriched Tukey-Anscombe plot for the nitrogen oxyde example.*

Points outside the framed rectangle are approached to the latter by a nonlinear transformation.

l **Reference line.** In Figure 4.2.k , an additional declining straight line appears. It connects points for which the values of the target variable Y are constant (and equal to the mean of the Y_i s). It will prove useful as a reference line in 4.4.m, but it is not drawn by other programs displaying the plot.

m **(b) Equal variances.** These ideas can be adapted to examining the equality of variances, by plotting a "**smoothed standard deviation**", added to the smoother upwards and downwards. More precisely, the residuals $\widetilde{R}_i$ from the smoother discussed above are calculated. Then, the same smoothing method is applied to the positive new residuals,

and a respective line is added to the smoother and shown in the graph, and analogously with the negative new residuals. This leads to the two curves in the upper and lower part of Figure 4.2.k.

n **Scale-location plot.** An alternative possibility consists of plotting the absolute values $|R_i|$ of the residuals against the $\widehat{y}_i$. The `plot` method for `lm` objects in R shows a scatterplot of square root transformed $|R_i|$ against fitted values $\widehat{y}_i$ (Figure 4.2.n (i)), called scale-location plot. In the example, the curve in this plot is very flat and therefore, the assumption of equal variances appears to be clearly fulfilled. .

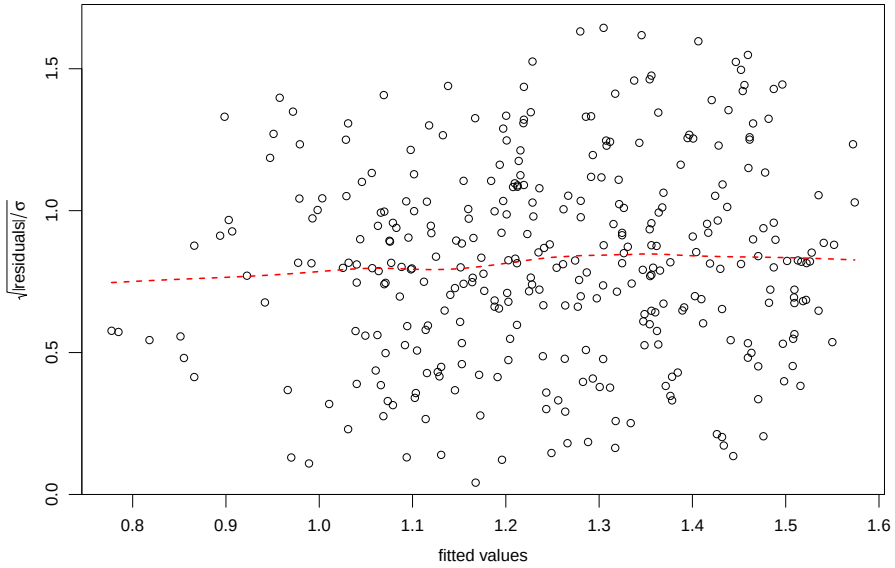

Figure 4.2.*n* (*i*): *Scale-location plot*: *Square root of absolute residuals* $|R_i|$ *against fitted values in the nitrogen oxyde example.*

Figure 4.2.n (ii) shows a slightly different version, implemented in the context of `regr`. Here, the $|\widetilde{R}_i|$ (4.2.m) are shown without transformation, since the appearent symmetry in the foregoing plot may lead to confusion. Furthermore, simulated curves are again generated in the way described for the Tukey-Anscombe plot. They show even more clearly that in the example, the slight increase for the actual residuals is well within the range of random variation.

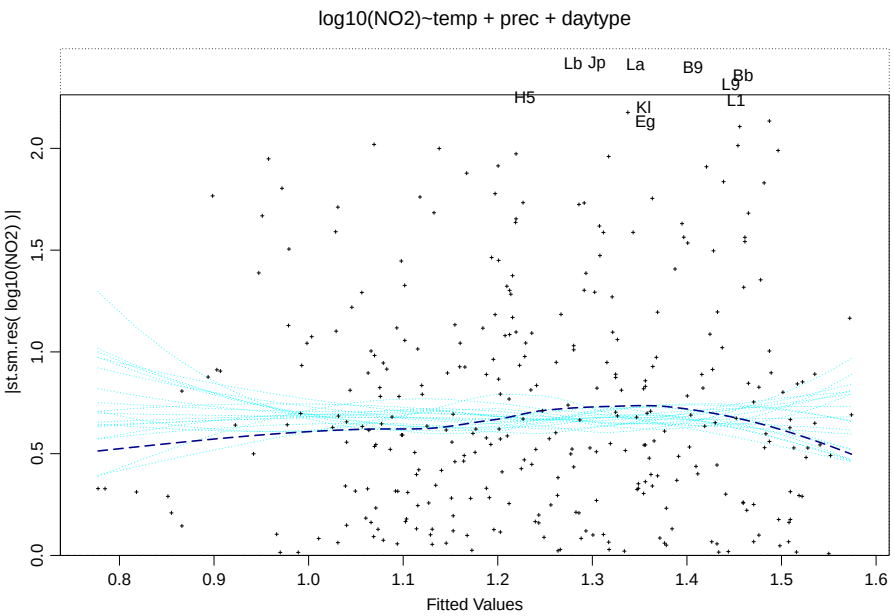

Figure 4.2.*n* (*ii*): *Alternative version of the scale-location plot*: *Absolute residuals* $|\widetilde{R}_i|$ (*siehe* 4.2.*m*) *against fitted values in the nitrogen oxyde example*

4.3 Distribution of random deviations

a **Histogram.** The assumption of a normal distribution for the random deviations ((c) in 4.1.a) can be checked by graphical means. Unfortunately, we do not know the random deviations E_i, but at least we have the **residuals** R_i. The histogram of residuals may be supplemented by a corresponding normal density curve to see the agreement (Figure 4.3.a). It is determined by an expection of 0 and the empirical variance of the residuals.

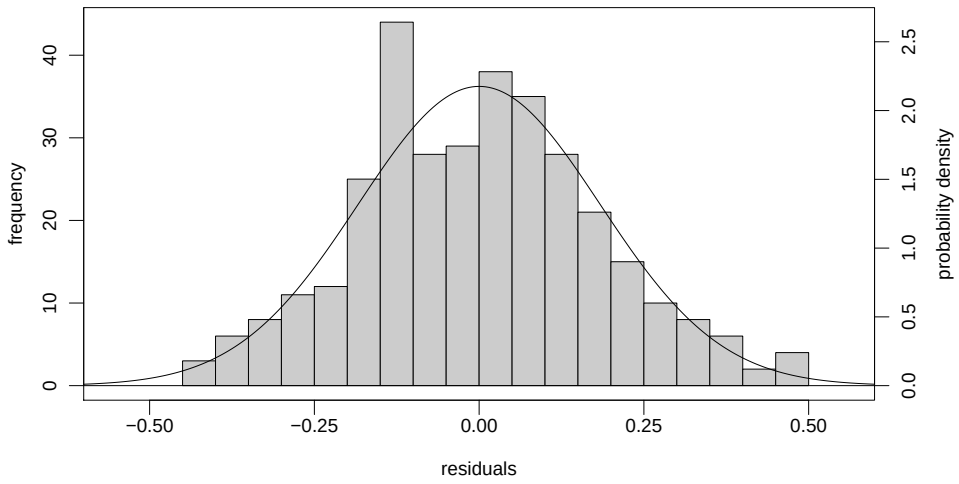

Figure 4.3.*a*: *Histogram of the residuals in the nitrogen oxyde example*

* The empirical variance of the residuals differs from the estimated variance $\widehat{\sigma}^2$ of the random deviations. Instead, it equals $(\sum R_i^2)/(n-1) = \widehat{\sigma}^2(n-p)/(n-1)$.

b **Simulated histograms.** In the example, the histogram shows good agreement with the normal distribution, except for a strange looking peak below -0.1. The question

whether such a peculiar feature may be due to chance can be answered by comparing this histogram with some others that show the natural random variability of histograms. Figure 4.3.b shows 6 such histograms obtained by simulated data according to the model, in the same way as in 4.2.i The second instance displays a similar peak as the histogram for the real data (4.3.a).

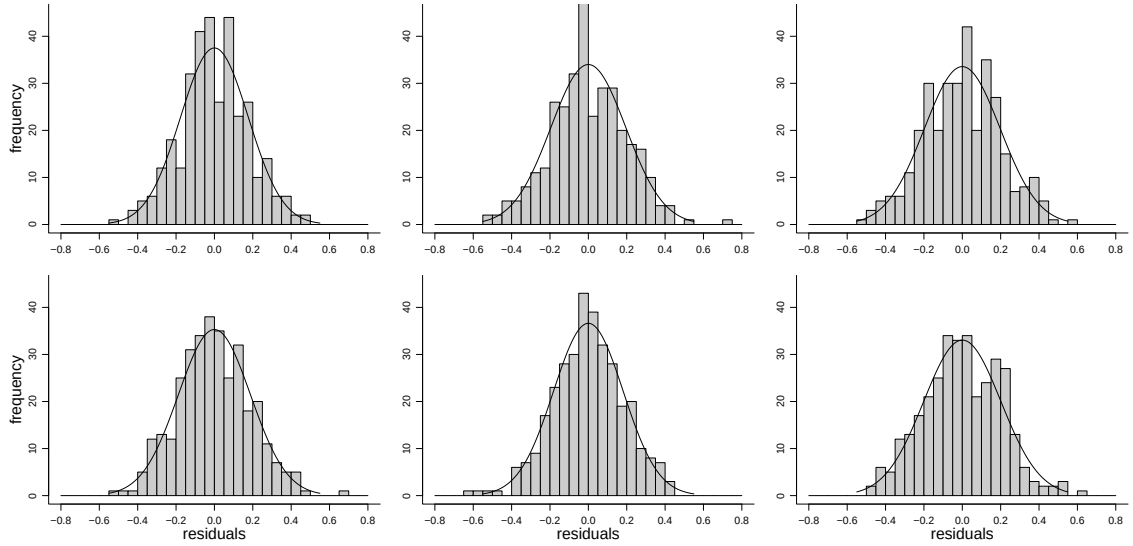

Figure 4.3.*b*: *Histograms of residuals from* 6 *simulated sets of* Y *values in the blasting example*

c Note that checking an assumption of a normal distribution for the values of the target variable itself is meaningless, since the Y_i have different expected values.

d **Normal Plot.** Another graphical check for a normal distribution is possible by generating a so-called normal plot. It is based on comparing the empirical **quantiles** with the quantiles of the (standard) normal distribution, see Figure 4.3.d. For explanations, consult other sources.

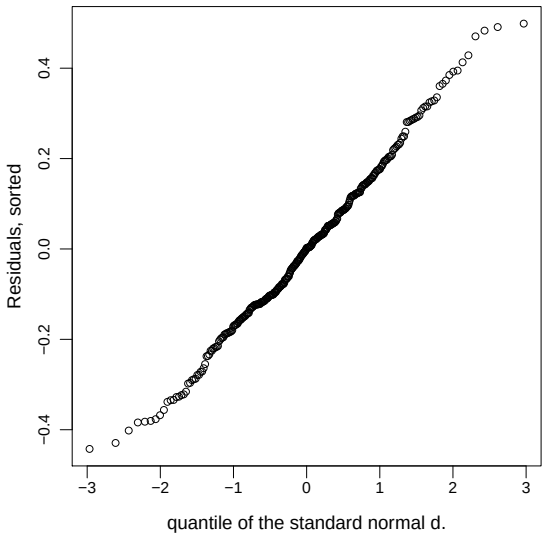

Figure 4.3.*d*: *Normal plot of residuals in the nitrogen oxyde example*

e **Goodness of fit tests.** As shown in Figure 4.3.b, the histograms show random variations even if they come from data corresponding perfectly to the model, and similarly, normal plots can show their random peculiarities. How far can they go?

This question can be answered by statistical tests, called "goodness of fit tests". Each of these tests examines a particular type of potential discrepancies. We do not discuss these tests here for reasons, since as for all tests, these cannot prove the null hypothesis that the distribution is correct.

As in 4.2.i, the judgement if any discrepancies in the normal plot may be due to chance can be eased by adding 19 normal plots from simulating data according to the model.

f **Random deviations and residuals.** In the foregoing arguments, we have cheated a bit, as indicated at the beginning. Instead of examining the distribution of the random deviations E_i, we have used the residuals R_i. This is not the same. The difference can be studied with the aid of matrix algebra in a straightforward way. Here is the main result.

If the random deviations are normally distributed, so are the residuals from a fit by Least Squares. But they have **different theoretical variances**, even if the random deviations have identical ones. In fact, $\mathrm{var}\langle R_i \rangle$ depends on $[x_i^{(1)}, x_i^{(2)}, \dots]$!

(Does the notion of variance for *one* residual confuse you? Each R_i is a random variable with a theoretical variance, which should not be confused with the empirical variance of all the residuals together.)

We have

$$\mathrm{var}\langle R_i \rangle = (1 - H_{ii})\,\sigma^2 \ .$$

The quantity H_{ii} is a function of all the $x_i^{(j)}$. It is called the **leverage** and often denoted as h_i.

g The levarages have some intuitive interpretations:

- If the value of Y_i is increased by Δy_i, then $H_{ii}\Delta y_i$ is the change in the corresponding $\widehat{y}_i$. Thus, if H_{ii} is large, the ith observation "forces" the estimated regression function to come near to it. It has a high leverage to influence the regression function in its own favor, whence the name.
- This makes the result on the variances qualitatively plausible: If the ith observatrion attracts the regression function forcefully, the deviation R_i tends to get smaller, and thus its variance also becomes smaller.
- Leverage points in physics are those that are far from the center of rotation. In our context this means being far from the bulk of the observations regarding the x variables.

* For simple regression, the H_{ii} equal $(1/n) + (x_i - \bar{x})^2/\mathrm{SSQ}^{(X)}$. Thus, they are a simple function of the squared distance from their center of gravity $\bar{x}$. In multiple regression, they are an equally simple function of the so-called Mahalanobis distance H_{ii}

- The leverages are between 0 and 1. Their average always equals p/n.

h **Standardized residuals.** For the residuals to get the same distribution, they therefore have to be standardized,

$$R_i^* = R_i \Big/ \left(\widehat{\sigma}\sqrt{1 - H_{ii}}\right) \quad .$$

These standardized residuals should be used for examining the distribution of the random deviations – as well as for the scale-location plot, which is concerned directly with variances.[4]

Often, however, the differences in the variances $\operatorname{var}\langle R_i \rangle$ are small and ordinary residuals are good enough for residual analyses. The distinction becomes essential in weighted regression.

4.4 Should the target variable be transformed?

a **Diagnosis.** After examining some diagnostic tools, we can now proceed to study some syndroms and derive corresponding therapies. To this end, we reverse the path and go from a known disease to the corresponding syndrom.

▷ In the **nitrogen oxyde example**, the target variable $\mathtt{NO}_2$ was logarithmically transformed after eyeballing the data in order to get a chance to fulfill the assumptions. How would the graphical diagnostics look like if we had not taken logarithms? Figure 4.4.a shows it! ◁

b **A syndrom.** The pattern is most salient in the Tukey-Anscombe plot, particularly in its enriched form shown in Figure 4.4.b. It features

- a curve bended upwards,
- a scatter that increases like a funnel open towards the right side,
- a skewed distribution of residuals.

In the scale-location plot, the increasing scatter towards the right also becomes clear – even more so if residuals $\widetilde{R}_i$ from the smoother in the Tukey-Anscombe plot (siehe 4.2.m) are used instead of the R_i from the (wrong) model.

The distribution of residuals is also clearly seen to be skewed, in the histogram as well as in the normal plot.

c These three symptoms form a **syndrom** that calls for a **transformation**

$$\widetilde{Y} = g\langle Y \rangle$$

of the target variable that reduces the skewness.

In the present example, we already know the therapy: If the target variable is logarithmized, the model fits well.

The logarithmic transformation is not the only one that reduces the skewness. All monotone increasing function bending downwards (being concave) may help. The most popular after the logarithm is the square root transformation. It reduces the skewness less

[4] Often, the quantities R_i/σ are called standardized residuals. However, this simply rescales them and does not make their distribution more equal. Therefore, plots will look the same as with the R_i .

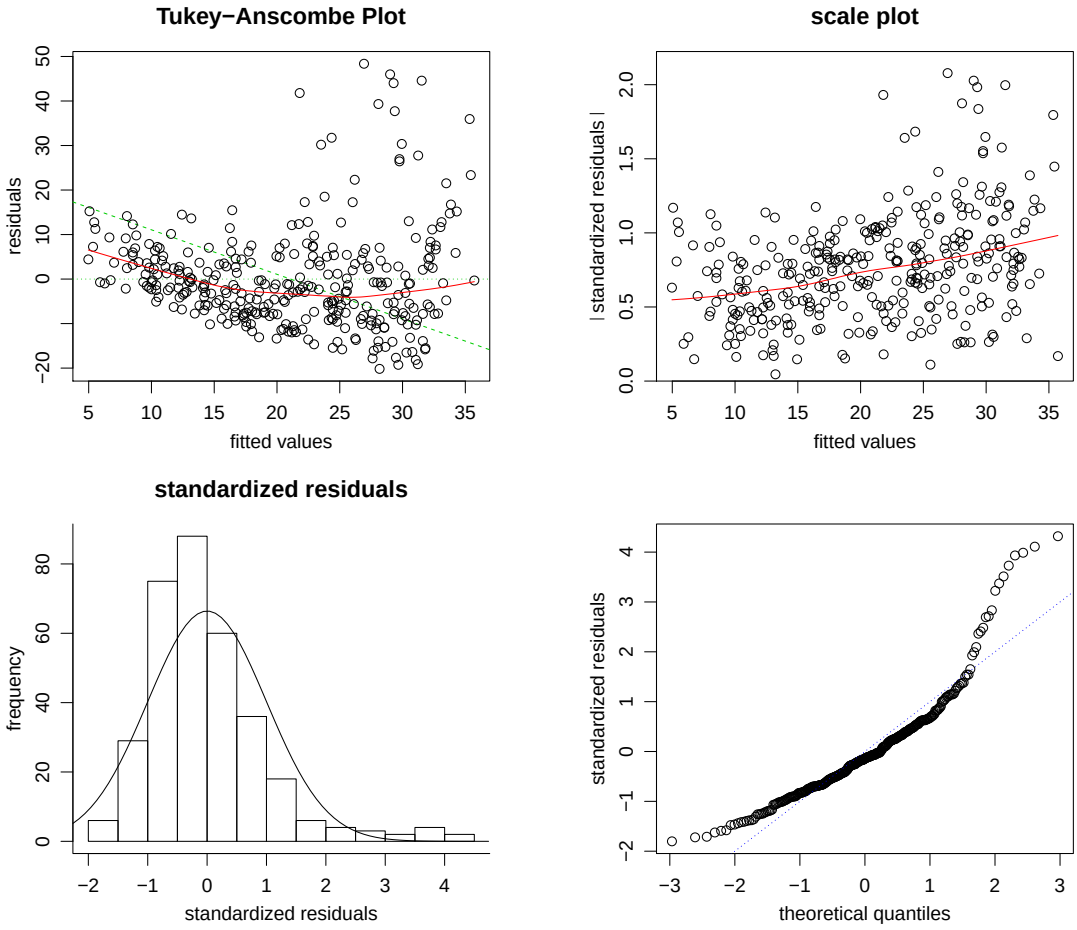

Figure 4.4.*a*: *Tukey-Anscombe plot, scale-location plot and normal plot of standardized residuals for untransformed target variable in the nitrogen oxyde example*

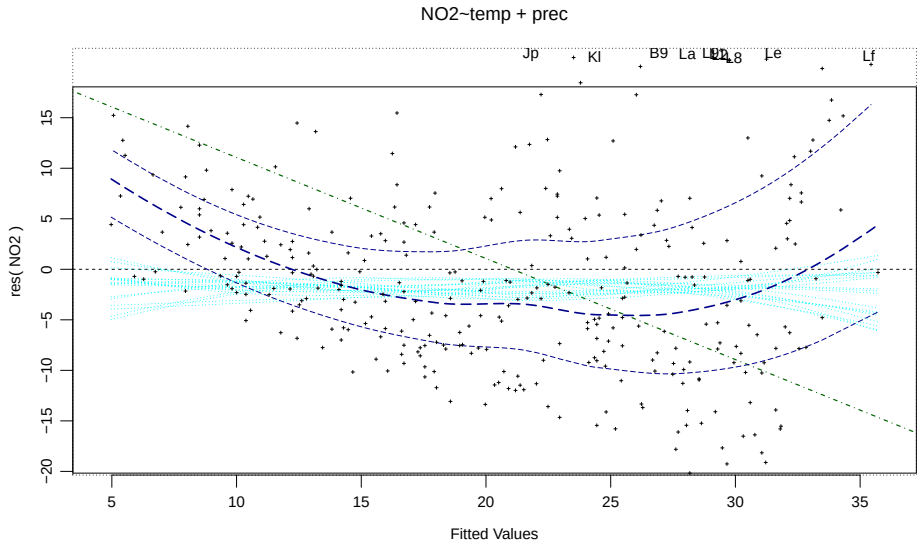

Figure 4.4.*b*: *Enriched Tukey-Anscombe plot for the model with untransformed target variable in the nitrogen oxyde example*

than the former.

Only monotone functions, which are invertible, are appropriate for our purpose here. If a function was used which has the same value for two or more values of the argument, the nature of the relationship of the target variable to the input variables would be fundamentally changed. This goes beyond our intention to modify the model for better correspondence with the assumptions.

We will shortly come back to the choice of a suitable transformation.

d ▷ In the **example of alkaic soils**, the Tukey-Anscombe plot (Figure 4.4.d) shows an analogous pattern as the nitrogen oxyde example with untransformed target variable, but in reversed direction and weaker: The smoother bens slightly downwards, the scatter decreases (for $\widehat{y} > 5$) towards the right and the distribution of residuals is skewed to the uncommon side.

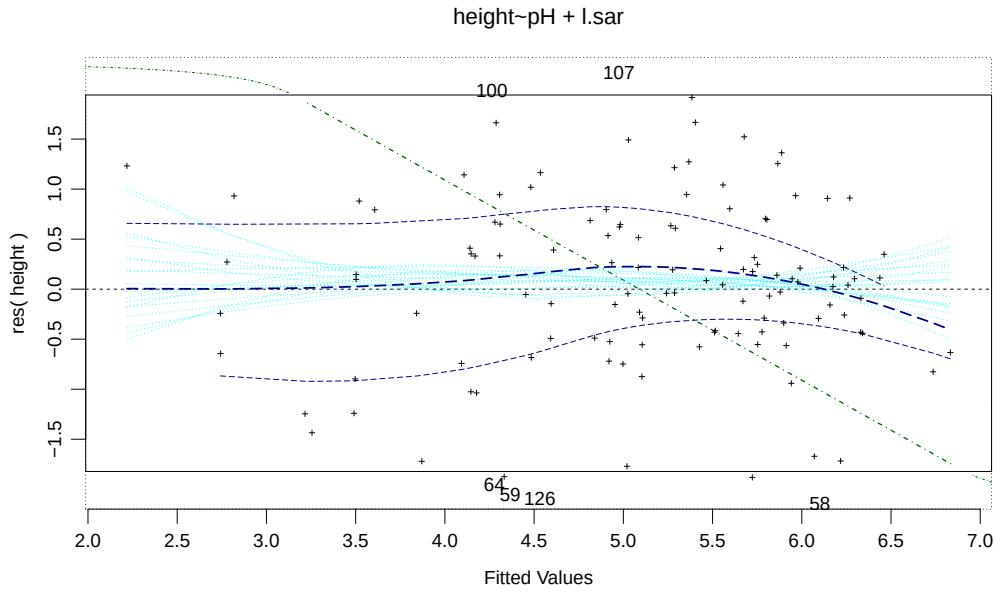

Figure 4.4.*d*: *Tukey-Anscombe plot for the example of alkaic soils*

Here we need a transformation to reduce the negative skewness, that is, its function must be curved upwards (conves). Experience as well as trial and error leads to $\widetilde{Y} = Y^2$ in this case. After applying it, the Tukey-Anscombe plot does not show any discrepancies any more, and the residuals display a symmetric distribution.

The transformation $\widetilde{Y} = Y^2$ is not often helpful. It is not the only one that would lead to an acceptable reslt. In this case, there is a plausibility to it, since the square of the tree's height may be roughly proportional to the area of its leaves. ◁

e **Transformations help often!** All discrepancies could be cured by applying a transformation to the target variable – this must be a very fortunate case! Surprisingly, this case occurs quite frequently in practice. (If you like, you may add philosophical thoughts to this statement – which was however not studied by empirical investigations, as far as I can tell.)

f **Which transformations** should be considered in order to cure a certain syndrom? The following recommendations do not rely on hard facts or theories but rather reflect experience of applied statistics, plausibility, simplicity, and further "soft" arguments.

g **First aid-transformations.** The following transformations are commonly useful:

- the logarithmic transformation, for **concentrations and amounts**, that is, for random variables that only admit positive values;
- the square-root tranformation, for **counts**, and
- the so-called arcsine transformation, for **proportions or percentages / 100**.

These transformations have been called first aid transformations by J. W. Tukey. They should always be applied as a first choice and even be preferred to not using any transformation, unless there are counter arguments.

* Tukey recommended the log transformation also for counts. This causes a problem if zero counts occur – an easy way out is to add 1 to all observed counts or to use the procedure below (4.4.i). If the target variable is a count, the best solution is to use a so-called **Poisson regression**, see other sources for an explanation.

If ordinary linear regression is preferred or counts occur among the input variables, there are usually two reasons for using the square-root transformation:
– Larger (expected) counts lead to larger scattering;
– For small (expected) counts, their distribution is skewed.

Both discrepancies disappear almost completely upon a square-root transfomation if the random deviations have a Poisson distribution.

h **Back transformation of the logarithm.** ... has been studied when treating simple regression.

i The **logarithmic transformation** thus plays a prominent role. From a data analytic view, it applies when the standard deviation is aprroximately proportional to the fitted values. There is, however, a problem if 0 is a possible value. In this case, the transformation rule has to be slightly modified. If it was applied anyway, the function in `R` and other programs would turn zeros into missing data, and it would be misleading to drop all observations with the minimal value, 0, from fitting a model.

The simplest formula for a modified logarithm is $\widetilde{Y} = \log\langle Y + c\rangle$ with a chosen c. Often, $c = 1$ is chosen for simplicity. However, the effect of adding 1 depends heavily on the values Y_i of the original target variable (try it for large and for small numbers!). Therefore, it should depend on the non-zero values of Y – unless Y is a count, in which case $c = 1$ makes sense.

If the non-zero values had a log-normal distribution, then $c = \text{med}\langle Y_k\rangle / s^{2.9}$ with $s = \text{med}\langle Y_k\rangle / q_{0.25}\langle Y_k\rangle$ would be an estimator of the 2.5% quantile ($q_{0.25}$ is the lower quartile). This constant is therefore in the range of the smallest positive values. Its choice is still somewhat arbitrary, but at least it makes the effect of the transformation independent of the scale of the data.

* The R-package `regr0` provides the function `logst` (started log) that leaves the values that are larger than c unaffected and applies a linear function to the smaller ones (including 0) in a smooth way.

j* **Box-Cox-Transformations.** see other sources

k As the back-transformation of the logarithm (4.4.h) pinpoints, when transforming the target variable, the **regression function** for the original target variable **will also change its shape, to nonlinear.**. This may be inappropriate in some applications since the linear regression function (in the sense of linearity in the input variable(s)) for the original target variable may be rooted in some relevant theory in the field of application.

l **Transformations do not always help.** Even if no theory impedes transformations, it may well be that the fortunate case mentioned above does not happen, that is, a bending smoother in the Tukey-Anscombe plot, a dependence of the scatter on the fitted values, and a skewness of the residuals' distribution cannot be cured by just applying a transformation on the target variable.

For example, if (b) the variances appear to be equal and (c) the residuals are symmetrically distributed, but the regression function is not ok, then other remedies are needed. First, transformations of the input variables (see 4.5.a) may be tried. If this fails, non-linear regression should be applied.

m **Reference line in the Tukey-Anscombe plot.** Bended smoothers cannot be cured by transforming the target variable in the following situation, formulated here for simple regression: If the true model would be quadratic and the square term is missing in the formula, a curved smoother will be the result. If the quadratic function has its extremum in the range of the input variable's observed values, then this does not go away by transforming the target variable (by a monotone function).

A monotone transformation of the target variable can only turn the relationship with an input variable into a straight line if this relation itself is monotone. Whether this is the case is easily seen in a plot of the target variable against the input variable, in simple regression, or the fitted values, in multiple regression (Figure 4.4.m (A)). We had reasons for preferring the Tukey-Anscombe plot to this plot. In order allow the user to make the same distinction, a reference line is drawn in the Tukey-Anscombe plot. It connects **points with equal Y values** (the sum of the two coordinates, $\widehat{y}$ and R, equals Y and is constant on this line), as already mentioned in 4.2.l . See diagram (B).

Comparing the diagrams (A) and (B), we conclude: **A monotone transformation of the target variable can only help if the smooth curve in the Tukey-Anscombe plot is nowhere steeper than the reference line,** since that indicates a monotone increase (or decrease) in Y .

4.5 Outliers and long tailed distributions

... some text awaits translation

a* **Robust methods.** Not yet translated.

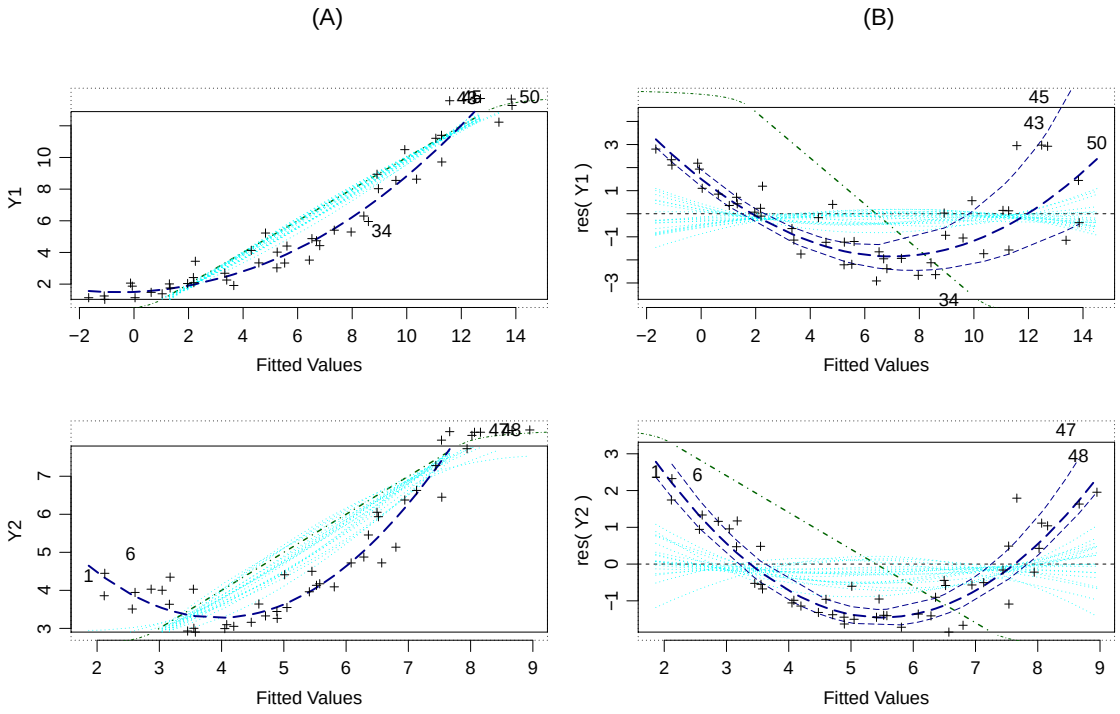

Figure 4.4.*m*: *Target variable* (*A*) *and residuals* (*B*) *against fitted values in two artificial examples. Only the first one can be "cured" by a transformation of* Y *. Based on the reference line, both cases can be distinguished not only in* (*A*)*, but also in* (*B*)*.*

4.6 Residuals and input variables

a The Tukey-Anscombe plot can show discrepancies of the regression function as well as unequal variances. Similar patterns can emerge when an **input variable** is used as teh horizontal axis rather than th fitted values.

▷ Figure 4.6.a displays such plots for the two continuous input variables in the **blasting example**. Again, a smooth line for the actual points and 19 simulated ones are shown. ◁

Figure 4.6.*a*: *Residuals against input variables with smoother* (– – –) *and reference line* (– · – · –)

b **Reference line.** As in the Tukey-Anscombe plot, a **reference line** is shown that should connect points with equal Y values. Since Y_i no longer is the sum of (a multiple of) the horizontal coordinate $x_i^{(j)}$ and the residual R_i, the precise statement defining the reference line is more complicated: it connects points for which the sum of the estimated effect of the input variable $X^{(j)}$ and the residual is constant,

$$\widehat{\beta}_j x_i^{(j)} + R_i = \text{const} \ .$$

The first term is called **component effect**. The sum, called **partial residual**, can also be writtten as $Y_i - \sum_{\ell \neq j} \widehat{\beta}_\ell x_i^{(\ell)}$, that is, the observed target value, "corrected for the effects of the other regressors in the model".

An unpleasant feature is that a **positive effect** of $X^{(j)}$ on the target value leads to a **decreasing line** and vice versa.

[original input variable is used rather than transformed v.]

c The smoother of the residuals in the plots should be flat, up to random variation. Deviations in form of a curved line can often be achieved by transforming the input variable under consideration. As for the Tukey-Anscombe plot, this can only be successful if the smoother does not become steeper than the reference line (cf. 4.2.l).

d **Quadratic term, spline.** If no transformation of $X^{(j)}$ is successful, an additional **quadratic term** $X^{(j)2}$ may help. A simple linear regression then becomes a quadratic one.

Splines, see other sources

e The plots of residuals against input variables may also show that the **scatter of the residuals** depend on such a variable. Then, **weighted regression** is the method to use. See other sources.

4.7 Independence

a **Time sequence.** The last assumption to be checked is the **independence** of the random deviations. If the observations have a natural, in particular, a **temporal order**, the residuals R_i should be plotted in this order.

▷ In the **blasting example** (Figure 4.7.a) a downward tendency towards the end may be spotted. This may arguably be attributed to chance. ◁

b When correlations – temporal, spatial or others – are affecting a dataset, then p-values are often clearly too low and confidence intervals, too short. The well-known method that deals adequately with such correlations is called **Generalized Least Squares**. The topic is a subject of **regression of time series**.

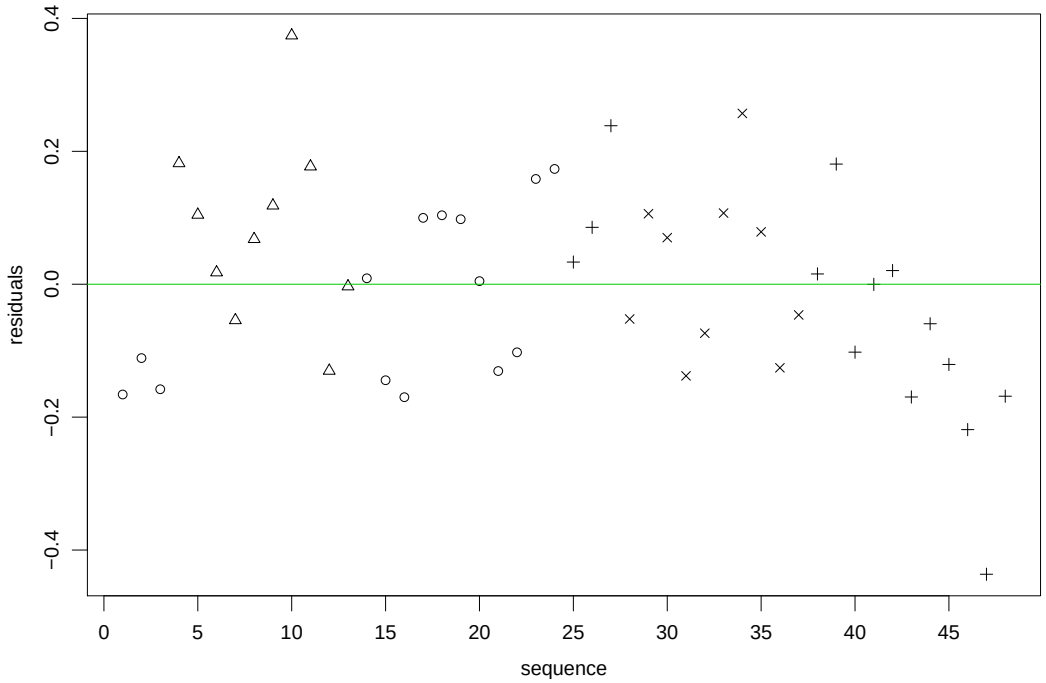

Figure 4.7.*a*: *Residuals against time order for the blasting example. The different locations are marked by different symbols.*

4.8 Influential observations

[some paragraphs to be translated]

a **Leverage plot.** A substanial part of "influence diagnostics" may fortunately be displayed in a single scatterplot, called **leverage plot**. It shows the residuals R_i against the leverages H_{ii} (4.3.g).

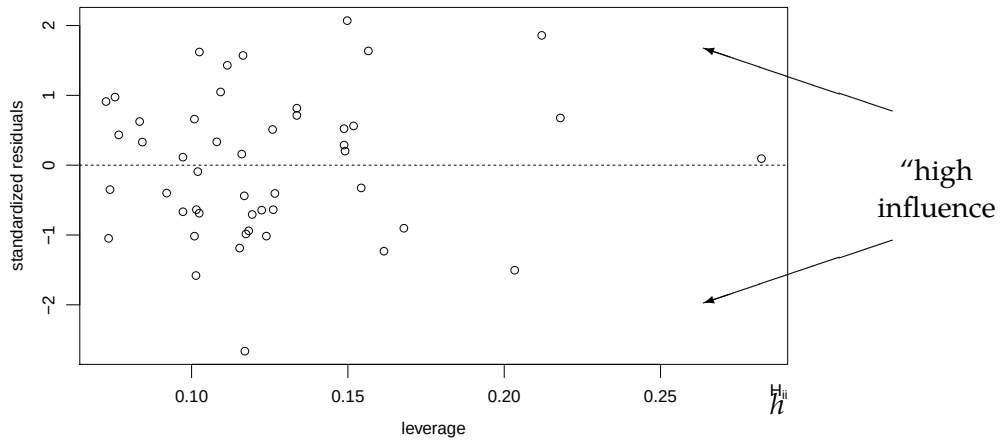

Figure 4.8.*a*: *Leverage plot for the blasting example*

The influence diagnostics increase with $|R_i|$ and H_{ii}. In the plot, the "dangerous" observations show up on the right hand side, in the upper and lower corners. (Figure 4.8.a). There are no well founded limits beyond which "danger" is acute.

▷ In the example, the largest leverage is critically high and the two residuals at $H_{ii} > 0.3$ are remarkably large. It would be useful to repeat the analysis without them and compare the results and their interpretation – well documenting that two observations have been dropped if the analysis continues that way. ◁

b* Cook's distance summarizes the changes of all fitted values $\widehat{y}_i$ caused by dropping the i th observation. (The summary is // $(\underline{\widehat{y}}_{(-i)} - \underline{\widehat{y}})^T (\underline{\widehat{y}}_{(-i)} - \underline{\widehat{y}})$, divided by $p\widehat{\sigma}^2$). It can be written as

$$d_i^{(C)} = \frac{R_i^2 H_{ii}}{p\widehat{\sigma}^2 (1 - H_{ii})^2} = (1/p) R_i^{*2} H_{ii}/(1 - H_{ii}) ,$$

and thus is again a function of R_i^*, H_{ii}, and p.

In `R`, contours of constant Cook's distance are shown in the leverage plot to indicate more precisely the "dangerous" regions. However, there are again no well justified limits.

4.R R Functions

a **Function `plot`** shows, if applied to the result of fitting a regression model, several plots that serve an analysis of residuals. The most important one is the Tukey-Anscombe plot (residuals against fitted values, siehe 4.2.f). Furthermore, there is usually a scale-location plot siehe 4.2.n) for examining the equality of the variances and a normal plot of the residuals (4.3.d) A fourth plot is the levarage plot (residuals against H_{ii}, siehe 4.8.a).

b **Function `plot`** for `regr` objects. If the model was fitted by `regr`, the function `plresx` is subsequently called for all input variables in the model. The argument `plotselect` can be used to ask for a plot of the response against the fitted values, or to select only a subset of the four plots usually shown. If the argument `sequence` is `TRUE`, the residuals are plotted against their sequence. There are many more arguments to control the output of this function.

The aim of this `plot` method is to provide all plots that are commonly useful for a thorough residual analysis. Experience suggests that most users restrict their attention to the plots generated by the call to `plot`, but if they do this for an `lm` object, they do not get what is needed.

c **Function `termplot`.** "Partial" residuals (residuals + component effects) are plotted against regressors, and the estimated effect is drawn. Unless the argument `partial.resid=TRUE` is set, this will usually result in a single straight line shown in the plot.

d **Function `plresx`** for `regr` objects. It plots residuals against input variables, including a smooth line and 19 simulated versions of it. A reference line is drawn which corresponds to *minus* the component effect. It helps to find a suitable transformation of the input variable if needed.

e **Documentation.** The functions for `regr` objects call, for each plot, the function `stamp` that documents the display by adding a line in small print in the lower corner of the right margin. The stamp includes the date and potentially a project and step title set by `userOptions(project=projecttitle, step=stepname)` .

f **Splines.** Residual analysis may lead one to desire a smooth function of an input variable as a new regressor. Such a function may be specified by using "regression splines" as provided by the function `bs` of the package `splines` . The revised call may then contain a formula like `(log10(NO2) ~ bs(temp, df=5) +daytype)` .