7 Linear Regression Models

Regression analysis concerns the study of relationships between quantitative variables with the object of identifying, estimating, and validating the relationship. It is the statistical methodology for predicting values of one or more response (dependent) variables from a collection of predictor (independent) variable values. It can also be used for assessing the effects of the predictor variables on the responses.

We first show an example of simple linear regression model for the prediction of a single response with one single explanatory variable. This model is then generalized to handle the prediction of one dependent variable with several independent variables.

Example (Basel, p. 1-6)**.** Figure 1.5, p. 1-15, shows a strong linear relationship between the annual heating degree days and annual mean temperature for Basel. To find an appropriate model we can use the method of (simple) linear regression.

Example (Weather Report, p. 1-6)**.** We want to examine the linear relationship between the altitude and the annual mean temperature for different stations in Switzerland (see Figure 7.1). The slope of the line can be interpreted as the lapse rate, which is the decrease of temperature with height. The environmental lapse rate, is the rate of decrease of temperature with altitude in the stationary atmosphere at a given time and location. As an average, the International Civil Aviation Organization (ICAO) defines an international standard atmosphere with a temperature lapse rate of 6.49 °C/1000 m from sea level to 11 km. Source: Wikipedia.

7.1 Multiple Linear Regression

Example (Basel, p. 1-6)**.** We are interested to find a linear model for the annual cloudiness in Basel. Figure 1.5, p. 1-15, supports the assumption that there is a linear relationship between annual cloudiness as dependent variable on one hand and the annual sunshine duration and annual precipitation as explanatory variables on the other hand. We consider the time period 1980-2000.

The classical linear regression model states that Y is composed of a mean, which depends in a continuous manner on the z_i's, and a random error ε, which accounts for measurement error and the effects of other variables not explicitly considered in the model. The values of the predictor variables recorded from the experiment or set by the investigator are treated as fixed. The error is viewed as a random variable whose behavior is characterized by a set of distributional assumptions. Specifically, the linear regression model with a single response takes the form

$$\begin{array}{ccccc} Y & = & \beta_0 + \beta_1 z_1 + \ldots + \beta_r z_r & + & \varepsilon \\ \text{response} & & \text{mean (depending on } z_1, \ldots, z_r) & & \text{error} \end{array} \tag{7.1}$$

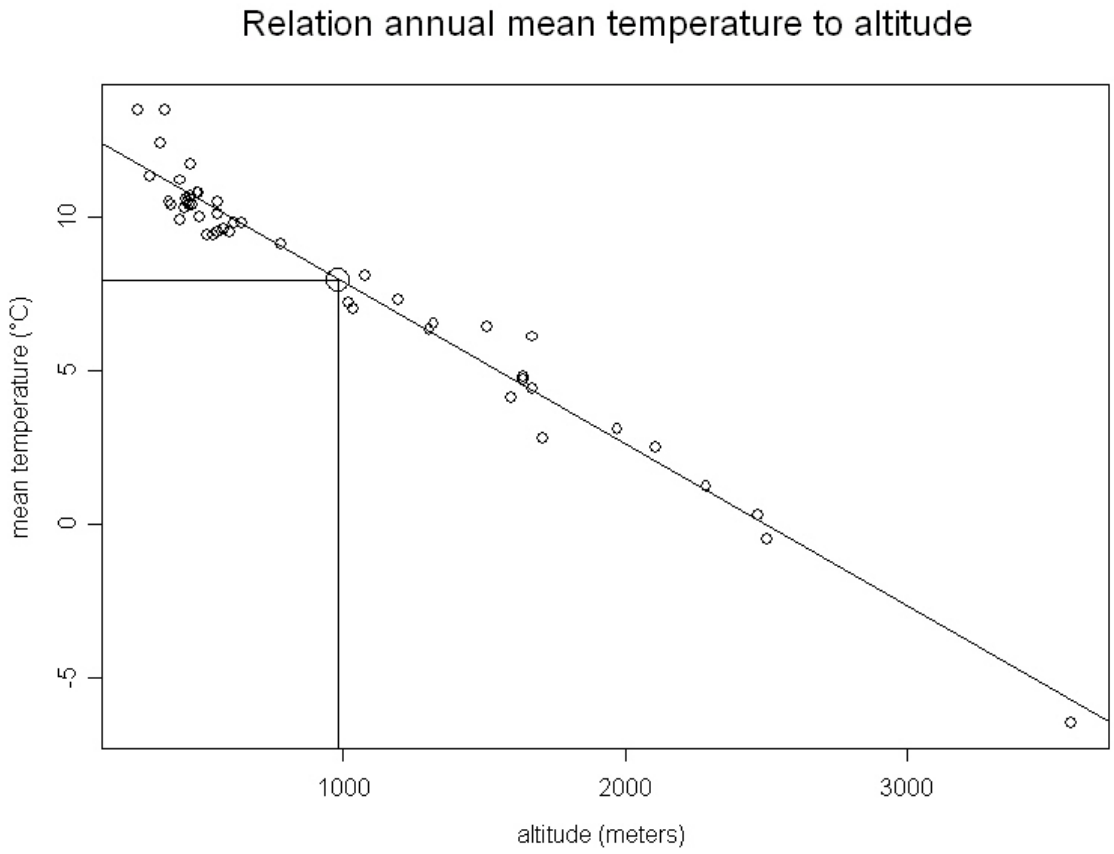

Figure 7.1: Plot of the annual mean temperatures as a function of the altitude for the different Swiss stations. Data set: Weather Report, p. 1-6.

Remark. The term linear refers to the fact that the mean is a linear function of the unknown parameters $\beta_0, \ldots, \beta_r$. The predictor variables may or may not enter the model as first-order terms.

Remark. Figure 7.2 shows the notation that will be used in this chapter.

With n independent observations on Y and the associated values of z_i, the complete model becomes

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 z_{11} + \ldots + \beta_r z_{1r} + \varepsilon_1 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 z_{n1} + \ldots + \beta_r z_{nr} + \varepsilon_n \end{aligned} \tag{7.2}$$

where the error terms are assumed to have the properties

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2, \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j. \tag{7.3}$$

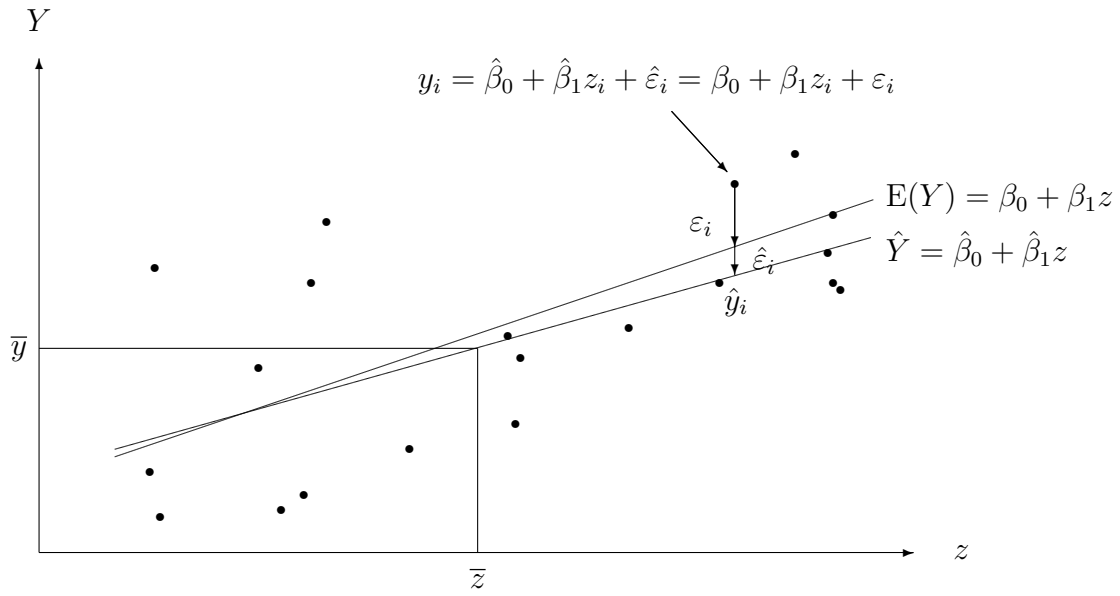

Figure 7.2: Notation for the linear regression.

Definition 7.1.1. In matrix notation, the classical linear regression model (7.2) becomes

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & z_{11} & \cdots & z_{1r} \\ \vdots & & & \vdots \\ 1 & z_{n1} & \cdots & z_{nr} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_r \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times(r+1))}{\mathbf{Z}} \underset{((r+1)\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

and the specifications in (7.3)

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \qquad \text{and} \qquad \mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I}. \tag{7.4}$$

$\boldsymbol{\beta}$ and σ^2 are unknown parameters and the design matrix $\mathbf{Z}$ has jth row $(z_{j0}, \ldots, z_{jr})$ with $z_{j0} = 1$.

7.2 Least Squares Estimation

One of the objectives of regression analysis is to develop an equation that will allow the investigator to predict the response for given values of the predictor variables. Thus, it is necessary to "fit" the model in Definition 7.1.1 to the observed y_i corresponding to the known values $1, z_{j1}, \ldots, z_{jr}$. That is, we must determine the values for the regression coefficients $\boldsymbol{\beta}$ and the error variance σ^2 consistent with the available data.

Let $\boldsymbol{b}$ be the trial values for $\boldsymbol{\beta}$. The method of least squares selects $\boldsymbol{b}$ so as to minimize the sum of squares of the differences:

$$\begin{aligned} S(\boldsymbol{b}) &= \sum_{j=1}^{n}(y_j - b_0 - b_1 z_{j1} - \ldots - b_r z_{jr})^2 \\ &= (\boldsymbol{y} - \mathbf{Z}\boldsymbol{b})'(\boldsymbol{y} - \mathbf{Z}\boldsymbol{b}). \end{aligned}$$

The coefficients $\boldsymbol{b}$ chosen by the least squares criterion are called least squares estimates of the regression parameter $\boldsymbol{\beta}$. They will henceforth be denoted by $\hat{\boldsymbol{\beta}}$ to emphasize their role as estimates of $\boldsymbol{\beta}$.

The coefficients $\hat{\boldsymbol{\beta}}$ are consistent with the data in the sense that they produce estimated (fitted) mean responses, $\hat{\beta}_0 + \hat{\beta}_1 z_{j1} + \ldots + \hat{\beta}_r z_{jr}$, the sum of whose squares of the differences from the observed y_i is as small as possible. The deviations

$$\hat{\varepsilon}_j = y_j - \hat{\beta}_0 - \hat{\beta}_1 z_{j1} - \ldots - \hat{\beta}_r z_{jr}, \qquad j = 1, \ldots, n \tag{7.5}$$

are called residuals. The vector of residuals

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}$$

contains the information about the remaining unknown parameter σ^2.

Proposition 7.2.1. *Let $\mathbf{Z}$ have full rank $r + 1 \leq n$. The least squares estimate of $\boldsymbol{\beta}$ is given by*

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{y}.$$

Let $\hat{\boldsymbol{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{H}\boldsymbol{y}$ denote the fitted values of $\boldsymbol{y}$, where $\mathbf{H} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'$ is called hat matrix. Then the residuals

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}$$

satisfy $\mathbf{Z}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ and $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = 0$. Also, the the residual sum of squares are

$$\sum_{j=1}^{n}(y_j - \hat{\beta}_0 - \hat{\beta}_1 z_{j1} - \ldots - \hat{\beta}_r z_{jr})^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\mathbf{Z}\hat{\boldsymbol{\beta}}.$$

Sum of Squares (SS) Decomposition

According to Proposition 7.2.1, we have $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = 0$ and therefore the total sum of squares $\boldsymbol{y}'\boldsymbol{y} = \sum_{j=1}^{n} y_j^2$ satisfies

$$\boldsymbol{y}'\boldsymbol{y} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}. \tag{7.6}$$

We find the basic decomposition of the sum of squares about the mean as

$$\boldsymbol{y}'\boldsymbol{y} - n\overline{y}^2 = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} - n(\overline{\hat{y}})^2 + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$$

or

$$\sum_{j=1}^{n}(y_j-\overline{y})^2 \quad = \quad \sum_{j=1}^{n}(\hat{y}_j-\overline{y})^2 \quad + \quad \sum_{j=1}^{n}\hat{\varepsilon}_j^2 \tag{7.7}$$

Total SS about mean Regression SS Residual (error) SS

The preceding sum of squares decomposition suggests that the quality of the models fit can be measured by the coefficient of determination

$$R^2 := 1 - \frac{\sum_{j=1}^{n}\hat{\varepsilon}_j^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2} = \frac{\sum_{j=1}^{n}(\hat{y}_j-\overline{y})^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}.$$

Remark. Interpretation: The quantity R^2 gives the proportion of the total variation in the y_j's explained by the predictor variables $z_1, \ldots, z_r$. Here $R^2 = 1$ if the fitted equation passes through all the data points, so that $\hat{\varepsilon}_j = 0$ for all j. At the other extreme, $R^2 = 0$ if $\hat{\beta}_0 = \overline{y}$ and $\hat{\beta}_1 = \ldots = \hat{\beta}_r = 0$. In this case, the predictor variables $z_1, \ldots, z_r$ have no influence on the response.

Sampling Properties of Classical Least Squares Estimators

The least squares estimator $\hat{\boldsymbol{\beta}}$ and the residuals $\hat{\boldsymbol{\varepsilon}}$ have the sampling properties detailed in the next proposition.

Proposition 7.2.2. *Under the general linear regression model in Definition 7.1.1, the least squares estimator* $\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{Y}$ *has*

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \qquad \textit{and} \qquad \mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}.$$

The residuals $\hat{\boldsymbol{\varepsilon}}$ *have the properties*

$$\mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0} \qquad \textit{and} \qquad \mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I}-\mathbf{H}).$$

Also $\mathrm{E}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) = (n-r-1)\sigma^2$, *so defining*

$$s^2 := \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-(r+1)} = \frac{\boldsymbol{Y}'(\mathbf{I}-\mathbf{H})\boldsymbol{Y}}{n-r-1}$$

we have

$$\mathrm{E}(s^2) = \sigma^2.$$

Moreover, $\hat{\boldsymbol{\beta}}$ *and* $\hat{\boldsymbol{\varepsilon}}$ *are uncorrelated.*

7.3 Inferences about the Regression Model

We describe inferential procedures based on the classical linear regression model in Definition 7.1.1 with the additional assumption that the errors $\boldsymbol{\varepsilon}$ have a normal distribution.

Inferences concerning the Regression Parameters

Before we can assess the importance of particular variables in the regression function

$$E(Y) = \beta_0 + \beta_1 z_1 + \ldots + \beta_r z_r$$

we must determine the sampling distributions of $\hat{\boldsymbol{\beta}}$ and the residual sum of squares, $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$. To do so, we shall assume that the errors $\boldsymbol{\varepsilon}$ have a normal distribution.

Proposition 7.3.1. *Let $\boldsymbol{Y} = \mathbf{Z}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where $\mathbf{Z}$ has full rank $r+1$ and $\boldsymbol{\varepsilon}$ is distributed as $N_n(\mathbf{0}, \sigma^2\mathbf{I})$. Then the maximum likelihood estimator of $\boldsymbol{\beta}$ is the same as the least squares estimator $\hat{\boldsymbol{\beta}}$. Moreover,*

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{Y} \quad \textit{is distributed as} \quad N_{r+1}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right)$$

and is distributed independently of the residuals $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}$. Further

$$n\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \quad \textit{is distributed as} \quad \sigma^2\chi^2_{n-r-1}$$

where $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2.

Remark. A confidence ellipsoid for $\boldsymbol{\beta}$ is easily constructed. It is expressed in terms of the estimated covariance matrix $s^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}$, where $s^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-r-1)$.

Proposition 7.3.2. *Let $\boldsymbol{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Z}$ has full rank $r+1$ and $\boldsymbol{\varepsilon}$ is distributed as $N_n(\mathbf{0}, \sigma^2\mathbf{I})$. Then a $(1-\alpha)$ confidence region for $\boldsymbol{\beta}$ is given by*

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{Z}'\mathbf{Z}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq (r+1)s^2F_{r+1,n-r-1}(\alpha)$$

where $F_{r+1,n-r-1}(\alpha)$ is the upper (100α)th percentile of an F-distribution with $r+1$ and $n-r-1$ d.f.

Also, simultaneous $(1-\alpha)$ confidence intervals for the β_i are given by

$$\hat{\beta}_i \pm \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}\sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}, \qquad i = 0, 1, \ldots, r,$$

where $\widehat{\mathrm{Var}}(\hat{\beta}_i)$ is the diagonal element of $s^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}$ corresponding to $\hat{\beta}_i$.

Remark. Practitioners often ignore the "simultaneous" confidence property of the interval estimates in Proposition 7.3.2. Instead, they replace $(r+1)F_{r+1,n-r-1}(\alpha)$ with the one-at-a-time t value $t_{n-r-1}(\alpha/2)$ and use the intervals

$$\hat{\beta}_i \pm t_{n-r-1}(\alpha/2)\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}.$$

7.3.1 Likelihood Ratio Tests for the Regression Parameters

Part of regression analysis is concerned with assessing the effects of particular predictor variables on the response variable. One null hypothesis of interest states that certain of the z_i's do not influence the response Y. These predictors will be labeled $z_{q+1}, z_{q+2}, \ldots, z_r$. So

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_r = 0 \qquad \text{or} \qquad H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0} \tag{7.8}$$

where $\boldsymbol{\beta}'_{(2)} = (\beta_{q+1}, \beta_{q+2}, \ldots, \beta_r)$.

Setting

$$\underset{(n\times(r+1))}{\mathbf{Z}} = \left(\underset{(n\times(q+1))}{\mathbf{Z}_1} \;\middle|\; \underset{(n\times(r-q))}{\mathbf{Z}_2} \right), \quad \underset{((r+1)\times 1)}{\boldsymbol{\beta}} = \left(\underset{((q+1)\times 1)}{\boldsymbol{\beta}_{(1)}} \;\middle|\; \underset{((r-q)\times 1)}{\boldsymbol{\beta}_{(2)}} \right)' \tag{7.9}$$

we can express the general linear model as

$$\boldsymbol{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = (\mathbf{Z}_1|\mathbf{Z}_2) \begin{pmatrix} \boldsymbol{\beta}_{(1)} \\ \boldsymbol{\beta}_{(2)} \end{pmatrix} + \boldsymbol{\varepsilon} = \mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}. \tag{7.10}$$

Proposition 7.3.3. *Let $\boldsymbol{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Z}$ has full rank $r+1$ and $\boldsymbol{\varepsilon}$ be distributed as $N_n(\mathbf{0}, \sigma^2\mathbf{I})$. Then the null hypothesis $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$ is rejected if*

$$\frac{(SS_{res}(\mathbf{Z}_1) - SS_{res}(\mathbf{Z}))/(r-q)}{s^2} > F_{r-q,n-r-1}(\alpha),$$

where

$$\textit{Extra SS} = SS_{res}(\mathbf{Z}_1) - SS_{res}(\mathbf{Z}) = (\boldsymbol{y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)})'(\boldsymbol{y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)}) - (\boldsymbol{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}),$$

$\hat{\boldsymbol{\beta}}_{(1)} = \left(\mathbf{Z}_1'\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1'\boldsymbol{y}$ *and $F_{r-q,n-r-1}(\alpha)$ is the upper (100α)th percentile of an F-distribution with $r-q$ and $n-r-1$ d.f.*

Remark. To test whether all coefficients in a subset are zero, fit the model with and without the terms corresponding to these coefficients. The improvement in the residual sum of squares is compared to the residual sum of squares for the full model via the F-ratio.

7.4 Inferences from the Estimated Regression Function

Once an investigator is satisfied with the fitted regression model, it can be used to solve two prediction problems. Let $\boldsymbol{z}_0' = (1, z_{01}, \ldots, z_{0r})$ be selected values for the predictor variables. Then $\boldsymbol{z_0}$ and $\hat{\boldsymbol{\beta}}$ can be used to estimate

- the regression function $\beta_0 + \beta_1 z_{01} + \ldots + \beta_r z_{0r}$ at $\boldsymbol{z_0}$ and
- the value of the response Y at $\boldsymbol{z_0}$.

Example (Weather Report, p. 1-6)**.** We have found a strong linear relationship between the altitude and the annual mean temperature based on different stations in Switzerland. Now we have two different kinds of prediction: In the first case we are interested in estimating the mean annual temperature for the population consisting on all points at a given altitude. In the second case we are interested to a single station and we want to predict the specific mean annual temperature of this station. The prediction is still determined from the fitted line. However, the standard error of the prediction is larger, because a single observation is more uncertain than the mean of the population distribution.

Estimating the Regression Function at z_0

Let Y_0 denote the value of the response when the predictor variables have values $\boldsymbol{z}_0' = (1, z_{01}, \ldots, z_{0r})$. According to the model in Definition 7.1.1, the expected value of Y_0 is

$$\mathrm{E}(Y_0|\boldsymbol{z}_0) = \beta_0 + \beta_1 z_{01} + \ldots + \beta_r z_{0r} = \boldsymbol{z}_0'\boldsymbol{\beta}.$$

Its least squares estimate is $\boldsymbol{z}_0'\hat{\boldsymbol{\beta}}$.

Proposition 7.4.1. *For the linear regression model in Definition 7.1.1,* $\boldsymbol{z}_0'\hat{\boldsymbol{\beta}}$ *is the unbiased linear estimator of* $\mathrm{E}(Y_0|\boldsymbol{z}_0)$ *with minimum variance,* $\mathrm{Var}(\boldsymbol{z}_0'\hat{\boldsymbol{\beta}}) = \boldsymbol{z}_0' \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \boldsymbol{z_0}\, \sigma^2$. *If the errors* $\boldsymbol{\varepsilon}$ *are normally distributed, then a* $(1-\alpha)$ *confidence interval for* $\mathrm{E}(Y_0|\boldsymbol{z}_0) = \boldsymbol{z}_0'\boldsymbol{\beta}$ *is given by*

$$\boldsymbol{z}_0'\hat{\boldsymbol{\beta}} \pm t_{n-r-1}(\alpha/2)\sqrt{(\boldsymbol{z}_0' \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \boldsymbol{z}_0)s^2}$$

where $t_{n-r-1}(\alpha/2)$ *is the upper* $100(\alpha/2)$*th percentile of a t-distribution with* $n-r-1$ *d.f.*

Forecasting a New Observation at z_0

Prediction of a new observation, such as Y_0, at $\boldsymbol{z}_0' = (1, z_{01}, \ldots, z_{0r})$ is more uncertain than estimating the expected value of Y_0. According to the model in Definition 7.1.1,

$$\begin{array}{ccccc} Y_0 & = & \boldsymbol{z}_0'\boldsymbol{\beta} & + & \varepsilon_0 \\ \text{new response } Y_0 & & \text{expected value of } Y_0 \text{ at } \boldsymbol{z}_0 & & \text{new error} \end{array}$$

where ε_0 is distributed as $N(0, \sigma^2)$ and is independent of $\boldsymbol{\varepsilon}$ and, hence, of $\hat{\boldsymbol{\beta}}$ and s^2. The errors $\boldsymbol{\varepsilon}$ influence the estimators $\hat{\boldsymbol{\beta}}$ and s^2 through the responses $\boldsymbol{Y}$, but ε_0 does not.

Proposition 7.4.2. *Given the linear regression model in Definition 7.1.1, a new observation Y_0 has the unbiased predictor*

$$\boldsymbol{z}_0'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 z_{01} + \ldots + \hat{\beta}_r z_{0r}.$$

The variance of the forecast error $Y_0 - \boldsymbol{z}_0'\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(Y_0 - \boldsymbol{z}_0'\hat{\boldsymbol{\beta}}) = \sigma^2 \left(1 + \boldsymbol{z}_0' \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \boldsymbol{z}_0\right).$$

When the errors $\boldsymbol{\varepsilon}$ have a normal distribution, a $(1-\alpha)$ prediction interval for Y_0 is given by

$$\boldsymbol{z}_0'\hat{\boldsymbol{\beta}} \pm t_{n-r-1}(\alpha/2)\sqrt{s^2\left(1 + \boldsymbol{z}_0' \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \boldsymbol{z}_0\right)},$$

where $t_{n-r-1}(\alpha/2)$ is the upper $100(\alpha/2)$th percentile of a t-distribution with $n-r-1$ d.f.

Remark. Prediction is an important goal of most modeling efforts. To reduce the chance of overfitting the data the K-fold cross-validation procedure is recommended:

1. Divide the data into K subsets of approximately equal size. This is usually done randomly. If there is concern about ensuring that each subset is similar according to some criterion, a statistical design can be used to partition the data. The number of model parameters is a factor in selecting K because the desire is to maintain as many degrees as possible for estimating the error. A split that results in fewer than 30 observations in the training set is not recommended.

2. Each subset is removed from the sample data and a prediction made using the remaining data. Thus, there are K models. Corresponding to each model is a data set not used in that estimation.

3. Estimate the response variable in the corresponding omitted data set for each model.

4. The prediction error is an average of results made from each of the k "omitted subsets."

A special case of this procedure is the leave-one-out cross validation, which uses n training sets of size $n-1$.

7.5 Model Checking and other Aspects of Regression

Model evaluation is critical. It is necessary to know if the model is satisfactory for the job asked of it. Every statistical model and estimation technique carries with it a set

of assumptions. In linear regression, these assumptions include normally distributed independent errors with constant variance. When assumptions fail to hold, statistical tests of significance may be invalid and model estimates can be seriously biased. A challenge in model evaluation is that all assumptions are not of equal importance. With experience one learns that some assumptions may be relaxed without doing significant harm to model results, whereas others can be critical. Methods to evaluate a model include an examination of statistics produced by the model-fitting procedure. This investigation involves formal hypothesis testing and examination of descriptive statistics and graphs. Graphs are essential. One needs to view a variety of model evaluation tools because no single statistic or graph gives complete information on the quality of a model.

Does the Model fit?

Assuming that the model is "correct", we have used the estimated regression function to make inferences. Of course, it is imperative to examine the adequacy of the model before the estimated function becomes a permanent part of the decision-making apparatus.

All the sample information on lack of fit is contained in the residuals

$$\begin{aligned}
\hat{\varepsilon}_1 &= y_1 - \hat{\beta}_0 - \hat{\beta}_1 z_{11} - \ldots - \hat{\beta}_r z_{1r} \\
\hat{\varepsilon}_2 &= y_2 - \hat{\beta}_0 - \hat{\beta}_1 z_{21} - \ldots - \hat{\beta}_r z_{2r} \\
&\vdots \\
\hat{\varepsilon}_n &= y_n - \hat{\beta}_0 - \hat{\beta}_1 z_{n1} - \ldots - \hat{\beta}_r z_{nr}
\end{aligned}$$

or

$$\hat{\boldsymbol{\varepsilon}} = \left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\right)\boldsymbol{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}.$$

If the model is valid, each residual $\hat{\varepsilon}_j$ is an estimate of the error ε_j, which is assumed to be a normal random variable with mean zero and variance σ^2. Although the residuals $\hat{\boldsymbol{\varepsilon}}$ have expected value $\mathbf{0}$, their covariance matrix $\sigma^2(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}') = \sigma^2(\mathbf{I} - \mathbf{H})$ is not diagonal. Residuals have unequal variances and nonzero correlations. Fortunately, the correlations are often small and the variances are nearly equal.

Residuals should be plotted in various ways to detect possible anomalies. For general diagnostic purposes, the following are useful graphs:

1. Plot the residuals $\hat{\varepsilon}_j$ against the predicted values $\hat{y}_i$. Departures from the assumptions of the model are typically indicated by two types of phenomena:

   - A dependence of the residuals on the predicted value (see Figure 7.3 (a)). The numerical calculations are incorrect, or a β_0 term has been omitted from the model.
   - The variance is not constant. The pattern of residuals may be funnel shaped, as in Figure 7.3(b), so that there is large variability for large $\hat{y}$ and small

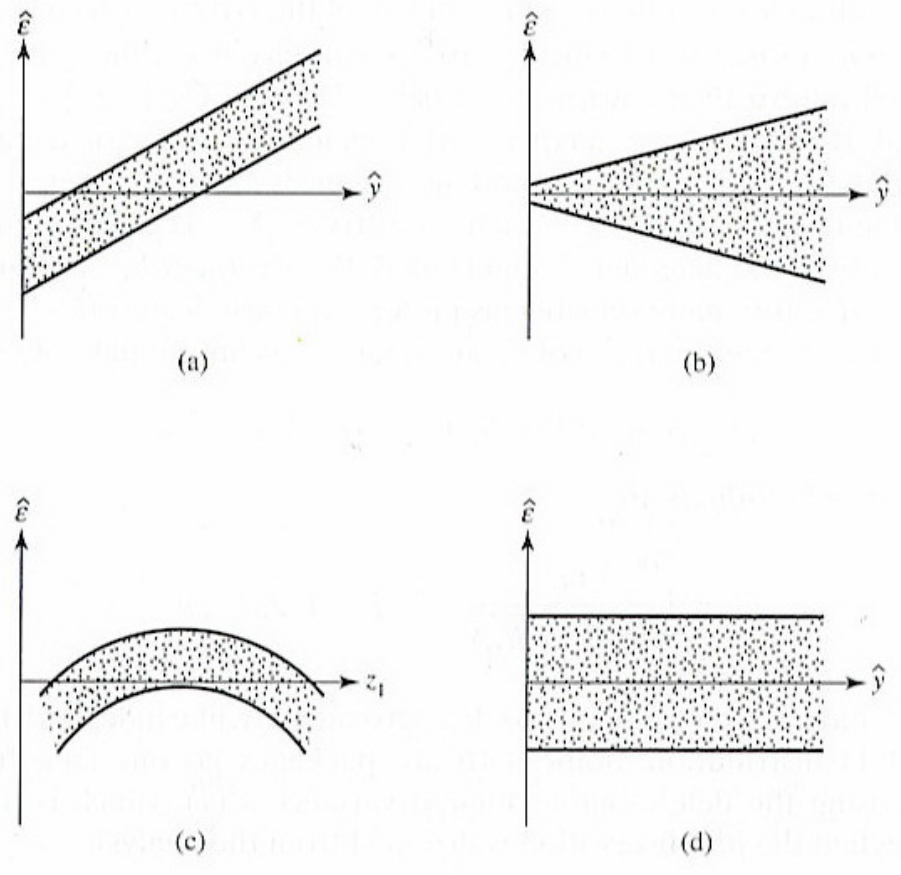

Figure 7.3: Residual plots. Source: Johnson and Wichern (2007).

variability for small $\hat{y}$. If this is the case, the variance of the error is not constant, and transformations or a weighted least squares approach (or both) are required. In Figure 7.3(d), the residuals form a horizontal band. This is ideal and indicates equal variances and no dependence on $\hat{y}$.

2. Plot the residuals $\hat{\varepsilon}_j$ against a predictor variable, such as z_1, or products of predictor variables, such as z_1^2 or $z_1 z_2$. A systematic pattern in these plots suggests the need for more terms in the model (see Figure 7.3(c)).

3. QQ-plots and histograms: Do the errors appear to be normally distributed? The QQ-plots and histograms help to detect the presence of unusual observations or severe departures from normality that may require special attention in the analysis.

4. Plot the residuals versus time: the assumption of independence is crucial, but hard to check. If the data are naturally chronological, a plot of the residuals versus time may reveal a systematic pattern. For instance, residuals that increase over time indicate a strong positive dependence.

Example (MeteoSchweiz 2007, p. 1-6)**.** Consider the data set in Chapter 1.3.4 and the Figure 7.1. How good is the fitted linear regression model? Figure 7.4 shows several plots to answer this question:

- Plot of the response against the fitted values: gives a good idea of how well the model has captured the broad outlines of the data.
- Plot of the residuals against the fitted values: often reveals unexplained structure left in the residuals, which in a strong model should appear as nothing but noise.
- Square root of absolute residuals against fitted values: useful in identifying outliers and visualizing structure in the residuals.
- Normal quantile plot of residuals: provides a visual test of the assumption that the model's errors are normally distributed.
- Residual-Fit spread plot: compares the spread of the fitted values with the spread of the residuals. Since the model is an attempt to explain the variation in the data, you hope that the spread in the fitted values is much greater than that in the residuals.
- Cooks distance plot: measure of the influence of individual observations on the regression coefficients.

Example (Basel, p. 1-6)**.** The report (see Table 7.1) and the residual plots (see Figure 7.5) of the multiple linear regression model confirms that the multiple linear regression model is appropriate for modelling the cloudiness as a function of the annual sunshine duration and annual precipitation.

Remark. When a multiple regression model is constructed, variables based on theory, results of exploratory data analysis, and intuition are often included. It is common that not all of these variables will be statistically significant. Before a model is put in service, should these nonsignificant variables be eliminated? Here are some guidelines:

- If the sample size is large, omit variables that are not statistically significant.
- If the sample size is small, err on the side of retaining variables.
- If theory suggests that a variable should be present, err on the side of retaining it.

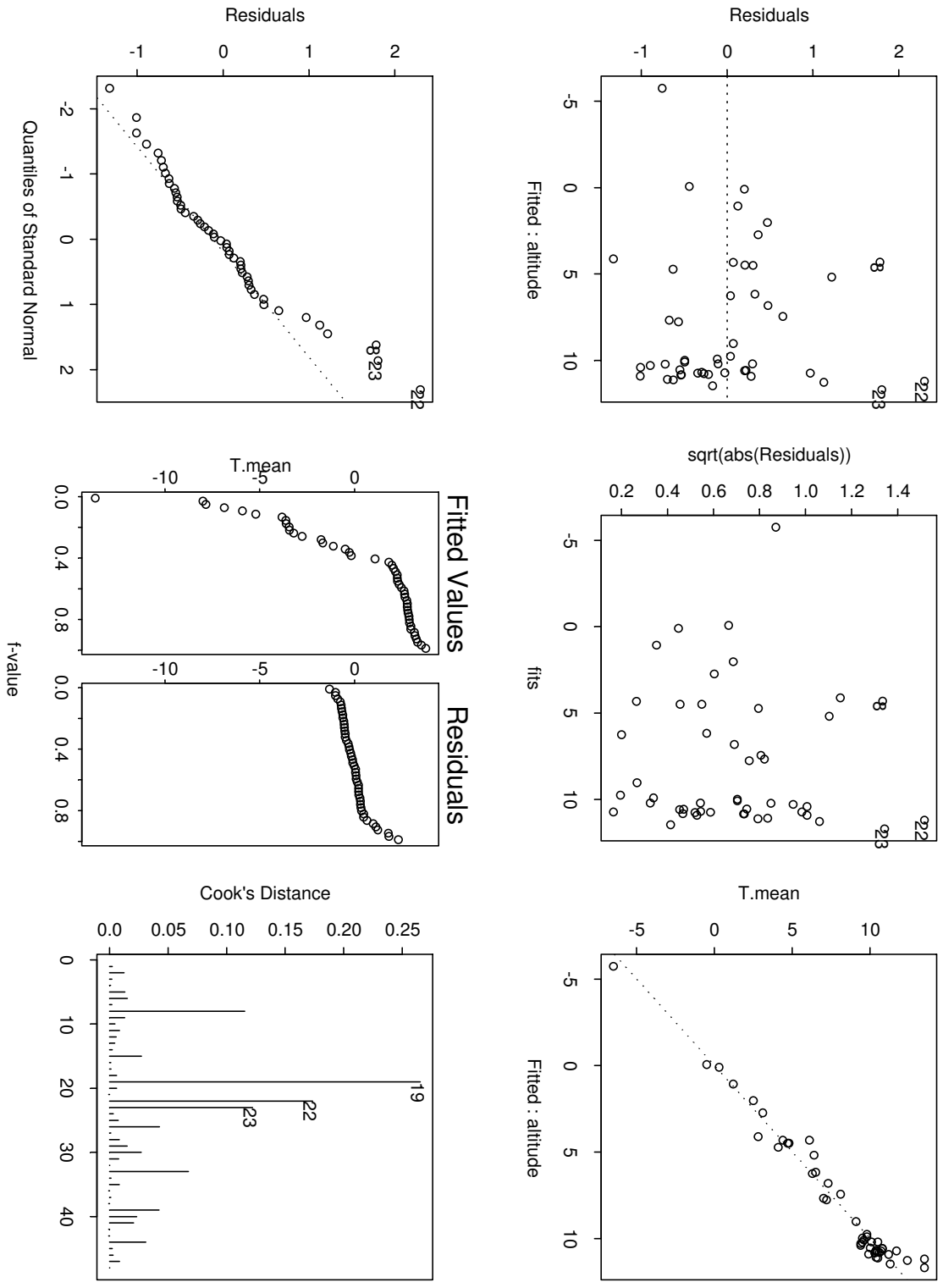

Figure 7.4: Residual plots of the annual mean temperatures as a function of the altitude of the different Swiss stations. Data set: Weather Report, p. 1-6.

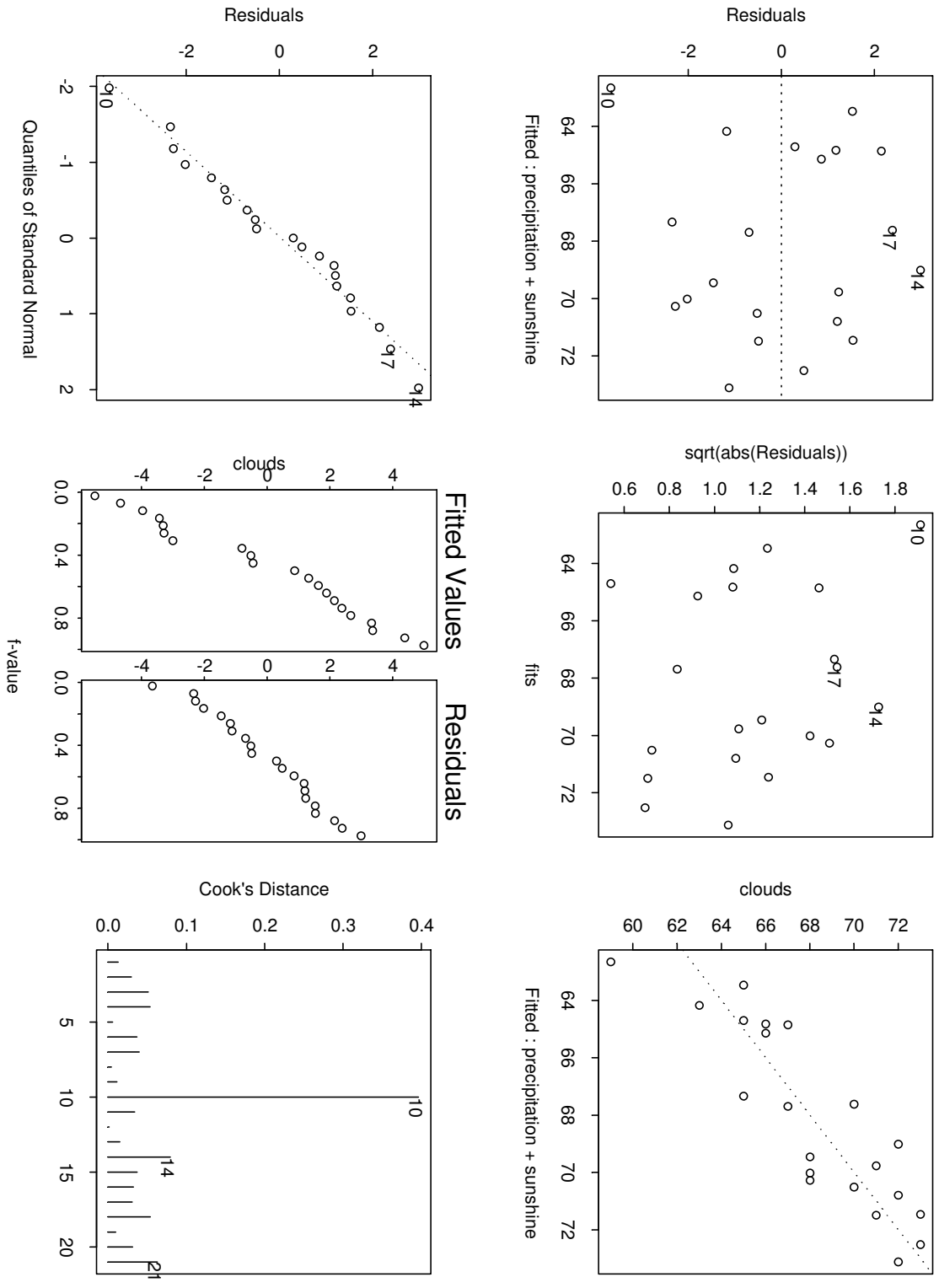

Figure 7.5: Residual plots of the annual cloudiness for Basel as a function of the annual sunshine duration and annual precipitation. Data set: Basel, p. 1-6.

Table 7.1: Summary of the multiple linear regression of the annual cloudiness for Basel as a function of the annual sunshine duration and annual precipitation. Data set: Basel, p. 1-6.

```
Call: lm(formula = clouds ~ precipitation + sunshine, data = daten, na.action = na.omit)
Residuals:
    Min     1Q Median    3Q   Max
 -3.658 -1.177  0.292 1.226 2.983

Coefficients:
                Value Std. Error  t value Pr(>|t|)
  (Intercept)  87.2684   6.5377    13.3484   0.0000
precipitation   0.0096   0.0036     2.6632   0.0158
     sunshine  -0.0167   0.0030    -5.6428   0.0000

Residual standard error: 1.872 on 18 degrees of freedom
Multiple R-Squared: 0.7633      Adjusted R-squared: 0.737
F-statistic: 29.03 on 2 and 18 degrees of freedom, the p-value is 2.33e-006

-------------------------------------------------------------------------------

       Shapiro-Wilk Normality Test

W = 0.9722, p-value = 0.7809

-------------------------------------------------------------------------------

       One sample Kolmogorov-Smirnov Test of Composite Normality

ks = 0.1255, p-value = 0.5
alternative hypothesis: True cdf is not the normal distn. with estimated parameters
sample estimates:
      mean of x standard deviation of x
 -2.114711e-017                1.77608

-------------------------------------------------------------------------------
```