6 Multivariate Regression

6.1 The Model

a In multiple linear regression, we study the relationship between several input variables or regressors and a continuous target variable. Here, **several target variables** are considered simultaneously.

▷ **Fossil Example.** From fossils that are found in different layers of the sea bed, we intend to determine environmental conditions (temperature, salinity) of the corresponding time period.

To that end, measurements were made on "coccoliths" of the type *Gephyrocapsa* in the top deposits at various points in the oceans and set in relation to model environmental conditions. In Fig. 6.1.a are shown the relationships between the individual environmental variables and the shape features of the fossils.

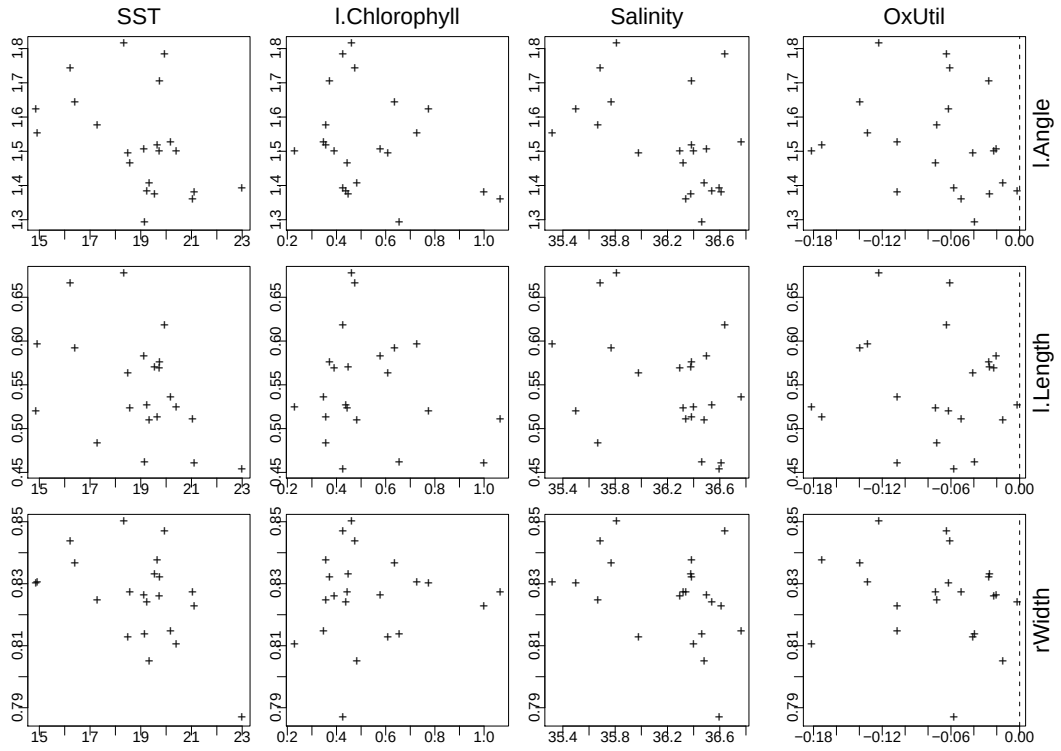

Figure 6.1.a: Environmental Variables and Shape Features in the Fossil Example

The corresponding model will be used later to find out environmental conditions for earlier time periods based on coccoliths from the corresponding deeper layers. For this conclusion we have to assume that these relationships have not changed since then. For details, see Bollmann, Henderiks and Brabec (2002). ◁

b **Model.** The model in linear regression with a single target variable was $Y_i = \beta_0 + \sum_k \beta_k x_i^{(k)} + E_i$. If now we study the relationship of multiple target variables $Y^{(j)}$, $j = 1, ..., m$, with the input variables (or explanatory variables) $X^{(k)}$, we can begin by setting up each model as

$$Y_i^{(j)} = \beta_0^{(j)} + \sum\nolimits_k \beta_k^{(j)} x_i^{(k)} + E_i^{(j)}$$

This should again be summarized with matrices,

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{E}\ .$$

We get the individual models in matrix notation if we choose the jth column of each of $\boldsymbol{Y}$, $\boldsymbol{\beta}$ and $\boldsymbol{E}$: $\underline{Y}^{(j)} = \boldsymbol{X}\,\underline{\beta}^{(j)} + \underline{E}^{(j)}$. As in the matrix notation for univariate regression, i.e. regression with an individual variable of interest, the **intercept** β_0 no longer appears in the matrix form; it is taken into account by including a column of ones in the design matrix $\boldsymbol{X}$.

c **Input variables, regressors, and terms.** In general, with regression models we study a relationship between target variables and input variables. The input variables are often called explanatory variables, which is justified if there is a cause-effect relationship. Since regression is also reasonable if this can not be postulated, the neutral expression **input variables instead of explanatory variables** should be used. The also common name "independent variables" is avoided, since the adjective "independent " is only confusing: The X variables do not have to be statistically independent from each other. Some may even be functions of others, as long as linear functions are avoided.

Input values often do not go into the regression model in their original form, but instead are initially transformed – individually, for example with a log transformation, or together, for example if one is expressed as a percentage of another. These transformed values, that enter into the model as X variables, are called **regressors**. The target variable can be handled analogously. The transformed target variables can then be called "regressands". This distinction is less important for the target variable: In residual analysis, the untransformed input variables play a role, but the untransformed target variables do not. In addition, since "regressand" sounds too similar to "regressor", we stick with the expression "target variable", that will also apply for transformed target variables.

In determining the model, the concept of "**term**" additionally comes into play. The dummy variables that correspond to a factor (see below, 6.1.f) or an interaction between variables each form a term. In choosing the model, each term is included or omitted.

d **Random Deviations.** The assumptions on the distribution the random deviations $E_i^{(j)}$ consist of the obvious generalization of the assumptions in the case of an individual target variable. We let $\underline{E}_i$ be the ith row of $\boldsymbol{E}$, i.e. the vector of the random deviations

of all target variables for the observation i. The assumptions are:

- Expected value $\mathcal{E}\langle \underline{E}_i \rangle = \underline{0}$. This definition is necessary for the identifiability of $\underline{\beta}$ and also indicates that the (linear) regression function is correct.
- The random deviations $E_i^{(j)}$ have variances σ_j^2, which are generally different for all of the target variables j. The $E_i^{(j)}$ for different target variables can be correlated. Both variances and covariances are characterized by the covariance matrix $\mathrm{var}\langle \underline{E}_i \rangle = \boldsymbol{\Sigma}$, which we assume to be equal for observations i.
- The random deviations of the *different observations* are independent (or at least uncorrelated), $\mathcal{E}\langle \underline{E}_h \underline{E}_i^T \rangle = \mathbf{0}$ if $h \neq i$.
- The random deviations are jointly normally distributed.

We can summarize all that with

$$\underline{E}_i \sim \mathcal{N}_m\langle \underline{0}, \boldsymbol{\Sigma} \rangle \,, \quad \text{independent} \,.$$

The model with the joint distribution of the error terms is the model of **multivariate regression**. It is also a *multiple regression*, in that it includes multiple regressors $X^{(j)}$.

e To begin we have just formally written the models for the individual target variables in a simple way in a single matrix formula. Through the assumption of a joint normal distribution of the error terms they now have a relationship in content.

The fact that the design matrix $\boldsymbol{X}$ is the same for all target variables does not necessarily specify a substantive relationship: If the coefficient matrix $\boldsymbol{\beta}$ contains in each row only one element different from zero, then the target variables react to completely different regressors, which we have only formally condensed into one matrix.

f **Analysis of Variance.** As in univariate regression, the input variables can also be factors (nominal or categorial variables), which enter as "dummy variables" in the design matrix $\boldsymbol{X}$. Multivariate analysis of variance with fixed effects (MANOVA) can thus be treated as a special case of regression. As with a single target variable, there are some additional interesting methodological aspects, which will not be treated here.

6.2 Estimation and Tests

a **Estimation of the Coefficients.** The columns of $\boldsymbol{\beta}$ can be estimated separately with least squares, so each with a (univariate) multiple regression calculation. However, this can be summarized and written as

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

The fitted values and the residuals are as defined in the univariate case and are summarized in the matrix $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\,\widehat{\boldsymbol{\beta}}$ and the residual matrix $\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$.

The **estimation of the covariance matrix** of the random deviations occurs via the empirical covariance matrix of the residuals, taking the number $n - p$ of degrees of

freedom into account,

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-p}\boldsymbol{R}^T\boldsymbol{R}$$

b **Distribution of the Estimated Coefficients.** As expected, the coefficients are unbiased and normally distributed. The covariance matrix of the estimated coefficients will have to be calculated somehow; we put our confidence in the programs!

* Would you like to know about it a bit more specifically? Here we encounter a difficulty in the notation: The estimated coefficients form a random matrix. We need not only the distribution of each individual element of this matrix, but also the joint distribution of the elements. In particular we are interested in the covariances between the elements. They are $\operatorname{cov}\langle\widehat{\beta}_h^{(j)}, \widehat{\beta}_k^{(\ell)}\rangle = ((\boldsymbol{X}^T\boldsymbol{X})^{-1})_{hk}\boldsymbol{\Sigma}_{j\ell}$. This can not be written directly as a matrix, since it varies over four indices! Apart from that, the derivation is not difficult: We set $\boldsymbol{C} = \boldsymbol{X}^T\boldsymbol{X}$ and calculate

$$\begin{aligned}\operatorname{cov}\langle\underline{\widehat{\beta}}^{(j)}, \underline{\widehat{\beta}}^{(k)}\rangle &= \operatorname{cov}\langle\boldsymbol{C}^{-1}\boldsymbol{X}^T\underline{Y}^{(j)}, \boldsymbol{C}^{-1}\boldsymbol{X}^T\underline{Y}^{(k)}\rangle = \boldsymbol{C}^{-1}\boldsymbol{X}^T\operatorname{cov}\langle\underline{Y}^{(j)}, \underline{Y}^{(k)}\rangle(\boldsymbol{C}^{-1}\boldsymbol{X}^T)^T \\ &= \boldsymbol{C}^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{j\ell}\boldsymbol{X}(\boldsymbol{C}^{-1})^T = \boldsymbol{\Sigma}_{j\ell}\boldsymbol{C}^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{C}^{-1} = \boldsymbol{\Sigma}_{j\ell}\boldsymbol{C}^{-1}\end{aligned}$$

c ▷ **In the fossil example** the results for the individual target variables are summarized in Table 6.2.c. They are not encouraging; the overall test does not give a significant result for any target variable! ◁

	l.Angle		l.Length		rWidth	
	coef	p-value	coef	p-value	coef	p-value
(Intercept)	447.787	0.347	1.6590	0.471	0.59265	0.283
SST	-0.721	0.800	-0.0102	0.463	-0.00493	0.147
l.Chlorophyll	-19.202	0.155	-0.0765	0.238	0.00208	0.890
Salinity	-10.756	0.452	-0.0242	0.726	0.00888	0.590
OxUtil	-23.770	0.662	0.0433	0.869	-0.04368	0.489
R^2	0.285	0.260	0.2571	0.198	0.24889	0.255

Table 6.2.c: Regression coefficients and coefficients of determination with p values for the individual shape variables as target variables in the fossil example

d **Joint Tests.** For each individual target variable $Y^{(j)}$ we know how we test whether it depends on the input variables. From the joint analysis also comes the **joint null hypothesis, that none of the target variables depend on a regressor** $X^{(k)}$, so that $\beta_k^{(j)} = 0$ for all j or, more comprehensively, that no relationship exists between any target variable and any regressor, so that all $\beta_k^{(j)} = 0$. In between lie, as in univariate regression, the comparisons of hierarchically nested models.

The most obvious way to test such a hypothesis is the application of the corresponding likelihood ratio test. This test statistic is called Wilks' Λ (capital greek lamda).

In univariate regression, the test statistic of the F-test essentially creates a relationship between the "between group sum of squares" and the "within group sum of squares". In the multivariate case both values become "sum of squares and cross products" *matrices*, denoted as $\boldsymbol{B}$ and $\boldsymbol{W}$. The decisive factor in the univariate case is the value of their

ratio. Useful test statistics here are, in analogy, functions of the eigenvalues λ_k of $\boldsymbol{W}^{-1}\boldsymbol{B}$,

- Wilks: $\prod_\ell 1/(1+\lambda_\ell)$
- Pillai: $\sum_\ell \lambda_\ell/(1+\lambda_\ell)$
- Lawley-Hotelling: $\sum_\ell \lambda_\ell$
- Roy (union-intersection): λ_1 (resp. $\lambda_1/(1+\lambda_1)$)

In the multivariate case there are thus **several common tests** which convert to the usual F-test (or t-test) for the case of an *individual* target variable. If the influence of an individual continuous variable is tested, all these tests give the same result (* since $\boldsymbol{B}$ has only one degree of freedom and thus is only the first eigenvalue λ_1 different from 0).

e ▷ **I**n the **fossil example** the global test, which checks the null hypothesis that no relationship exists between the shape and environment variables, shows no significance! The case thus appears hopeless. – Closer analysis yielded the possibility of identifying three groups from the distribution of the angle and the length and to introduce the proportions of these groups in the samples taken at each location as a new target variable. Table 6.2.e shows in the row labeled as ".total." that the environmental variables have a significant influence on the fractions of these groups. The other test statistics lead to p values of 0.0388 (Pillai), 0.0163 (Hotelling-Lawley) and 0.00381 (Roy).

	Df	Wilks	approx F	num Df	den Df	p value
SST	1	0.564	5.405	2	14	0.0182
l.Chlorophyll	1	0.886	0.905	2	14	0.4271
Salinity	1	0.847	1.267	2	14	0.3122
OxUtil	1	0.890	0.863	2	14	0.4431
.total.	4	0.417	1.922	8	28	0.0961
Residuals	15					

Table 6.2.e: Overall tests for the influence of the individual regressors on the group proportions, as well as for all regressors together in the fossil example

According to the table, the variable SST has a significant influence. However, as in univariate regression, it is conceivable that the higher p values of the other variables is caused by collinearity. To check this, we calculate as before the model without the least significant variable, so without `OxUtil`. The variable `Salinity` then obtains a p value of 0.177 – still insignificant. ◁

f **Relevance of Multivariate Regresssion.** For the interpretation, mostly the coefficients $\boldsymbol{\beta}$ are of interest.
Estimate and confidence interval for a $\beta_k^{(j)}$ from multivariate regression are identical to those from multiple regression of $Y^{(j)}$ on the regressors – the other target variables have no influence.
If we run a program for multivariate regression, we mainly get what we would for m runs of a program for multiple regression. Additionally we get:

- the covariance matrix of the random error. The correlation between the random deviations $E^{(j)}$ and $E^{(\ell)}$ of the linear regressions of $Y^{(j)}$ and $Y^{(\ell)}$ on the regressors $X^{(k)}$, $k = 1, ..., p$ also called **partial correlations** between $Y^{(j)}$ and $Y^{(\ell)}$, given the X variables, see below (Section **??**).
- *joint* tests for the question whether the target variables taken together show relationships with certain regressors.

g **Residual Analysis.** Residual analysis for checking the model assumptions is, as for all regression models, an indispensable part of a serious data analysis. To start, the regressions for all target variables should be checked with the known methods.

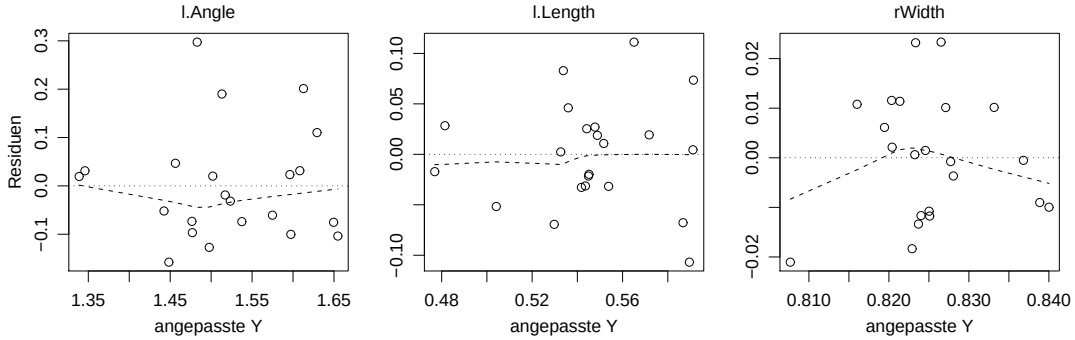

Figure 6.2.g (i): Tukey-Anscombe plots for the fossil example

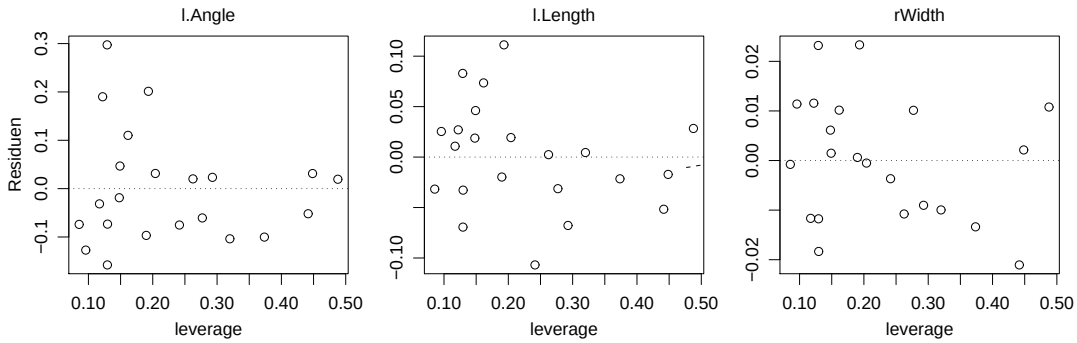

Figure 6.2.g (ii): Scatter plot of the residuals against leverage values for the fossil example

Fig. 6.2.g (i) shows for the example the compilation of the Tukey-Anscombe plots which serve for the overall model checking and especially for hints about the usefulness of a

transformation of the target variable. From the scatter plots of the residuals versus the *leverages,* Fig. 6.2.g (ii), we might recognize influential observations. The scatterplots of the residuals versus the input variables (Fig. 6.2.g (iii)) primarily give hints about nonlinearities in the input variables. – Aside from a skewed error distribution for `l.Angle` the example shows barely anything serious to consider.

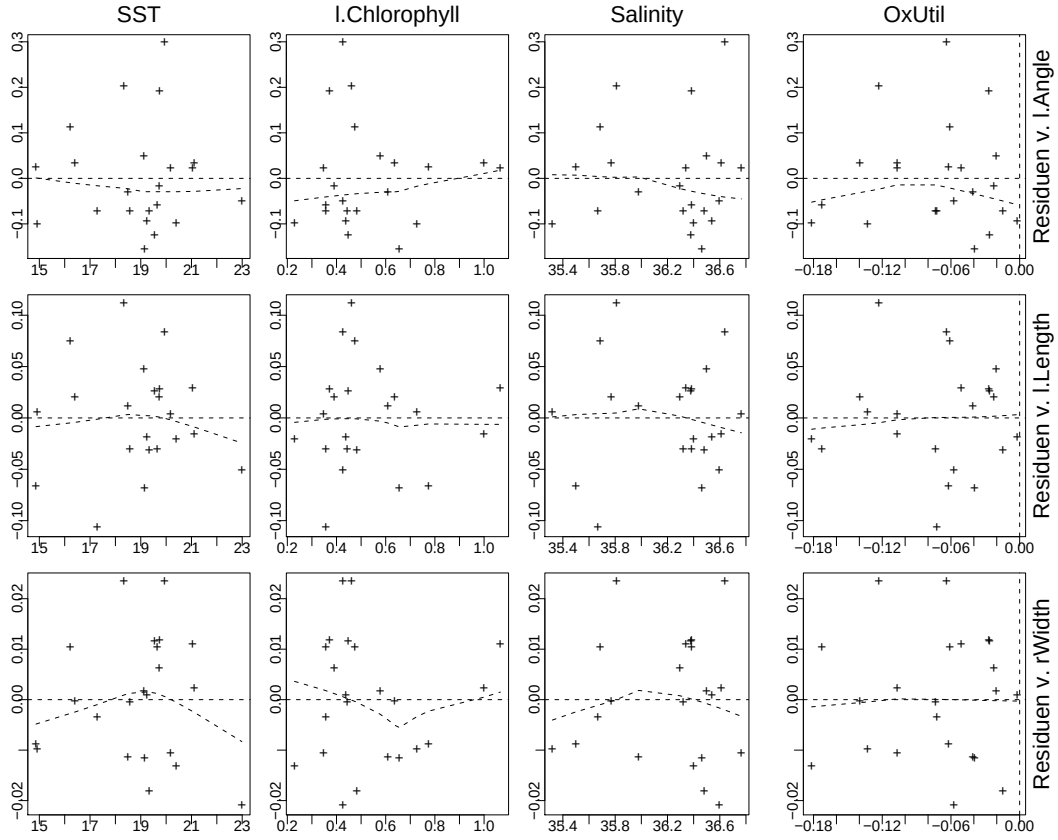

Figure 6.2.g (iii): Scatterplots of the residuals against the input variables for the fossil example

h To supplement the residual analysis for each target variable, it is worthwhile to consider a **scatterplot matrix of the residuals** (Fig. 6.2.h (i)). In the example, (at least) two extreme points with large residuals stand out. If more observations were available, we could repeat the calculation without these two points.

From the residuals and their estimated covariance matrix we get in the earlier discussed way (3.2.o) the **Mahalanobis values**, which with the help of a quantile-quantile diagram we can compare with the corresponding distribution, the "Root chi-squared distribution" (Fig. 6.2.h (ii)). With this we check an aspect of the assumption of the multivariate normal distribution of the error $\underline{E}_i$.

i* **Prediction.** For a given set $\underline{x}_0$ of values of the regressors, a new observation $\underline{Y}_0$ should be created. What can we say about the distribution of $\underline{Y}_0$ in advance?

For known parameters the problem is trivial: The total distribution of the new observation is given via the regression model 6.1.b. The best prediction is the expected value of $\underline{Y}_0$,

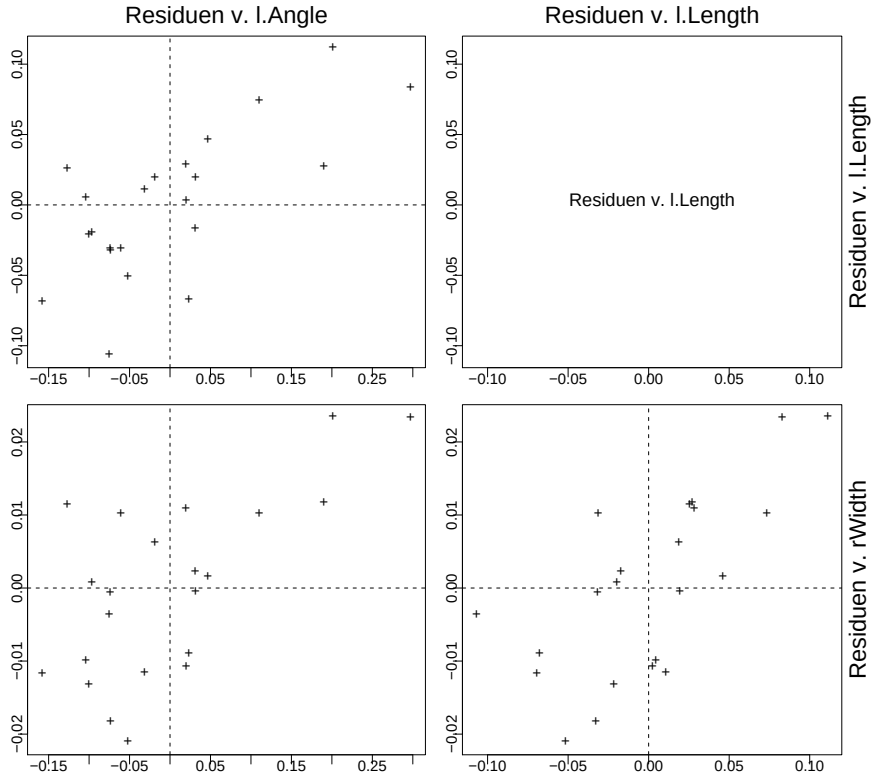

Figure 6.2.h (i): Scatterplot matrix of the residuals for the fossil example

$\mathcal{E}\langle \underline{Y}_0^T \rangle = \underline{x}_0^T \boldsymbol{\beta}$ (written transposed).

In reality, the relationship of $\underline{x}$ and $\underline{Y}$ must be estimated from the "training data" $\boldsymbol{X}$, $\boldsymbol{Y}$. This leads to the estimation $\widehat{\boldsymbol{\beta}}$, which we substitute in place of $\boldsymbol{\beta}$. The best prediction is then $\widehat{\underline{Y}}_0^T = \underline{x}_0^T \widehat{\boldsymbol{\beta}}$. The distribution of the prediction can be, as in the one dimensional case, derived from the distribution of the estimated coefficients.

j* **Prediction Interval.** In the univariate case, the prediction interval should contain the future *observation* with a predetermined probability (while a confidence interval contains a *parameter* with such a probability). Analogously to that case, the sum of the covariance matrices for the estimated expected value and for the random error, $\mathrm{var}\langle \widehat{\underline{Y}} \rangle + \widehat{\boldsymbol{\Sigma}}$, determine the size and shape of the desired region, which turns out to be an ellipsoid.

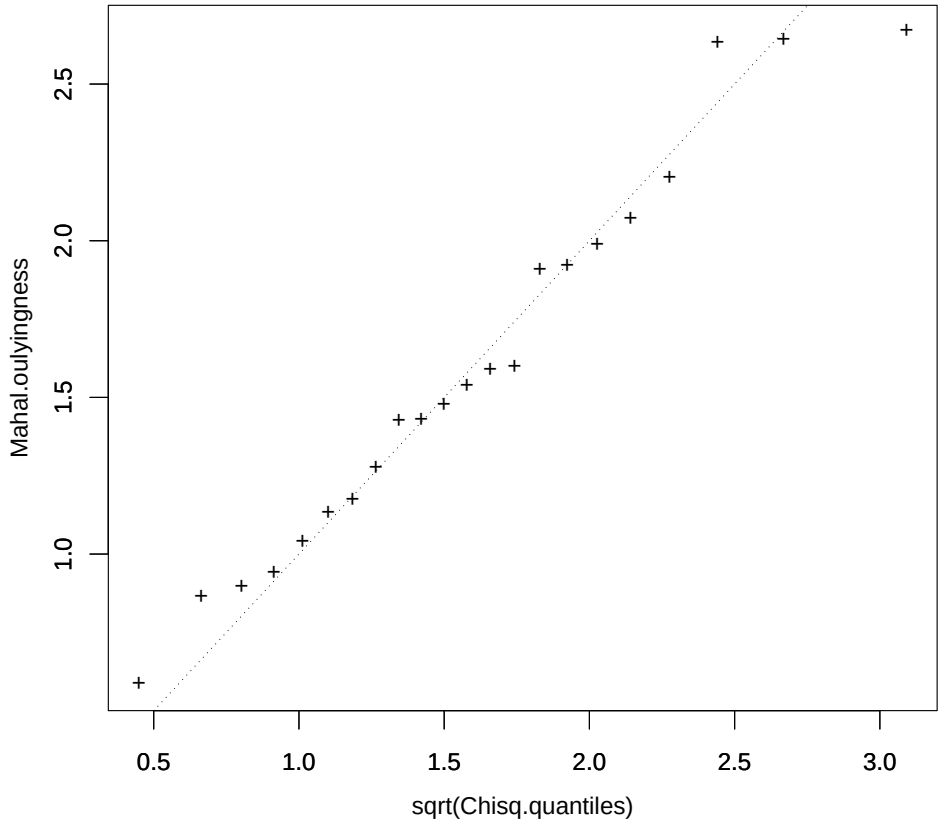

Figure 6.2.h (ii): Q-Q diagram of the lengths of the multivariate standardized residuals for the fossil example

6.3 Partial Correlation and Conditional Independence

a The purpose of many scientific studies is to find and study causal relationships. However, one of the basic messages of applied statistics says that it is impossible to infer such relationships on the basis of correlations in observational studies. A correlation constitutes a symmetric relationship, and either variable may be a cause for the other, or both of them may be caused by third variables. Causal inference therefore needs (well-designed) experiments.

* Note that this is even a philosophical insight: The word "cause" only makes philosophical sense in connection with human manipulation of processes. Natural processes will evolve, and it is not possible to say what is the cause of what.

b Nevertheless, hints can be obtained, most importantly on the **lack of a relationship**: If two variables are *independent*, then it is plausible to conclude that none of them is the cause of the other (in the sense that, if it is manipulated, the other variable will change). As a simplification, one searches for pairs of *uncorrelated* rather than independent variables.

c **Multiple Regression.** It is usually discussed in a course on linear regression that indirect relationships often lead to simple correlations between two variables: Families have more children in rural areas, and storks are also more abundant – but the resulting correlation between storks and babies is usually not interpreted in a causal way.

Such indirect relationships are excluded by including all potential "links" between two variables in a multiple regression model. In fact, this is the major advantage of modelling regression relationships with a multiple model instead of many simple regression models (or describing the relationships by simple correlations).

d **Partial correlation.** In fact, the basic idea is as follows: We are interested in the (potentially causal) relationship between $Y^{(1)}$ and $Y^{(2)}$, excluding indirect effects from $X^{(1)}, ..., X^{(p)}$. The exclusion of indirect effects is obtained by exploiting all the information that $X^{(1)}, ..., X^{(p)}$ contain about both $Y^{(1)}$ and $Y^{(2)}$. Again simplifying to linear relationships, this is achieved by "predicting" $Y^{(1)}$ and $Y^{(2)}$ from $X^{(1)}, ..., X^{(p)}$ using a linear regression model and then studying what is *not* "explained" by these models – the residuals of the two target variables $Y^{(1)}$ and $Y^{(2)}$. These are exactly the residuals from the *multivariate* regression of $Y^{(1)}$ and $Y^{(2)}$ (and possibly further Y's) on $X^{(1)}, ..., X^{(p)}$, which we have just met in the discussion about residual analysis (6.2.h).

e* More generally, we would be interested in the **joint conditional distribution** of $Y^{(1)}$ and $Y^{(2)}$, given $X^{(1)}, ..., X^{(p)}$. The partial correlation is an aspect of it. If the X's and Y's follow jointly a multivariate normal distribution, then that conditional distribution is bivariate normal, and the partial correlation characterizes the relationship fully.

f Given a set of variables, there is a huge number of ways in which partial correlations may be determined. The most "extreme" of these correlations are those calculated for each pair of variables, conditional on all the other variables under study. If such a partial correlation is high (significant and relevant), then this is a hint to a causal relationship in either direction. On the other hand, if it is negligible, this points to a lack of direct influence among this pair of variables.

g **Graphs.** There is a nice and intuitive way of visualizing the relevant partial correlations in a set of variables, consisting of a **graph** in the mathematical sense of the word: The variables are represented as points ("nodes"), between which lines ("edges") are drawn: A line connects two points if a partial correlation exists between the two variables. The positions of the points are not determined. They should be chosen such that the lines intersect as little as possible and are short.

The graph showing all the extreme partial correlations mentioned above ("partialling out" all other variables) is called the **conditional independence graph**.

In an attempt to generate graphs that are even more informative for causal relationships than this graph, a theory has been developed ...

A keyword is **directed acyclic graphs**

6.S S-Functions

a For performing multivariate analysis of variance and regression, the functions `lm` and `manova` are helpful.

```
> t.r  <- lm( cbind(Sepal.Length,Sepal.Width,Petal.Length,
Petal.Width)~Species, data=iris)
```

shows, since `lm` is called here with multiple target variables, an object of the class `mlm`, for which

```
> summary(t.r)
```

lists the results of the univariate regressions for all target variables successively.

b For multivariate tests we replace `lm` with `manova` in the previous call. The function `summary(t.r, test="Wilks")` then carries out the test. If the model includes several terms (factors, input values), a corresponding number of tests are carried out – Caution! These are "Type I" tests, which, for the model constructed by entering each term sequentially in the order given in the table, check whether the next term contributes a significant improvement of the model.

c Since R is, in this respect, currently incomplete, the author makes available some functions. These are available from the website `r-forge.r-project.org`.

d **Function `regr`.** The function `regr`, which can fit many regression models, also allows for multivariate regression and provides the information that is recommended here if we print the result or feed it to `plot`.

e The function `plot.regr`, which is activated with `plot(t.r)` for a `regr` result, delivers a comprehensive residual analysis. If a few target variables and input variables are in the model, this produces many pages of graphical output! In short, this function shows:

- Scatterplots of the residuals against the fitted values for all target variables. These scatter plots help to check the general shape of the regression function and, in particular, gives a hint about possible transformations of the target variable.
- Scatterplots of the absolute values of the residuals against the fitted values. We may discover deviations from the assumption of equal variance for all observations.
- Normal distribution qq-plots.
- Scatterplots of the residuals against the leverages. These show heavily influential points.
- Scatterplot matrix of the residuals for the different target variables.
- Scatterplot matrix of the residuals against the input variables in the model. This can give clues about deviations from linearity assumptions and provide possibilities for improvement through transformations of the input variables.

f The function `regr` calls more functions that can also be used individually for results of `lm` and `anova` (?).

g **Function `drop1.mlm`.** There is a function `drop1.mlm`, that we can call with

```
> drop1(t.r)
```

It gives the "Type III" tests and thus checks whether the individual terms of the model can be left out without significantly worsening the fit.

h **Function `summary.mreg`.** A summary of the coefficients in the form of a table, which corresponds to the matrix $\boldsymbol{\beta}$ and thus includes all target variables, can be obtained with

```
> summary.mreg(t.r)
```

The function also gives an analogous table for the standard error and the p-values that indicate whether an individual coefficient is significantly different from 0 (and thus whether the corresponding input variable can be left out of the model for a certain target variable – a univariate consideration).