Regression

Lecture notes for the Spring 2014 course

by Prof. Nicolai Meinshausen

Written by Professor Hansruedi Künsch, using
the transcript of Professor Frank Hampel's course
by Rolf Steiner

Translated by Charles Mitchell

Seminar for Statistics
ETH Zurich

Februrary 2014

Contents

Chapter 1

Linear Regression

1.1 Introduction: underlying question

1.1.1 Examples and historical background

Before modern times a widely-held but short-sighted assumption was that variations in repeated measurement must be due to one measurement being entirely **correct** and all others entirely **wrong**. If for example the position of a star is measured five times and this yields five different results, it was thought that the measurement had only been performed correctly once, and incorrectly on the other four occasions.

This way of thinking only changed about 400 years ago, when the concept of a **"random error"** was introduced. This new concept arose from the idea that **all measured data** contain a small **error** that makes them deviate from the truth, but that they all are **approximately true**. Thus stochastics was born.

This opened up the possibility of investigating approximate, stochastic relationships between variables – which is the subject of this course.

The method of least squares (or LS for short) was published by Legendre in 1805. Gauss also discussed this method in a book published in 1809 – and there mentioned that he had been using this method since 1795. This statement cannot be proven, and so it is not clear who has the honour of having first discovered least squares.

This method was originally used to solve problems of celestial mechanics, where data were fit to orbits determined by theory.

Astronomical example (Ceres): On the basis of the Titius-Bode law (discovered by Titius in 1766), which provides an empirical description of the pattern of planetary distances from the Sun, it was thought that the space between Mars and Jupiter must be inhabited by a further planet. On Jan 1st, 1801, Giuseppe Piazzi found the missing body and named it Ceres. It is the largest asteroid.

Ceres moves quite rapidly and was soon lost once more. It subsequently was Gauss who used the method of least squares to compute a sufficiently exact orbit from the few available data, so that Ceres could be rediscovered.

Later on, this method was used in a much more general way, including in the social sciences. Yule, for example, carried out an investigation in 1899 into whether poor people were best served by being put into poorhouses or by being supported in their usual surroundings. For this purpose he used the regression equation

$$\Delta Paup = a + b \cdot \Delta Out + c \cdot \Delta Old + d \cdot \Delta Pop + error.$$

Here Δ denotes the changes between two successive censuses, "Paup" is the number of people receiving poor relief, "Out" is the ratio of people getting poor relief outside of poorhouses to the number of people in poorhouses, "Old" is the proportion of over-65s in the general population and "Pop" is the total population. This regression equation was fitted using data from two censuses for each of a number of administrative districts (so-called "unions"). The districts had largely autonomous social policies. Yule formed 4 categories of such "unions" (rural, mixed, urban and metropolitan), and for each category the coefficients a, b, c and d were estimated separately. Thus the censuses of 1871 und 1881 yielded a estimated coefficient of $b = 0.755$ for metropolitan districts. In other words, an increase in the variable *Out* goes hand in hand with an increase in the number of poor people – even when other influential factors such as "Old" are accounted for. This led Yule to the conclusion that keeping people in their usual surroundings and supporting them there actually leads to even more poverty.

This is not a conclusive argument, however. It has been shown that those districts with more efficient administration also built more poorhouses at that time. At the same time, efficient administration leads to a reduction in poverty. i.e. the effects of efficient administration and of the establishment of poorhouses cannot be separated. We call these two variables "confounded". Other economic factors are also potential confounders. Generally speaking, a lot of care must be taken when attempting to conclude causality from an observed association. In particular, we cannot draw conclusions about interventions (changes in the variables) on the basis of the regression equation. For further discussion of this point, I refer to D. Freedman's article "From association to causation: Some remarks on the history of statistics", Statistical Science 14 (1999), 243-258, from which the example above is taken.

Another (fictitious) example: is there a link between the number of storks and the rate of human births? Some corresponding data are given in Figure 1.1.

The statistics show a highly significant connection between the number of storks and the birth rate. From this one might wrongly conclude that babies are brought by storks. ("Cause and effect, causal link").

In this case the confounding variable is time. This happens quite often (**meaningless correlation of time series**). In this example it is quite obvious that this asso-

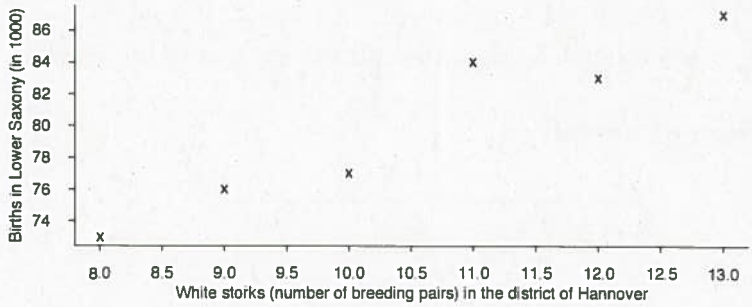

Figure 1.1: Connection between birth rates and stork numbers

ciation does not imply a causal relationship. However, if we look at the amount of brutality in TV shows and the rate of violent crime, we are quite sure to find a statistical link (both increase with time), and a causal relationship is possible a priori – though such a relationship is not easy to prove!

1.1.2 Linear model and examples

Multiple regression:
Given: a single dependent variable (target variable) which up to measurement errors (or random fluctuations) depends on several "independent" or "explanatory" variables (or experimental conditions).
Wanted: the parameter values that describe this linear dependence, and the error variance.

The same model expressed in formulae:

$$Y_i = \theta_1 x_{i1} + \ldots + \theta_p x_{ip} + \varepsilon_i \qquad (i = 1, \ldots, n)$$

Terminology:

- The numbers $(Y_i; i = 1, \ldots, n)$ form the vector $\mathbf{y}$ of realizations of the **dependent variable** (also known as the "target variable" or "response").
- The numbers $(x_{ij}; i = 1, \ldots, n)$ form the vector $\mathbf{x}^{(j)}$ of realizations of the j-th **independent (explanatory) variable** (experimental condition) $(j = 1, \ldots, p)$.
- The $(x_{ij}; j = 1, \ldots, p)$ form the vector $\mathbf{x}_i$ of explanatory variables (experimental conditions) of the i-th observation $(i = 1, \ldots, n)$.
- The numbers $(\theta_j; j = 1, \ldots, p)$ form the **unknown parameter** vector $\boldsymbol{\theta}$.
- The numbers $(\varepsilon_i; i = 1, \ldots, n)$ form the vector $\boldsymbol{\varepsilon}$ of (unknown) **errors**, which we shall assume to be random.

- n is the number of observations (the sample size), while p is the number of explanatory variables.

While the θ_j and ε_i are unknown quantities, all the x_{ij} and y_i are known.

Vectorial notation of model:

$$\boxed{Y_i = \mathbf{x}_i^T \boldsymbol{\theta} + \varepsilon_i \qquad (i = 1, \ldots, n)}$$

Matrix notation of model:

$$\boxed{\begin{array}{ccccccc} \mathbf{Y} & = & X & \times & \boldsymbol{\theta} & + & \boldsymbol{\varepsilon} \\ n \times 1 & & n \times p & & p \times 1 & & n \times 1 \end{array}}$$

where X is an $(n \times p)$ matrix with rows $\mathbf{x}_i^T$ and columns $\mathbf{x}^{(j)}$.

The first explanatory variable is usually a constant, i.e. $x_{i1} = 1$ for all i. The model thus contains an intercept. To give an interpretation to the parameter θ_1 in such a case, we assume the errors ε_i to have mean zero. In other situations such an assumption is also generally made.

We furthermore assume that there are more observations than covariates ($p < n$) and that the matrix X has the maximum rank possible, p, i.e. that the p columns of X are linearly independent. If this were not the case, the parameters would not be identifiable (different choices of parameter may yield the same model). Sometimes models with linear dependence in the columns are used all the same, and the identifiability forced by auxiliary conditions.

A word about the notation: we endeavour to consistently write vectors in bold type. On the other hand we are less consistent in distinguishing between random variables (in upper case) and their realizations (in lower case) as we often switch between these two interpretations, and as fixed matrices are also written in upper case.

Examples:

(1) **The location model:**

$$p = 1, \qquad X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \theta_1 = \mu.$$

(2) **The 2–sample model:**

$$p = 2, \qquad X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}, \qquad \theta = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$$

Here the most common questions are "Is $\mu_1 = \mu_2$ plausible?", and "How large is their difference?", respectively. From the introductory course, we can already handle these using the 2–sample t-test or the 2–sample Wilcoxon test.

Statistically speaking, the 2-sample model is the simplest one imaginable. (In practice it is mostly simpler than the 1-sample (location) model, as systematic and semi-systematic errors often cancel out.)

(3) **One-way (simple) analysis of variance with k levels (groups)** This is a generalization of the previous example from 2 to k groups. In this case, $p = k$, the parameters are the group means μ_j $(1 \leq j \leq k)$ and the matrix X looks similar to above. Another parametrization is also commonly used, namely

$$\mu_j = \mu + \alpha_j,$$

where μ is the overall mean and α_j is the j-th group effect. In this setup, the $k+1$ parameters cannot all be determined and the columns of X are linearly dependent. The identifiability of these parameters is forced by introducing a condition such as $\sum \alpha_j = 0$.

This type of model is treated at greater depth in the course Analysis of Variance ("Angewandte Varianzanalyse und Versuchsplanung").

(4) **Regression through the origin:** $Y_i = \beta x_i + \varepsilon_i$ $(i = 1, \ldots n)$.

$$p = 1, \qquad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad \theta_1 = \beta.$$

(5) **Simple linear regression:** $Y_i = \alpha + \beta x_i + \varepsilon_i$ $(i = 1, \ldots n)$.

$$p = 2, \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

(6) **Quadratic regression:** $Y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i$ $(i = 1, \dots n)$.

$$p = 3, \qquad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \qquad \theta = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}.$$

Conventional interpretation of quadratic regression: We fit a parabola to the two-dimensional point cloud $(x_1, y_1), \dots, (x_n, y_n)$, cf. Fig. 1.2, left side.

Alternative explanation: A fixed parabola is given by the pairs $(x_1, x_1^2), \dots, (x_n, x_n^2)$. In the third dimension (that is, y), we now look for a suitable plane containing it, cf. Fig. 1.2, right side.

Bottom line: The function we are fitting is quadratic in the known covariates, but **linear** in the unknown coefficients (and thus it is a linear model).

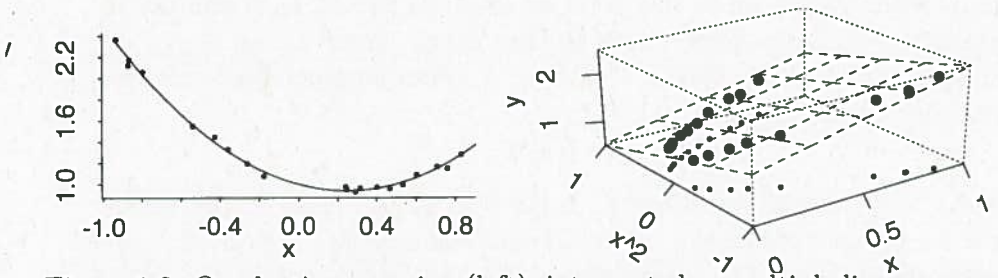

Figure 1.2: Quadratic regression (left), interpreted as multiple linear regression (right).

(7) **Power or exponential dependence:**
A dependence of the form $Y_i = \alpha x_i^\beta + \varepsilon_i$, or $Y_i = \alpha \exp(\beta x_i) + \varepsilon_i$, where α and β are unknown parameters, does not fit our model. However, the deterministic part does not change when we take logarithms on both sides. This leads us to the related model

$$\log(Y_i) = \log(\alpha) + \beta \log(x_i) + \varepsilon_i \ .$$

This is an example of the general linear model with target variables $\log(Y_i)$ and

$$p = 2 \qquad X = \begin{pmatrix} 1 & \log(x_1) \\ 1 & \log(x_2) \\ \dots & \\ 1 & \log(x_n) \end{pmatrix} \qquad \theta = \begin{pmatrix} \log(\alpha) \\ \beta \end{pmatrix}.$$

Inverting this transformation, we obtain

$$Y_i = \alpha x_i^\beta \cdot \eta_i \ ,$$

where $\eta_i = \exp(\varepsilon_i)$. In other words, the errors on the original scale are **multiplicative** instead of **additive**. When the dependence is of power or exponential type, multiplicative errors are usually more plausible, as then the error size is proportional to the scale of the target variables.

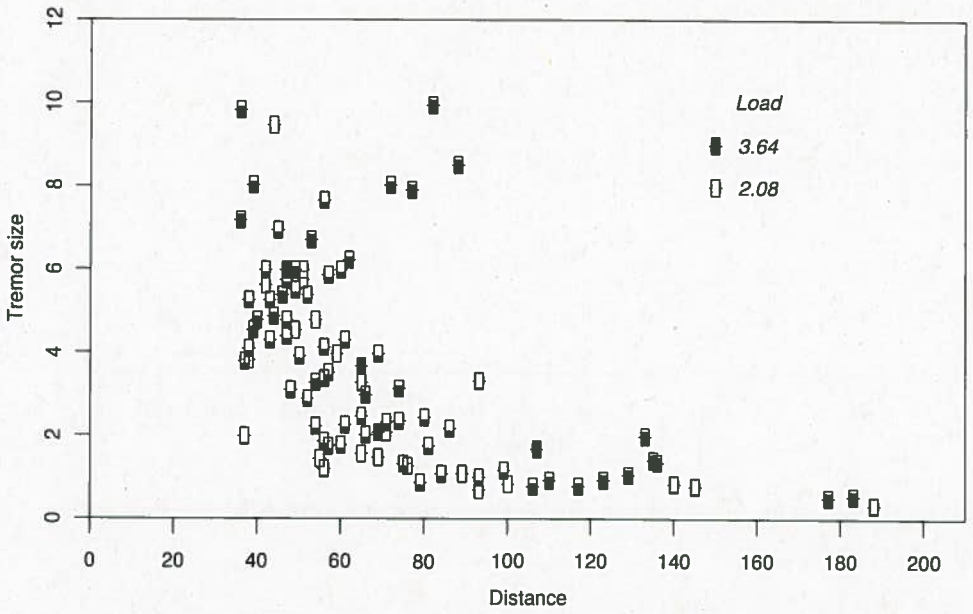

Figure 1.3: Dependence of tremor size on distance for various loadings

One concrete example of this is a dataset whose target variable is the size of the tremor caused by a controlled explosion, and the explanatory variables are the explosive load and the distance of the explosion from the location of the measurement. The corresponding data are plotted in Figure 1.3. It is clearly visible that the effect of distance is non-linear.

The physical expectation is that the tremor size should be inversely proportional to the squared distance. In this case, we have a power model with a known parameter β. Figure 1.4 plots the logarithmic tremor sizes against the logarithmic distances for a fixed load size. They have an approximately linear relationship with an roughly constant variation. One immediate question is whether or not the slope of the linear fit really is -2, the value postulated by theoretical physical considerations. This will be one of the questions we shall look into during this course.

These examples have shown us two important principles:

- A model is called linear if it is linear in its parameters. The original explanatory variables can be subjected to arbitrary transformations.
- We can often obtain a linear model simply by transforming both sides of a deterministic relationship. However, we must then think about whether or not the errors can plausibly be additive on the transformed scale.

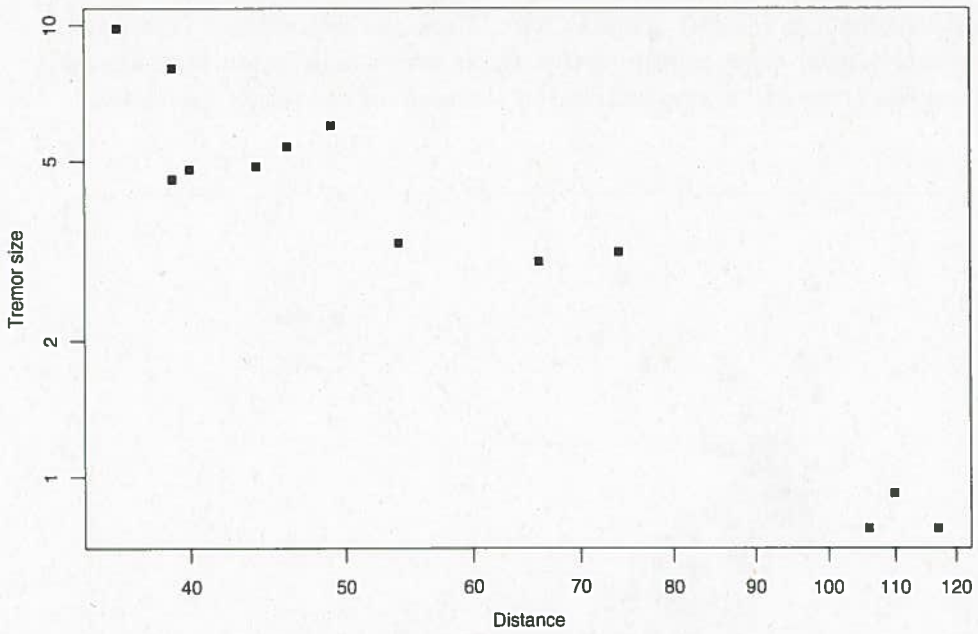

Figure 1.4: Distance and tremor size for explosions with explosive loading 3.12. Logarithmic axes are used

Our goals in regression analysis:

- **Producing a best "fit".** Fitting a (hyper)plane over the covariates through the response points and minimizing deviations from these responses. The standard procedure for this is the method of least squares, but the deviations can also be sized up in other ways.
- **Good parameter estimates.** These answer the question: How does the response change when an explanatory variable changes?
- **Good predictions.** These answer the question: What response can we expect to get under a new set of experimental conditions?
- **An indication of the uncertainty underlying the three previous problems** using tests and confidence intervals.
- **Developing a simple and functioning model.** This is usually the result of an iterative process.

1.2 Prerequisites for the linear model

For a linear model fit using least squares to be meaningful, we have to make certain assumptions. These are also needed for the validity of the statistical tests and confidence intervals we shall derive. Before we list these conditions

in descending order of importance, let us note that the model places no prerequisites on the explanatory variables. These can be continuous or discrete, and they may be transformed and combined in an arbitrary manner. Furthermore there is no difference in principle between the deterministic fixing of the values of explanatory variables by the experimenter, and their being realizations of random variables themselves. The theory derived in the following sections will always regard the explanatory variables as deterministic. That is to say: if these variables are random, our statements are to be understood as conditional ones given the values of the explanatory variables.

As we shall see later on, these assumptions can partly be checked statistically.

1. **The data are useful for gaining the information sought ("representative", "meaningful"). They are a random sample from the population under investigation.**

   - If this assumption is incorrect, the whole analysis is worthless from the outset. The data might just as well be discarded.
   - To judge the usefulness of the data we require some insight into the problem. The real underlying situation is the deciding factor – and that cannot be decided by using statistical methods.
   - Our actual target variables might not be measurable in any precise way. How does one measure intelligence, for example? We attempt this using tests and their evaluation, and we then define intelligence by the result of intelligence tests. The connection of these two quantities remains an unanswered question, though. A further example of this: the state measures its citizens' wealth by the assets declared on their tax forms.
     Transferring results from measurable quantities to the actual underlying quantities of interest is a whole problem unto itself.

2. **The regression equation is correct.** That is:

$$\mathbf{E}[\varepsilon_i] = 0 \quad \forall i$$

   Specifically: no significant explanatory variables should be missing and the relationship between the target variable and the explanatory variables should be a linear one (after suitable transformations).

3. **The errors are uncorrelated.** This means (under Assumption 2):

$$\mathbf{E}[\varepsilon_i \varepsilon_j] = 0 \quad \forall i, j \quad (i \neq j)$$

   If the errors are correlated, the least squares fit is still of some use, but its **precision** is not what we think it is. The levels we obtain for tests, and the confidence intervals we compute, are wrong. We will discuss this in more detail later.

4. **All covariates $\mathbf{x}_i$ are exact.**

   That is, the covariates $\mathbf{x}_i$ are all known without any errors. If, however, they do contain measurement or observational errors, the method of least squares makes systematic errors. There are ways of correcting for this if at least the ratio of error variances is known for each variable. In the literature, this is discussed under the heading of "errors in variables models".

   At first this may seem to contradict the previous statement that the $\mathbf{x}_i$ may be random. However, the assumption here is that we know exactly which value of $\mathbf{x}_i$ led to the corresponding y_i. This is a different question to how $\mathbf{x}_i$ came about.

5. **The error variance is constant ("homoscedasticity").** That is:

$$\mathbf{E}\left[\varepsilon_i^2\right] = \sigma^2 \quad \forall i$$

   All measurements should have the same precision. (In particular, there should be no "bad errors" with a much higher variance.) A constant error variance can often be reached by a simple transformation of the target variable. If the homoscedasticity assumption does not hold, the method of least squares quickly becomes imprecise (compared to other methods). This will also be discussed at greater depth later.

6. **The errors $(\varepsilon_i; i = 1, \ldots n)$ follow a joint normal distribution.** (The same then also holds for the Y_i.) Such a normal distribution of the errors is often a consequence of the general properties of the normal distribution (cf. Central Limit Theorem in the Appendix). It should however not be assumed without question.

Assumptions 2, 3, 5 and 6 can partly be checked using statistical tools. We shall discuss suitable methods for this later in these notes.

The conditions listed above are usually only satisfied in an approximate way. The art of statistics is to develop a feel for which deviations from the assumptions are significant, and which statements and methods are still meaningful when the model is false.

If for instance $(X_1, X_2, \ldots X_p, Y)$ is a $(p+1)$-dimensional random vector that follows an arbitrary distribution, and we fit a linear model using n independent realizations of that random vector, then we are actually estimating the coefficients of the best linear predictor, defined as

$$\arg\min_{\theta_0,\ldots\theta_p} \mathbf{E}\left[(Y - \theta_0 - \sum_{j=1}^{p} \theta_j X_j)^2\right] .$$

Thus least squares can nearly always be used if prediction is our only goal. Some care must be taken in interpreting the parameters and specifying the precision.

Finally, here is an example where Assumption 3 is not true (and neither are Assumptions 1 and 2). The dependent variable here is the number of live births in Switzerland since 1930, and the explanatory variable is time (as well as certain functions of time, if e.g. quadratic trends are also of interest).

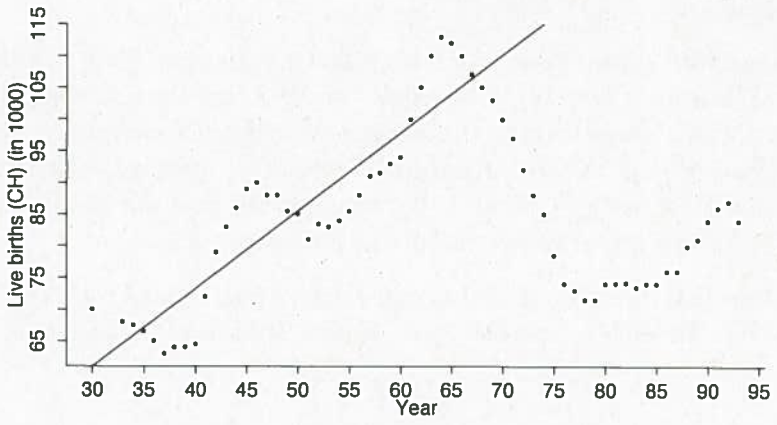

Figure 1.5: Effect of the contraceptive pill

As we see in Figure 1.5, the data on live births in Switzerland following World War II have an approximately linear trend up to 1964. A closer look, though, reveals that the data are not symmetrically distributed around the regression line, but that they form "groups" on either side of it; successive maxima and minima are about twenty years apart (one generation).

Furthermore, the years before 1964 are not "representative" for the following years – that is, the model is no longer valid. It is generally quite dangerous to extrapolate a fitted linear model to an area where no observations of the explanatory variables are available.

1.3 The least squares estimator

Consider the following model: $\boxed{\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}}$

We would like the "best possible" estimate of $\boldsymbol{\theta}$. The least squares estimate $\widehat{\boldsymbol{\theta}}$ is defined as the quantity that minimizes the L_2–norm of the error:

$$\|\mathbf{y} - X\widehat{\boldsymbol{\theta}}\| = \min_{\boldsymbol{\theta}} \|\mathbf{y} - X\boldsymbol{\theta}\|.$$

Thus we minimize the Euclidean distance of the error $\mathbf{y} - X\boldsymbol{\theta}$ from the vector zero.

1.3.1 Normal equations

To compute the least squares estimate, we calculate the partial derivatives of $\|\mathbf{y} - X\boldsymbol{\theta}\|^2$ by $\boldsymbol{\theta}$ (which form a vector) and require them to be zero, thus

obtaining the equation

$$(-2)\ X^T(\mathbf{y} - X\widehat{\boldsymbol{\theta}}) = \mathbf{0} \qquad ((p \times 1) - \text{dimensional zero vector}),$$

which is the same as: $\boxed{X^T X\widehat{\boldsymbol{\theta}} = X^T\mathbf{y}}$.

These are the **normal equations**. We have p linear equations for p unknowns (note that X^TX is a $p \times p$ matrix). The entries of X^TX are the scalar products of the columns of X. Thus solving the normal equations is especially simple when the columns $\mathbf{x}^{(j)}$ of X are orthogonal. Another interpretation can be found by writing X^TX as $\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$: this way, we see that X^TX is n times the matrix of empirical second moments of the covariates $(\mathbf{x}_i)$.

If we now assume that the matrix X has full rank (which means rank p), then X^TX is invertible. In such a case the least squares solution is unique and can be written as

$$\boxed{\widehat{\boldsymbol{\theta}} = (X^TX)^{-1}X^T\mathbf{y}.}$$

This is a useful formula for theoretical considerations, but its susceptibility to rounding errors make it unsuitable for numerical computation. One numerically stable algorithm uses QR decomposition and Givens rotations.

Special case: simple linear regression $y_i = \alpha + \beta x_i + \varepsilon_i$. Then

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

and plugging this into the normal equations gives us

$$\boxed{\begin{array}{lcl} n\alpha + (\sum_{i=1}^n x_i)\cdot\beta & = & \sum_{i=1}^n y_i \\ (\sum_{i=1}^n x_i)\alpha + (\sum_{i=1}^n x_i^2)\cdot\beta & = & \sum_{i=1}^n x_iy_i \end{array}}.$$

To solve this system of equations in a simple manner we employ the "orthogonalization" technique, i.e. we introduce the new variable

$$x \longrightarrow \widetilde{x} := x - \overline{x} \text{ (where } \overline{x} := \frac{1}{n}\sum_{i=1}^n x_i \text{ is the "arithmetic mean")}$$

Writing $\widetilde{\alpha} = \alpha + \beta\overline{x}$, we then get $y_i = \widetilde{\alpha} + \beta\widetilde{x}_i$. As

$$\sum_{i=1}^n \widetilde{x}_i = \sum_{i=1}^n (x_i - \overline{x}) = \sum_{i=1}^n x_i - n\overline{x} = n\overline{x} - n\overline{x} = 0,$$

we immediately obtain:

$$\widehat{\widetilde{\alpha}} = \frac{\sum_{i=1}^n y_i}{n} = \overline{y}\ ; \qquad \widehat{\beta} = \frac{\sum_{i=1}^n \widetilde{x}_iy_i}{\sum_{i=1}^n \widetilde{x}_i^2}.$$

A simple transformation back then finally gives us the desired quantities:

$$\boxed{\widehat{\beta} = \frac{\sum_{i=1}^n (x_i - \overline{x})y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \; ; \qquad \widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}}$$

In a multiple linear regression setup that includes an intercept term (thus making the corresponding column in X contain only the number 1), we can orthogonalize in a similar way:

$$y_i = \alpha + \sum_{j=2}^{p} \theta_j \widetilde{x}_{ij} \; .$$

Here $\widetilde{x}_{ij} = x_{ij} - \overline{x}_j$, with $\overline{x}_j = \sum_{i=1}^n x_{ij}/n$ (for $j > 1$) and $\alpha = \theta_1 + \sum_{j=2}^p \theta_j \overline{x}_j$. From this we can conclude that $\overline{y} = \widehat{\theta}_1 + \sum_{j=2}^p \widehat{\theta}_j \overline{x}_j$, i.e. the point $(\overline{x}_2, \ldots \overline{x}_p, \overline{y})$ lies on the fitted plane.

1.3.2 Geometric interpretation

We can interpret this estimation procedure by looking at rows or at columns. In the former case, if the model contains an intercept (i.e. $x_{i1} \equiv 1$), then we have n points $(y_i, x_{i2}, \ldots x_{ip})$ randomly spread around a $(p-1)$-dimensional hyperplane in a p-dimensional space. (If our model lacks an intercept, we have n points spread around a hyperplane through the origin in $(p+1)$-dimensional space.) However, this random spread only occurs parallel to the y-axis. The least squares estimator therefore estimates the parameters of the hyperplane so as to minimize the sum of squared distances of the points from the plane *parallel to the y-axis.*

More mileage is to be had by interpreting the column vectors in the model. Here the vector $\mathbf{y}$ of observations is a single point in the n-dimensional space $\mathbb{R}^n$. If we vary the value of the parameter $\boldsymbol{\theta}$, the product $X\boldsymbol{\theta}$ describes a p-dimensional subspace of $\mathbb{R}^n$, i.e. a p-dimensional hyperplane through the origin. Then an obvious way to estimate $\boldsymbol{\theta}$ is to make $X\boldsymbol{\theta}$ minimize $\mathbf{y}$ on this hyperplane. Choosing the L_2 norm to yield our metric in $\mathbb{R}^n$ amounts to choosing the Euclidean distance, which geometrically implies that we are performing an orthogonal projection of $\mathbf{y}$ onto this hyperplane. In particular, the least squares estimator is characterized by the property that $\mathbf{r} = \mathbf{y} - X\widehat{\boldsymbol{\theta}}$ (the vector of residuals) is orthogonal to all the columns of X:

$$(\mathbf{y} - X\widehat{\boldsymbol{\theta}})^T X = 0.$$

This amounts to a geometric interpretation of the normal equations (cf. Figure 1.6).

Now the orthogonal projection $X\widehat{\boldsymbol{\theta}}$ is the estimate of $\mathbf{E}[\mathbf{y}]$ given by our model. Its components are called the **fitted values**, and they are generally denoted by $\widehat{\mathbf{y}}$.

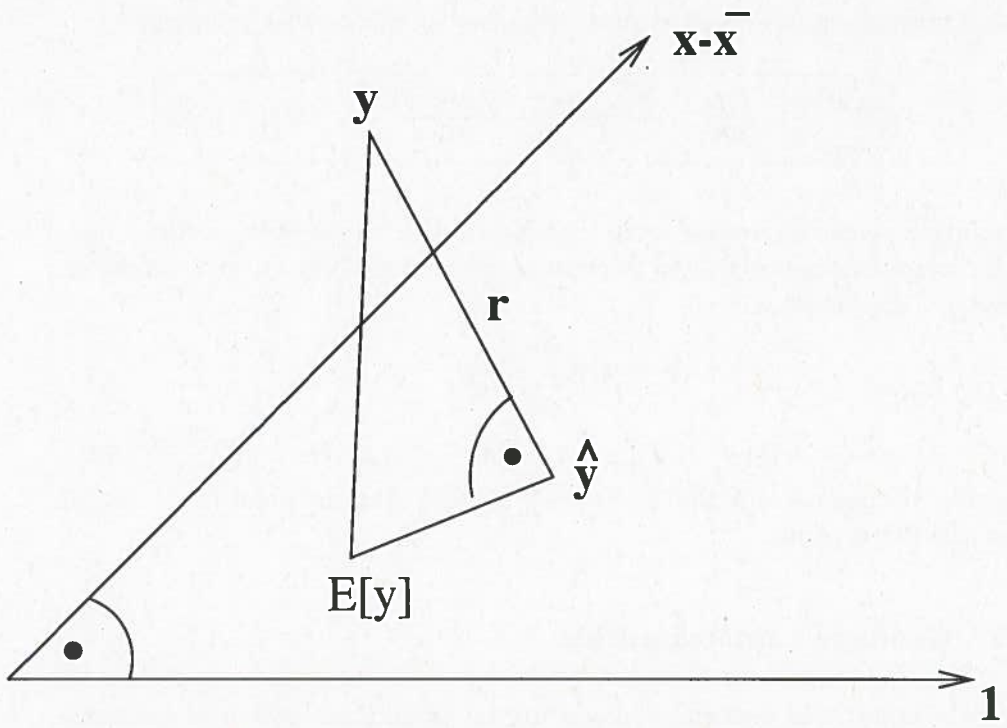

Figure 1.6: The vector $\mathbf{r}$ of residuals is orthogonal to the hyperplane spanned by the vectors $\mathbf{1}$ and $\mathbf{x}$.

To compute $\widehat{\mathbf{y}}$, we can use the formula

$$\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\theta}} = \underbrace{X(X^TX)^{-1}X^T}_{=:P}\mathbf{y} = P\mathbf{y} \qquad \text{and thus:} \qquad \boxed{\widehat{\mathbf{y}} = P\mathbf{y}}$$

It is easy to check that the matrix P has the following properties:

$$P^T = P, \quad P^2 = P$$

and

$$\sum_i P_{ii} = tr(P) = tr(X(X^TX)^{-1}X^T) = tr((X^TX)^{-1}X^TX) = tr(I_{p\times p}) = p.$$

These are necessary and sufficient conditions for P to be an orthogonal projection from $\mathbb{R}^n$ to $\mathbb{R}^p$.

It is quite evident that the matrix P only depends on the explanatory variables (the experimental conditions), but not on the target variables (the responses). It is also known as the hat matrix, as it "puts the hat on" $\mathbf{y}$. Furthermore, the diagonal entries P_{ii} tell us how much influence the observation y_i (at the point $\mathbf{x}_i$) has over the fitted value $\widehat{y}_i$.

The residuals $\mathbf{r}$, also denoted by $\widehat{\boldsymbol{\varepsilon}}$, can be written in a manner similar to the fitted values. That is:

$$\mathbf{r} = \widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}} = \underbrace{(I - P)}_{=:Q}\mathbf{y} = Q\mathbf{y} \qquad \text{and thus:} \qquad \boxed{\mathbf{r} = Q\mathbf{y}}$$

Q is an orthogonal projection yet again; it is orthogonal to P and has values in the residual space:

$$Q^T = Q^2 = Q, \quad PQ = QP = 0 \ \ (0 \text{ as an } n \times n \text{ matrix}), \quad tr(Q) = n - p.$$

1.3.3 Link to the MLE for the normal distribution

Assuming that the explanatory variables are given, the assumptions of our model (independence and normality of the errors) imply that the conditional density of $y_1, ..., y_n$ is

$$L_{\mathbf{y},X}(\boldsymbol{\theta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma} \varphi((y_i - \sum_{j=1}^{p} \theta_j x_{ij})/\sigma).$$

When performing maximum likelihood estimation, we consider the density to be a function of the parameters σ and $\boldsymbol{\theta}$ (fixing y_i and x_{ij}). This function is the so-called likelihood function, and we estimate the parameters $\boldsymbol{\theta}$ and σ so that they maximize this likelihood function (or equivalently, we maximize its logarithm). The result of this process is called the Maximum Likelihood Estimator (MLE).

It is immediately obvious that maximizing with respect to $\boldsymbol{\theta}$ does not depend on the value of σ, and that it amounts to minimizing $\|\mathbf{y} - X\boldsymbol{\theta}\|^2$. If all the errors ε_i are i.i.d. and they all have the distribution $\mathcal{N}(0, \sigma^2)$, the MLE of $\boldsymbol{\theta}$ is exactly the same as the least squares estimator.

The MLE for σ^2, on the other hand, is

$$\widehat{\sigma}^2_{ML} = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n}.$$

This is not the estimator generally used, however. Instead, it is scaled to become unbiased. We shall see that the correct scaling factor here is $n/(n-p)$. In other words, the estimator of error variance we shall use is

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n - p}.$$

1.3.4 Why not regress on each variable individually ?

The following (artificial) example shows why multiple regression cannot simply be replaced by several simple regression procedures.

Let there be 2 covariates x_1, x_2, and assume that we have the following observations:

x_1	0	1	2	3	0	1	2	3
x_2	−1	0	1	2	1	2	3	4
y	1	2	3	4	−1	0	1	2

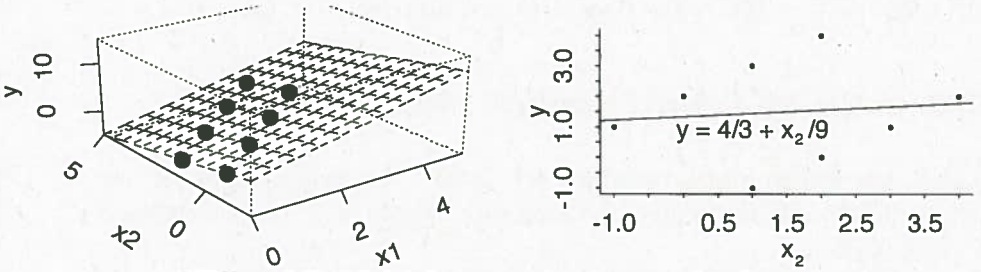

Figure 1.7: Multiple regression versus simple regression

Left side of Fig. 1.7: We plot the values of y against the corresponding values of the covariates x_1 und x_2. Subsequently, we find a plane that fits these 8 points **exactly** (in three-dimensional space):

$$\boxed{y = 2x_1 - x_2 \qquad (\widehat{\sigma}^2 = 0)}$$

The coefficients (2 and -1, respectively) tell us how y changes if we change exactly one of x_1 or x_2 and fix the other one.

We conclude that y decreases as x_2 increases (x_2 larger $\Rightarrow$ y smaller).

Right side of Fig. 1.7: We simply regress y on x_2 and forget about the values of x_1 (they are not kept fixed). The ensuing regression line is:

$$\boxed{y = \frac{1}{9}x_2 + \frac{4}{3} \qquad (\widehat{\sigma}^2 = 1.72)}$$

This line tells us how changes x_2 affect y if x_1 is allowed to vary.

We conclude that y increases as x_2 increases (x_2 larger $\Rightarrow$ y larger).

The reason for this difference in the behaviour of y depending on x_2 is that the covariates x_1 and x_2 are **strongly correlated**. That is: when x_2 increases, so does x_1.

In summary:

Combining several simple regressions (each using the method of least squares) generally only gives us the same result as a multiple regression if the explanatory variables are orthogonal.

1.4 Properties of LS estimation

Let us first take an intuitive look at the precision of the regression plane. We assume a known linear model and simulate a random point cloud from this model. This cloud of points we can now use to fit a regression plane using the

method of least squares. If we now take a second (or third, ...) point cloud, we will generally get a different estimated regression plane – even though the underlying model, and thus the theoretical plane, are the same (cf. Fig. 1.8). In other words, the parameter estimates and the fitted regression plane are random ! Because of this, we do need some idea of their precision.

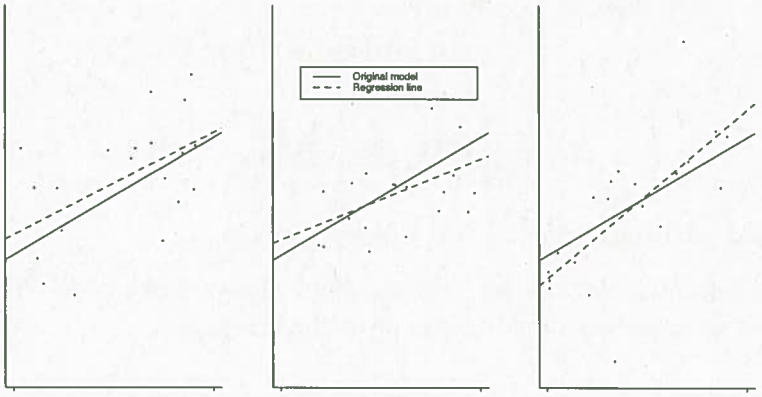

Figure 1.8: Three different estimated regression lines for the same underlying model.

1.4.1 Moments without any normality assumptions

The next results do not require the errors ε_i to be normally distributed. The **assumptions in this section** are the following:

Usual model: $\boxed{\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}}$ **with** $\boxed{\begin{array}{l}\mathbf{E}[\boldsymbol{\varepsilon}] = \mathbf{0} \\ \mathrm{Cov}[\boldsymbol{\varepsilon}] = \mathbf{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] = \sigma^2 I_{n\times n}\end{array}}$

Results:

(i) $\mathbf{E}\left[\widehat{\boldsymbol{\theta}}\right] = \boldsymbol{\theta}$, as
$\mathbf{E}\left[\widehat{\boldsymbol{\theta}}\right] = \mathbf{E}\left[(X^TX)^{-1}X^T\mathbf{y}\right] = \mathbf{E}\left[(X^TX)^{-1}X^T(X\boldsymbol{\theta}+\boldsymbol{\varepsilon})\right] = \boldsymbol{\theta} + \mathbf{0} = \boldsymbol{\theta}.$

(ii) $\mathbf{E}[\widehat{\boldsymbol{\varepsilon}}] = \mathbf{0}$, $\mathbf{E}[\widehat{\mathbf{y}}] = \mathbf{E}[\mathbf{y}] = X\boldsymbol{\theta}$.

(iii) $\mathrm{Cov}[\widehat{\boldsymbol{\theta}}] = \sigma^2(X^TX)^{-1}$, as
$\mathbf{E}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right] = \mathbf{E}\left[(X^TX)^{-1}X^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^TX(X^TX)^{-1}\right] = \sigma^2(X^TX)^{-1}.$

(iv) $\mathrm{Cov}[\widehat{\mathbf{y}}] = \mathrm{Cov}[P\mathbf{y}] = \sigma^2PP^T = \sigma^2P$ (because P is a projection matrix).

(v) $\mathrm{Cov}[\widehat{\boldsymbol{\varepsilon}}] = \sigma^2Q$ (similarly).

(vi) $\mathrm{Cov}[\widehat{\boldsymbol{\varepsilon}}, \widehat{\mathbf{y}}] = 0$, (as $QP = 0$).

The covariance matrices in (iv) and (v) are only positive definite. We can see from (v) that the residuals $r_i = \widehat{\varepsilon}_i$ are correlated – unlike the true errors – and

that their variance

$$\mathrm{Var}[\widehat{\varepsilon}_i] = \sigma^2(1 - P_{ii})$$

is not constant. We can furthermore conclude that

$$\begin{aligned} \mathbf{E}\left[\sum_{i=1}^{n} r_i^2\right] &= \sigma^2 \sum_{i=1}^{n} (1 - P_{ii}) \\ &= \sigma^2(n - \mathrm{tr}(P)) = \sigma^2(n-p). \end{aligned}$$

Therefore

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} r_i^2}{n-p} = \frac{\|\mathbf{y} - X\widehat{\boldsymbol{\theta}}\|^2}{n-p}$$

is an **unbiased estimate** of σ^2, as we already claimed.

Note that we cannot make any statements about the variance of $\widehat{\sigma}^2$. For this we would need to know the fourth moment of the errors ε_i.

1.4.2 Distribution under the assumption of normality

Assumptions in this section:

Usual model: $\boxed{\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}}$ **now assuming** $\boxed{\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0,\ \sigma^2 I_{n\times n})}$

Results:

(i) $\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_p(\boldsymbol{\theta},\ \sigma^2(X^TX)^{-1})$ (as $\widehat{\boldsymbol{\theta}}$ is a linear combination of normally distributed quantities and therefore itself follows a normal distribution).

(ii) $\widehat{\mathbf{y}} \sim \mathcal{N}_n(X\boldsymbol{\theta},\ \sigma^2 P)$, $\widehat{\boldsymbol{\varepsilon}} \sim \mathcal{N}_n(0,\ \sigma^2 Q)$ (for the same reason as above).

iii) $\widehat{\mathbf{y}}$ and $\widehat{\boldsymbol{\varepsilon}}$ are **independent** (as they are uncorrelated and both normally distributed).

(iv)

$$\frac{\sum_{i=1}^{n} r_i^2}{\sigma^2} \sim \chi^2_{n-p}.$$

(see below).

(v) $\widehat{\sigma}^2$ is independent of $\widehat{\boldsymbol{\theta}} = (X^TX)^{-1}X^T\mathbf{y}$ (This is a consequence of iii), as $\widehat{\boldsymbol{\theta}} = (X^TX)^{-1}X^T\widehat{\mathbf{y}}$).

Proof of iv): Regard a coordinate system with an orthogonal basis, such that the first p vectors in the basis span the column space of X. Denote the corresponding transformation matrix by A; in other words, the columns of A contain the coordinates of the new basis vectors in the old coordinate system. Then A is orthogonal, and if we use stars to denote the new coordinate system, we have $\mathbf{y}^* = A^T\mathbf{y}$, $\boldsymbol{\varepsilon}^* = A^T\boldsymbol{\varepsilon}$ etc.. By construction,

$$\begin{aligned} \widehat{\mathbf{y}}^* &= (y_1^*, y_2^*, \ldots y_p^*, 0, \ldots 0)^T, \\ \widehat{\boldsymbol{\varepsilon}}^* &= (0, \ldots, 0, \varepsilon_{p+1}^*, \ldots \varepsilon_n^*)^T \end{aligned}$$

(which can easily be checked, noting that the last $n-p$ rows of A^TX are all identically equal to zero). Because of the orthogonality of A, we find that

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} \widehat{\varepsilon}_i^{*2} = \sum_{i=p+1}^{n} \varepsilon_i^{*2}.$$

From this the claim follows, as the orthogonality of A means that ε^* also follow a normal distribution $\mathcal{N}_n(\mathbf{0}, \sigma^2 I_{n\times n})$.

1.4.3 Asymptotic normality

The above results on the distribution of estimators are the key to subsequent statements about uncertainty, i.e. confidence intervals or tests. One might therefore ask how decisive the assumption of normal errors is. It has been seen that the results are still approximately correct if the errors are not normally distributed. This is investigated mathematically by looking at the limit behaviour of the distribution when the number of observations goes to infinity.

Consider the following situation: We have n data $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ which satisfy the linear model. Here each $\mathbf{x}_i$ is a p-dimensional column vector (i.e. $\mathbf{x}_i^T$ is the i-th row of X). We assume that the errors ε_i are i.i.d. but not necessarily normally distributed, and look at the limit case $n \longrightarrow \infty$.

For the asymptotic approximation to hold, we need some weak conditions on the explanatory variables $\mathbf{x}_i$:

- The smallest eigenvalue of $X^TX = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$, namely $\lambda_{\min,n}$, converges to ∞.
- $\max_j P_{jj} = \max_j \mathbf{x}_j^T(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T)^{-1}\mathbf{x}_j$ converges to zero.

The first condition states that increasing n always yields more information, while the second condition prohibits any $\mathbf{x}_j$ from dominating the others.

Theorem 1.4.1. *If the errors ε_i are i.i.d. with mean 0 and variance σ^2, and if $(\mathbf{x}_i)$ satisfies the conditions just given, then the LS estimators $\widehat{\boldsymbol{\theta}}$ are consistent (for $\boldsymbol{\theta}$), and the distribution of*

$$(X^TX)^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

converges weakly to $\mathcal{N}_p(\mathbf{0}, \sigma^2 I)$.

Proof: The i-th component $\widehat{\theta}_i$ is unbiased and has variance $\sigma^2((X^TX)^{-1})_{ii}$, which converges to zero by the first assumption. Then consistency follows from Chebyshev's inequality.

To show the weak convergence of a random vector in $\mathbb{R}^p$, it suffices to show the weak convergence of the distributions of all its linear combinations (Theorem of Cramér and Wold, for which see the literature). So we consider

$$\mathbf{c}^T((X^TX)^{1/2})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{c}^T((X^TX)^{-1/2})X^T\boldsymbol{\varepsilon} = \mathbf{a_n}^T\boldsymbol{\varepsilon} = \sum_{i=1}^{n} a_{ni}\varepsilon_i \ ,$$

where

$$\mathbf{a_n} = X((X^TX)^{-1/2})^T\mathbf{c}.$$

We have a sum of n independent, but not identically distributed terms $a_{ni}\varepsilon_i$. Furthermore, the distribution of these summands changes with n. This is precisely the situation for which Lindeberg's Theorem was established (see the introductory class for details). We have

$$\mathrm{Var}\sum_{i=1}^{n} a_{ni}\varepsilon_i = \sigma^2\sum_{i=1}^{n} a_{ni}^2 = \sigma^2\mathbf{a_n}^T\mathbf{a_n} = \sigma^2\mathbf{c}^T\mathbf{c}.$$

Without loss of generality, we can assume this variance to be 1. We now need only check the condition

$$\sum_{i=1}^{n} a_{ni}^2\mathbf{E}\left[\varepsilon_i^2 1_{[|\varepsilon_i|>\eta/a_{ni}]}\right] \to 0$$

for all $\eta > 0$. As all ε_i have the same distribution and a finite second moment, it follows that

$$\mathbf{E}\left[\varepsilon_i^2 1_{[|\varepsilon_i|>d]}\right] = \mathbf{E}\left[\varepsilon_1^2 1_{[|\varepsilon_1|>d]}\right] \overset{d\to\infty}{\longrightarrow} 0.$$

As $\sum_i a_{ni}^2 = 1$ also holds, it suffices to show that $\max_i |a_{ni}|$ converges to zero. From the Schwarz inequality, we obtain

$$a_{ni}^2 \le \|\mathbf{c}\|^2\,\mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i.$$

Thus the claim follows straight from the second condition. □

We can also show that $\widehat{\sigma}^2$ is consistent. However, to show the asymptotic normality of $\widehat{\sigma}^2$ we require the existence of a fourth moment for the ε_i. The asymptotic variance would then depend quite strongly on the exact value of this variance.

Conclusions:

The tests and confidence intervals for $\boldsymbol{\theta}$ and for the means $\mathbf{E}[\mathbf{y}]$, which we shall derive using the assumption of normal errors, still have roughly the correct level if this normality assumption does not hold. However, we know nothing about the efficiency of these methods in such a case, and the confidence intervals for σ found in the literature usually have a very wrong level in non-normal situations.

1.5 Tests and confidence intervals

1.5.1 Basic test statistics

Assume the linear model with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_{n\times n})$. As seen in the last section, these assumptions imply that $\widehat{\boldsymbol{\theta}}$ **exactly follows an** $\mathcal{N}_p(\boldsymbol{\theta}, \sigma^2(X^TX)^{-1})$**–distribution**, and $\widehat{\sigma}^2$ is **independent** of $\widehat{\boldsymbol{\theta}}$. Some of the consequences of this are:

(a) For each individual parameter θ_i, we have

$$\frac{\widehat{\theta}_i - \theta_i}{\widehat{\sigma}\sqrt{((X^TX)^{-1})_{ii}}} \sim t_{n-p}.$$

(b) The entire parameter vector $\boldsymbol{\theta}$ gives rise to an F-distribution as follows:

$$\frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(X^TX)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{p\widehat{\sigma}^2} \sim F_{p,n-p}$$

(c) Similarly for each linear transformation $\boldsymbol{\vartheta} = B\boldsymbol{\theta}$ by a $(q \times p)$ matrix B:

$$\frac{(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})^TV^{-1}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})}{q\widehat{\sigma}^2} \sim F_{q,n-p},$$

where $V = B(X^TX)^{-1}B^T$.

(d) For the expectation of the i-th observation (that is, the true location of the hyperplane for the i–th experimental condition), we have

$$\frac{\widehat{y}_i - \mathbf{E}[y_i]}{\widehat{\sigma}\sqrt{p_{ii}}} \sim t_{n-p} \quad \text{where } p_{ii} := (P)_{ii}$$

(e) For the expectation of a new observation under an arbitrary experimental condition $\mathbf{x}_0$ (the true location of the hyperplane for the new experimental condition $\mathbf{x}_0$), we have

$$\frac{\widehat{y}_0 - \mathbf{E}[y_0]}{\widehat{\sigma}\sqrt{\mathbf{x}_0^T(X^TX)^{-1}\mathbf{x}_0}} \sim t_{n-p}$$

(f) A random new observation $y_0 = y_0(\mathbf{x}_0)$ under the experimental condition $\mathbf{x}_0$ satisfies

$$\frac{y_0 - \widehat{y}_0}{\widehat{\sigma}\sqrt{1 + \mathbf{x}_0^T(X^TX)^{-1}\mathbf{x}_0}} \sim t_{n-p}.$$

These facts allow us to carry out statistical tests in the usual way (using the quantities on the left side as test statistics) and to compute confidence intervals for individual parameters, for linear combinations of them or for the unknown true location of the hyperplane at some $\mathbf{x}_0$. Furthermore, statement (f) enables the construction of prediction intervals for future observations.

We illustrate this by the tremor size example that we introduced earlier. Table 1.1 shows the computer output when we take the logarithm of tremor size as our response variable and the logarithms of distance and load size as explanatory variables.

The output contains the estimated coefficients as well as the residual standard error $\widehat{\sigma}\sqrt{((X^TX)^{-1})_{ii}}$ and the results of tests of the null hypotheses $\theta_i = 0$.

Coefficients:	Value	Std. Error	t value	Pr(> |t|)
(Intercept)	2.8323	0.2229	12.71	0.000
log10(dist)	-1.5107	0.1111	-13.59	0.000
log10(ladung)	0.8083	0.3042	2.66	0.011

Residual standard error: 0.1529 on 45 degrees of freedom
Multiple R-Squared: 0.8048
F-statistic: 92.79 on 2 and 45 degrees of freedom
p-value 1.11e-16

Table 1.1: Computer output for the tremor size example

The other numbers here will be explained in the subsequent sections. Now the t distribution with 45 degrees of freedom is very close to the standard normal distribution. Thus the numbers in Table 1.1 show us that the true coefficient of the logarithm of distance is less than 2 (in absolute terms), while the true coefficient of the logarithm of load size may well be 1.

1.5.2 Confidence band for the entire hyperplane

We can also compute a 95% confidence set, for instance, in which the true hyperplane lies. Before we do this, we will first mention an obvious strategy which fails. For any point $\mathbf{x}_0$ (at which the fitted hyperplane has the expected response value $\widehat{y}(\mathbf{x}_0)$), we can construct an interval around $\widehat{y}(\mathbf{x}_0)$ as above, inside which we expect the value on the true hyperplane to lie with 95% confidence. If we do this for each possible value of $\mathbf{x}_0$, do we then get a 95% confidence set for the true hyperplane?

The answer is "no", of course. At each individual $\mathbf{x}_0$, the probability of the true hyperplane passing through this confidence set is exactly 95%, but for *two* such intervals, the probability of the hyperplane passing through both of them is at least 90% and at most 95%.

Similarly in extreme cases: 10 points $\Rightarrow$ 50 - 95 %
20 points $\Rightarrow$ 0 - 95 %

This is not a good way to get a confidence set for the true hyperplane!

There is a better way, and it is as follows: The Schwarz inequality for the scalar product $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T (X^T X)^{-1} \mathbf{b}$ implies that

$$\begin{aligned} |\widehat{y}_0 - \mathbf{E}[y_0]| &= |\mathbf{x}_0^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})| = |\mathbf{x}_0^T (X^T X)^{-1} (X^T X)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})| \\ &\leq (\mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0)^{1/2} ((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T (X^T X)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}))^{1/2}. \end{aligned}$$

Using statement (b) above tells us that with probability $1 - \alpha$, we have

$$(\widehat{y}_0 - \mathbf{E}[y_0])^2 \leq \widehat{\sigma}^2 \, (\mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0) \, p \, F_{p,n-p}(1 - \alpha)$$

simultaneously for all $\mathbf{x}_0$. This gives us the simultaneous confidence set we are looking for. Its shape is that of a hyperboloid, and it is the envelope of all hyperplanes whose parameters are compatible with the data according to b).

1.5.3 Comparison of nested models, analysis of variance

Prerequisites:

"Basic hypothesis"	$H : \mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ $(X : n \times p,\ rank(X) = p,\ \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0,\ \sigma^2 I))$.
"Special null hypothesis"	H_0: the above, and additionally $B\boldsymbol{\theta} = \mathbf{b}$ (where the dimensions of B are $(p-q) \times p$, and $rank(B) = p - q < p$).

Example:

$$B = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix}, \quad \mathbf{b} = \mathbf{0}.$$

This null hypothesis can be written in words as "The first $p - q$ coefficients θ_i are all zero." We are testing whether the first $p - q$ variables are superfluous to the model.

By statement (c) of Section 1.5.1,

$$\frac{(B\widehat{\boldsymbol{\theta}} - \mathbf{b})^T (B(X^T X)^{-1} B^T)^{-1} (B\widehat{\boldsymbol{\theta}} - \mathbf{b})}{(p-q)\widehat{\sigma}^2}$$

is a suitable test statistic for this null hypothesis. Under this hypothesis, its distribution is $F_{p-q,n-p}$. However, we can use the following geometric argument to obtain a different shape and interpretation of this test statistic.

We assume that $\mathbf{b} = \mathbf{0}$ (this is not a significant restriction, as we can replace the original observations by new ones $\mathbf{y} - X\boldsymbol{\theta}$, taking some $\boldsymbol{\theta}$ that satisfies $B\boldsymbol{\theta} = \mathbf{b}$). Then we can project $\mathbf{y}$, first into the p–dimensional space spanned by the columns of X, and from there into the q–dimensional subspace defined by the additional condition $B\boldsymbol{\theta} = \mathbf{0}$.

Let the corresponding sums of squares of the residuals (under H and H_0) be SSE and SSE_0.

We know that:

- $SSE/(n-p)$ always is an **unbiased** estimator of σ^2, assuming the basic hypothesis H **and** the null hypothesis H_0.
- SSE and $SSE_0 - SSE$ are sums of squares in orthogonal subspaces, and **under the null hypothesis** H_0, we know that $(SSE_0 - SS_e)/(p-q)$ is an **unbiased** estimate of σ^2. If only the basic hypothesis H is true, the expectation of this difference is greater than σ^2.

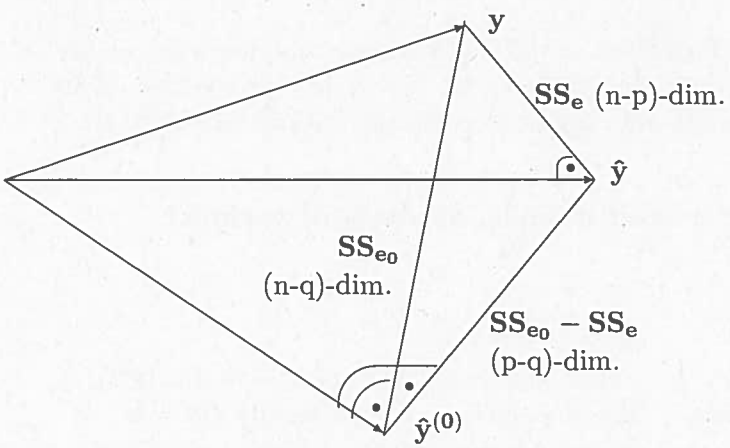

Figure 1.9: Comparison of models

- Because of the orthogonality of the subspaces, we have

$$\left\|\mathbf{y}-\widehat{\mathbf{y}}^{(0)}\right\|^2 = \|\mathbf{y}-\widehat{\mathbf{y}}\|^2 + \left\|\widehat{\mathbf{y}}-\widehat{\mathbf{y}}^{(0)}\right\|^2$$

Thus given $\mathbf{H_0}$:

$$\frac{(SSE_0 - SSE)/(p-q)}{SSE/(n-p)} = \frac{\left\|\widehat{\mathbf{y}}-\widehat{\mathbf{y}}^{(0)}\right\|^2/(p-q)}{\|\mathbf{y}-\widehat{\mathbf{y}}\|^2/(n-p)} \sim F_{p-q,n-p}$$

and we can use the expression on the left side as a test statistic for H_0.

At first glance, the two test statistics we have just derived are different: having the same distribution does not make them identical. However, the following lemma shows that both these expressions are in fact identical.

Lemma 1.5.1. *The least squares estimator $\widehat{\boldsymbol{\theta}}_{(0)}$ under the supplementary condition $B\boldsymbol{\theta} = \mathbf{b}$ is*

$$\widehat{\boldsymbol{\theta}}_{(0)} = \widehat{\boldsymbol{\theta}} - (X^TX)^{-1}B^T(B(X^TX)^{-1}B^T)^{-1}(B\widehat{\boldsymbol{\theta}} - \mathbf{b}).$$

Furthermore,

$$SSE_0 = SSE + (B\widehat{\boldsymbol{\theta}} - \mathbf{b})^T(B(X^TX)^{-1}B^T)^{-1}(B\widehat{\boldsymbol{\theta}} - \mathbf{b}).$$

Proof: We introduce a vector $\boldsymbol{\lambda}$ containing the $p-q$ Lagrange multipliers for the $p-q$ supplementary conditions. Our task is now to minimize

$$(y - X\boldsymbol{\theta})^T(y - X\boldsymbol{\theta}) + (B\boldsymbol{\theta} - \mathbf{b})^T\boldsymbol{\lambda}$$

over $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. This gives us the equations

$$X^T(y - X\widehat{\boldsymbol{\theta}}_{(0)}) + B^T\boldsymbol{\lambda} = 0, \quad B\widehat{\boldsymbol{\theta}}_{(0)} = \mathbf{b}.$$

It is easy to check that $\widehat{\theta}_{(0)}$ satisfies these conditions. Moreover, Pythagoras' Theorem implies that

$$(y - X\widehat{\theta}_{(0)})^T(y - X\widehat{\theta}_{(0)}) = SSE + (X(\widehat{\theta} - \widehat{\theta}_{(0)}))^T(X(\widehat{\theta} - \widehat{\theta}_{(0)})).$$

Plugging in, we obtain the second claim. □

Just as multiple regression cannot merely be replaced by simple regressions on individual variables, the test of the null hypothesis $\beta_1 = \beta_2 = 0$ may yield completely different results than the two tests of the null hypotheses $\beta_1 = 0$ and $\beta_2 = 0$. For instance, it may happen that the latter null hypotheses are both accepted without any problem, but the combined null hypothesis $\beta_1 = \beta_2 = 0$ is thrown out quite clearly. This means that either of the explanatory variables can be left out – but not both. The solution to this apparent paradox lies in the heavy correlation of the two variables. Thus either of them can replace the other quite simply.

It frequently happens that an explanatory variable is categorical (place of origin, type, colour, sex, ...). In the tremor size example, explosions were carried out at six different locations; this may impact the result by means of variation in ground consistency. Such a variable is also called a **factor**. The simplest model for this merely postulates a different intercept for each category and assumes that all other coefficients are the same for all categories. This model is written by introducing indicator variables for each category and using these as additional explanatory variables. For the matrix X to still have full rank, we must omit either the first column $x_{ij} \equiv 1$ or the indicator variable for the first category. For such a categorical variable, one meaningful null hypothesis is that the coefficients of *all* indicators are zero; this can be testes using an F-test. The results for the tremor size example are given in Table 1.2. The third row compares the full model to the model without the explanatory variables given by the location factor "St". It shows the extreme significance of the location.

	Df	Sum of Sq	RSS	F Value	Pr(F)
log10(dist)	1	2.79	5.07	108	0
log10(loading)	1	0.59	2.86	23	7.62e-06
St	5	2.10	4.38	16	0

Table 1.2: Tests of the effects of individual terms in the tremor size example

The decomposition

$$\left\|\mathbf{y} - \widehat{\mathbf{y}}^{(0)}\right\|^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 + \left\|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(0)}\right\|^2$$

is also known as analysis of variance, and thus Table 1.2 is known as an Analysis of Variance (ANOVA) table.

1.5.4 Coefficient of determination

One particularly special case of the preceding results is the following: testing to see whether the response actually depends on the covariates $\mathbf{x}$.

$$X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 1 & \dots & \dots & x_{np} \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \quad (p-1)\times p$$

Under the null hypothesis H_0 : $\widehat{\boldsymbol{\theta}}_{(0)} = \begin{pmatrix} \overline{y} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \widehat{\mathbf{y}}^{(0)} = \begin{pmatrix} \overline{y} \\ \overline{y} \\ \vdots \\ \overline{y} \end{pmatrix} =: \overline{\mathbf{y}} \quad (n\times 1)$

$$\begin{aligned} SSE_0 &= \|\mathbf{y} - \widehat{\mathbf{y}}^{(0)}\|^2 = \|\mathbf{y} - \overline{\mathbf{y}}\|^2 \\ &= \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 + \|\widehat{\mathbf{y}} - \overline{\mathbf{y}}\|^2 \end{aligned}$$

Analysis of variance (ANOVA) table:

	Sum of squares	Degrees of freedom	Mean square	$\mathbf{E}$ [Mean square]
Regression	$\|\widehat{\mathbf{y}} - \overline{\mathbf{y}}\|^2$	$p-1$	$\|\widehat{\mathbf{y}} - \overline{\mathbf{y}}\|^2/(p-1)$	$\sigma^2 + \frac{\|\mathbf{E}[\mathbf{y}] - \mathbf{E}[\overline{\mathbf{y}}]\|^2}{p-1}$
Error	$\|\mathbf{y} - \widehat{\mathbf{y}}\|^2$	$n-p$	$\|\mathbf{y} - \widehat{\mathbf{y}}\|^2/(n-p)$	σ^2
Total around overall mean	$\|\mathbf{y} - \overline{\mathbf{y}}\|^2$	$n-1$	–	–

We test the significance of the dependence on covariates by means of the test statistic

$$F = \frac{\|\widehat{\mathbf{y}} - \overline{\mathbf{y}}\|^2/(p-1)}{\|\mathbf{y} - \widehat{\mathbf{y}}\|^2/(n-p)}.$$

Under H_0, this follows an $F_{p-1,n-p}$ distribution. In the class Analysis of Variance ("Angewandte Varianzanalyse und Versuchsplanung"), similar (but more complex) decompositions of the sum of squares $\|\mathbf{y} - \overline{\mathbf{y}}\|^2$ are analyzed in special linear models with a range of F-tests.

We could also examine the distribution of the test statistic F under the alternative hypothesis $H \cap (\neg H_0)$, and thus study the power of the F-test. The distribution we obtain is the so-called non-central F distribution $F_{p-1,n-p,\delta^2}$ with **non-centrality parameter** $\delta^2 = \|\mathbf{E}[\mathbf{y}] - \mathbf{E}[\overline{\mathbf{y}}]\|^2$ (although the literature defines this non-centrality parameter in more than one way). See the literature for more details.

One important quantity is the quotient

$$R^2 := \frac{\|\widehat{\mathbf{y}} - \overline{\mathbf{y}}\|^2}{\|\mathbf{y} - \overline{\mathbf{y}}\|^2}.$$

This is the **coefficient of determination**, the proportion of variance explained by the model. It measures the **goodness of fit** of the model with explanatory variables $\mathbf{x}^{(j)}$. It is not difficult to see that R^2 is also the maximum squared correlation of $\mathbf{y}$ with an arbitrary linear combination of the columns $\mathbf{x}^{(j)}$. The coefficient of determination is also equal to the square of the **multiple correlation coefficient** between $\mathbf{y}$ and the $\mathbf{x}^{(j)}$. The linear combination maximizing the correlation with $\mathbf{y}$ is the least squares estimate $\widehat{\mathbf{y}}$ itself.

Remark: R^2 and F are **at first** the **most important** numbers in the computer output.

1.6 Simple linear regression

1.6.1 Results for the special case of simple linear regression

We have already derived the least squares estimators. Here we merely give explicit formulæ for the most important test statistics and confidence intervals:

Test of the null hypothesis $\beta = \beta_0$ at level γ: Reject the null if

$$\frac{|\widehat{\beta} - \beta_0|}{\widehat{\sigma}/\sqrt{SS_X}} > t_{n-2;1-\gamma/2},$$

where

$$SS_X = \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Correspondingly, the confidence interval for β is

$$\widehat{\beta} \pm t_{n-2;1-\gamma/2} \cdot \frac{\widehat{\sigma}}{\sqrt{SS_X}}.$$

The confidence interval for the expectation of a new observation at x_0 (i.e. the value of the regression line at x_0) is:

$$\widehat{\alpha} + \widehat{\beta} x_0 \pm t_{n-2;1-\gamma/2} \cdot \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{SS_X}}.$$

The confidence interval for the entire regression line (simultaneously for all x) is

$$\widehat{\alpha} + \widehat{\beta} x \pm \sqrt{2F_{2,n-2;1-\gamma}} \cdot \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{SS_X}}.$$

Naturally, the simultaneous confidence interval is better than the individual one. Finally, the prediction interval for a new observation at x_0:

$$\widehat{\alpha} + \widehat{\beta} x_0 \pm t_{n-2;1-\gamma/2} \cdot \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{SS_X}}.$$

The prediction interval is wider than the confidence interval. All three intervals have boundaries that form hyperbolae.

1.6.2 Regression and correlation

The concept of **correlation** used to be applied more often than regression was.

Let Y **and** X be random variables, i.e. the data $x_1, \ldots, x_n$ are no longer considered to be fixed.

Definition 1.6.1. *The correlation ("**Pearson product moment correlation coefficient**") is defined as:*

$$\rho = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}} \quad (\textit{if } \mathrm{Var}(X) \neq 0,\ \mathrm{Var}(Y) \neq 0)$$

Properties of correlation:

(i) $-1 \leq \rho \leq +1$ (Schwarz inequality)

(ii) $|\rho| = 1 \Leftrightarrow$ The joint distribution of X and Y is concentrated on a line (and the sign of ρ matches the sign of this lines gradient).

(iii) If $\rho = 0$, X and Y are said to be uncorrelated.

(iv) ρ can be estimated by

$$r = \widehat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

and for this estimate $\widehat{\rho}$, we have:

- $-1 \leq \widehat{\rho} \leq 1$
- $|\widehat{\rho}| = 1 \Leftrightarrow$ all the points lie on a single line
- sign $(\widehat{\rho})$ = sign $(\widehat{\beta})$

In Figure 1.10, we can see some typical scatterplots using a variety of correlation coefficients.

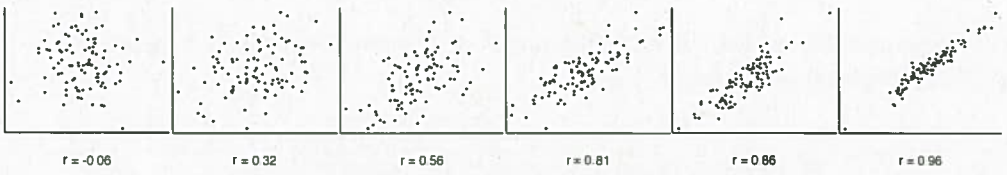

Figure 1.10: Scatterplots with various correlation coefficients.

The z transformation ("variance-stabilizing transformation for the correlation coefficient") (Fisher). Let (X,Y) be jointly normally distributed. Define

$$z := \tanh^{-1}(\widehat{\rho}) = \frac{1}{2}\log\left(\frac{1+\widehat{\rho}}{1-\widehat{\rho}}\right).$$

Then for an arbitrary value of ρ, we have the following good approximation (for about $n > 10$)

$$z \sim \mathcal{N}\left(\tanh^{-1}(\rho), \frac{1}{n-3}\right).$$

Graphical interpretation of the z transformation:

- If the true value of ρ is near 0, the variance of $\widehat{\rho}$ is **high**.
- If the true value of ρ lies near ± 1, the variance of $\widehat{\rho}$ is **small**.

The z transformation rescales so as to make the variance constant (i.e. it "compresses in the middle" und "stretches at the edges").

If we want to test $\rho = 0$ against $\rho \neq 0$, we have 3 tests to choose from:

1. Table or diagram (see Figure 1.11)
2. t- or F-test of $\beta = 0$
3. $\tanh^{-1}$ transformation.

The first and third methods also allow testing for any other fixed value of ρ (and thus the construction of confidence intervals).

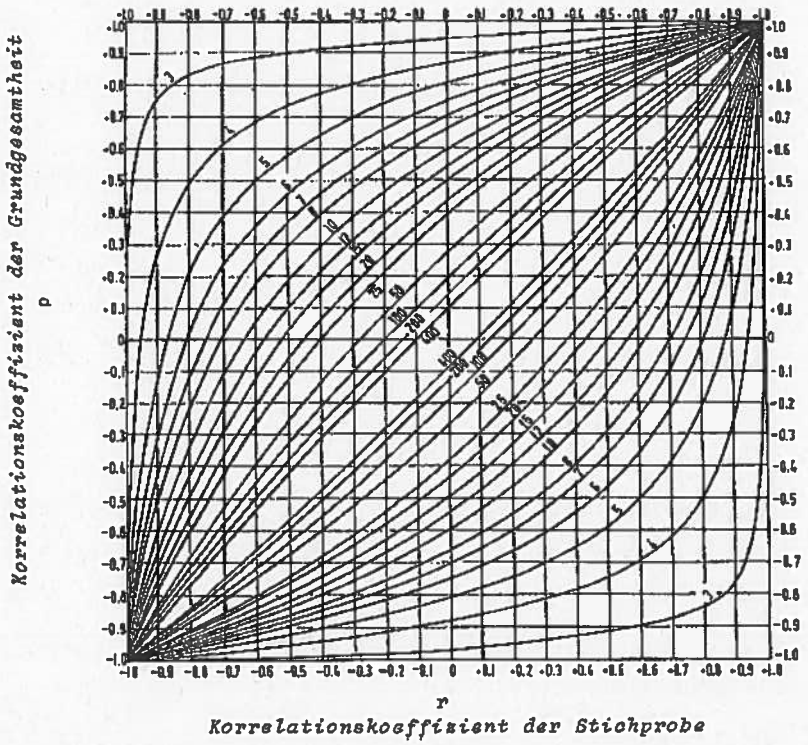

Figure 1.11: Confidence limits of the correlation coefficient. On the horizontal axis: the correlation coefficient r of the sample; on the vertical axis: the true correlation coefficient ρ. The labels on the curves denote the sample size (from F.N. DAVID: Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples, The Biometrika Office, London 1938)

Rank correlation: Since Pearson's correlation is not robust towards outliers (see Figure 1.12), some sort of **rank correlation** is often used. There are two types: that of Spearman and that of Kendall:

"Spearman's rank correlation" is simply the Pearson correlation of the **ranks**

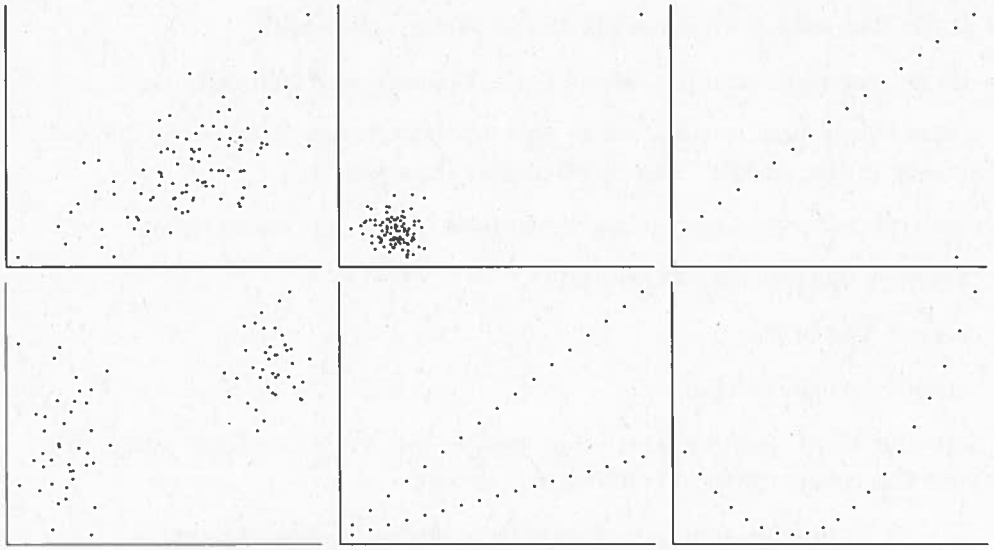

Figure 1.12: Various types of point clouds with correlation r = 0.7

of the X_i with those of the Y_i. As the sum of ranks (or of their squares) is fixed, the formulæ can be simplified thus:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \qquad D_i := Rg(X_i) - Rg(Y_i).$$

Kendall's rank correlation is defined as

$$r_K = 2 \cdot \frac{T_k - T_d}{n(n-1)},$$

where $\begin{array}{lll} T_k & = \#\ \textbf{concordances} & = \#\ \text{pairs with}\ \ (x_i - x_j)(y_i - y_j) > 0 \\ T_d & = \#\ \textbf{discordances} & = \#\ \text{pairs with}\ \ (x_i - x_j)(y_i - y_j) < 0 \end{array}$

Addendum: partial correlations

Let X, Y and Z be random variables. Then the partial correlation between X and Y given Z is defined as:

$$\rho_{XY.Z} := \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{(1-\rho_{XZ}^2)(1-\rho_{YZ}^2)}}, \quad \text{or estimated as:} \quad r_{XY.Z} := \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}}$$

This measures the strength and direction of the linear dependence between X and Y after accounting for the linear dependence of X and Y on Z.

1.6.3 Switching X and Y; regression to the mean

If both X and Y are considered to be random, we can write the least squares estimation line as follows:

$$y - \overline{y} = \widehat{\rho}\frac{\widehat{\sigma}_Y}{\widehat{\sigma}_X}(x - \overline{x}).$$

Since $\widehat{\rho}$ is always less than 1 in absolute terms, the prediction for Y is always closer to the mean than the corresponding value of X, if we measure distances in terms of standard deviations. If for instance $\widehat{\rho}$ is positive and our observation of X lies e.g. 1 standard deviation above the mean, the prediction of the corresponding value of Y will be **less** than one standard deviation above the mean. In other words, we always predict a regression (return) to the mean; this has given rise to the name "regression".

This phenomenon is continually being rediscovered, and is often interpreted at length in the framework of cultural pessimism. However, as our formulae show, this phenomenon is a very general one that occurs continually and requires no special interpretation. It lies in the very nature of prognoses that as more observations are available in the middle, any prognosis tends towards the mean.

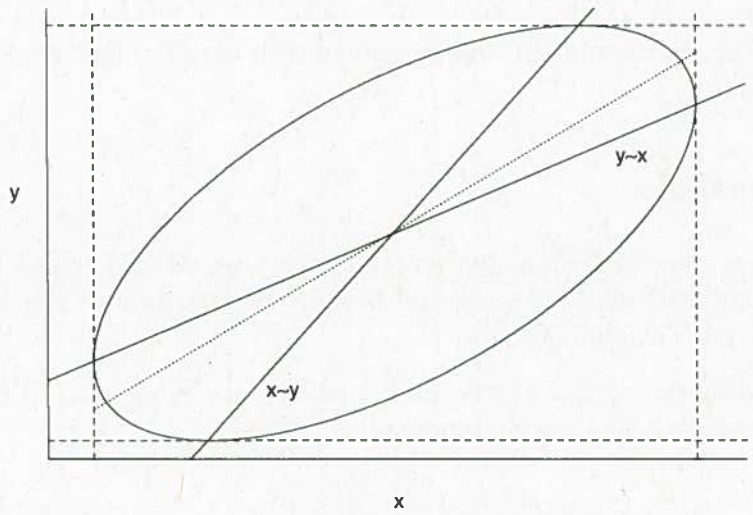

Figure 1.13: Regression lines "y versus x" und "x versus y"

Thus such a regression to the mean is nothing special. This becomes even clearer if we switch the roles of X and Y. Due to symmetry, the regression line for X by Y looks like

$$x - \overline{x} = \widehat{\rho}\frac{\widehat{\sigma}_X}{\widehat{\sigma}_Y}(y - \overline{y}).$$

In other words, if we look back and ask what value of X led to Y being one standard deviation above the mean, the answer will be "less than one standard deviation". Now we might be tempted to see this as a sign of progress rather than regression!

Drawing both regression lines in the same plot, we can see how the gradients

$$\widehat{\rho}\frac{\widehat{\sigma}_Y}{\widehat{\sigma}_X} \text{ and } \frac{1}{\widehat{\rho}}\frac{\widehat{\sigma}_Y}{\widehat{\sigma}_X}$$

differ. These two regression lines are quite evidently not the same; instead, we have the scissor-like gap shown in Figure 1.13).

If (X, Y) follows a two-dimensional normal distribution, the contours of the joint density are ellipses. The regression lines intersect a contour at points in which it has vertical or horizontal tangents, respectively, as these points are maxima of the conditional density of X given $Y = y$ or Y given $X = x$, respectively.

1.7 Analyzing residuals, verifying the model and dealing with breaches of assumptions

Residual analysis is the process of graphically (and some times also numerically) analyzing the residuals, i.e. the error estimates

$$r_i := \widehat{\varepsilon}_i = y_i - \widehat{y}_i,$$

in order to verify the assumptions on the model after fitting it, and to develop a better model.

1.7.1 Normal plot

Assumptions on the distribution can quite generally be checked with a quantile/quantile plot (QQ plot). If we are specifically checking against the normal distribution, we call it a normal plot.

First we introduce the normal plot for i.i.d. random variables $X_1, \ldots, X_n$. The "empirical cumulative distribution function" is defined as

$$u = F_n(x) = \frac{1}{n}\#\{X_i \leq x\}.$$

This is a step function which approaches the true distribution function when n becomes large (Lemma of Glivenko and Cantelli). In particular, we have

$$F_n(x) \longrightarrow \Phi(\frac{x-\mu}{\sigma})$$

if the X_i follow a normal distribution. Thus if we set

$$z := \Phi^{-1}(F_n(x)),$$

then we obtain

$$z \approx \frac{x-\mu}{\sigma}$$

for sufficiently large n. In a normal plot, we plot x against z at selected points. If X_i really does follow a normal distribution, the normal plot will roughly exhibit a straight line whose intercept and slope are μ and σ, respectively. However, the random fluctuations of the data do lead to some deviation away from an exact line. We can get an idea of the size of such fluctuations by performing simulations – see Figure 1.14.

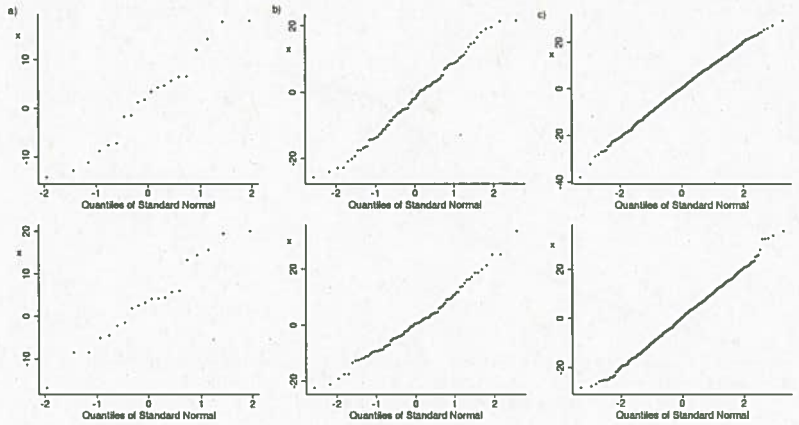

Figure 1.14: QQ plots for a normally distributed random variable X, from which a) 20, b) 100 and c) 1000 samples are taken.

If the assumption of normality is incorrect, the normal plot shows systematic deviations from a straight line. Some typical cases are given in Figure 1.15. However, the interpretation of this plot is not always clear, as the boundaries between the different types of situations are continuous. For instance, the case illustrated in Figure 1.16 can be interpreted either as a mixture of two groups or as a short-tailed distribution.

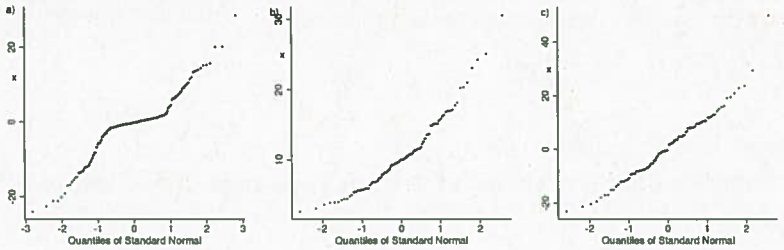

Figure 1.15: a) Heavy-tailed distribution b) Skew distribution c) Outliers

There is also a formal test of normality based on the normal plot, the "Shapiro-Wilks test". It essentially measures the correlation of the point cloud seen in the normal plot.

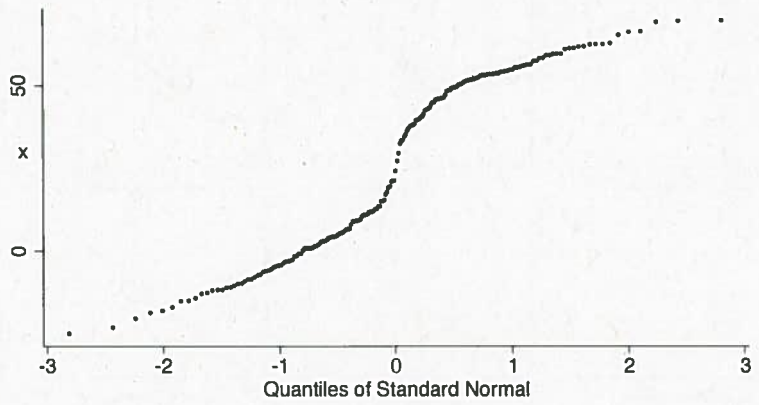

Figure 1.16: Two groups or short-tailed distribution?

Making these plots

- For small n (ca. $n \leq 100$):
  Plot all observations individually along the vertical axis. This automatically gives us the order statistics (ordered observations) $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. By definition, we have $F_n(X_{(i)}) = i/n$, but the effects of the jumps of F_n are generally toned down by plotting a slightly modified quantity on the horizontal axis, such as $\Phi^{-1}(\frac{i-1/2}{n})$ (or $\Phi^{-1}(\frac{i}{n+1})$, $\Phi^{-1}(\frac{i-(3/8)}{n+(1/4)})$ or $\Phi^{-1}(\frac{i-(1/3)}{n+(1/3)})$.

- For large n (ca. $n \geq 100$):
  Choose some equidistant values of x from the range of the sample (horizontal axis).

- The horizontal axis is frequently labelled $u = \Phi(z)$. This non-linear scale ensures that the plot of (u, x) is a line. Thus the points $F_n^{-1}(u)$ on the vertical axis can easily be plotted without any calculations. In the time before computers (when special "probability graph paper" was used, this was especially important.

- The axes are often switched around, i.e. z is plotted against x.

We have so far not paid any attention to the errors ε_i while performing regression. However, we can use the normal plot of the residuals $\widehat{\varepsilon}_i$ or the standardized residuals $\widehat{\varepsilon}_i/\sqrt{1-P_{ii}}$. Recall that the residuals $\widehat{\varepsilon}_i$ neither have constant variance, nor are they uncorrelated. Standardizing makes their variance constant again, though this effect is usually negligible.

In summary: the **normal plot** checks the normality of the residuals against possible skewness, heavy- (or short-)tailedness, outliers and other peculiarities.

1.7.2 Tukey-Anscombe plot

The Tukey-Anscombe plot is a plot of the residuals against the fitted values $\widehat{y}_i$. We always have $\sum r_i \widehat{y}_i = 0$, i.e. the Tukey-Anscombe plot always has sample correlation zero. If this plot exhibits a non-linear structure, this is an indication that the model assumptions are broken. If the residuals were plotted against the y_i, their correlation would make any interpretation more difficult.

For a simple linear regression, this is (essentially) equivalent to plotting the r_i against the x_i (unlike multiple regression, for which the plot of r_i against $\widehat{y}_i$ is more informative than the component-wise plots of r_i against x_{ij}). Likewise, this plot is quite similar to the original scatterplot of y_i against x_i for a simple linear regression. It is, however, easier to recognize deviations from a horizontal line than deviations from a sloping line.

The ideal look of a Tukey-Anscombe plot is shown in Figure 1.17. One frequent deviation from the assumption of constant error variance is a variance that increases as the target variable does. The effect of this on the Tukey-Anscombe plot can be seen in Figure 1.18 a)–c). If the Tukey-Anscombe plot contains some structure in the form of a "trend", this is an indication that the regression function has not been specified correctly (i.e. the mean error is not zero.) Figure 1.18d is a typical example of a case in which a quadratic term is most likely missing.

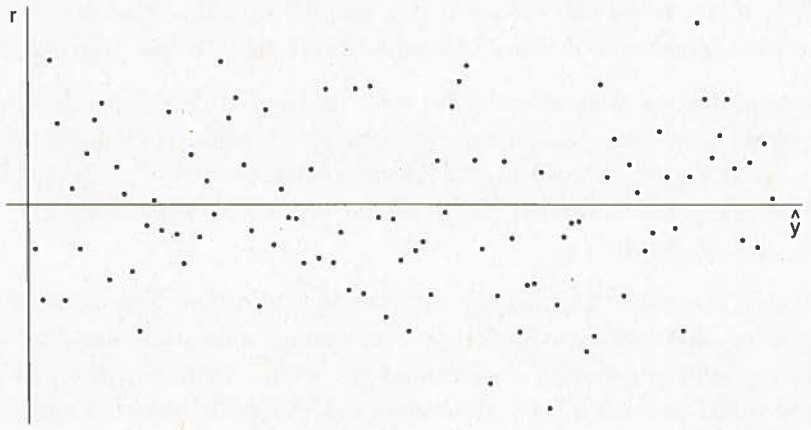

Figure 1.17: The Tukey-Anscombe Plot for an example in which the model assumptions hold true.

If the Tukey-Anscombe plot shows some sort of connection between the error variance and $\widehat{y}_i$ or the variables in $\mathbf{x}$, the target variables should be transformed or a "weighted regression" performed (see Section 1.7.5). If the spread of the errors increases linearly with the fitted values, a logarithmic transformation will stabilize the variance; and if the error spread is proportional to the square root of the fitted values, taking the square root of the target variable stabilizes the variance. (This can be shown using Taylor expansions.)

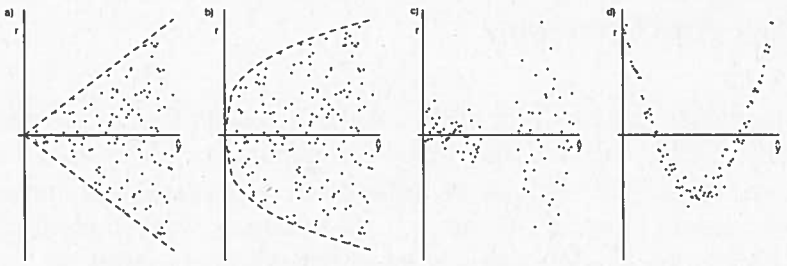

Figure 1.18: a) Linear increase in standard deviation, b) non-linear increase in standard deviation, c) 2 groups with differing variance, d) missing quadratic term.

1.7.3 Time series plot, Durbin-Watson test

If the errors are dependent, the levels of the tests and confidence intervals are no longer correct. This can be seen quite easily: If $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \Sigma)$, a simple calculation shows that

$$\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_p(\boldsymbol{\theta}, (X^T X)^{-1}(X^T \Sigma X)(X^T X)^{-1}).$$

The size of the correlation effects between the errors also depends on the covariates X and on the shape of the covariance matrix Σ. In many cases, this effect is considerable.

We cannot get far without making some assumptions about the dependence structure: even if we knew the errors ε_i, we cannot estimate the $n(n+1)/2$ entries of the covariance matrix from n data without some extra information.

If the observations form a time series, their covariance often is a (usually monotonely decreasing) function of the time between observations. Dependence of this type can be detected by plotting the residuals r_i against the observation times t_i. Where these are unknown, the position $k(i)$ of the observation in the series may be used instead.

If the points vary randomly around the horizontal axis in the time series plot, everything is fine. However, if adjacent r_i are similar, this indicates that the errors may be serially correlated. Sometimes we even observe a jump in the level of the residuals. In such a case, the model has evidently changed suddenly at a particular point in time.

It is possible to test independence against an alternative of serial correlation. Two such tests are:

(i) The **run test**, which counts the number of continuous sub-sequences ("runs") in which the residuals have identical signs. When independence is assumed, there should not be too many or too few runs.

(ii) The **Durbin-Watson test**, which uses the test statistic

$$T = \frac{\sum_{i=1}^{n-1}(r_{i+1} - r_i)^2}{\sum_{i=1}^{n} r_i^2}.$$

Expanding this, we find that

$$T \approx 2\left(1 - \frac{\sum_{i=1}^{n-1} r_i r_{i+1}}{\sum_{i=1}^{n} r_i^2}\right).$$

(The deviation stemming from this approximation is marginal.) The quotient in this formula is an estimate of the correlation of ε_i and ε_{i+1} (assuming that all the ε_i have the same variance). If the ε_i are independent, T is approximately 2; lower values of T indicate positive dependence.

Finding the critical values for this test is made all the more difficult by the dependence of the distribution of r_i, and thus the distribution of T, on the experimental design, i.e. on the chosen $\mathbf{x}_i$. The test only looks at the extremes over all designs, which leads to two different tabulated values (see e.g. Sen und Srivastava (1990), p. 326). If T is smaller than the lower tabulated value, the null hypothesis of "independence" is rejected. If T is larger than the upper tabulated value, the null hypothesis is kept, and between these two critical values the situation depends on the $\mathbf{x}_i$, i.e. the test "abstains".

The drawback of the Durbin-Watson test is its exclusive focus on the correlation between observations that immediately follow each other.

Recall that the consequences of serial correlation also depend on the experimental conditions (the values of the covariates). If we can freely choose the order of our observations, then – in a univariate case – it would be tempting to make our observations in ascending order of x, as this often is the easiest to do. In such an experimental setup, however, the effect of positive error correlation on the slope estimate is particularly large – and thus it is not a good choice. In other words, error trends in time are mixed with the effect of the covariate.

It is much better to choose values of x that are as "orthogonal to time" as possible, i.e. such that the x_i and the t_i (or $k(i)$) are uncorrelated. Then error trends will cancel each other out during slope estimation (linear trends will cancel out exactly, non-linear ones approximately) and positive error correlation will thus only have a small effect on the variance of the slope estimate. The same is not true for the intercept, though: there the variance of the arithmetic mean is large when the observations have high correlation. The easiest way to obtain near-orthogonality of covariates to time is by randomizing the order of the values of x.

1.7.4 Interior analysis

"Interior analysis" is the local estimation of error variance using replicates or "near replicates". "Replicates" are repeated measurements at the same $\mathbf{x}_i$, and "near replicates" are measurements made close to $\mathbf{x}_i$ (cf. Figure 1.19). "Interior analysis" checks for a "lack of fit", i.e. a systematic error in the class of fitted models. In other words, it checks the assumption $\mathbf{E}[\varepsilon_i] = 0$.

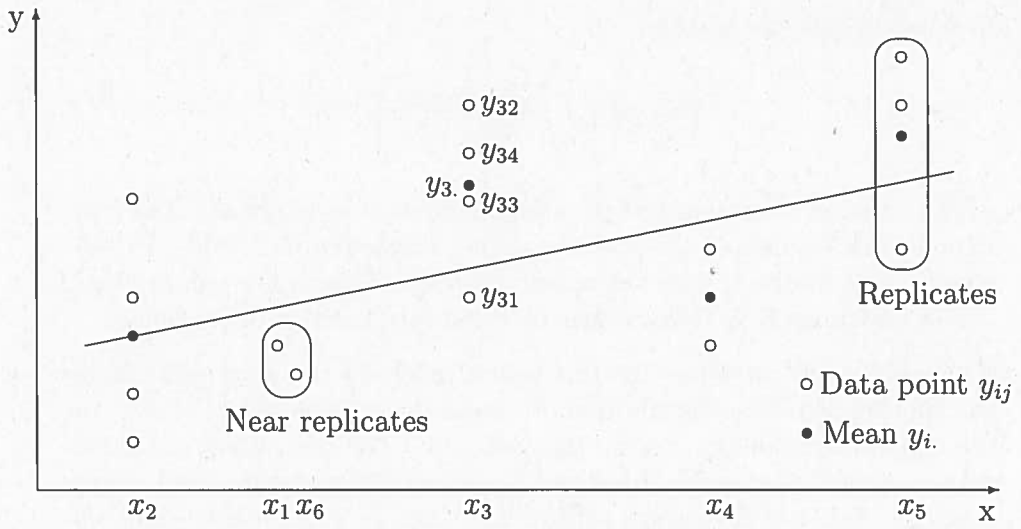

Figure 1.19: Replicates

The model with n_i replicates at $\mathbf{x}_i$ can be formulated thus:

$$Y_{ij} = \mathbf{x}_i^T\theta + \varepsilon_{ij} \quad (i = 1, \ldots, k; j = 1, \ldots, n_i),$$

where ε_{ij} *i.i.d.* $\sim \mathcal{N}(0, \sigma^2)$. The sample size is $n = \sum_{i=1}^{k} n_i$. We can compare this model to the larger model where Y and X are connected by an arbitrary function f. Introducing the quantities $\mathbf{E}[Y_{ij}] = f(\mathbf{x}_i) = \mu_i$ as independent parameters, we have:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad (i = 1, \ldots, k; j = 1, \ldots, n_i).$$

The least squares estimate of μ_i is simply the (arithmetic) mean of the observations at $\mathbf{x}_i$:

$$\widehat{\mu}_i = y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}/n_i.$$

Thus we have two nested models and can perform the usual F-test. The orthogonal decomposition for the ANOVA table is:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \widehat{y}_i)^2 = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - y_{i\cdot})^2}_{(1)} + \underbrace{\sum_{i=1}^{k} n_i(y_{i\cdot} - \widehat{y}_i)^2}_{(2)}$$

- (1) measures the random error, and has $\sum_{i=1}^{k}(n_i - 1) = n - k$ degrees of freedom
- (2) measures the random error and the "lack of fit", and has $k - p$ degrees of freedom

If we have two near replicates, we can correct them by moving each of them parallel to the regression line so that they both have the average of their covariates $\mathbf{x}$ (cf. Figure 1.20). This then gives us two replicates. When p is large, however, it can be hard to find near replicates.

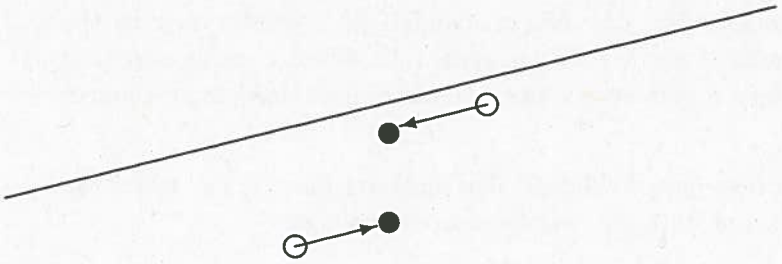

Figure 1.20: Correcting near replicates

1.7.5 Generalized least squares, weighted regression

Model: $\boxed{\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}}$ as previously, **but now with** $\boxed{\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\ \sigma^2\Sigma)}$

.

We assume Σ to be known and σ^2 unknown – i.e. the error covariance matrix is given up to a multiplicative constant. We furthermore assume that Σ is **positive definite**. Then there exists a regular matrix A for which $AA^T = \Sigma$, i.e. A is a square root of Σ (cf. Appendix A).

Reduction to the standard model: We compute the transform

$$\tilde{\mathbf{y}} := A^{-1}\mathbf{y} = A^{-1}(X\boldsymbol{\theta} + \boldsymbol{\varepsilon}) = \underbrace{A^{-1}X}_{\tilde{X}}\boldsymbol{\theta} + \underbrace{A^{-1}\boldsymbol{\varepsilon}}_{\tilde{\boldsymbol{\varepsilon}}} = \tilde{X}\boldsymbol{\theta} + \tilde{\boldsymbol{\varepsilon}}$$

Then we have

$$\begin{aligned} \mathbf{E}[\tilde{\boldsymbol{\varepsilon}}] &= \mathbf{E}\left[A^{-1}\boldsymbol{\varepsilon}\right] = A^{-1}\mathbf{E}[\boldsymbol{\varepsilon}] = \mathbf{0} \\ \operatorname{Cov}[\tilde{\boldsymbol{\varepsilon}}] &= \operatorname{Cov}[A^{-1}\boldsymbol{\varepsilon}] = A^{-1}\operatorname{Cov}[\boldsymbol{\varepsilon}](A^{-1})^T \\ &= A^{-1}\sigma^2(AA^T)(A^{-1})^T = \sigma^2 I. \end{aligned}$$

In other words, the "tilde model" obtained by the linear transformation using A^{-1} satisfies the conditions for the standard multiple regression model we already know. The key point here is that A is invertible (which is ensured by the positive definiteness of Σ).

Applying the known theory to the "tilde model":

For the "tilde model", we estimate $\boldsymbol{\theta}$ using least squares, that is, we minimize

$$\|\tilde{\mathbf{y}} - \tilde{X}\boldsymbol{\theta}\|^2 = (\mathbf{y} - X\boldsymbol{\theta})^T A^{-T}A^{-1}(\mathbf{y} - X\boldsymbol{\theta}) = (\mathbf{y} - X\boldsymbol{\theta})^T\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\theta}).$$

This is the same as performing a least squares estimation for the original data $(\mathbf{y}, X)$ using a different scalar product. The estimate obtained by this is

$$\widehat{\boldsymbol{\theta}} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{\mathbf{y}} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{y},$$

which is the so-called **generalized least squares estimate** of $\boldsymbol{\theta}$. Its distribution is

$$\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_p(\boldsymbol{\theta},\ \sigma^2(X^T\Sigma^{-1}X)^{-1}).$$

Tests, confidence intervals, etc. can be constructed in a similar way to the standard case. The results of Section 1.9 will show that if $\Sigma \neq I$, the generalized least squares method has a smaller variance than the standard least squares method.

One important special case occurs when Σ is a diagonal matrix, i.e. when the errors are uncorrelated, but have varying degrees of precision:

$$\Sigma = \begin{pmatrix} v_1 & & 0 \\ & v_2 & \\ & & \ddots \\ 0 & & v_n \end{pmatrix} \qquad (v_i > 0 \quad \forall i)$$

In such a case we introduce weights w_i which are proportional to $\frac{1}{v_i}$, i.e. we minimize $\sum_i w_i r_i^2$. The more precise an observation, the greater a weight is given to it by the generalized least squares procedure.

It is, however, a fairly rare occurrence that we would actually know the error covariance matrix up to a constant factor. If we do not have such information, we often first use standard least squares, estimate a covariance matrix $\widehat{\Sigma}$ from the residuals, and then use this estimated covariance matrix to perform generalized least squares. This procedure is frequently used when the errors exhibit correlation in time, and it is usually referred to as the Cochrane-Orcutt procedure.

1.8 Model selection

We assume that our observations have been generated by the model

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \qquad (i = 1, \ldots, n),$$

where ε_i are i.i.d. with $\mathbf{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Moreover, we model the regression function as

$$f(\mathbf{x}_i) \approx \sum_{j=1}^{p} \theta_j x_{ij}.$$

Here we always assume an intercept to be present, i.e. $x_{i1} = 1$ for all i.

Question: Which variables should we include in the model?

Naive answer: "More is always better!".

At a general level, this answer is wrong, as it is quite possible that some of these p variables are "superfluous" or make only a very marginal contribution to the explanatory power of the model. At the same time, estimating more coefficients raises the random error of our parameter estimates and predictions. Thus this answer is not correct.

Our choice of variables may be determined by theoretical considerations in the our field of work (such as in physics). This is not model selection in a

statistical sense. Furthermore, we do not consider transformations of y and $\mathbf{x}$, unlike before. The only question is whether or not a variable (that may have been transformed earlier on) is included or omitted.

Our search for the "best model" depends on the question we are trying to answer:

(i) Using regression to find an explanatory model

(ii) Using regression to make predictions

We shall first discuss stepwise methods, and then we shall look at methods which evaluate all of the 2^{p-1} possible models according to some suitable criterion and choose the "best" one. Stepwise methods of course take less effort.

1.8.1 Model selection using "stepwise regression"

(1) **Forward stepwise regression:**

Start with the model containing only the intercept, i.e. with only the constant x_{i1}:

$$y_i = \theta_1 + \varepsilon_i.$$

(Of course the estimate of θ_1 is then merely the average of all observations). Now the variables are included one by one (and "stepwise"); at each step the model with the most significant F–value over the previous model is included.

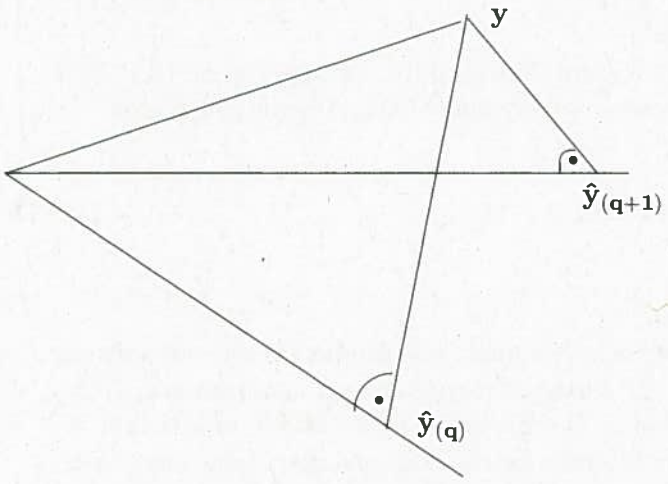

Figure 1.21: Previous model (q-dimensional) and new enclosing model ($(q+1)$-dimensional, i.e. with one more variable).

Stopping condition: Repeat until none of the F–statistics are significant (for a given significance level). As we are carrying out repeated tests, we should take care not to misinterpret this level.

(2) **Backwards stepwise regression:**

Start with the full model:

$$y_i = \theta_1 + \theta_2 x_{i2} + \ldots + \theta_p x_{ip} + \varepsilon_i$$

Now the variables are taken out of the model one by one (and "stepwise"). At each step, we remove the variable whose F–value in the comparative test is smallest, and continue until all such F are significant.

Discussion of stepwise regression:

- "Backward" selection requires more effort, and under some circumstances it may even incur more numerical problems (if e.g. $p \geq n$, we cannot even carry out this procedure). However, it gives us the greatest certainty of finding a good model.
- "Forward" and "backward" selection are often combined (using two different significance levels, so as to avoid entering a loop of removing and including the same $\mathbf{x}^{(j)}$).
- The stopping rule does not necessarily gives us a "best" model according to the criteria we shall discuss in the next section.
- The sequence of included or removed variables should not be regarded as some kind of rank of the importance of the variables.
- "Forward" and "backward" selection may yield entirely different solutions.

Example of this last point: We choose three explanatory variables such that

- X_1 and X_2 each have only very little correlation with Y, but Y is (almost) a linear combination of X_1 and X_2.
- X_3 correlates strongly with Y.

Forward selection will first select X_3 and then stop (or perhaps choose $\{X_1, X_3\}$ or $\{X_2, X_3\}$), while backward selection will choose $\{X_1, X_2\}$ and then stop.

1.8.2 Model selection criteria

Mallows' C_p statistic

The C_p statistic is an estimate of the mean quadratic prediction error of a fitted model which is averaged over the observed experimental conditions $\mathbf{x}_i$ ($i = 1, \ldots, n$). It also takes into account the bias of a badly-fitting model, but it requires a good (bias-free and sufficiently exact) estimate of σ^2 (e.g. one taken from a "full" model that might contain far too many variables for its fit to be used, one estimated using replicates or "near replicates", or one known from experience). It automatically "penalizes" superfluous variables and can be used as a quality measure (estimated) of a model.

By the assumptions made at the beginning of this section, the variables y_i are independent with mean $\mathbf{E}[y_i] = f(\mathbf{x}_i) = \mu_i$ and variance $\mathrm{Var}(y_i) = \sigma^2$. Each model is described by the subset $M \subset \{1, 2, \ldots, p\}$ of variables included in it. Here we assume that $x_{i,1} \equiv 1$ for each index i, and that each M contains the

index 1. We denote the corresponding matrix of covariates by X^M (a submatrix of X), i.e.

$$X^M = (x_{ij}; 1 \leq i \leq n, j \in M).$$

Next we use least squares to estimate the parameter of the model M by

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

and the mean vector $\mu_i = \mathbf{E}[y_i]$ by

$$\widehat{\mathbf{y}}^M = X^M \widehat{\boldsymbol{\theta}}^M.$$

So we a fit a linear model M with $|M|$ variables – which may or may not be correct. For instance, the fit may have the shape of a straight line, even though the expectations $\mathbf{E}[y_i]$ lie on a parabola; cf. Figure 1.22.

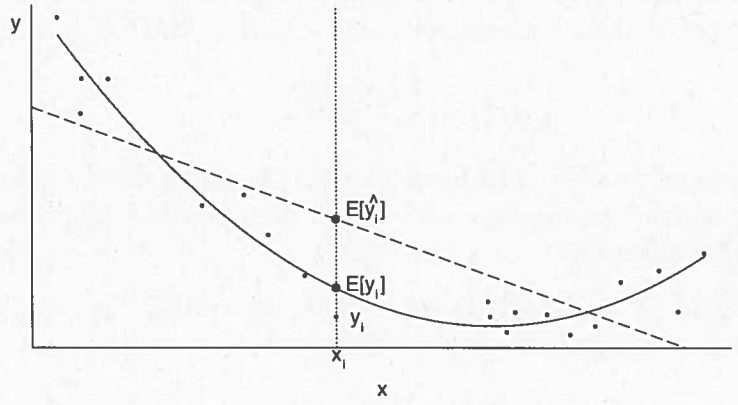

Figure 1.22: Theoretical best linear fit (dashed line) for a model that is actually quadratic (solid line).

If the model is correct, we have

$$\mathbf{E}[\widehat{\mathbf{y}}^M] = X^M((X^M)^T X^M)^{-1}(X^M)^T \boldsymbol{\mu},$$

which amounts to the best approximation of $\boldsymbol{\mu}$ by the variables in the model. Of course, this approximation gets better and better as we include more variables in our model (it usually even improves quite clearly, as in practice the influence of a variable is rarely zero, even if it is small). Furthermore, we have

$$\mathrm{Cov}(\widehat{\mathbf{y}}^M) = \sigma^2 X^M((X^M)^T X^M)^{-1}(X^M)^T,$$

i.e. the random fluctuations are as if the model were true. In particular, we have

$$\sum_{i=1}^{n} \mathrm{Var}(\widehat{y}_i^M) = \sigma^2 tr(X^M((X^M)^T X^M)^{-1}(X^M)^T) = |M|\sigma^2.$$

The more variables we include in our model, the greater the sum of variances of $\widehat{y}_i^M$.

Our quality measure for a model is the sum of mean squared errors $\widehat{y}_i^M$ from the true values μ_i:

$$SMSE = SMSE(M) = \mathbf{E}\left[||\widehat{\mathbf{y}}^M - \boldsymbol{\mu}||^2\right] = \sum_{i=1}^{n} \mathbf{E}\left[(\widehat{y}_i^M - \mu_i)^2\right].$$

Now for any random variable Z and each constant c , we know that

$$\mathbf{E}\left[(Z-c)^2\right] = \mathbf{E}\left[((Z - \mathbf{E}[Z]) + (\mathbf{E}[Z] - c))^2\right] = \mathrm{Var}(Z) + (\mathbf{E}[Z] - c)^2 + 2 \cdot 0,$$

and thus

$$SMSE = \sum_{i=1}^{n} \mathrm{Var}(\widehat{y}_i^M) + \sum_{i=1}^{n} (\mathbf{E}\left[\widehat{y}_i^M\right] - \mu_i)^2 = |M|\sigma^2 + \sum_{i=1}^{n} (\mathbf{E}\left[\widehat{y}_i^M\right] - \mu_i)^2.$$

The first of these summands is small for models containing few variables, while the second is small for models containing many variables. $SMSE$ is often scaled by σ^2, so that

$$\Gamma_p(M) = \frac{SMSE(M)}{\sigma^2}$$

becomes the term of interest. The inequality $\Gamma_p(M) \geq |M|$ always holds, with equality being attained exactly when the model M is bias-free (but potentially contains superfluous terms).

We can also regard $\widehat{y}_i^M$ as a prediction for a new observation $Y_{n+i} = \mu_i + \varepsilon_{n+i}$. In this case the sum of prediction square errors is

$$SPSE = \sum_{i=1}^{n} \mathbf{E}\left[(Y_{n+i} - \widehat{y}_i^M)^2\right] = \sum_{i=1}^{n} \mathbf{E}\left[(Y_{n+i} - \mu_i)^2\right] + \sum_{i=1}^{n} \mathbf{E}\left[(\widehat{y}_i^M - \mu_i)^2\right] = n\sigma^2 + SMSE.$$

(To be more precise, we should actually refer to this as "sum of mean squared prediction errors".)

Thus minimizing $SMSE$ or Γ_p or $SPSE$ always leads to the same model. However, we cannot compute any of these quantities without knowledge of σ and $\boldsymbol{\mu}$. One naive estimate of $SPSE$ is the sum of squared errors

$$SSE(M) = ||\mathbf{y} - \widehat{\mathbf{y}}^M||^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i^M)^2,$$

i.e. the criterion used in least squares estimation. However, this quantity always decreases when additional variables are included in the model, and it underestimates $SPSE$:

$$\begin{aligned} \mathbf{E}\left[||\mathbf{y} - \widehat{\mathbf{y}}^M||^2\right] &= \sum_{i=1}^{n} \mathrm{Var}(y_i - \widehat{y}_i^M) + \sum_{i=1}^{n} (\mathbf{E}[y_i] - \mathbf{E}\left[\widehat{y}_i^M\right])^2 \\ &= (n - |M|)\sigma^2 + \sum_{i=1}^{n} (\mathbf{E}[y_i] - \mu_i)^2 = SPSE(M) - 2|M|\sigma^2. \end{aligned}$$

It is therefore better to estimate $SPSE$ by $SSE(M)+2|M|\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ denotes an estimate of σ^2 (e.g. from the full model $M = \{1, 2, \ldots p\}$), and to then choose the model minimizing this estimate of $SPSE$. Similarly, we can use

$$C_p(M) := \frac{SSE(M)}{\widehat{\sigma}^2} - n + 2|M|,$$

as an estimate of Γ_p. Random fluctuations mean that C_p can become smaller than $|M|$ (or even negative), unlike Γ_p.

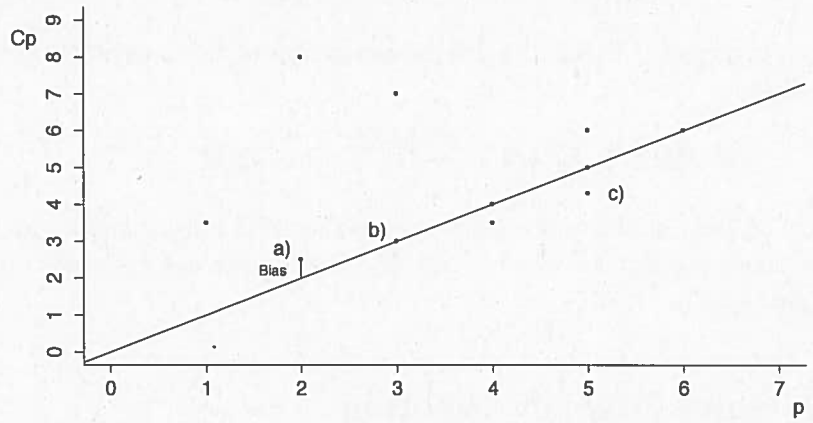

Figure 1.23: C_p plot: a) best prediction estimate, b) best bias-free prediction estimate, c) Point at which the random fluctuation of C_p causes it to be below the model size.

Akaike's information criterion AIC

For an arbitrary model (not necessarily a regression model) with k parameters, we define

$$AIC(\alpha) = -2\widehat{\ell}_k + \alpha\, k$$

where $\widehat{\ell}_k$ denotes the maximum log–likelihood in the model (i.e. the log-likelihood at the MLE). Thus the first term measures the goodness of fit, and the second penalizes the complexity of the model. The multiplicative constant in the penalty term is usually taken to be 2. When selecting between different candidate models, we can now choose the one which **minimizes the AIC**.

In particular, if we have a linear model with normally distributed errors and a selection $M \subset \{1, 2, \ldots p\}$ of explanatory variables, the log-likelihood is

$$\log f_M(\mathbf{y}, \boldsymbol{\theta}) = -\frac{n}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{1}{2\sigma^2}(\mathbf{y} - X^M\boldsymbol{\theta}^M)^T(\mathbf{y} - X^M\boldsymbol{\theta}^M).$$

Assuming σ^2 to be given (e.g. by experience or by prior investigation), we then see that

$$-2\widehat{\ell}_M = \frac{1}{\sigma^2}\underbrace{(\mathbf{y} - X^M\widehat{\boldsymbol{\theta}}^M)^T(\mathbf{y} - X^M\widehat{\boldsymbol{\theta}}^M)}_{SSE(M)},$$

(up to a constant), i.e. that $AIC(2)$ is equal to C_p up to a constant (whose exact value is irrelevant for model selection) when both use the same σ.

If furthermore σ is replaced by its maximum likelihood estimate

$$\widehat{\sigma}^2(M) = \frac{SSE(M)}{n},$$

Akaike's criterion is

$$AIC(\alpha) = n \log(\widehat{\sigma}^2(M)) + \alpha\,|M|$$

(again up to a constant). Using the Taylor expansion of the logarithm at σ^2, we conclude that

$$AIC(2) \approx n \log(\sigma^2) + \frac{SSE(M)}{\sigma^2} - n + 2|M|.$$

Thus we see that even in this more general case the AIC is very similar to C_p – at least for those models, in which $SSE(M)/n$ lies near the estimate of σ^2 used to compute C_p.

1.9 The Gauss-Markov theorem

This theorem states the "optimality" of the least-squares estimate in a certain sense. There is a version under the assumption of normality, and one without this assumption. The difference between these two is not only in the assumptions; there are also significant differences between the statements they make!

First we give the result that does not assume normality.
Theorem 1.9.1 (Gauss–Markov). *Let*

$$\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon} \qquad \mathbf{E}[\boldsymbol{\varepsilon}] = \mathbf{0} \qquad \mathrm{Cov}[\boldsymbol{\varepsilon}] = \sigma^2 I \qquad \mathrm{rank}[X] = p.$$

Furthermore, let $\mathbf{c}$ *be an arbitrary* p*-dimensional vector, and* $\widehat{\boldsymbol{\theta}}$ *the least squares estimator. Then* $\mathbf{c}^T\widehat{\boldsymbol{\theta}}$ *has minimal variance amongst all linear unbiased estimators of* $\mathbf{c}^T\boldsymbol{\theta}$.

Because of this, we also call least squares estimators "BLUE" ("best linear unbiased estimators").

The version assuming normality is as follows:
Theorem 1.9.2. *Let furthermore* $\boldsymbol{\varepsilon}$ *be normally distributed. Then* $\mathbf{c}^T\widehat{\boldsymbol{\theta}}$ *has minimal variance amongst all unbiased estimators of* $\mathbf{c}^T\boldsymbol{\theta}$.

Thus we say that least squares estimators are "UMVU" ("uniformly minimum variance unbiased"). Uniformly, as this holds for arbitrary values of $\boldsymbol{\theta}$ and σ^2.)

It is this theorem (not the Gauss–Markov theorem) that best justifies using least squares (besides the simplicity of this method). However, unbiasedness is not always absolutely necessary, and does not hold e.g. in Bayesian regression and in "ridge regression". We shall not discuss these methods here, though.

The omission of the normality assumption in the Gauss–Markov theorem is paid for by the restriction of this result to linear estimators. Crucially, though, **all linear** estimators are **less efficient** (i.e. have a higher variance), even if the deviation from normality is relatively small.

Illustrative examples: If we attempt to approximate the true error distributions of high-quality data by a t distribution with ν degrees of freedom, we find $\nu = 5 - 9$ degrees of freedom for data appearing to be "quite normal", but often only $\nu = 3$ degrees of freedom – and even $\nu = 1$ (Cauchy distribution) can occasionally occur. Now the asymptotic efficiency (which is e.g. the inverse ratio of the required sample sizes needed to obtain the same precision) of the least squares estimator compared to an asymptotically best estimator (such as maximum likelihood) under a t distribution with ν degrees of freedom is $= 1 - 6/(\nu(\nu+1))$ (for $\nu \geq 2$). Thus the actual efficiency of the least squares method is 80-93% for "quite normal" data ($t_5 - t_9$), and ca. 50% for "fairly normal" data (t_3). If we look at $\widehat{\sigma}^2$, the situation is much worse still !

Proof of the Gauss–Markov theorem Let $\mathbf{a}$ be an n–dimensional vector and a_0 a constant for which $\mathbf{a}^T\mathbf{y} + a_0$ constitutes an unbiased estimate of $\mathbf{c}^T\boldsymbol{\theta}$. Then we have

$$\mathbf{E}\left[\mathbf{a}^T\mathbf{y} + a_0\right] = \mathbf{a}^T X\boldsymbol{\theta} + a_0 = \mathbf{c}^T\boldsymbol{\theta}$$

for all $\boldsymbol{\theta}$. From this we conclude that $a_0 = 0$ and $\mathbf{a}^T X = \mathbf{c}^T \Leftrightarrow X^T\mathbf{a} = \mathbf{c}$.

The vector $\mathbf{a_{KQ}} = X(X^TX)^{-1}\mathbf{c}$ from the least-squares estimate is a special solution of $X^T\mathbf{a} = \mathbf{c}$. Furthermore, $\mathbf{a_{KQ}}$ is orthogonal to all solutions $\mathbf{a_h}$ of the homogeneous system of linear equations, $X^T\mathbf{a} = \mathbf{0}$, as $\mathbf{a_{KQ}}^T\mathbf{a_h} = \mathbf{c}^T(X^TX)^{-1}X^T\mathbf{a_h} = 0$. As $\mathrm{Cov}[\mathbf{Y}] = \sigma^2 I$, we thus can see that

$$\mathrm{Var}((\mathbf{a_{KQ}} + \mathbf{a_h})^T\mathbf{Y}) = \mathrm{Var}(\mathbf{a_{KQ}}^T\mathbf{Y}) + \mathrm{Var}(\mathbf{a_h}^T\mathbf{Y}) \geq \mathrm{Var}(\mathbf{a_{KQ}}^T\mathbf{Y}).$$

□

Proof of the variant Gauss–Markov theorem (using the multidimensional Cramér-Rao inequality, which is proven in the class on Mathematical Statistics):

We regard the following general situation: Let $(f_{\boldsymbol{\eta}}(\mathbf{y}))$ be a parametric family of strictly positive densities in $\mathbb{R}^n$. Let $\boldsymbol{\eta}$ be a variable parameter with values in an open subset of $\mathbb{R}^k$, and let $f_{\boldsymbol{\eta}}(\mathbf{y})$ be differentiable wrt $\boldsymbol{\eta}$. Our parameter of interest is $g(\boldsymbol{\eta})$, where g is an arbitrary real-valued function of $\boldsymbol{\eta}$. Then we have

Theorem 1.9.3 (Cramér-Rao). *If $T(\mathbf{y})$ is an arbitrary unbiased estimate of $g(\boldsymbol{\eta})$, i.e.*

$$\mathbf{E}_{\boldsymbol{\eta}}[T(\mathbf{y})] = g(\boldsymbol{\eta}) \quad \forall\boldsymbol{\eta},$$

then g is differentiable and

$$\mathrm{Var}_{\boldsymbol{\eta}}(T(\mathbf{y})) \geq \frac{\partial g}{\partial \boldsymbol{\eta}}^T I(\boldsymbol{\eta})^{-1} \frac{\partial g}{\partial \boldsymbol{\eta}},$$

where $I(\eta)$ denotes the so-called Fisher information matrix:

$$I(\eta) = \mathbf{E}_\eta \left[\frac{\partial \log f_\eta(\mathbf{Y})}{\partial \eta} \frac{\partial \log f_\eta(\mathbf{Y})}{\partial \eta}^T \right].$$

We now apply this to

$$f_\eta(\mathbf{y}) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp(-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\theta})^T(\mathbf{y} - X\boldsymbol{\theta})),$$

$$\eta = (\sigma^2, \boldsymbol{\theta}^T)^T$$

and

$$g(\eta) = \mathbf{c}^T\boldsymbol{\theta}.$$

After some calculations, we obtain the Fisher information

$$I(\eta) = \begin{pmatrix} \frac{n}{2\sigma^4} & 0 \\ 0 & \frac{1}{\sigma^2}X^TX \end{pmatrix},$$

i.e. the least squares estimate $\mathbf{c}^T\widehat{\boldsymbol{\theta}}$ attains the Cramér-Rao lower bound. It thus obviously has minimal variance. □

Chapter 2

Nonlinear and nonparametric methods

2.1 Robust methods

We generally call a statistical procedure for a parametric model robust if its properties do not change much when slight deviations from the model occur. For the linear model, the method of least squares is not robust, as heavy-tailed error distributions have serious ramifications ! The distribution of the estimates is fairly stable (cf. Section 1.4.3), but for comparatively small deviations from normality, there are better estimators than least squares. Consequently, the level of tests and confidence intervals based on least squares is robust, but their power is not.

One related effect is that least squares estimates and the subsequent tests and confidence intervals are extremely sensitive to individual outliers – and observations that look like outliers are produced quite often when the underlying error distribution is heavy-tailed.

We will now first take a closer look at the effects of an outlier on the least squares method; then we shall discuss more robust alternatives.

2.1.1 Influence of individual observations on the LSE

First we examine the effect that omitting or adding observations has on the least squares estimate. The following lemma of Gauss will be useful here:

Lemma 2.1.1. *Let A be an invertible matrix of dimension $p \times p$, and let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors of dimension p for which $\mathbf{b}^T A^{-1}\mathbf{a} \neq -1$. Then $A + \mathbf{a}\mathbf{b}^T$ is also invertible, and we find that*

$$(A - \mathbf{a}\mathbf{b}^T)^{-1} = A^{-1} + \frac{1}{1 - \mathbf{b}^T A^{-1}\mathbf{a}} A^{-1}\mathbf{a}\mathbf{b}^T A^{-1}.$$

Proof: exercise.

We denote by $\widehat{\boldsymbol{\theta}}^{(-i)}$ the least squares estimate when the i-th observation is omitted, and we furthermore use the shorthand notation

$$A = X^T X = \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T, \quad \mathbf{c} = X^T y = \sum_{j=1}^{n} y_j \mathbf{x}_j^T.$$

Then the above lemma tells us that

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{(-i)} &= (A - \mathbf{x}_i \mathbf{x}_i^T)^{-1}(\mathbf{c} - y_i \mathbf{x}_i) \\
&= A^{-1}\mathbf{c} - y_i A^{-1}\mathbf{x}_i + \frac{1}{1 - \mathbf{x}_i^T A^{-1}\mathbf{x}_i} A^{-1}\mathbf{x}_i \mathbf{x}_i^T A^{-1}(\mathbf{c} - y_i \mathbf{x}_i) \\
&= \widehat{\boldsymbol{\theta}} - y_i A^{-1}\mathbf{x}_i \left(1 + \frac{\mathbf{x}_i^T A^{-1}\mathbf{x}_i}{1 - \mathbf{x}_i^T A^{-1}\mathbf{x}_i}\right) + \mathbf{x}_i^T \boldsymbol{\theta} A^{-1}\mathbf{x}_i \frac{1}{1 - \mathbf{x}_i^T A^{-1}\mathbf{x}_i},
\end{aligned}$$

and thus

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}} &= -\frac{1}{1 - \mathbf{x}_i^T (X^T X)^{-1}\mathbf{x}_i} (X^T X)^{-1}\mathbf{x}_i (y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\theta}}) \\
&= -\frac{r_i}{1 - P_{ii}} (X^T X)^{-1}\mathbf{x}_i.
\end{aligned}$$

We see that the influence of the i-th observation depends on the i-th residual and also on the diagonal entry P_{ii} in the "hat matrix" (when using all observations, including the i-th). Thus a plot of the residuals r_i against the P_{ii} is often used to detect influential observations.

The difference of the parameter estimates is somewhat difficult to interpret, as it is a whole vector that moreover depends on how the covariates are scaled. One scalar invariant can be obtained by computing the length of $\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}}$ with respect to the metric defined by the estimated covariance matrix of $\widehat{\boldsymbol{\theta}}$:

$$D_i = \frac{(\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}})^T (X^T X)(\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}})}{p\widehat{\sigma}^2} = \frac{1}{p} \frac{r_i^2}{\widehat{\sigma}^2(1 - P_{ii})} \frac{P_{ii}}{1 - P_{ii}}.$$

We refer to D_i as Cook's distance. It is a simple function of P_{ii} and the square of the studentized residual $r_i/(\widehat{\sigma}\sqrt{1 - P_{ii}})$, and is frequently used as a diagnostic tool. Observations yielding a much larger value of D_i than the others should be looked at more closely or omitted altogether.

We can likewise regard the change that occurs when an observation is added at an arbitrary location $(y, \mathbf{x})$:

$$\Delta\widehat{\boldsymbol{\theta}} = \frac{1}{1 + \mathbf{x}^T (X^T X)^{-1}\mathbf{x}} (X^T X)^{-1}\mathbf{x}(y - \mathbf{x}^T \widehat{\boldsymbol{\theta}}).$$

We see that the least squares estimate can be changed arbitrarily much by a single new observation – in other words, the least squares estimator is not robust.

Moreover, this effect depends quite strongly on the location of the new observation. This formula can be made a little clearer if we assume that the $\mathbf{x}_i$ are chosen randomly and are i.i.d.. We then obtain the first-order approximation

$$\Delta\widehat{\boldsymbol{\theta}} \sim \frac{1}{n}(\mathbf{E}\left[\mathbf{x}_i\mathbf{x}_i^T\right])^{-1}\mathbf{x}(y - \mathbf{x}^T\boldsymbol{\theta})$$

for $n \to \infty$.

The discovery – and subsequent special treatment – of influential observations using Cook's distance does have two drawbacks, however: First, the effect of omitting two or more observations is not merely the sum of the individual effects (as one influential observation can mask others). Secondly, omitting influential observations leads to question marks over the validity of tests and confidence intervals based on the remaining data.

2.1.2 Huber and L_1 regression

The reason for the large influence that individual observations can have on the least squares estimate is that large residuals have a high weight when a quadratic criterion is being used. To avoid this, we can instead look at the L_1 estimator:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^T\boldsymbol{\theta}|.$$

Historically speaking, this method is even older than least squares: it was suggested and discussed by Boscovich in 1760 and by Laplace in 1789 !

In the location model, i.e. when $p = 1$ and $x_i \equiv 1$, the solution is the median of the data, an estimator which for normally distributed data is substantially less precise than the arithmetic mean (i.e. the least squares estimator): to reach the same precision, the median requires 50% more observations.

One compromise between minimizing the L_2 distance and minimizing the L_1 distance is given by Huber regression:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho_c(y_i - \mathbf{x}_i^T\boldsymbol{\theta}),$$

where

$$\rho_c(u) = \frac{1}{2}u^2 \ (|u| \le c), \quad \rho_c(u) = c(|u| - \frac{c}{2}) \ (|u| \ge c),$$

cf. Figure 2.1. Choosing $c = 0$ leads to L_1 regression. If we compute derivatives and set them to zero, we obtain the equations

$$\sum_{i=1}^{n} \psi_c(y_i - \mathbf{x}_i^T\boldsymbol{\theta})\mathbf{x}_i = 0,$$

where $\psi_c(u) = \rho_c(u)' = \text{sign}(u)\min(|u|, c)$.

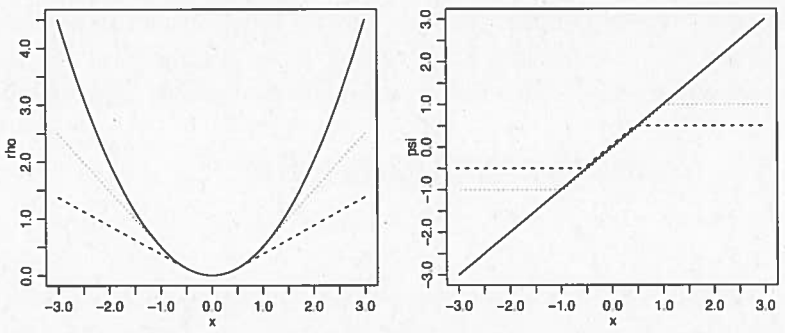

Figure 2.1: The Huber function and its derivative for various values of c

However, Huber regression only makes sense if the "corner" at c is chosen in relation to the variance of the residuals. Thus the estimators generally used are:

$$\begin{aligned} \sum_{i=1}^{n} \psi_c\left(\frac{y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\theta}}}{\widehat{\sigma}}\right) x_i &= 0, \\ \sum_{i=1}^{n} \chi\left(\frac{y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\theta}}}{\widehat{\sigma}}\right) &= 0. \end{aligned}$$

The function $\chi(u)$ is chosen either to be $\psi_c(u)^2 - \beta$ or to be $\chi(u) = \text{sign}(|u| - \beta)$. The constant β is fixed by the condition

$$\int \chi(u) \exp(-u^2/2) du = 0,$$

which ensures that for normally distributed errors $\widehat{\sigma}$ is a consistent estimate of the standard deviation. The first choice, χ, is Huber's so-called Proposal 2, while the second choice is $1/\beta$ times the median of the absolute residuals.

Closed-form computation of the L_1 and Huber estimators is no longer possible, but efficient algorithms for them are now known. The computational problem of L_1 regression can even be reduced to a linear optimization problem that "interior point" methods solve more quickly than least squares. Huber regression is performed by iterating the weighted least squares procedure using the weights

$$w_i \propto \frac{\psi_c((y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\theta}})/\widehat{\sigma})}{y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\theta}}} \propto \min(1, \frac{c\widehat{\sigma}}{|y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\theta}}|})$$

until they stabilize.

Furthermore, no closed-form expression exists for the distribution of the L_1 or Huber regression estimators. Asymptotic methods are thus used to show that when the covariates $\mathbf{x}_i$ are random, independent and identically distributed, the standardized vector $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ approximately has a normal distribution with expectation zero and covariance matrix

$$\frac{\mathbf{E}\left[\psi_c(\varepsilon_i/\sigma)^2\right]}{P[|\varepsilon_i| \le c\sigma]^2} \sigma^2 \mathbf{E}\left[\mathbf{x}_i \mathbf{x}_i^T\right]^{-1}.$$

Up to a constant factor, this covariance matrix is the same for least squares. When c lies in the interval $[1, 1.5]$, this factor is much less than 1 for heavy-tailed distributions and only slightly larger than 1 for the normal distribution. When comparing numerical estimates, however, we should remember that for non-normal errors ε_i, the parameter σ is no longer the error standard deviation, but the solution of

$$\mathbf{E}\left[\chi(\varepsilon_i/\sigma)\right] = 0.$$

This asymptotic approximation is also the basis of tests and confidence intervals, whose details we omit here.

Unfortunately Huber regression does not solve all the problems we have with influential observations. The exact effect of adding or omitting an observation can no longer be specified. We can, however, approximate the difference in the estimator caused by the addition of an observation at $(\mathbf{x}, y)$ by

$$\Delta\widehat{\boldsymbol{\theta}} \sim \frac{1}{nP[|\varepsilon_i| \leq c\sigma]}(\mathbf{E}\left[\mathbf{x}_i\mathbf{x}_i^T\right])^{-1}\mathbf{x}\ \psi_c(\frac{y - \mathbf{x}^T\boldsymbol{\theta}}{\sigma})\sigma.$$

Thus the influence of large values of y is bounded when $\mathbf{x}$ is fixed, but by varying $\mathbf{x}$ we can nonetheless increase this influence arbitrarily. More refined methods lacking this weakness will be discussed in the next two sections.

Huber regression is frequently replaced by estimators that assume ψ to be an odd function and χ an even function. Such estimators are known as M-estimators. Particularly popular choices of ψ include those that converge to zero for $|r|$ large, as they then remove large outliers entirely. However, this usually comes at the price of non-unique solutions to the defining equations. The solution actually found then depends on the algorithm – and especially on the choice of starting value.

2.1.3 Regression estimators with restrictions on influence

To restrict the influence of both y and $\mathbf{x}$, we regard estimators defined by the following type of equations:

$$\sum_{i=1}^{n} \eta\left(\mathbf{x}_i, \frac{y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\theta}}}{\widehat{\sigma}}\right)\mathbf{x}_i = 0.$$

To compute $\widehat{\sigma}$, we use an equation similar to that in Huber regression. Several kinds of $\eta(\mathbf{x}, r)$ are used here, including

$$\eta(\mathbf{x}, r) = \min(1, \frac{a}{||A\mathbf{x}||})\psi_c(r) \quad \text{(Mallows)}$$

and

$$\eta(\mathbf{x}, r) = \frac{1}{||A\mathbf{x}||}\psi_c((||A\mathbf{x}||)r) \quad \text{(Schweppe)}.$$

The matrix A is to be chosen in such a way that $||A\mathbf{x}||$ expresses the deviation of $\mathbf{x}$ from the set of covariates $(\mathbf{x}_i)_{1\leq i\leq n}$. This can be achieved e.g. by setting

$$||A\mathbf{x}||^2 = \text{const} \cdot \mathbf{x}^T (X^T X)^{-1} \mathbf{x},$$

though this choice in turn can be influenced quite strongly by an observation that has an unusual $\mathbf{x}_i$, since $X^T X$ can be rewritten as $\sum_i \mathbf{x}_i \mathbf{x}_i^T$. Thus $X^T X$ is replaced by a similar, yet robust, quantity. This leads to additional equations for A which we shall not discuss any further here.

Mallows' choice of η always chooses lower weights for observations with strongly deviating explanatory variables. To better understand Schweppe's choice, it helps to look at the identity $\psi_c(dr)/d = \psi_{c/d}(r)$. Thus Schweppe's suggestion merely lowers the corner in the Huber function, thus enabling an observation with strongly deviant covariates to nonetheless have full weight, if the corresponding residual is close to zero.

For this procedure, the difference in estimates after adding an observation at $(\mathbf{x}, y)$ is approximately given by

$$\Delta\widehat{\boldsymbol{\theta}} \sim \frac{1}{n} \left(\mathbf{E} \left[\frac{\partial}{\partial r} \eta(\mathbf{x}_i, \frac{\varepsilon_i}{\sigma}) \mathbf{x}_i \mathbf{x}_i^T \right] \right)^{-1} \mathbf{x}\, \eta \left(\mathbf{x}, \frac{y - \mathbf{x}^T \boldsymbol{\theta}}{\sigma} \right) \sigma.$$

Thus the influence of both $\mathbf{x}$ and y is restricted under both choices of η.

However, these estimators are still not satisfactory, as their *"breaking point"* is no greater than $1/p$. The breaking point is defined as the maximum proportion of outliers an estimator can withstand without diverging.

2.1.4 Regression estimators with high breaking point

We can obtain estimators whose breaking point does not depend on the dimension by replacing the arithmetic mean by the median in

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2 = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2$$

to obtain

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \text{median}((y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2)$$

(Least median of squares; Hampel 1975, Rousseeuw 1984). In other words: amongst pairs of parallel hyperplanes sandwiching 50% of all observations $(y_i, \mathbf{x}_i)$, we look for the pair whose distance along the y-axis is minimal. This procedure is illustrated in the left half of Figure 2.2.

It is intuitively clear (and provable) that this procedure can tolerate outliers in roughly 50% of all observations without diverging. Computing this estimator is a much greater problem, however, as the target function $\text{median}((y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2)$ generally has many local minima (cf. Figure 2.2, right side). Thus we need to

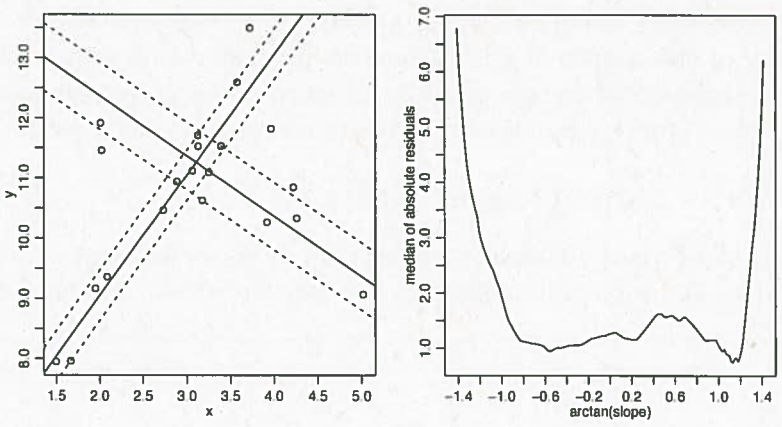

Figure 2.2: Least median of squares for simple regression. On the left, two lines with a band containing 50% of the observations. The intercept is chosen so as to minimize the band's diameter along the y-axis. On the right, the diameter of the band as a function of its slope.

search the entire space to find the global minimum, and this quickly becomes too costly as we move to higher dimensions. In general, stochastic algorithms are used to select $p+1$ data points at random, fit a plane through them and then compute the value of the target function for the corresponding θ.

One further drawback of this method is its lack of efficiency in the normal case: the estimator then only converges at the rate $n^{-1/3}$. A better convergence rate can be obtained e.g. by replacing the median by a truncated mean of the $((y_i - \mathbf{x}_i^T\theta)^2)$, omitting the αn (for $\alpha < 0.5$) largest squared residuals. Usually this estimator is used as the starting point from which to perform a Newton iteration to solve the estimator equations for an M-estimator using a function ψ that converges to zero. The result is then called the MM-estimator.

Developing robust regression estimators that exhibit good statistical and algorithmic properties is still an on-going topic for research.

2.2 Nonlinear least squares

In this chapter, we discuss methods for estimating θ in models of the form

$$y_i = f(\mathbf{x}_i, \theta) + \varepsilon_i.$$

Here f is a known function of the experimental conditions and of the parameters. The key assumptions on f that we shall make are that it is nonlinear in the parameters θ, and that we cannot (or do not want to) transform it into a linear model. Now p, the dimension of the parameter, need no longer be the dimension of the explanatory variables. For the vector ε of errors, we make the same assumptions as in the linear model, i.e. $\mathbf{E}[\varepsilon] = 0$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$.

Many applied problems are of this type, and the shape of f usually follows from the theory of the science in which a particular application arises. As an example, the description of the cumulative oxygen usage y of microorganisms in samples of river water as a function of incubation time x is usually performed by the model

$$f(x, \theta) = \theta_1(1 - \exp(-\theta_2 x)).$$

Thus the parameter θ_1 is the saturation point, and $\theta_1 \cdot \theta_2$ is the slope at $x = 0$. Some sample data and a possible regression function are shown in Figure 2.3.

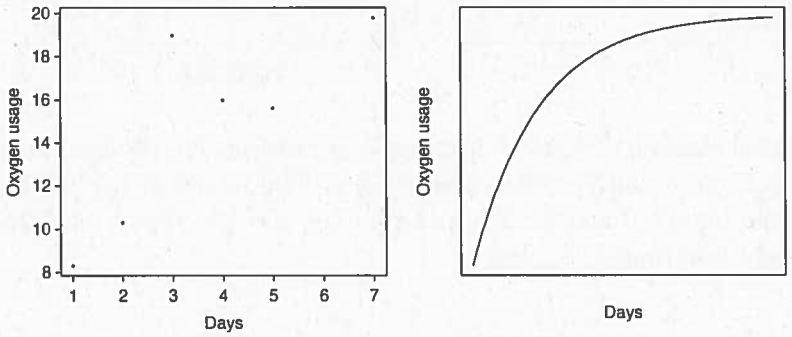

Figure 2.3: Oxygen usage data as a function of incubation time (left), and a typical regression function (right).

Another, similar, example is given by the so-called Michaelis-Menten model, which describes the dependence of reaction speed y on substrate concentration x. In this model, we have

$$f(x, \theta) = \frac{\theta_1 x}{\theta_2 + x}.$$

The transformation $y \to 1/y$, $x \to 1/x$ turns this into a linear regression model. However, the data shown in Figure 2.4 no longer have constant variance once this transformation has been applied; thus nonlinear least squares provide a much better fit.

In many applications the quantities $\mathbf{x}_i$ are times or locations at which a variable is observed whose development satisfies an ordinary or partial differential equation. In such a case, the parameters $\boldsymbol{\theta}$ are the parameters of the differential equation (including boundary conditions, if required), and $f(\mathbf{x}, \boldsymbol{\theta})$ denotes the solution of the differential equation for parameter $\boldsymbol{\theta}$ at $\mathbf{x}$.

The least squares estimate is defined to be

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} S(\boldsymbol{\theta}),$$

where

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2.$$

This carries the following geometric interpretation: If we vary $\boldsymbol{\theta}$, the points $(f(\mathbf{x}_1, \boldsymbol{\theta}), \ldots, f(\mathbf{x}_n, \boldsymbol{\theta}))^T$ describe a p-dimensional curved surface in $\mathbb{R}^n$, the so-

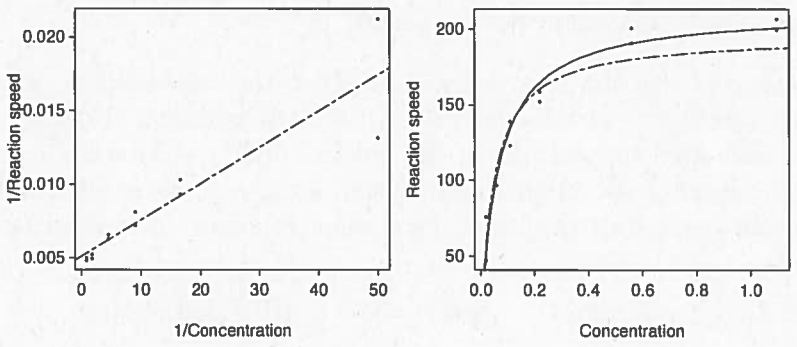

Figure 2.4: Reaction time as a function of substrate. On the left, a fitted line after a linearizing transformation; on the right, fitted functions on the untransformed scale estimated in the untransformed variables (solid line) and the transformed variables (–·–), respectively.

called response surface. We now seek the point on the response surface that is closest to the observation $(y_1, \ldots, y_n)^T$.

The solution does not generally have a closed form, and thus iterative methods (Gauss-Newton, Levenberg-Marquardt) are used. In practice, having good starting values is crucial.

The error variance σ^2 can be estimated in a way similar to that for linear regression:

$$\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}))^2.$$

2.2.1 Asymptotic confidence intervals and tests

In non-linear models, even the assumption of normal errors is not enough to allow exact tests and confidence intervals. We must thus rely on asymptotics.

The asymptotic properties of $\widehat{\boldsymbol{\theta}}$ can be obtained by means of a Taylor approximation around the true parameter $\boldsymbol{\theta}_0$. Thus we approximate the response surface in ia neighbourhood of the true parameter by a plane:

$$f(\mathbf{x}_i, \boldsymbol{\theta}) \approx f(\mathbf{x}_i, \boldsymbol{\theta}_0) + \mathbf{a}(\boldsymbol{\theta}_0)_i^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

where

$$\mathbf{a}(\boldsymbol{\theta})_i = \left(\frac{\partial}{\partial \boldsymbol{\theta}_j} f(\mathbf{x}_i, \boldsymbol{\theta}); j = 1, \ldots p\right)^T.$$

This means that we approximately have a linear model with explanatory variables $\mathbf{a}(\boldsymbol{\theta}_0)_i$ near the true parameter. Furthermore, it can be shown that (under certain technical conditions) the distribution of $\widehat{\boldsymbol{\theta}}$ is asymptotically the same as in the approximating model, i.e.

$$\widehat{\boldsymbol{\theta}} \overset{\text{asymptotically}}{\sim} \mathcal{N}(\boldsymbol{\theta}_0, \sigma^2(A(\boldsymbol{\theta}_0)^T A(\boldsymbol{\theta}_0))^{-1}).$$

Here the matrix $A(\boldsymbol{\theta})$ consists of the rows $\mathbf{a}(\boldsymbol{\theta})_i^T$.

These approximated distributions do not yet allow the construction of tests and confidence intervals, as both σ and $A(\boldsymbol{\theta}_0)$ are still unknown. However, we can plug in $\widehat{\sigma}$ and $A(\widehat{\boldsymbol{\theta}})$ instead. As in the linear model, we generally use the t distribution instead of the normal distribution here, and the F distribution instead of the chi-squared distribution. Thus the confidence interval for θ_k is

$$\widehat{\theta}_k \pm t_{n-p;1-\alpha/2}\, se(\widehat{\theta}_k), \qquad se(\widehat{\theta}_k) = \widehat{\sigma}\sqrt{((A(\widehat{\boldsymbol{\theta}})^T A(\widehat{\boldsymbol{\theta}}))^{-1})_{kk}}.$$

2.2.2 More precise tests and confidence intervals

In a similar way to the F test used in linear regression, we can test two nested models by means of the differences in sums of squared deviations. To test the null hypothesis $B\boldsymbol{\theta} = \mathbf{b}$, we also need the least squares estimate under the null hypothesis:

$$\widehat{\boldsymbol{\theta}}_0 = \arg \min_{\boldsymbol{\theta}; B\boldsymbol{\theta}=\mathbf{b}} S(\boldsymbol{\theta}).$$

Then we compute the test statistic

$$T = \frac{(S(\widehat{\boldsymbol{\theta}}_0) - S(\widehat{\boldsymbol{\theta}}))/q}{S(\widehat{\boldsymbol{\theta}})/(n-p)},$$

where q denotes the rank of B. In the linear model, this statistic was identical to the one computed from joint normal distribution of $\widehat{\boldsymbol{\theta}}$ and had an $F_{q,n-p}$ distribution under the null hypothesis. In the nonlinear setup, these two statistics differ, and they are only approximately F distributed. This approximation, however, is often significantly better the the normal approximation of $\widehat{\boldsymbol{\theta}}$.

In particular, we can use this to test the null hypothesis $\theta_k = \theta_k^*$ for any arbitrary but fixed value of θ_k^*. The test statistic then becomes

$$T_k(\theta_k^*) = \frac{S(\widehat{\boldsymbol{\theta}}^{(-k)}) - S(\widehat{\boldsymbol{\theta}})}{S(\widehat{\boldsymbol{\theta}})/(n-p)} = \frac{S(\widehat{\boldsymbol{\theta}}^{(-k)}) - S(\widehat{\boldsymbol{\theta}})}{\widehat{\sigma}^2},$$

where

$$\widehat{\boldsymbol{\theta}}^{(-k)} = \widehat{\boldsymbol{\theta}}^{(-k)}(\theta_k^*) = \arg \min_{\boldsymbol{\theta}; \theta_k = \theta_k^*} S(\boldsymbol{\theta}).$$

(Thus $\widehat{\boldsymbol{\theta}}^{(-k)}$ is the least squares estimate under the null hypothesis, i.e. it is equal to $\widehat{\boldsymbol{\theta}}_0$ in the above notation). Since an F distribution with one degree of freedom in the denominator is simply the distribution of the square of a t-distributed variable, we can also regard the test statistic

$$\tau_k(\theta_k^*) = \frac{\operatorname{sign}(\theta_k^* - \widehat{\theta}_k)}{\widehat{\sigma}} \sqrt{S(\widehat{\boldsymbol{\theta}}^{(-k)}) - S(\widehat{\boldsymbol{\theta}})}.$$

Under the null hypothesis, this follows a t distribution with $n-p$ degrees of freedom. By reversing the test, we obtain the following confidence interval for θ_k:

$$\left\{\theta_k^* \mid \sqrt{S(\widehat{\boldsymbol{\theta}}^{(-k)}) - S(\widehat{\boldsymbol{\theta}})} \leq t_{n-p;1-\alpha/2}\,\widehat{\sigma}\right\}.$$

If we apply a monotone transformation to θ_k, this interval can simply be transformed in the same way, an advantage not shared by the interval $\widehat{\theta}_k \pm t_{n-p;1-\alpha/2}\, se(\widehat{\theta}_k)$. The differences between these two intervals show us the effects of the nonlinearity in the model. Thus we often plot τ_k against θ_k^* and use the evident curvature to gain an impression of the degree of nonlinearity present.

When $p = 2$, we can of course also plot the contours of $S(\theta_1, \theta_2)$. In the linear case, these are ellipses; thus the degree of nonlinearity can be seen in the amount by which these contours differ from elliptical shapes. Simultaneous confidence sets are bounded precisely by these contour lines. The further away they are from being circular, the greater the dependence between the two parameters being estimated. In the same figure we can plot the so-called profile traces $\widehat{\theta}_2^{(-1)}(\theta_1^*)$ versus θ_1^* and $\widehat{\theta}_1^{(-2)}(\theta_2^*)$ versus θ_2^*. They intersect at $\widehat{\boldsymbol{\theta}}$, and their angle shows how strongly the estimated parameters depend on each other. Furthermore, the profile traces intersect the contours in points where the latter have horizontal or vertical tangents, respectively, since the gradient is perpendicular to the contours. If the model is a linear one, S is quadratic and the profile traces are straight lines (cf. Lemma 1.5.1). Thus we obtain a diagram like that in Figure 1.13 (and the two regression lines there are exactly the two profile traces). The contours and profile traces for the data in Figures 2.3 and 2.4 are shown in Figure 2.5. We can see that the effects of nonlinearity are comparatively harmless in the reaction speed example, while they are fairly extreme in the oxygen use example.

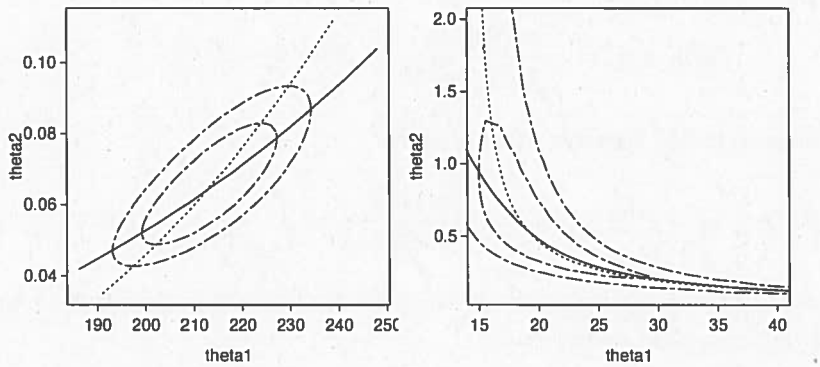

Figure 2.5: Contours and profile traces for the examples of reaction speed (left) and oxygen use (right). The underlying data are shown in Figures 2.4 and 2.3.

Once $p > 2$, it is much more difficult to visualize the effect of nonlinearities and of the dependence between the parameters. We could in principle compute the

contours of the so-called profile likelihood

$$\min_{\boldsymbol{\theta};\theta_j=\theta_j^*,\theta_k=\theta_k^*} S(\boldsymbol{\theta})$$

for all pairs (j,k), but this is often too expensive computationally. Instead, we can at least plot pairs of profile traces, plotting $\widehat{\theta}_j^{(-k)}(\theta_k^*)$ against θ_k^* and $\widehat{\theta}_k^{(-j)}(\theta_j^*)$ against θ_j^* for all $j<k$. The curvature of these profile traces indicates the extent of nonlinearity, while their angle shows the level of dependence between the corresponding parameter estimates. Furthermore, the profile traces once more intersect the contours of the profile likelihood at points where the tangents are horizontal or vertical, respectively. This gives us an impression of where the contour lines should be.

2.3 Generalized linear models

2.3.1 Logistic regression

Many applications have a binary response variable Y (e.g. in medicine, patients may be healed or dead), and the success probability $P[Y=1]$ depends on explanatory variables. Models that have this probability as a linear function of the explanatory variables are rarely meaningful, as the success probability must always lie between 0 and 1. Logistic regression uses the model

$$\log\left(\frac{P_{\boldsymbol{\theta}}[Y_i=1]}{P_{\boldsymbol{\theta}}[Y_i=0]}\right)=\sum_{j=1}^{p} x_{ij}\theta_j=\mathbf{x}_i^T\boldsymbol{\theta}$$

(if the covariates $\mathbf{x}_i$ are also random, the LHS contains conditional probabilities given $\mathbf{x}_i$). Solving for $P_{\boldsymbol{\theta}}[Y_i=1]$ gives us

$$P_{\boldsymbol{\theta}}[Y_i=1]=\frac{\exp(\mathbf{x}_i^T\boldsymbol{\theta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\theta})}=P[U\geq-\mathbf{x}_i^T\boldsymbol{\theta}],$$

where U has a so-called logistic distribution:

$$P[U\leq u]=P[U\geq -u]=\frac{\exp(u)}{1+\exp(u)}=\int_{-\infty}^{u}\frac{e^t}{(1+e^t)^2}dt.$$

This can be interpreted as a model in which latent variables Z_i satisfy a linear relationship with logistic errors ε_i:

$$Z_i=\mathbf{x}_i^T\boldsymbol{\theta}+\varepsilon_i.$$

Instead of Z_i, it is the indicator $Y_i=1_{[Z_i\geq 0]}$ which is observed. Taking normally distributed errors instead of logistically distributed ones gives us the so-called probit model. These models differ only very slightly, but the logistic model is computationally simpler.

The parameters in such a model are nearly always estimated using maximum likelihood. It is easy to see that for any $y \in \{0,1\}$, we have

$$P_{\boldsymbol{\theta}}[Y_i = y] = \left(\frac{P_{\boldsymbol{\theta}}[Y_i = 1]}{P_{\boldsymbol{\theta}}[Y_i = 0]}\right)^y P_{\boldsymbol{\theta}}[Y_i = 0] = \exp(y \cdot \mathbf{x}_i^T\boldsymbol{\theta} - \log(1 + \exp(\mathbf{x}_i^T\boldsymbol{\theta}))).$$

Thus (assuming the independence of the observations) we have the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i\mathbf{x}_i^T\boldsymbol{\theta} - \log(1 + \exp(\mathbf{x}_i^T\boldsymbol{\theta}))).$$

This is a concave function whose maximum is given by

$$\sum_{i=1}^{n}(y_i - P_{\widehat{\boldsymbol{\theta}}}[Y_i = 1])\mathbf{x}_i = 0.$$

These equations are generally solved numerically using iterative methods. If the experimental condition $\mathbf{x}_i$ has several observed response values $(y_{ij}; 1 \le j \le n_i)$, then $\ell(\boldsymbol{\theta})$ only depends on their sum $y_{i+} = \sum_j y_{ij}$ and the total n_i.

Confidence intervals and tests in this model depend on the asymptotic normal approximation

$$\widehat{\boldsymbol{\theta}} \overset{\text{asymptotically}}{\sim} \mathcal{N}(\boldsymbol{\theta}, V(\boldsymbol{\theta})).$$

The asymptotic covariance matrix $V(\boldsymbol{\theta})$ of $\widehat{\boldsymbol{\theta}}$ is the inverse of the *Fisher information* (cf. Section 1.9 or Mathematical Statistics):

$$V(\boldsymbol{\theta})^{-1} = I(\boldsymbol{\theta}) = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\mathbf{E}\left[(y_i - P_{\boldsymbol{\theta}}[Y_i = 1])^2\right] = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\frac{\exp(\mathbf{x}_i^T\boldsymbol{\theta})}{(1 + \exp(\mathbf{x}_i^T\boldsymbol{\theta}))^2}.$$

Comparisons of two nested models with dimensions p and $q < p$ are possible by using double the log-likelihood quotient

$$2(\ell(\widehat{\boldsymbol{\theta}}^{(p)}) - \ell(\widehat{\boldsymbol{\theta}}^{(q)})),$$

which is known to asymptotically follow a chi-squared distribution with $(p - q)$ degrees of freedom.

2.3.2 General case

In the general case we have independent observations Y_i and a density or distribution function of the form

$$p_{\beta_i}(y_i) = \exp(y_i\beta_i + c(\beta_i))h(y_i)$$

(a so-called exponential family). For such a model, we have

$$\mathbf{E}[Y_i] = \mu(\beta_i) = -c'(\beta_i)$$

(which can be seen by taking the derivative of $\int p_\beta(y)dy = 1$ with respect to β and switching the integration and differentiation steps). Many common distributions take on the form of such an exponential family, including the normal, binomial and Poisson distributions. For the normal distribution, we have

$$p(y) = \exp(y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2}).$$

Thus if σ is known, we have an exponential family with parameters

$$\beta = \frac{\mu}{\sigma^2}, \quad c(\beta) = -\frac{1}{2}\sigma^2\beta^2.$$

The binomial(n, p) distribution can be written as

$$p(y) = \left(\frac{p}{1-p}\right)^y (1-p)^n \binom{n}{y}.$$

This is an exponential family with

$$\beta = \log\left(\frac{p}{1-p}\right), \quad c(\beta) = -n\log(1+e^\beta).$$

Similarly, the Poisson(λ) distributions form an exponential family with

$$\beta = \log\lambda, \quad c(\beta) = -e^\beta.$$

In a generalized linear model, the effect of the explanatory variables $\mathbf{x}_i$ on the observation Y_i is described by a link between $\mathbf{x}_i$ and the parameter β_i for the i-th observation. That is, we want there to be a function g for which

$$g(\mu(\beta_i)) = \mathbf{x}_i^T\boldsymbol{\theta},$$

i.e. a suitable transformation of β_i is to be linear in the explanatory variables. The function g is called a *link function*. If g is exactly μ^{-1}, we call it the *canonical* link function. The linear model with normal errors, and logistic regression, are examples of generalized linear models with a canonical link function.

The general case can be treated in much the same way as the logistic case, i.e. taking the maximum likelihood estimate and using the asymptotic normality of this estimate (or the asymptotic chi-squared distribution of the likelihood quotient) to construct tests and confidence intervals. For details, see the literature.

2.4 Cox regression

In medical and technical applications, the response variable is often a survival or failure time. We could in principle use a generalized linear model with exponential or Gamma distributions. For practical purposes, though, the Cox model has emerged as the front runner – and it is not bound to any specific family of distributions. Let F be a distribution on the positive real numbers,

and denote its density by f. Then the failure rate (also known as hazard or risk function) is defined as

$$\lambda(t) = \lim_{h \to 0} \frac{1}{h} P[t \leq T \leq t+h | T \geq t] = \frac{f(t)}{1 - F(t)} = -\frac{d}{dt} \log(1 - F(t)).$$

Thus the failure rate determines the distribution, namely by

$$F(t) = 1 - \exp(-\int_0^t \lambda(u) du).$$

The Cox model now postulates that the shape of the i-th failure rate as a function of the explanatory variables $\mathbf{x}_i$ is

$$\lambda_i(t) = \exp(\mathbf{x}_i^T \boldsymbol{\theta}) \lambda_0(t),$$

where λ_0 is a base rate. Of course this model does not allow an intercept to be present, as a constant could be absorbed into λ_0. Raising the j-th component of the explanatory variables by one unit leads to multiplying the failure rate by the factor $\exp(\boldsymbol{\theta}_j)$ (homogeneously for all times). Because of this property, this model is also known as "proportional hazard model". It is easy to see that any strictly monotonous and differentiable transformation of the survival times transforms a Cox model into another Cox model with the same parameters and a different λ_0. If λ_0 is not specified, this merely means that the choice of a timescale remains open.

The base rate λ_0 naturally appears in the likelihood function, too, which means that we cannot simply employ the maximum likelihood estimator. We shall instead estimate the parameter $\boldsymbol{\theta}$ by means of the so-called partial likelihood, defined as

$$\prod_{i=1}^{n} \frac{\exp(\mathbf{x}_i^T \boldsymbol{\theta})}{\sum_{j; t_j \geq t_i} \exp(\mathbf{x}_j^T \boldsymbol{\theta})}.$$

The i-th factor is the conditional probability of failure for the i-th observed unit in the interval $[t_i, t_i + dt)$, given the failure of one of the units that was still working just before time t_i. Thus the only information on failure times used is the order in which the failures occur.

Nearly all data of this type suffer from the added complication of censored observations, whose failure times T_i are not known exactly, but it is only known that they are greater that some observed censoring time C_i. The reasons for this include the termination of the study before the failure of all units, or patients moving away or dying of other causes. In such situations the partial likelihood can be defined in a similar way: We compute the product over all uncensored observations, but in the denominator, we only add up the uncensored observations with $t_j \geq t_i$ and the censored observations with $c_j \geq t_i$.

Once again, tests and confidence intervals depend on asymptotic considerations that we shall not elaborate here.

2.5 Nonparametric regression

Over the past 20-30 years much attention has been focused on procedures that make no parametric assumptions about the shape of f in the model

$$y_i = f(\mathbf{x}_i) + \varepsilon_i,$$

and only ask for f to be a smooth function. Weakening the assumptions in this way is of course a great advantage, especially in explorative investigation, where we aim at being as open as possible towards the data.

We separate the deterministic case where the $\mathbf{x}_i$ are fixed and the stochastic case where the $\mathbf{x}_i$ are realizations of random variables that follow a distribution G, irrespectively of the properties of the errors ε_i. A more general model assumes that the data $(\mathbf{X}_i, Y_i)$ are i.i.d random vectors with some arbitrary joint distribution, and that our intention is to estimate

$$f(\mathbf{x}) = \mathbf{E}\left[Y_i \mid \mathbf{X}_i = \mathbf{x}\right].$$

We can show that for any function g, the inequality

$$\mathbf{E}\left[(Y_i - f(\mathbf{X}_i))^2\right] \leq \mathbf{E}\left[(Y_i - g(\mathbf{X}_i))^2\right]$$

holds (as long as the second moments do in fact exist). This means that $f(\mathbf{X}_i)$ is the best prediction of Y_i from $\mathbf{X}_i$ in terms of mean squared error. The conditional error variance

$$\mathbf{E}\left[(Y_i - f(\mathbf{x}_i))^2 \mid \mathbf{X}_i = \mathbf{x}_i\right],$$

which corresponds to $\mathrm{Var}[\varepsilon_i]$, then depends on $\mathbf{x}_i$ in general.

All nonparametric estimates of $f(\mathbf{x})$ essentially average the responses y_i for those i whose $\mathbf{x}_i$ are in close proximity to $\mathbf{x}$. There are, however, large differences in the method of averaging and in how to determine proximity to $\mathbf{x}$.

2.5.1 Some procedures for the one-dimensional case

Kernel estimators:

As an estimate of $f(x)$, we take the weighted mean of the y_i in such a way that the weight of y_i is monotonely decreasing in the distance $|x - x_i|$. The weights are determined by a so-called kernel K and a bandwidth $h > 0$. A kernel is a probability density which is symmetric around 0 and either has support $[-1, 1]$ or is very rapidly decreasing (as is e.g. the normal density). There are two versions that differ considerably in the stochastic case. The Nadaraya-Watson version is defined as

$$\widehat{f}(x) = \frac{\sum_{i=1}^{n} y_i K((x - x_i)/h)}{\sum_{i=1}^{n} K((x - x_i)/h)}.$$

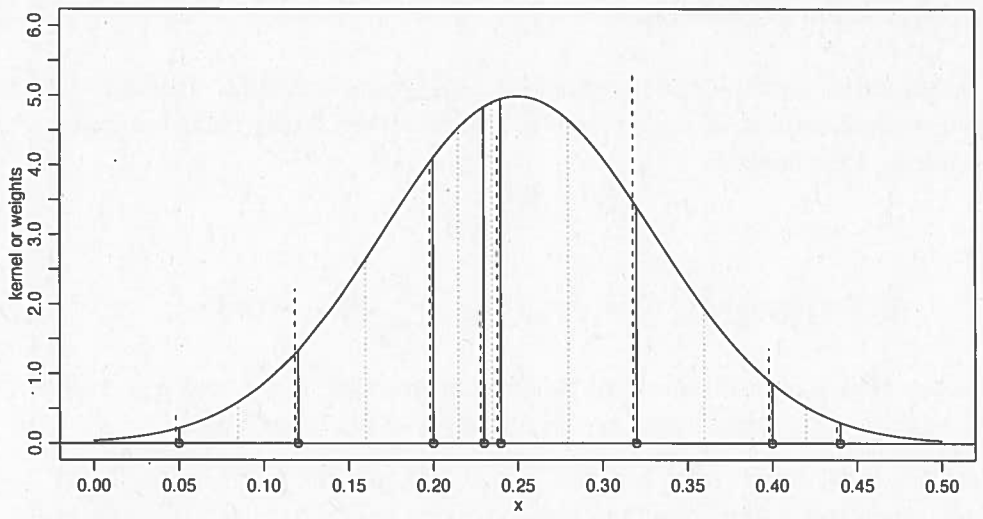

Figure 2.6: Weights of the Nadaraya-Watson (solid lines) and Gasser-Müller (dashed lines) versions. The locations x_i of the observations are marked by circles on the x-axis, and the boundary points s_i by the dotted lines. All weights have been scaled by a constant factor.

For the Gasser-Müller version we assume that we have ordered covariates

$$0 \leq x_1 < x_2 < \ldots < x_n \leq 1,$$

and we define $s_0 = -\infty$, $s_i = (x_i + x_{i+1})/2$ for $0 < i < n$ and $s_n = +\infty$. Our estimate is then

$$\widehat{f}(x) = \sum_{i=1}^{n} y_i \int_{s_{i-1}}^{s_i} \frac{1}{h} K((x-u)/h) du.$$

The differences between these two estimators are quite important if the the covariates x_i are distributed unevenly, as we can see in Figure 2.6. We can see that the Gasser-Müller version gives higher weights to observations whose corresponding x_i have an isolated location.

In both versions, the bandwidth h regulates the smoothness of $\widehat{f}$: The larger h is chosen to be, the smoother the estimate, but the less it can adapt to the data.

If the covariates x_i are chosen at random, the number of observations having a significantly non-zero weight can vary quite strongly as x varies. For the first estimator (Nadaraya-Watson), there may even be none of them. To avoid such an occurrence, we can choose a variable bandwidth, e.g. to fix the number of observations for which $x - h \leq x_i \leq x + h$. The latter procedure leads to a so-called nearest neighbour estimator.

Kernel estimators especially struggle to estimate f at the fringes, i.e. for $x < h$ and $x > 1 - h$. As for such points the averaged observations nearly all lie on one side of x, systematic errors can occur.

Local polynomial regression:

Here we no longer assume the function f to be locally constant. Instead, we locally use a polynomial of degree $p > 0$, which is then fitted using weighted least squares. This leads to

$$\widehat{f}(x) = \widehat{\theta}_0(x),$$

where

$$\widehat{\theta}(x) = \arg\min_{\theta} \sum_{i=1}^{n} K((x - x_i)/h)(y_i - \sum_{j=0}^{p} \theta_j (x_i - x)^j)^2.$$

It turns out that p is best chosen to be odd. In practice, $p = 1$ and $p = 3$ are used most often. The advantages are seen quite clearly at the fringes.

Instead of a fixed bandwidth, we can once more use the nearest neighbour version that ensures a fixed number of observations in $[x - h, x + h]$. This is the key idea of the function `loess` in the statistical software packages S-Plus and R.

Smoothing splines:

A smoothing spline is defined implicitly to be the solution of the minimization problem

$$\arg\min_{f} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_0^1 f''(x)^2 dx.$$

The first term, a sum over i, indicates the goodness of fit of f to the observations. The second term, the L_2 norm of the second derivative, measures the smoothness of f, and the parameter λ regulates the compromise that governs the trade-off of the opposite aims of making both terms small simultaneously.

It can be shown that the solution of the minimization problem is a cubical spline with nodes at the x_i, and that this solution is furthermore linear on the fringe intervals $[0, x_1]$ and $[x_n, 1]$. If we take $\lambda \to 0$, we obtain the spline that just interpolates the data, and for $\lambda \to \infty$ we obtain the least squares line. Thus the role of λ corresponds to that of the bandwidth h.

To compute the smoothing spline,we choose a basis of the vector space of splines with nodes x_i. Computing the coefficients of the solution with respect to this basis then leads to minimization of a quadratic function. If we are applying a numerically stable and fast procedure, a good choice of basis is crucial. Here the so-called B-splines have a good track record.

2.5.2 Bias-variance tradeoff

All non-parametric procedures contain a smoothing parameter that has a strong influence on their behavior. For splines, this is λ, and for kernel estimators, it is the bandwidth h. The purpose of this smoothing parameter becomes apparent

if we look at the bias and variance of such an estimator. We can show that for local polynomials with odd p,

$$\mathbf{E}\left[\widehat{f}(x)\right] - f(x) \sim const(K,p)\, h^{p+1}\, f^{(p+1)}(x)$$

and

$$\mathrm{Var}[\widehat{f}(x)] \sim const(K,p)\, \frac{\sigma_\varepsilon^2}{nh} \left(\frac{1}{nh}\sum K((x-x_i)/h)\right)^{-1}.$$

Here we assume the $\mathbf{x}_i$ to be fixed (or if they are not, we condition on them), and that the bandwidth $h = h_n$ satisfies $h_n \to 0$ and $nh_n \to \infty$. From the above formulæ, we see that h should be small to obtain a small (absolute) bias, and h should be large to obtain a small variance. Hence the title of this section.

One quantity that considers bias and variance simultaneously is the mean squared error:

$$\mathbf{E}\left[(\widehat{f}(x) - f(x))^2\right] = \mathrm{Var}[\widehat{f}(x)] + (\mathbf{E}\left[\widehat{f}(x)\right] - f(x))^2 = O(\frac{1}{nh}) + O(h^{2(p+1)}).$$

The order of this is minimal if both its summands have the same order, i.e. $h = O(n^{-1/(2p+3)})$. Such a choice leads to a mean squared error of the order $O(n^{-(2p+2)/(2p+3)})$. Looking closely, we would prefer to have p as large as possible; in practice, however, this is not quite right, as a larger value of p also enlarges the constants and requires stronger assumptions on f. In any case, the constants are the greatest problem when applying these results: they contain unknown terms such as the derivatives of f, and the optimal choice of h depends on the location x. Thus the data-dependent yet optimal choice of bandwidth is quite a difficult problem.

We could also find the optimal kernel K (by minimizing $const(K,p)$). It turns out, however, that the choice of K is of secondary importance, as almost all continuous kernels are nearly as good as each other.

The bias of the Nadaraya-Watson kernel estimators has a more complicated form, but their variance is asymptotically equivalent to that of the local polynomials with $p = 1$. For the Gasser-Müller version, the bias is the same as for local polynomials with $p = 1$, but instead the constant involved in the variance is 1.5 the size as the one obtained for local polynomials of degree $p = 1$. We even know the bias and variance for the smoothing splines; it is similar to that of local polynomials of degree $p = 3$.

2.5.3 Curse of dimensionality

We can in principle generalize kernel estimators and local polynomials to more than one dimension. In practice, though, they usually fail from dimensions 3 or 4 upwards. The reason for this is that high-dimensional spaces contain a large amount of space and are not well covered by a finite set of points. In other words: any two $\mathbf{x}_i$ nearly always lie very far apart, making a reasonable compromise on the bias-variance dilemma impossible. This "curse of dimensionality" is

quite well illustrated by the following example: If the points $\mathbf{x}_i$ are equally distributed inside the cube $[-1,1]^p$, the proportion of points lying in the unit sphere $\{\mathbf{x}; ||\mathbf{x}|| \leq 1\}$ is approximately the same as the probability of a $\mathbf{x}_i$ lying in this unit sphere, i.e. the volume of the unit sphere times 2^{-p}. If $p = 2$, this probability is 79%, for $p = 5$ it is still 16%, and for $p = 10$ only 0.25% ! On the other hand, the ratio of diameters of sphere and cube is $1 : \sqrt{p}$, thus for $p = 10$, we roughly have the ratio $1 : 3$. So the assumption that f is constant or linear on the unit sphere does not differ significantly from the assumption that f is constant or linear on the entire cube – yet we still nearly always have too few observations inside the unit sphere to estimate a constant or linear function well.

Appendix A

Results from probability theory

A.1 Computation of moments

Let the following be given:

- Random variables X, Y, Z
- Constants a, b

Computation of means:

- $\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$.
- $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$.
- $\mathbf{E}[XY] = \mathbf{E}[X] \cdot \mathbf{E}[Y]$, if X and Y are independent or at least uncorrelated.

Computation of variance and covariance:

- $\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$, and thus: $\sigma[aX + b] = |a|\sigma[X]$
- $\mathrm{Cov}[X, Y] := \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] = \mathrm{Cov}[Y, X]$.
- $\mathrm{Var}[X + Y] = \mathrm{Var}\, X + \mathrm{Var}\, Y + 2\,\mathrm{Cov}[X, Y]$
- In particular: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$, if X and Y are independent or at least uncorrelated.
- $\mathrm{Cov}[aX + bY, Z] = a\,\mathrm{Cov}[X, Z] + b\,\mathrm{Cov}[Y, Z]$.

Moments of random vectors:

Let $\mathbf{Y}$ be a $n \times 1$-dimensional random vector. We define $\mathbf{E}[\mathbf{Y}]$ to be the vector with components $\mathbf{E}[y_i]$, and the covariance matrix of $\mathbf{Y}$ to be

$$\mathrm{Cov}[\mathbf{Y}] = (\mathrm{Cov}(y_i, y_j)_{ij}) = \begin{pmatrix} \mathrm{Var}[y_1] & \mathrm{Cov}[y_1, y_2] & \dots & \mathrm{Cov}[y_1, y_n] \\ \mathrm{Cov}[y_2, y_1] & \mathrm{Var}[y_2] & \dots & \mathrm{Cov}[y_2, y_n] \\ \dots & \dots & \dots & \dots \\ \mathrm{Cov}[y_n, y_1] & \mathrm{Cov}[y_n, y_2] & \dots & \mathrm{Var}[y_n] \end{pmatrix}$$

Thus a covariance matrix is always symmetric.

Let A be a fixed matrix of dimension $m \times n$, and let $\mathbf{b}$ be a fixed vector of dimension $m \times 1$. Then as in the scalar case, we have:

- $\mathrm{Cov}[\mathbf{Y}] = \mathbf{E}[\mathbf{Y}\mathbf{Y}^T] - \mathbf{E}[\mathbf{Y}]\,\mathbf{E}[\mathbf{Y}]^T$.
- $\mathbf{E}[A\mathbf{Y} + \mathbf{b}] = A\mathbf{E}[\mathbf{Y}] + \mathbf{b}$,
- $\mathrm{Cov}[A\mathbf{Y} + \mathbf{b}] = A \cdot \mathrm{Cov}[\mathbf{Y}] \cdot A^T$.

In particular, the latter rule tells us that for an arbitrary vector $\mathbf{a}$ of dimension $n \times 1$:

$$0 \leq \mathrm{Var}[\mathbf{a}^T\mathbf{Y}] = \mathrm{Cov}[\mathbf{a}^T\mathbf{Y}] = \mathbf{a}^T\,\mathrm{Cov}[\mathbf{Y}]\mathbf{a}.$$

In other words, every covariance matrix is **positive semidefinite** (and usually even **positive definite**).

If Σ is an arbitrary positive semidefinite matrix, there exist matrices A satisfying $AA^T = \Sigma$. We call each such matrix A a root of Σ, and write $A = \Sigma^{1/2}$. Note that $\Sigma^{1/2}$ is only defined up to multiplication by an orthogonal matrix. Numerically speaking, the easiest way to find $\Sigma^{1/2}$ uses the Cholesky decomposition, which yields a lower triangular matrix. Some care must also be taken when using $\Sigma^{1/2}$ in computations, as e.g. $(\Sigma^{-1})^{1/2} = (\Sigma^{1/2})^{-T}$.

If $\mathbf{Y}$ is a random vector with covariance matrix Σ and $A = \Sigma^{1/2}$, the rules listed above imply that the covariance matrix of the random vector $\mathbf{X} = A^{-1}\mathbf{Y}$ is precisely the identity matrix, i.e. that the components of $\mathbf{X}$ are uncorrelated and have variance 1.

A.2 The normal distribution

A.2.1 Univariate normal distribution

a) **Density of the "standard normal distribution"** $\mathcal{N}(0,1)$**:**

$$\boxed{\varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}$$

(This is a probability density, i.e. the integral of $\varphi(x)$ over the real axis is 1). The cumulative density function

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)dy$$

has no closed-form description, but it can be looked up in tables.

b) **3 reasons why the normal distribution is important (as an ideal model):**

(i) **Simplicity** (and the beauty of the ensuing theory).

(ii) **Central limit theorem** (and the **elementary error hypothesis**): An error in measurement consists of many tiny yet independent elementary errors which **combine additively**; thus their sum is approximately **normally distributed**. A **multiplicative combination** of elementary errors gives us a **logarithmic normal distribution**, which transforms into a normal distribution when logarithms are taken.

(iii) **Experience** Many data sets are approximately normally distributed (possibly only after a suitable transformation has been applied).

c) **Why is the normal distribution so simple in some sense?**

If $f(x)$ is a density on $\mathbb{R}$ that satisfies

$$\frac{d\log f(x)}{dx} = \frac{f'}{f}(x) = ax + b\,,$$

it follows that $f(x) = e^{\frac{1}{2}ax^2+bx+c}$. This quantity f'/f is quite important in statistics. Thus in a certain sense, the normal distribution really is the most simple continuous distribution on the whole real line!

d) **Linear transformations of the standard normal distribution $\mathcal{N}(0,1)$:**

Regard $X \sim \mathcal{N}(0,1)$ and the transformation $x \longmapsto y := \mu + \sigma x$, taking an arbitrary μ and some $\sigma > 0$. Then the distribution of $Y := \mu + \sigma X$ is the general normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Computing the density of Y: From

$$f_X(x)dx = f_Y(y)dy \text{ with } dy = \sigma dx,$$

we conclude that

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx = \underbrace{\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}}_{\text{density of } Y:=\mu+\sigma X}\,dy.$$

e) **Moments of the normal distribution:**

Standard normal distribution: $X \sim \mathcal{N}(0,1)$.

$$\begin{aligned}
\mathbf{E}[X] &= \int_{-\infty}^{\infty} x\varphi(x)dx = 0 \\
\mathrm{Var}[X] &= \int_{-\infty}^{\infty} x \cdot x\varphi(x)dx = \underbrace{-x\varphi(x)\,|_{-\infty}^{+\infty}}_{0} + \underbrace{\int_{-\infty}^{\infty} \varphi(x)dx}_{1} = 1 \\
\mathbf{E}[X^3] &= 0 \text{ (symmetry)} \\
\mathbf{E}[X^4] &= \int_{-\infty}^{\infty} x^3 \cdot x\varphi(x)dx = -x^3\varphi(x)\,|_{-\infty}^{+\infty} + \int_{-\infty}^{\infty} 3x^2\varphi(x)dx = 3.
\end{aligned}$$

General normal distribution: $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Mean:	$\mathbf{E}[Y]$	$=$	μ
Variance:	$\mathrm{Var}[Y]$	$=$	σ^2
Skewness (standardized 3rd moment):	γ_1	$=$	$\mathbf{E}\left[(Y-\mu)^3\right]/\sigma^3 = 0$
Kurtosis, excess (stand. 4th moment):	γ_2	$=$	$\mathbf{E}\left[(Y-\mu)^4\right]/\sigma^4 - 3 = 0$

f) Shape of the normal distribution

Figure A.1 gives a graphical representation of the density of the standard normal distribution $\mathcal{N}(0,1)$.

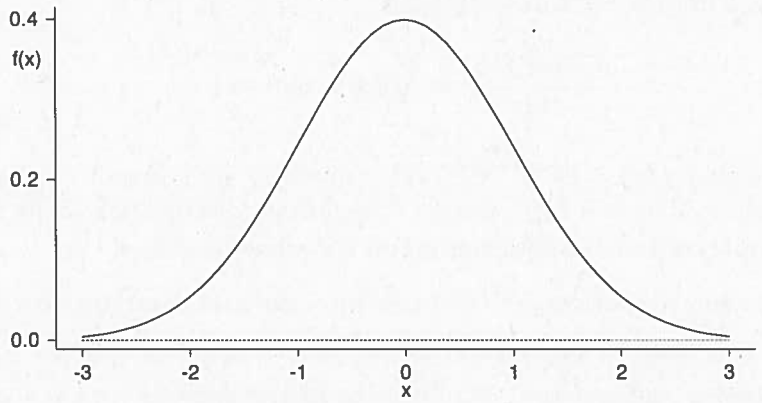

Figure A.1: Density of the standard normal distribution

Some values of the cumulative distribution function:

x	0	0.6745	1	1.64	1.96	2.58	3.3
$\Phi(x)$	$\frac{1}{2}$	$\frac{3}{4}$	84 % $\approx \frac{5}{6}$	95 %	97.5	99.5	99.95 %

The density $f(x)$ of the normal distribution goes to 0 more and more quickly in relative terms when $x \longrightarrow \pm\infty$, i.e. $|f'|/f \longrightarrow \infty$. Although the normal distribution has a **positive density** all the way from $-\infty$ to $+\infty$, it is in practice a very **"short-tailed"** distribution that all but disappears beyond $\mu \pm 3\sigma$ or $\mu \pm 4\sigma$. This contradicts most empirical distributions of measurement errors. Thus the **"normal approximation"** is usually

only helpful **in the middle** of the distribution, e.g. up to $\mu \pm 2\sigma$ or $\mu \pm 2.5\sigma$.

A.2.2 Multivariate normal distribution

Let $Y_1, Y_2 \ldots Y_n$ be independent standard normal random variables. Then their joint density is simply the product of their individual densities, that is,

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-n/2} \exp(-\tfrac{1}{2}\mathbf{y}^T\mathbf{y}).$$

This is the n-dimensional standard normal distribution. It exhibits spherical symmetry. The general n-dimensional normal distribution is defined as the distribution of a vector $\mathbf{X}$ that is obtained by a linear transformation of a standard normally distributed n-dimensional vector:

$$\mathbf{X} = A\mathbf{Y} + \boldsymbol{\mu}.$$

Here $\boldsymbol{\mu}$ is an $(n \times 1)$-dimensional vector, and A is an $(n \times n)$-dimensional matrix. If A is a singular matrix, we call the resulting distribution degenerate. We shall first discuss the non-degenerate case, however. In this case the distribution of $\mathbf{X}$ has the density

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= (2\pi)^{-n/2}(|\det A|)^{-1} \exp(-\tfrac{1}{2}(A^{-1}(\mathbf{x}-\boldsymbol{\mu}))^T(A^{-1}(\mathbf{x}-\boldsymbol{\mu}))) \\ &= (2\pi)^{-n/2}(\det \Sigma)^{-1/2} \exp(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})), \end{aligned}$$

where $\Sigma = AA^T$. According to the computational laws for means, we have

$$\mathbf{E}[\mathbf{X}] = A\mathbf{E}[\mathbf{Y}] + \boldsymbol{\mu} = \boldsymbol{\mu}$$

and

$$\mathrm{Cov}[\mathbf{X}] = \mathbf{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right] = \mathbf{E}\left[A\mathbf{Y}\mathbf{Y}^TA^T\right] = A\mathbf{E}\left[\mathbf{Y}\mathbf{Y}^T\right]A^T = AA^T = \Sigma.$$

Thus as in the univariate case, the normal distribution is determined entirely by the first two moments, and we therefore write $\mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$. The covariance matrix Σ can be an arbitrary positive definite matrix. The joint density has its maximum at $\boldsymbol{\mu}$ and is constant on similar ellipsoids centred on $\boldsymbol{\mu}$ (the principal axes of ellipsoids are given by the eigenvectors of Σ).

The following results both follow straight from this definition:

Theorem A.2.1. *Uncorrelated and jointly normally distributed random variables are independent. More precisely: if* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ *and if* $\Sigma_{ij} = 0$ *for all* $i \in I$ *and* $j \in J$ *for two disjoint index sets* $I, J \subset \{1, \ldots, n\}$*, then the random vectors* $(X_i, i \in I)$ *and* $(X_j, j \in J)$ *are independent of each other.*

Proof The joint density can be decomposed as the product of the individual densities.

Theorem A.2.2. *Standard normally distributed random variables keep their distribution when subjected to orthogonal transformations.*

With some more effort, we can even show the following:
Theorem A.2.3. *Linear combinations of jointly normally distributed random variables are themselves jointly normally distributed.*

Proof For linear combinations $A\mathbf{X}+\mathbf{b}$ that use an $n \times n$-dimensional matrix A, the claim follows immediately from the definition of the n-dimensional normal distribution. If less than n linear combinations are involved, we assume without loss of generality that $\mathbf{X}$ has a standard normal distribution. For the sake of simplicity, we take a single linear combination $\mathbf{a}^T\mathbf{X} = \sum_{i=1}^{n} a_i X_i$ and assume the length of $\mathbf{a}$ is 1. Then we choose an orthogonal matrix A whose first row is equal to $\mathbf{a}^T$, i.e. $\mathbf{a}^T\mathbf{X}$ is the first component of $A\mathbf{X}$. This makes it clear that $\mathbf{a}^T\mathbf{X}$ is normally distributed, as the components of $A\mathbf{X}$ are independent and normally distributed.

Thus we can understand the structure of the degenerate normal distribution:
Theorem A.2.4. *If $\mathbf{X}$ has a degenerate n-dimensional normal distribution, then it has r components $(X_{i_1}, X_{i_2}, \ldots X_{i_r})$ which have a non-degenerate r-dimensional distribution, while the remaining $n-r$ components can be written as linear combinations of these. Here r is the rank of Σ.*

We can furthermore use Theorem A.2.3 to show that in particular:
Corollary A.2.1. *The n marginal distributions of an n-dimensional normally distributed random variable themselves all follow normal distributions.*

The converse of this corollary is not true, however. Even if all marginal distributions of a random variable are normal, the joint distribution need not necessarily be a normal distribution!

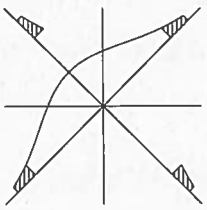

Figure A.2: Despite normal marginal distributions, the joint distribution need not be a normal one.

Consider the following example:
Let U be a univariate standard normal random variable. Let $X = Y = U$, which implies that the entire distribution of (X, Y) is concentrated on the diagonal and the marginal distributions are certainly normal. If we now cut away part of the joint distribution in a symmetric way (as illustrated in Figure A.2) and place the cut-off piece on the other diagonal, we **no longer** have a **joint normal distribution**, while the projections that gives us the marginal distributions are still the same, i.e. they are still normally distributed!

We should furthermore note that

$$\begin{aligned} &\text{Cut-off point} \longrightarrow \infty \quad \Longrightarrow \text{Correlation } +1 \\ &\text{Cut-off point} \longrightarrow 0 \quad \Longrightarrow \text{Correlation } -1, \end{aligned}$$

and between these two extremes the correlation varies in a continuous manner, including taking on the value 0 at some point. However, the two variables here are always dependent. Thus we also see that without joint normality, lack of correlation does not mean independence.

A.2.3 Chi-squared, t and F distributions

These distributions are derived from the normal distribution and have an important role in various tests in regression. Let

$$X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n \text{ independent } \sim \mathcal{N}(0,1).$$

Then the distribution of

$$Z_m = \sum_{i=1}^{m} X_i^2$$

is called the **chi-squared distribution** with m degrees of freedom (written χ^2_m). In particular, we have $\mathbf{E}[Z_m] = m$, $Var[Z_m] = 2m$ and

$$\mathcal{L}\left(\frac{Z_m - m}{\sqrt{2m}}\right) \overset{m\to\infty}{\longrightarrow} \mathcal{N}(0,1).$$

($\mathcal{L}$ stands for (distribution) law, and by convergence we mean weak convergence.)

The distribution of

$$V_n = \frac{X_1}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}}$$

is called t **distribution** with n degrees of freedom (written t_n). In particular, t_1 is the (standard) "Cauchy distribution", and

$$\mathcal{L}(V_n) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1).$$

The distribution of

$$W_{m,n} = \frac{\frac{1}{m}\sum_{i=1}^{m} X_i^2}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}$$

is called F **distribution** with m degrees of freedom in the enumerator and n degrees of freedom in the denominator (and is written $F_{m,n}$). In particular, we have

$$\mathcal{L}(W_{m,n}) \to \frac{1}{m}\chi^2_m \quad (m \text{ fest}, n \to \infty)$$

and

$$\mathcal{L}(W_{m,n}) \to 1, \quad (m \to \infty, n \to \infty).$$

On the computation of the densities of these distributions
We start with χ^2_1; that is, let $X \sim \mathcal{N}(0,1)$ be given and $\mathcal{L}(X^2)$ sought. The solution:

$$P[X^2 \le c] = P[|X| \le \sqrt{c}] = P[-\sqrt{c} \le X \le \sqrt{c}] = \Phi(\sqrt{c}) - \Phi(-\sqrt{c})$$

Differentiating gives us the corresponding density.

The formula for the density of χ^2_m can be obtained by repeated application of the convolution formula (density of the sum of independent random variables).

The formula for the t and F distributions is based on the following line of thought: let U and $V > 0$ be two independent random variables with densities f_U and f_V. Then we have the identity

$$P[\frac{U}{V} \leq x] = \int \int_{\{u \leq xv\}} f_U(u) f_V(v) du dv = \int_0^\infty f_V(v) F_U(xv) dv.$$

Taking derivatives by x on the right-hand side, we then obtain the density of U/V:

$$f_{U/V}(x) = \int_0^\infty f_V(v) f_U(xv) v dv.$$

As actually computing these integrals is a time-consuming process that provides no new insights, we omit it here.

Finally we mention a result than can sometimes be quite useful:

Lemma A.2.1. *If a random vector* $\mathbf{X}$ *has a* $\mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ *distribution, then if follows that* $(\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})$ *has a chi-squared distribution with* n *degrees of freedom.*

Proof We can write $\mathbf{X}$ as $\boldsymbol{\mu} + A\mathbf{Y}$, where $\mathbf{Y}$ follows a $\mathcal{N}_n(0, 1_n)$ distribution and $AA^T = \Sigma$. Then the quadratic form in $\mathbf{X}$ is exactly $\mathbf{Y}^T\mathbf{Y} = \sum Y_i^2$, and the claim thus follows by the definition of the chi-squared distribution.