Regression and Time Series Analysis

Claudia Redenbach

RPTU Kaiserslautern-Landau
Summer Term 2023

Contents

Chapter 1

Basic Notions of Probability Theory and Statistics

1.1 Random Variables, Distributions and their Parameters

Definition 1.1.1
Let Z be a real-valued random variable. We define

a) the distribution function of Z via

$$F(z) = P(Z \leq z), \ -\infty < z < \infty.$$

b) the density (if it exists) of F as $p(z) = F'(z)$ almost everywhere. Then

$$P(Z \in B) = \int_B p(z)\ dz \quad \text{for all} \quad B \in \mathcal{B} \text{ (Borel sets)}$$

c) the expectation or mean of Z (if it exists) as

$$E[Z] = \int zp(z)\ dz$$

d) the variance of Z as

$$\text{var}[Z] = E\left[(Z - E[Z])^2\right] = \int (z - E[Z])^2 p(z)\ dz\ (\leq \infty)$$

e) the standard deviation of Z as

$$\sigma(Z) = \sqrt{\text{var}Z}$$

Definition 1.1.2
A real-valued random variable Z is normally distributed or Gaussian with mean μ and variance σ^2 if it has the density

$$\varphi_{\mu,\sigma^2}(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(z-\mu)^2}{\sigma^2}}$$

Notation: $\mathcal{L}(Z) = \mathcal{N}(\mu, \sigma^2)$ $\quad$ ($\mathcal{L}(X)$ = law or distribution of X)

Definition 1.1.3
Let $Z_1, \ldots, Z_N$ be real-valued random variables. They are independent if

$$P(Z_1 \in B_1, \ldots, Z_N \in B_N) = \prod_{j=1}^{N} P(Z_j \in B_j) \text{ for all } B_1, \ldots, B_N \in \mathcal{B}.$$

If $Z_1, \ldots, Z_N$ have densities $p_1, \ldots, p_N$, then an equivalent condition for independence is the factorization of the joint density $p(z_1, \ldots, z_N)$ of $Z = (Z_1, \ldots, Z_N)^T \in \mathbb{R}^N$, i.e.

$$p(z_1, \ldots, z_N) = \prod_{j=1}^{N} p_j(z_j) \text{ for all } z_1, \ldots, z_N \in \mathbb{R}.$$

If $Z_1, \ldots, Z_N$ are independent and have the same distribution they are independent, identically distributed (i.i.d.).

Definition 1.1.4
Let X, Z be real-valued random variables with $E[X^2], E[Z^2] < \infty$. The

$$\begin{aligned} &\text{covariance of } X \text{ and } Z \text{ is} && \operatorname{cov}(X, Z) = E\big[(X - EX)(Z - EZ)\big] \\ &\text{correlation of } X \text{ and } Z \text{ is} && \operatorname{corr}(X, Z) = \frac{\operatorname{cov}(X, Z)}{\sigma(X)\sigma(Z)} \end{aligned}$$

Lemma 1.1.5

a) $-1 \leq corr(X, Z) \leq +1$

b) $corr(X, Z) = 1(-1)$ *iff* $Z = b_1 + b_2 X$ *almost surely for some* $b_1 \in \mathbb{R}, b_2 > 0 (< 0)$.

c) X, Z *independent implies* $cov(X, Z) = 0$, *i.e.* X, Z *are uncorrelated.*

d) X, Z *jointly normally distributed with* $cov(X, Z) = 0$ *implies that* X, Z *are independent.*

Definition 1.1.6
Real random variables $Z_1, \ldots, Z_d$ are jointly normally distributed if the random vector $Z = (Z_1, \ldots, Z_d)^T \in \mathbb{R}^d$ has a multivariate normal distribution with mean vector $\mu = (EZ_1, \ldots, EZ_d)^T$ and positive definite covariance matrix $\Sigma = (\operatorname{cov}(Z_i, Z_j))_{i,j=1,\ldots,d}$,
i.e. Z has the density

$$p(z) = p(z_1, \ldots, z_d) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(z-\mu)^T \Sigma^{-1} (z-\mu)}$$

Notation: $\mathcal{L}(Z) = \mathcal{N}_d(\mu, \Sigma)$, where $\Sigma = \operatorname{cov}(Z) = E[(Z - \mu)(Z - \mu)^T]$

Corollary 1.1.7
Let $\mathcal{L}(Z) = \mathcal{N}_d(\mu, \Sigma)$. *Then,* $Z_1, \ldots, Z_d$ *are independent iff* Σ *is diagonal, i.e.* $cov(Z_i, Z_j) = 0$ *for all* $i \neq j$.

Lemma 1.1.8
Let $Z = (Z_1, \ldots, Z_d)^T$ be a random vector in $\mathbb{R}^d$ with $EZ = \mu$, $cov(Z) = \Sigma$ and let A be an $m \times d$-matrix and $b \in \mathbb{R}^m$. Consider $X = AZ + b \in \mathbb{R}^m$.

a) $EX = A\mu + b, \quad cov(X) = A\Sigma A^T$

b) If $\mathcal{L}(Z) = \mathcal{N}_d(\mu, \Sigma)$, *then* $\mathcal{L}(X) = \mathcal{N}_m(A\mu + b,\ A\Sigma A^T)$.
In particular, for $m = 1,\ A = (a_1, \ldots, a_d) = a^T,\ b = 0$

$$\mathcal{L}\left(\sum_{k=1}^{d} a_k Z_k\right) = \mathcal{N}(a^T\mu, a^T\Sigma\ a).$$

Definition 1.1.9
Let $X_0, X_1, \ldots, X_n, Z_1, \ldots, Z_m$ be i.i.d. $\mathcal{N}(0,1)$.

a) $\mathcal{L}\left(\sum_{j=1}^{n} X_j^2\right) = \chi_n^2$ Chi-square distribution with n degrees of freedom (d.f.)

b) $\mathcal{L}\left(\frac{\sqrt{n}X_0}{\sqrt{X_1^2+\ldots+X_n^2}}\right) = t_n$ t-or Student distribution with n d.f.

c) $\mathcal{L}\left(\frac{\frac{1}{n}(X_1^2+\ldots+X_n^2)}{\frac{1}{m}(Z_1^2+\ldots+Z_m^2)}\right) = F_{n,m}$ F-or Fisher distribution with (n, m) d.f.

Definition 1.1.10
Let Z be a real random variable with distribution function F and density p, $0 < \alpha < 1$. Then, $q = F^{-1}(\alpha)$ is the α-quantile of $\mathcal{L}(Z)$, i.e.

$$P(Z \leq q) = \alpha.$$

1.2 Statistical Models, Estimators and Tests

In statistics, a data vector $z = (z_1, \ldots, z_N)^T$ is modelled as a random vector $Z = (Z_1, \ldots, Z_N)^T \in \mathbb{R}^N$. If the model assumes the distribution $\mathcal{L}(Z)$ to be known up to a finite-dimensional parameter $\vartheta \in \Theta \subseteq \mathbb{R}^d$, then we speak of a parametric statistical model:

$$\mathcal{L}(Z) \in \{P_\vartheta,\ \vartheta \in \Theta\}$$

where P_ϑ are probability distributions on $\mathbb{R}^N$.

Definition 1.2.1
Let $\mathcal{L}(Z) \in \{P_\vartheta,\ \vartheta \in \Theta\},\ \Theta \subseteq \mathbb{R}^d$.

a) An estimator $\hat{\vartheta}_N = T_N(Z)$ of ϑ is given by any measurable mapping $T_N : \mathbb{R}^N \to \Theta$.

b) T_N is consistent if $\hat{\vartheta}_N \xrightarrow[p]{} \vartheta$ for $N \to \infty$, i.e.

$P_\vartheta(||\hat{\vartheta}_N - \vartheta|| > \varepsilon) \xrightarrow[N \to \infty]{} 0$ for all $\varepsilon > 0$. (convergence in probability)

c) The mean-squared error (mse) of T_N is given by

$$\mathrm{mse}_\vartheta(T_N) = E_\vartheta\left[||T_N(Z) - \vartheta||^2\right] = E_\vartheta\left[||\hat{\vartheta}_N - \vartheta||^2\right]$$

The bias of T_N is given by

$$\mathrm{bias}_\vartheta(T_N) = E_\vartheta\left[T_N(Z)\right] - \vartheta = E_\vartheta\left[\hat{\vartheta}_N\right] - \vartheta$$

T_N is called unbiased if $E_\vartheta\left[T_N(Z)\right] = E_\vartheta\left[\hat{\vartheta}_N\right] = \vartheta$ for all $\vartheta \in \Theta$, i.e. $\mathrm{bias}_\vartheta(T_N) = 0$.

The notation $P_\vartheta, E_\vartheta, \mathrm{mse}_\vartheta, \ldots$ refers to the quantities computed under the assumption that ϑ is the true parameter of $\mathcal{L}(Z)$. Each of these quantities can be interpreted as a function on Θ.

Lemma 1.2.2
a) For $d = 1$ $mse_\vartheta(T_N) = var_\vartheta[T_N] + [bias_\vartheta(T_N)]^2$.
b) If $mse_\vartheta(T_N) \longrightarrow 0$ for $N \to \infty$, then T_N is consistent.

Example 1.2.3
$Z_1, \ldots, Z_N$ i.i.d. with mean μ and variance $\sigma^2 > 0$. Estimators for μ and σ^2 are the sample mean

$$\hat{\mu}_N = \overline{Z}_N = \frac{1}{N}\sum_{j=1}^{N} Z_j$$

and the sample variances

$$\hat{\sigma}_N^2 = \frac{1}{N}\sum_{j=1}^{N}(Z_j - \overline{Z}_N)^2 \text{ or } \hat{s}_N^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Z_j - \overline{Z}_N)^2,$$

respectively.

$\hat{\mu}_N$ is an unbiased estimator of μ. Therefore,

$$\mathrm{mse}_{\mu,\sigma^2}(\hat{\mu}_N) = \mathrm{mse}_{\mu,\sigma^2}(\overline{Z}_N) = \mathrm{var}_{\mu,\sigma^2}\overline{Z}_N = \frac{1}{N}\,\sigma^2 \xrightarrow[N \to \infty]{} 0$$

i.e. $\hat{\mu}_N = \overline{Z}_N$ is consistent by Lemma 1.2.2. This fact also follows from the law of large numbers (LLN). Essentially also from LLN follows $\hat{\sigma}_N^2, \hat{s}_N^2 \xrightarrow[p]{} \sigma^2$ for $N \to \infty$. After some calculations, we further get

$$E_{\mu,\sigma^2}[\hat{s}_N^2] = \sigma^2 \text{ (unbiased)}, \; E_{\mu,\sigma^2}[\hat{\sigma}_N^2] = E_{\mu,\sigma^2}\left[\frac{N-1}{N}\hat{s}_N^2\right] = (1 - \frac{1}{N})\,\sigma^2$$

i.e. $\text{bias}_{\mu,\sigma^2}(\hat{\sigma}_N^2) = -\frac{1}{N}\sigma^2$.

Example 1.2.4
$(X_1, Z_1)^T, \ldots, (X_N, Z_N)^T$ i.i.d. Then the sample covariance

$$\hat{c}_N = \frac{1}{N}\sum_{j=1}^{N}(X_j - \overline{X}_N)(Z_j - \overline{Z}_N)$$

is a consistent estimator of $c = \text{cov}(X_1, Z_1)$.

Theorem 1.2.5
Assume $Z_1, \ldots, Z_N$ i.i.d. $\mathcal{L}(Z_k) = \mathcal{N}(\mu, \sigma^2)$

a) $\overline{Z}_N$ and $\hat{s}_N^2$ (or $\hat{\sigma}_N^2$) are independent.

b) $\mathcal{L}\left(\frac{(N-1)\hat{s}_N^2}{\sigma^2}\right) = \chi^2_{N-1}$

c) $\mathcal{L}\left(\frac{\sqrt{N}(\overline{Z}_N - \mu)}{\hat{s}_N}\right) = t_{N-1}$

Definition 1.2.6
Let $Z = (Z_1, \ldots, Z_N)^T$ be a random vector with $\mathcal{L}(Z) \in \{P_\vartheta, \vartheta \in \Theta\}$, $\Theta \subseteq \mathbb{R}$, $0 < \gamma < 1$. Let $A_N = A_N(Z) \leq B_N(Z) = B_N$ be functions of the data Z. The random interval $[A_N, B_N]$ is a γ-confidence interval for ϑ if

$$P_\vartheta(\vartheta \in [A_N, B_N]) \geq \gamma \quad \text{for all} \quad \vartheta \in \Theta.$$

Example 1.2.7
Let $Z_1, \ldots, Z_N$ i.i.d. $\mathcal{L}(Z_k) = \mathcal{N}(\mu, \sigma^2)$. Then Theorem 1.2.5 b) and c) yield the following confidence intervals:

a) Let $\chi^2_{\frac{1-\gamma}{2}}$ and $\chi^2_{\frac{1+\gamma}{2}}$ be the $\frac{1-\gamma}{2}$- and $\frac{1+\gamma}{2}$-quantile, respectively, of χ^2_{N-1}. Then, a γ-confidence interval for σ^2 is

$$\left[\frac{(N-1)\hat{s}_N^2}{\chi^2_{\frac{1+\gamma}{2}}}, \frac{(N-1)\hat{s}_N^2}{\chi^2_{\frac{1-\gamma}{2}}}\right].$$

b) Let $t_{\frac{1-\gamma}{2}}$ and $t_{\frac{1+\gamma}{2}}$ be the $\frac{1-\gamma}{2}$- and $\frac{1+\gamma}{2}$-quantile, respectively, of t_{N-1}. Then, $t_{\frac{1-\gamma}{2}} = -t_{\frac{1+\gamma}{2}}$ and a γ-confidence interval for μ is

$$\left[\overline{Z}_N - \frac{\hat{s}_N}{\sqrt{N}}t_{\frac{1+\gamma}{2}}, \overline{Z}_N + \frac{\hat{s}_N}{\sqrt{N}}t_{\frac{1+\gamma}{2}}\right]$$

Definition 1.2.8
Let $Z = (Z_1, \ldots, Z_N)^T \in \mathbb{R}^N$ be a random vector with $\mathcal{L}(Z) \in \{P_\vartheta, \vartheta \in \Theta\}$, where P_ϑ has a density $p_\vartheta(z) = p_\vartheta(z_1, \ldots, z_N)$ for all $\vartheta \in \Theta$.
The likelihood of ϑ is $L(\vartheta|Z) = p_\vartheta(Z)$.
The log likelihood of ϑ is $\ell(\vartheta|Z) = \log L(\vartheta|Z)$.
The maximum likelihood (ML) estimator of ϑ is

$$\hat{\vartheta}_N = \arg\max_{\vartheta\in\Theta} L(\vartheta|Z) = \arg\max_{\vartheta\in\Theta} \ell(\vartheta|Z)$$

Remark 1.2.9
If $Z = (Z_1, \ldots, Z_N)^T,\ Z_1, \ldots, Z_N$ are i.i.d. with density $f_\vartheta(x),\ x \in \mathbb{R}$, then the density of Z is

$$p_\vartheta(z) = \prod_{j=1}^{N} f_\vartheta(z_j)$$

and

$$\ell(\vartheta|Z) = \sum_{j=1}^{N} \log f_\vartheta(Z_j).$$

Example 1.2.10
Let $Z = (Z_1, \ldots, Z_N)^T,\ Z_1, \ldots, Z_N$ i.i.d. $\mathcal{L}(Z_k) = \mathcal{N}(\mu, \sigma^2),\ \vartheta = (\mu, \sigma^2)^T$. Then

$$\log f_\vartheta(x) = \log \varphi_{\mu,\sigma^2}(x) = -\log\sqrt{2\pi} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}$$

$$\ell(\vartheta|Z) = \text{ const.} - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{N}(Z_j - \mu)^2$$

Setting the partial derivatives of $\ell(\vartheta|Z)$ w.r.t. μ, σ^2 to 0, we get that

$$\hat{\mu}_N = \overline{Z}_N \text{ and } \hat{\sigma}_N^2 = \frac{1}{N}\sum_{j=1}^{N}(Z_j - \overline{Z}_N)^2$$

are the ML-estimators of μ, σ^2.

Definition 1.2.11
Let $Z \in \mathbb{R}^N$ be a random vector, $\mathcal{L}(Z) \in \{P_\vartheta, \vartheta \in \Theta\},\ \Theta \subseteq \mathbb{R}^d$. Let $\Theta_0 \cup \Theta_1 = \Theta,\ \Theta_0 \cap \Theta_1 = \emptyset$.

a) A test of the hypothesis $H_0 : \vartheta \in \Theta_0$ against the alternative $H_1 : \vartheta \in \Theta_1$ is given by an acceptance region $S \subseteq \mathbb{R}^N$. If $Z \in S$, then H_0 is not rejected; if $Z \notin S$, then H_0 is rejected in favour of H_1.

b) If $Z \notin S$, but $\vartheta \in \Theta_0$, then H_0 is falsely rejected. This is an error of 1st kind. If $Z \in S$, but $\vartheta \in \Theta_1$, then H_0 is falsely retained. This is an error of 2nd kind.

c) If $P(\text{error of 1st kind}) = P_\vartheta(Z \notin S) \leq \alpha$ for all $\vartheta \in \Theta_0$ then α is the level of the test. The power of the test is defined by 1−P(error of 2nd kind), i.e.

$$\beta(\vartheta) = P_\vartheta(Z \notin S),\ \vartheta \in \Theta_1\,,$$

i.e. $\beta(\vartheta)$ is the probability of detecting correctly that the true parameter ϑ is not from Θ_0.

Example 1.2.12 (t-Test)
Let $Z_1, \ldots, Z_N$ be i.i.d. $\mathcal{L}(Z_k) = \mathcal{N}(\mu, \sigma^2)$. $H_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0$ (σ^2 arbitrary and unknown). Let

$$T_N = T_N(Z) = \frac{\sqrt{N}(\overline{Z}_N - \mu_0)}{\hat{s}_N}$$

be the t-statistic. If H_0 is true, $\mathcal{L}(T_N) = t_{N-1}$ (Theorem 1.2.5 c). The t-test has the acceptance region $S = \{z \in \mathbb{R}^N, |T_N(z)| \leq c_\alpha\}$. If c_α is the $(1 - \frac{\alpha}{2})$-quantile of t_{N-1}, then the test has level α, as

$$P_{\mu_0}(|T_N| > c_\alpha) = \alpha$$

The t-test is an example of a likelihood ratio test.

Definition 1.2.13
Let $Z \in \mathbb{R}^N$ satisfy $\mathcal{L}(Z) \in \{P_\vartheta, \vartheta \in \Theta\}$, where $\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$. The likelihood ratio (LR) of testing $H_0 : \vartheta \in \Theta_0$ against $H_1 : \vartheta \in \Theta_1$ is

$$\lambda(Z) = \frac{\max_{\vartheta \in \Theta_0} L(\vartheta|Z)}{\max_{\vartheta \in \Theta} L(\vartheta|Z)} = \frac{\max_{\vartheta \in \Theta_0} L(\vartheta|Z)}{L(\hat{\vartheta}_N|Z)}$$

where $\hat{\vartheta}_N$ is the ML-estimator of ϑ. A likelihood ratio test has an acceptance region of the form $S = \{z \in \mathbb{R}^N : \lambda(z) \geq c\}$.

Chapter 2

Linear Regression

The starting point of regression analysis are data which come in pairs:

$$(x_1, y_1), \ldots, (x_N, y_N)$$

The goal is to explain the values y_k as well as possible by the corresponding x_k, i.e. to investigate the dependence of y_k on the *explanatory or predictor variable* x_k. y_k and x_k are also called the *dependent* and the *independent variable*, respectively. Here, we always consider the situation of one-dimensional or univariate $y_k \in \mathbb{R}$ whereas $x_k \in \mathbb{R}^p$ may be higher-dimensional or multivariate.

In regression analysis, we model the y_k as realizations of random variables Y_k. The explanatory variables x_k may be deterministic or random depending on the actual experiment resulting in the data. The whole stochastic analysis is conditional on x_k. Hence, even in the random case the x_k are treated as given numbers and there is no essential difference between both cases.

The classical regression approach assumes that the random variables Y_k depend on x_k only through their means. The deviations from the mean caused, e.g., by observation errors, are assumed to be independent and identically distributed. Hence, we consider models of the form

$$Y_k = m(x_k) + Z_k, \quad k = 1, \ldots, N, \quad Z_1, \ldots, Z_N \text{ i.i.d. with mean 0 and variance } \sigma^2 < \infty.$$

Hence, $Y_1, \ldots, Y_N$, are independent random variables, all having the same distribution except for their differing means $\mathrm{E}Y_k = m(x_k)$.

The first goal of regression analysis is to estimate the *regression function* $m(x)$ from the data. In parametric regression, we assume that $m(x) = m(x; b)$ is known up to finitely many parameters forming a *parameter vector* $b \in \mathbb{R}^d$, i.e. we only have to estimate b from the data. Examples are

i) $m(x) = b_1 + b_2 x$

ii) $m(x) = b_1 + b_2 x + b_3 x^2$

iii) $m(x) = b_1 \cos x + b_2 \sin x$

iv) $m(x) = b_1 + b_2 \exp(-\frac{x}{b_3})$.

i) - iii) are examples of linear regression where the regression function is a linear function of the unknown parameters. In iv), the parameter b_3 enters in a nonlinear manner.

We mainly consider linear regression in this course. The ideas, concepts and methods can, however, be transferred to general nonlinear parametric regression problems. The two main additional complications are:

- we no longer have explicit formulas for the estimator of the parameter vector b; it is defined implicitly as an extremum of some data-dependent function and has to be calculated numerically;
- for proving theoretical results we have to introduce an additional step which essentially consists in a linearization, e.g. by a Taylor expansion of order 1.

2.1 Preliminaries

In this section, we study some of the main ingredients of regression analysis in a particularly simple situation where the independent variable x_k is univariate and where the regression function $m(x; b)$ is not only linear in the parameter but also affine-linear in x.

Example 2.1.1
Performances in athletics have improved during the last century. In this example we study the development of high jump performances over time by the example of the height achieved by the Olympic high jump champion. We get the *scatterplot* $(x_k, y_k), k = 1, \ldots, N$, of the data shown in Figure 2.1.
The first impression is that the height increases roughly linearly with time. To quantify this impression, we need to find a straight line $y = b_1 + b_2 x$, that fits the points (x_k, y_k), $k = 1, \ldots, N$. The *mean squared deviation* we need to minimize is given by

$$D(b_1, b_2) = \sum_{k=1}^{N} (y_k - b_1 - b_2 x_k)^2 .$$

Setting the partial derivatives of D with respect to b_1 and b_2 to 0, we obtain the solution of the optimization problem, the *least-squares (LS) estimators*

$$\hat{b}_1 = \overline{y}_N - \hat{b}_2 \overline{x}_N \tag{2.1}$$

$$\hat{b}_2 = \frac{\sum_{k=1}^{N} (y_k - \overline{y}_N)(x_k - \overline{x}_N)}{\sum_{k=1}^{N} (x_k - \overline{x}_N)^2}, \tag{2.2}$$

where $\overline{x}_N$ and $\overline{y}_N$ are the sample means of $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$, respectively. Figure 2.2 shows the data together with the resulting *least-squares line* $y = \hat{b}_1 + \hat{b}_2 x$.

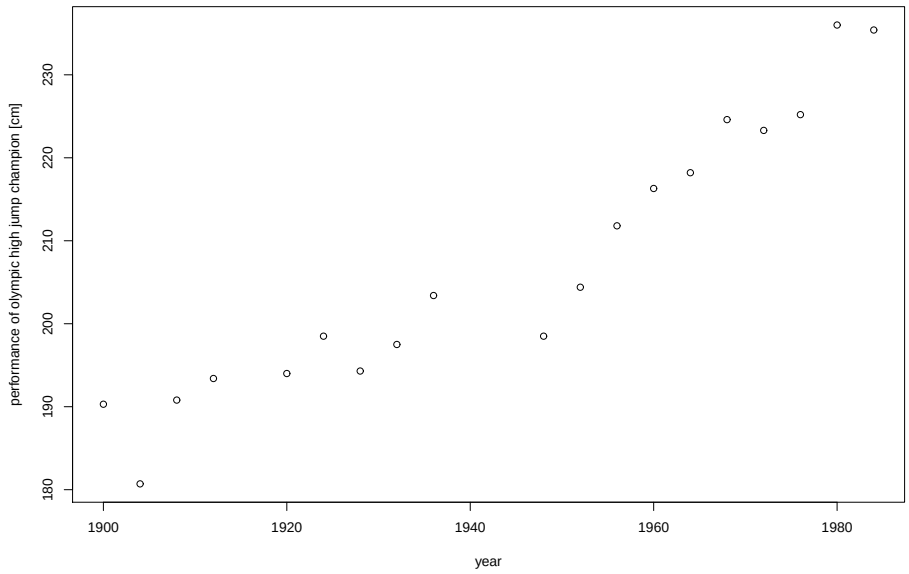

Figure 2.1: Performance of the Olympic high jump champion depending on the year.

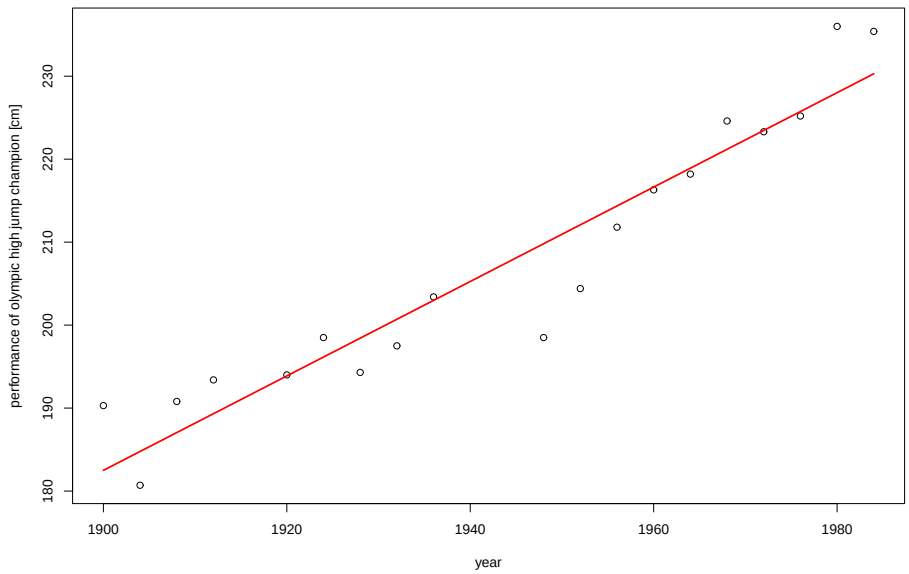

Figure 2.2: Least-squares line of height versus year.

Now, we can state the following questions:

a) How well does the relation $y = \hat{b}_1 + \hat{b}_2 x$ describe the dependence in the data?

b) Is it possible to describe the data as well by $y_k = \text{const.}$? In other words, is there any dependence at all between y_k and the independent variable x_k?

c) Is a parabola $y = b_1 + b_2x + b_3x^2$ (where the parameters can again be obtained by the least-squares approach) a much better model for the data? In other words: Can the complication caused by the additional parameter b_3 be justified?

To answer these questions, we need a statistical model for the data since the y_k are affected by the random measurement errors. As a starting point, we have the basic model for regression and analysis of variance.

Basic model:

1) $x_1, \ldots, x_N$ are known and deterministic.

2) The measurements $Y_k = m(x_k) + Z_k,\ k = 1, \ldots, N$, are realizations of independent real random variables.

3) $Z_1, \ldots, Z_N$ are i.i.d. with mean $\mathrm{E}Z_k = 0$.

We will often assume additionally that $Z_1, \ldots, Z_N$ are $\mathcal{N}(0, \sigma^2)$-distributed with unknown variance σ^2, i.e., $\mathcal{L}(Y_k) = \mathcal{N}(m(x_k), \sigma^2),\ k = 1, \ldots, N$.

Example 2.1.2 (Example 2.1.1 continued)
We model the data by

$$\text{(R1)} \qquad Y_k = b_1 + b_2x_k + Z_k,\ \ k = 1, \ldots, N, \quad Z_1, \ldots, Z_N \text{ i.i.d. } \mathcal{L}(Z_k) = \mathcal{N}(0, \sigma^2).$$

The Y_k are independent $\mathcal{N}(b_1 + b_2x_k, \sigma^2)$-distributed which allows for parameter estimation by maximum likelihood. The likelihood function is

$$L(\vartheta | y_1, \ldots, y_N) = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_k - b_1 - b_2x_k)^2}{2\sigma^2} \right)$$

with $\vartheta = (b_1, b_2, \sigma^2)$. Thus, the log–likelihood function is of the form

$$l(\vartheta | y_1, \ldots, y_N) = \text{ const } - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} D(b_1, b_2),$$

and it follows that the ML estimators for b_1, b_2 are identical to the least squares estimators $\hat{b}_1, \hat{b}_2$ of b_1, b_2.

Additionally, we obtain the ML estimator of the scale parameter σ^2

$$\hat{\sigma}^2 = \frac{1}{N} D(\hat{b}_1, \hat{b}_2) = \frac{1}{N} \sum_{k=1}^{N} (Y_k - \hat{b}_1 - \hat{b}_2x_k)^2. \tag{2.3}$$

Now the above questions b) and c) can be formulated as statistical decision problems or hypothesis tests:
b) Test $H_0 : b_2 = 0$ against $H_1 : b_2 \neq 0$.
c) In the model $Y_k = b_1 + b_2x_k + b_3x_k^2 + Z_k,\ k = 1, \ldots, N,$ where $Z_1, \ldots, Z_N$ are i.i.d.

$\mathcal{L}(Z_k) = \mathcal{N}(0, \sigma^2)$, test the hypothesis $H_0 : b_3 = 0$ against the alternative $H_1 : b_3 \neq 0$.

We assume model (R1) for the high jump data, and we want to check if height depends on time at all. For this purpose, we test the hypothesis $b_2 = 0$ against the alternative $b_2 \neq 0$. We want to apply the likelihood ratio (LR) test. Consider the data vector $Y = (Y_1, \ldots, Y_N)^T$. To compute the likelihood–ratio

$$\lambda(Y) = \frac{\max_{b_1,\sigma^2} L(b_1, 0, \sigma^2|Y)}{\max_{b_1,b_2,\sigma^2} L(b_1, b_2, \sigma^2|Y)}$$

we must first maximize the likelihood function over the hypothesis set $\mathbb{R} \times \{0\} \times (0, \infty)$ as well as over the set of all parameters $\mathbb{R}^2 \times (0, \infty)$. We have already shown that the ML estimators of b_1, b_2 are identical to the least squares estimators (2.1), (2.2) of these parameters, and the ML estimator of σ^2 is given by (2.3).

Under the hypothesis $b_2 = 0$, the data $Y_1, \ldots, Y_N$ are i.i.d $\mathcal{L}(Y_k) = \mathcal{N}(b_0, \sigma_0^2)$. Therefore, the likelihood–function $L(b_0, 0, \sigma_0^2|Y)$ becomes maximal when we use the well-known ML estimators

$$\hat{b}_0 = \overline{Y}_N, \;\; \hat{\sigma}_0^2 = \frac{1}{N}\sum_{k=1}^{N}(Y_k - \overline{Y}_N)^2$$

for the mean and variance of a sample of independent normally distributed random variables. The likelihood function for the model (R1) evaluated at its maximum is

$$\begin{aligned} L(\hat{b}_1, \hat{b}_2, \hat{\sigma}^2|Y) &= \prod_{k=1}^{N}(2\pi\hat{\sigma}^2)^{-\frac{1}{2}} \exp\left\{\frac{-(Y_k - \hat{b}_1 - \hat{b}_2 x_k)^2}{2\hat{\sigma}^2}\right\} \\ &= (2\pi\hat{\sigma}^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_{k=1}^{N}(Y_k - \hat{b}_1 - \hat{b}_2 x_k)^2\right\} \\ &= (2\pi\hat{\sigma}^2)^{-\frac{N}{2}} e^{-\frac{N}{2}} \end{aligned}$$

by (2.3). Analogously, under the hypothesis $L(\hat{b}_0, 0, \hat{\sigma}_0^2|Y) = (2\pi\hat{\sigma}_0^2)^{-\frac{N}{2}} e^{-\frac{N}{2}}$. Hence, we get the likelihood ratio as

$$\lambda(Y) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{N}{2}}.$$

We do not reject the hypothesis $b_2 = 0$ if $\lambda(Y) \geq c_\alpha$, i.e. if, in view of the data, the plausibility of the coarser model corresponding to the hypothesis is not much less than the plausibility of the full model (R1). To determine c_α such that the test has the chosen level α, we have to find the distribution of the likelihood ratio statistic or, equivalently, of a monotone transformation of $\lambda(Y)$.

For that purpose, we look at a geometric interpretation of the likelihood ratio and the related

test. We need the following notation:

$$\begin{aligned} M &= \{(a + bx_1, \ldots, a + bx_N)^T;\ a, b \in \mathbb{R}\} \subseteq \mathbb{R}^N \\ M_0 &= \{(a, \ldots, a)^T;\ a \in \mathbb{R}\} \subseteq M \\ Y &= (Y_1, \ldots, Y_N)^T \text{ data vector} \\ \widehat{\mu} &= (\hat{b}_1 + \hat{b}_2 x_1, \ldots, \hat{b}_1 + \hat{b}_2 x_N)^T \in M \\ \widehat{\mu}^{(0)} &= (\hat{b}_0, \ldots, \hat{b}_0)^T \in M_0. \end{aligned}$$

$M_0 \subset M$ both are linear subspaces of the sample space $\mathbb{R}^N$. In the regression model (R1) the expectation EY of the data vector lies in M; under the hypothesis $b_2 = 0$ it lies in M_0. $\widehat{\mu}$ is the estimator of EY that we obtain when the parameters are estimated with the least squares approach. It follows from the definition of the least squares estimators that

$$||Y - \widehat{\mu}||^2 = D(\hat{b}_1, \hat{b}_2) = \min_{\mu \in M} ||Y - \mu||^2 .$$

Hence, $\widehat{\mu}$ is the orthogonal projection of the data vector Y onto the subspace M of $\mathbb{R}^N$. Accordingly, $\widehat{\mu}^{(0)}$ is the orthogonal projection of Y on M_0. The ML variance estimators in the model (R1) and according to the reduced model under the hypothesis $b_2 = 0$

$$\hat{\sigma}^2 = \frac{1}{N}||Y - \widehat{\mu}||^2 \text{ and } \hat{\sigma}_0^2 = \frac{1}{N}||Y - \widehat{\mu}^{(0)}||^2$$

are, up to the factor $\frac{1}{N}$, the squared distances between Y and the spaces M and M_0, respectively. The likelihood ratio is consequently

$$\lambda(Y) = \left(\frac{||Y - \widehat{\mu}||}{||Y - \widehat{\mu}^{(0)}||} \right)^N \leq 1 ,$$

since, due to $M_0 \subseteq M$, we always have $||Y - \widehat{\mu}|| \leq ||Y - \widehat{\mu}^{(0)}||$. So the LR test of the hypothesis $b_2 = 0$ against the alternative $b_2 \neq 0$ rejects H_0 if the data vector is much closer to the subspace M of all possible expectation values EY under the full model (R1) than to M_0, the subspace of admissible expectation values of Y under the hypothesis.
Still, we need to find the distribution of $\lambda(Y)$. However, instead of $\lambda(Y)$, we use a monotone decreasing transformation

$$R(Y) = (N-2)\{\lambda(Y)^{-\frac{2}{N}} - 1\} = (N-2)\left\{ \frac{||Y - \widehat{\mu}^{(0)}||^2}{||Y - \widehat{\mu}||^2} - 1 \right\},$$

since the test statistic $R(Y)$ has a simpler distribution. Due to monotonicity, H_0 should then be rejected if $R(Y)$ is too large.
As $Y - \widehat{\mu} \perp M$, $\widehat{\mu} - \widehat{\mu}^{(0)} \in M$, we get from Pythagoras' Theorem

$$||Y - \widehat{\mu}||^2 + ||\widehat{\mu} - \widehat{\mu}^{(0)}||^2 = ||Y - \widehat{\mu}^{(0)}||^2 .$$

hence

$$R(Y) = (N-2)\frac{||\widehat{\mu} - \widehat{\mu}^{(0)}||^2}{||Y - \widehat{\mu}||^2}.$$

From the Fundamental Theorem of Linear Models (Theorem 2.5.1) we see that under the hypothesis $b_2 = 0$

$$P_{H_0}(R(Y) > f_\alpha) = \alpha,$$

where f_α is the $(1-\alpha)$-quantile of the Fisher distribution $F_{1,N-2}$. Hence, we reject H_0 if $R(Y) > f_\alpha$
We summarize:

Proposition 2.1.3
Let the regression model (R1) hold.
a) The ML estimators of b_1, b_2 are the LS estimators $\hat{b}_1, \hat{b}_2$ given by (2.1), (2.2), and the ML estimator $\hat{\sigma}^2$ for σ^2 is given by (2.3).
b) The likelihood ratio test statistic for $H_0 : b_2 = 0$ against $H_1 : b_2 \neq 0$ is

$$\lambda(Y) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{N}{2}} = \left(\frac{||Y - \widehat{\mu}||}{||Y - \widehat{\mu}^{(0)}||}\right)^N.$$

c) In the regression model (R1), under the hypothesis $H_0 : b_2 = 0$, the statistic

$$R(Y) = (N-2)\frac{||\widehat{\mu} - \widehat{\mu}^{(0)}||^2}{||Y - \widehat{\mu}||^2}$$

is $F_{1,N-2}$-distributed.

Example 2.1.4 (Example 2.1.1 continued)
For the highjump data, we obtain $R(Y) = 169.35$. The 0.95-quantile of the $F_{1,17}$-distribution is $f_{0.05} = 4.45$. Hence, $H_0 : b_2 = 0$ is rejected.

2.2 The General Linear Regression Model

Now we focus on the general linear regression model, in which the expectations are linear functions of d unknown parameters $b_1, \ldots, b_d$. The regression function is a function of d *factors* or *regressors* with known values $x_{k1}, \ldots, x_{kd}$ in the k-th random experiment. The factors may be separate explanatory variables or different functions of the same explanatory variable or combinations of both.

We assume the following *linear regression model with Gaussian residuals*

$$(LRG) \qquad Y_j = b_1 x_{j1} + \ldots + b_d x_{jd} + Z_j, \;\; j = 1, \ldots, N, \;\; Z_1, \ldots, Z_N \text{ i.i.d. } \mathcal{N}(0, \sigma^2).$$

In vector form, we have

$$(LRG) \qquad Y = \mathbf{X}\, b + Z, \;\; Z_1, \ldots, Z_N \text{ are i.i.d. } \mathcal{N}(0, \sigma^2).$$

where $Y = (Y_1, \ldots, Y_N)^T$ is the data vector,

$Z = (Z_1, \ldots, Z_N)^T$ the vector of residuals,

$b = (b_1, \ldots, b_d)^T$ the d-dimensional parameter vector and

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1d} \\ \vdots & & \vdots \\ x_{N1} & \ldots & x_{Nd} \end{pmatrix}$$

is the $N \times d$ design matrix. The design matrix describes for which combinations of the factors $x_{j1}, \ldots, x_{jd}$, $j = 1, \ldots, N$, data have been collected.

Example 2.2.1
If we have one explanatory variable x, assuming the value x_j in the j-th experiment, and if we fit a quadratic function $b_1 + b_2 x + b_3 x^2$, to the data, we get the special case of (LRG) with

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{pmatrix}.$$

The k-th column of $\mathbf{X}$ corresponds to the different values of the k-th factor in the sample, the j-th row of $\mathbf{X}$ corresponds to the values of all factors in the j-th experiment.

The least squares estimator $\hat{b}$ of the parameter vector is obtained by minimizing

$$D(b) = \sum_{j=1}^{N} \big(Y_j - b_1 x_{j1} - \ldots - b_d x_{jd}\big)^2 = ||Y - \mathbf{X}\, b||^2 \tag{2.4}$$

with respect to $b \in \mathbb{R}^d$. Again, the LS estimator coincides with the ML estimator in the Gaussian model. We formulate that result in the more general context of general nonlinear regression models.

Proposition 2.2.2
For $\mu(b) = (\mu_1(b), \ldots, \mu_N(b))^T$ known up to $b \in \mathbb{R}^d$, let

$$Y_j = \mu_j(b) + Z_j, \;\; j = 1, \ldots, N, \;\; Z_1, \ldots, Z_N \text{ i.i.d.} \;\; \mathcal{L}(Z_i) = \mathcal{N}(0, \sigma^2).$$

Then, the LS estimator $\hat{b}$ defined by minimizing

$$D(b) = \sum_{j=1}^{N} \big(Y_j - \mu_j(b)\big)^2 = ||Y - \mu(b)||^2$$

coincides with the ML estimator, and the ML estimator of σ^2 is given by $\hat{\sigma}^2 = \frac{1}{N} D(\hat{b})$.

Proof:
Since the Y_j are independent $\mathcal{N}(\mu_j(b), \sigma^2)$-distributed, the likelihood, i.e. the joint density at the data, is

$$L(b, \sigma^2 | Y_1, \ldots, Y_N) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(Y_j - \mu_j(b))^2}{2\sigma^2} \right) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(- \frac{D(b)}{2\sigma^2} \right).$$

Hence, for any σ^2, the likelihood is maximized w.r.t. b iff $D(b)$ is minimized. The form of $\hat{\sigma}^2$ follows from setting the derivative of $\log L(\hat{b}, \sigma^2 | Y_1, \ldots, Y_N)$ w.r.t. σ^2 to 0 and solving for σ^2. □

In (LRG), we have $\mu(b) = \mathbf{X}b$. Therefore, the minimizer of (2.4) is the ML estimator of b. The following theorem provides an explicit formula for this estimator and its distribution under the model (LRG). For sake of simplicity we limit ourselves to the case where the *design vectors*, i.e. the columns of the design matrix,

$$x^{(j)} = (x_{1j}, \ldots, x_{Nj})^T, \;\; j = 1, \ldots, d,$$

are linearly independent, hence the matrix $\mathbf{X}^T\mathbf{X}$ is invertible. Without this assumption at least one factor would always be a linear combination of the other factors and therefore redundant.

Theorem 2.2.3
Assume that $x^{(1)}, \ldots, x^{(d)}$ are linearly independent and consequently $\mathbf{X}^T\mathbf{X}$ is invertible.

a) The least squares estimator of the parameter vector b in the model (LRG) is

$$\hat{b} = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T Y,$$

and the ML estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{N} ||Y - \mathbf{X}\hat{b}||^2.$$

b) $\hat{b}$ is d-dimensional normally distributed with mean

$$E\,[\hat{b}] = b \qquad \textit{(unbiasedness)}$$

and covariance matrix

$$\Sigma_b = E\left[(\hat{b} - b)(\hat{b} - b)^T \right] = \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}.$$

Proof:
a) Observe that

$$\mathrm{E}[Y] = \mathbf{X}b = \sum_{k=1}^{d} b_k x^{(k)} \in Lin\{x^{(1)}, \ldots, x^{(d)}\} = \{\mathbf{X}c \,:\, c \in \mathbb{R}^d\} = M_X \subset \mathbb{R}^N$$

Therefore, if $\hat{b}$ minimizes $D(b) = ||Y - \mathbf{X}b||^2$, $\widehat{Y} = \mathbf{X}\hat{b}$ will be the orthogonal projection of the data vector Y onto M_X. Hence, $Y = \mathbf{X}\hat{b} + Y'$ where $Y' \perp M_X$, i.e. $\mathbf{X}^T Y' = 0$. This implies

$$\mathbf{X}^T Y = \mathbf{X}^T \mathbf{X}\hat{b}, \quad \text{i.e.} \quad \hat{b} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T Y.$$

The second part follows directly from Proposition 2.2.2.
b) Note that the data vector Y is N-variate normally distributed with mean $\mathbf{X}b$ and covariance matrix $\sigma^2 I_N$ where I_N denotes the $N \times N$- unit matrix. Then, by Lemma 1.1.8, $\hat{b}$ as a linear transformation of Y is again normally distributed with mean

$$E[\hat{b}] = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{X}b = b$$

and covariance matrix

$$\Sigma_b = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\sigma^2 I_N\;\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

□

Remark 2.2.4
Without the assumption of a normal distribution, Theorem 2.2.3 remains asymptotically ($N \to \infty$) valid if we only assume that the residuals $Z_1, \ldots, Z_N$ are independent with $EZ_j = 0$, $\operatorname{var} Z_j = \sigma^2 < \infty$, $j = 1, \ldots, N$. The LS estimator is, however, obviously no longer the ML estimator such that we do not know anything about its optimality. We shall see in Section 2.8 that it is still optimal in a restricted class of estimators. From practical experience, LS estimators are generally reasonable with one exception: if there are outliers in the data, e.g. due to a heavy-tailed distribution of the residuals or due to gross recording errors, the LS approach may lead to wrong results. In that situation, one has to take recourse to *robust regression* techniques.

2.3 Model check by analysis of sample residuals

We have a look at methods for checking the quality of a regression model. On the one hand, a good model should not be unnecessarily complicated. On the other hand, the model should describe the data generating mechanism adequately. One way to make the latter statement more precise is to require that the model is able to produce good forecasts of future data.

We first consider again the simple model (R1), and assume that the assumption of Theorem 2.2.3 is satisfied, i.e.

$$(x_1, \ldots, x_N)^T \neq c(1, \ldots, 1)^T \quad \text{for all } c. \tag{2.5}$$

Based on N observations, we estimate the regression parameters by the LS estimators $\hat{b}_1, \hat{b}_2$. Now, let us assume that we plan an additional experiment with outcome Y_{N+1} at $x_{N+1} = x$, and we want to predict the random variable

$$Y_{N+1} = E[Y_{N+1}] + Z_{N+1} = b_1 + b_2 x + Z_{N+1}.$$

As we do not know the parameters, we have to replace them by their estimates. That is, a feasible prediction would be

$$\widehat{Y}_{N+1} = \hat{b}_1 + \hat{b}_2 x.$$

The quality of that estimate can be measured by the *mean squared prediction error*

$$\mathrm{E}\left[(Y_{N+1} - \widehat{Y}_{N+1})^2\right] = \mathrm{E}\left[(Z_{N+1} + \mathrm{E}Y_{N+1} - \widehat{Y}_{N+1})^2\right] = \mathrm{E}[Z_{N+1}^2] + \mathrm{E}\left[(\{b_1 - \hat{b}_1\} + \{b_2 - \hat{b}_2\}x)^2\right],$$

where we have used that $Z_{N+1} = Y_{N+1} - \mathrm{E}[Y_{N+1}]$ has mean 0 and is independent of $\widehat{Y}_{N+1}$ as the latter depends only on the first N observations.

In practice, we do not know the true value Y_{N+1}. So the best we can do is compare prediction and observation for our data points. This motivates the definition of the *sample residuals*

$$\widehat{Z}_j = Y_j - \hat{Y}_j = Y_j - \hat{b}_1 - \hat{b}_2 x_j,\ j = 1, \ldots, N.$$

Proposition 2.3.1
Assume model (R1) and (2.5).

a) $\widehat{Z} = (\widehat{Z}_1, \ldots, \widehat{Z}_N)^T$ *is* N*-variate normally distributed with*

$$\begin{aligned} E[\widehat{Z}_j] &= 0, \qquad var[\widehat{Z}_j] = \sigma^2\left(1 - \frac{1}{N} - \frac{1}{N}\frac{(x_j - \overline{x}_N)^2}{\overline{x_N^2} - (\overline{x}_N)^2}\right) \\ cov(\widehat{Z}_j, \widehat{Z}_k) &= -\sigma^2\left(\frac{1}{N} + \frac{1}{N}\frac{(x_j - \overline{x}_N)(x_k - \overline{x}_N)}{\overline{x_N^2} - (\overline{x}_N)^2}\right), j \neq k, \end{aligned}$$

where we have used the abbreviation $\overline{x_N^2} = \frac{1}{N}\sum_{j=1}^{N} x_j^2$.

b) $\widehat{Z} \perp Lin(x^{(1)}, x^{(2)}) = Lin\big((1, \ldots, 1)^T, (x_1, \ldots, x_N)^T\big)$, *i.e.*

$$\sum_{j=1}^{N} \widehat{Z}_j = 0, \quad \sum_{j=1}^{N} x_j \widehat{Z}_j = 0.$$

Proof:
a) By Theorem 2.2.3, we have

$$\hat{Z} = Y - \mathbf{X}\hat{b} = Y - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = (I_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)Y.$$

The data vector Y is $\mathcal{N}_N(\mathbf{X}b, \sigma^2 I_N)$-distributed. Hence, Lemma 1.1.8 shows that $\hat{Z}$ is a normal random vector. For the mean, we have $E[\hat{Z}] = E[Y] - E[\mathbf{X}\hat{b}] = \mathbf{X}b - \mathbf{X}E[\hat{b}] = 0$. The form of the covariances between the coordinates of $\widehat{Z}$ follows by straightforward calculation again from Lemma 1.1.8.

b) $\widehat{Y} = \hat{b}_1 x^{(1)} + \hat{b}_2 x^{(2)}$ is the orthogonal projection of Y onto $\mathrm{Lin}(x^{(1)}, x^{(2)})$ as

$$\sum_{j=1}^{N} (Y_j - b_1 - b_2 x_j)^2 = ||Y - b_1 x^{(1)} - b_2 x^{(2)}||^2$$

is minimized by the least squares estimators $\hat{b}_1, \hat{b}_2$.Therefore, $\widehat{Z} = Y - \widehat{Y} \perp \mathrm{Lin}(x^{(1)}, x^{(2)})$. $\square$

Remark 2.3.2
If N is large, $\mathrm{var}\widehat{Z}_j \approx \sigma^2$ and $\mathrm{cov}(\widehat{Z}_j, \widehat{Z}_k) \approx 0, j \neq k$, i.e. if the model is correct, $\widehat{Z}_1, \ldots, \widehat{Z}_N$ should look similar to i.i.d. normal random variables with mean 0 and variance σ^2. We generalize this result to arbitrary linear regression models.

Proposition 2.3.3
Assume model (LRG), and let $\mathbf{X}^T\mathbf{X}$ *be invertible. Let* $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ *be the so-called* hat matrix.

a) $\widehat{Z} = Y - \mathbf{X}\hat{b} = (I_N - H)Y$ *is* N*-variate normal:*

$$\mathcal{L}(\widehat{Z}) = \mathcal{N}_N\big(0, \sigma^2(I_N - H)\big)$$

b) $\widehat{Z} \perp Lin(x^{(1)}, \ldots, x^{(d)})$*, where* $x^{(1)}, \ldots, x^{(d)}$ *are the columns of* $\mathbf{X}$*.*

Proof:
For any $c \in \mathbb{R}^d$, we have

$$(I_N - H)\mathbf{X}c = \mathbf{X}c - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}c = 0 \tag{2.6}$$

a) We just show the statement for the covariances. The rest follows as in Proposition 2.3.1 a).
$\mathrm{cov}(\widehat{Z}) = (I_N - H)\mathrm{cov}(Y)(I_N - H)^T = \sigma^2(I_N - H)^2$ as H is symmetric.
It remains to show that $(I_N - H)^2 = I_N - H$. For any $y \in \mathbb{R}^N$:
$(I_N - H)^2 y = (I_N - H)y - (I_N - H)Hy = (I_N - H)y$ as, by (2.6), $(I_N - H)\mathbf{X}\underbrace{(\mathbf{X}^T\mathbf{X}^{-1})\mathbf{X}^T y}_{=:c} = 0$.

b) $\mathrm{Lin}(x^{(1)}, \ldots, x^{(d)}) = \{\mathbf{X}c : c \in \mathbb{R}^d\} =: M_X \subseteq \mathbb{R}^N$
For any $u = \mathbf{X}c \in M_X$, we have from a), symmetry of H, and using again (2.6) for the last equation

$$\langle \widehat{Z}, u \rangle = \widehat{Z}^T u = Y^T(I_N - H)^T u = Y^T(I_N - H)u = 0.$$

□

Remark 2.3.4

- The name hat matrix stems from the fact that

  $$HY = X(X^TX)^{-1}X^TY = X\hat{b} = \hat{Y}.$$

  So application of H 'puts a hat' onto Y.

- Another name is influence matrix. It can be explained as follows. We have $\hat{Y}_i = \sum_j H_{ij}Y_j$, hence

  $$\frac{\partial \hat{Y}_i}{\partial Y_j} = H_{ij}.$$

  In particular, the diagonal elements H_{ii} measure the influence of the Y_i on their own prediction (leverage).

Looking at the *residual plots* is a simple method to detect rough deviations of the data from the postulated model. The true residuals $Z_1, \ldots, Z_N$, which are i.i.d under the model cannot be observed. However, we may estimate them by the sample residuals $\widehat{Z} = (\widehat{Z}_1, \ldots, \widehat{Z}_N)^T$ discussed above. As we have seen, $\widehat{Z}$ has mean value 0 and covariance matrix

$$\mathrm{E}\left[\widehat{Z}\widehat{Z}^T\right] = \sigma^2(I_N - H) = \sigma^2(I_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \ .$$

Under suitable and not too strong assumptions on the design matrix $\mathbf{X}$, it follows that

$$\text{var } \widehat{Z}_i \approx \sigma^2, \ \ \text{cov}(\widehat{Z}_i, \widehat{Z}_j) \approx 0, \quad i, j = 1, \ldots, N.$$

Therefore, $\widehat{Z}_1, \ldots, \widehat{Z}_N$ should look similar to i.i.d random variables with mean value 0. We can check this based on the residual plots, i.e. the graphic representation of the points

$$(\widehat{Y}_j, \widehat{Z}_j), \ \ j = 1, \ldots, N \ , \ \text{ with } \widehat{Y} = (\widehat{Y}_1, \ldots, \widehat{Y}_N)^T = \mathbf{X}\hat{b}.$$

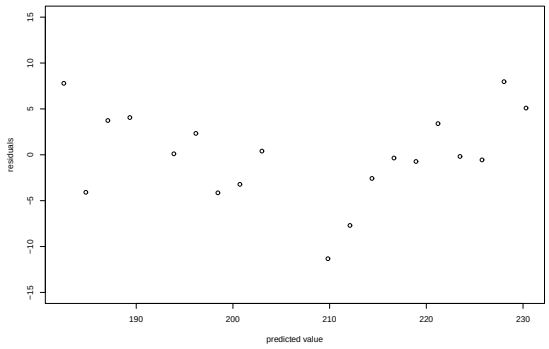

Figure 2.3: Residual plot for the highjump data.

2.4 Violations of the model assumptions

In the following, we will discuss several violations of the model assumptions as well as strategies for their treatment.

2.4.1 Wrong regression curve

Problem:
We have $Y_k = \mu(x_k) + Z_k$,but $\mu(x) \neq b_1 + b_2 x$ for any b_1, b_2.

Consequences: Sometimes negligible, sometimes very serious:

- Figure 2.4 left: Prediction of $\mu(x)$ is almost ok. Compared to the size of the random residuals Z_k, the difference $|\mu(x) - \hat{b}_1 - \hat{b}_2 x|$ is small (at least in the center of the plot).

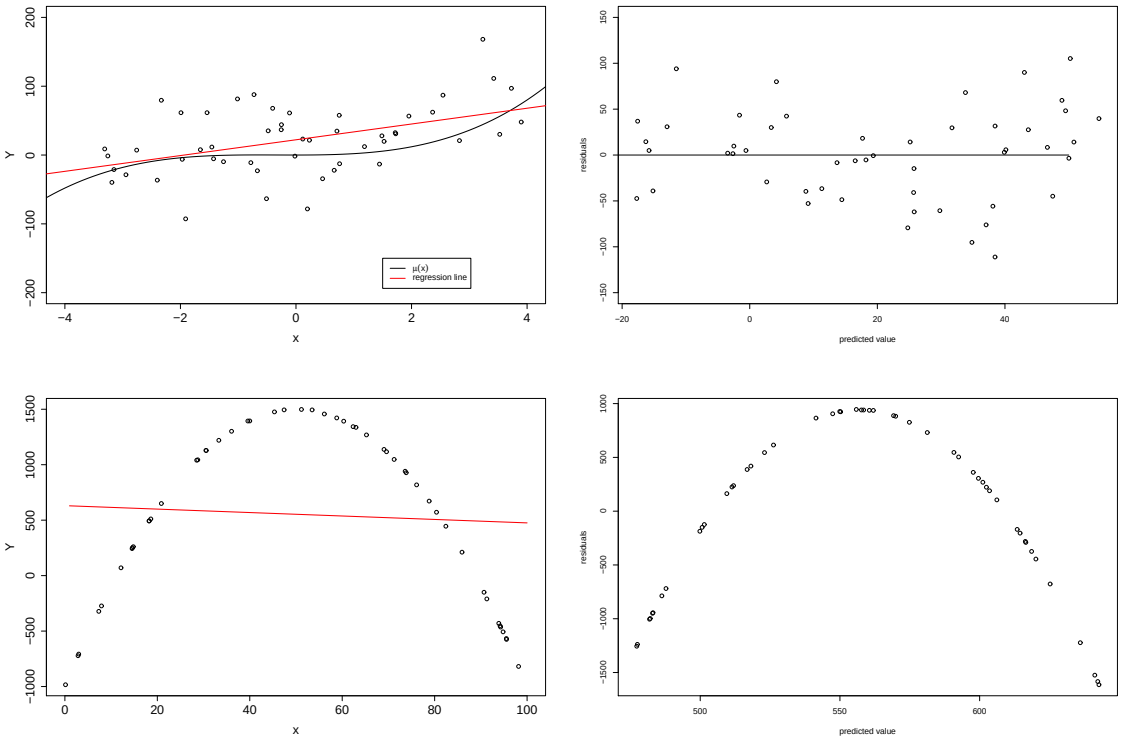

Figure 2.4: Examples where a regression line leads to the wrong model. Top: True mean (black) and regression line (red). Residual plots are shown on the right.

- Figure 2.4 right: The prediction may be highly erroneous. Since $\hat{b}_2 \approx 0$ we may even conclude that Y_k does not depend on x_k.

Remedies:

1. Extend the model, e.g., to $Y_k = b_1 + b_2 x_k + b_3 x_k^2 + Z_k$ (assuming that the data may be fitted by some low order polynomial).

2. Apply the model to transformed data, e.g., $Y_k^* = \log Y_k = b_1 + b_2 x_k + Z_k$ (see Figure 2.5).

3. "Automatic" model selection procedures (see Section 2.6).

4. If it is not clear how to formulate a parametric model, nonparametric regression techniques may be used.

2.4.2 Heteroskedasticity

Problem:
The variance $\text{var}[Z_j]$ depends on j, e.g., $\mathcal{L}(Z_j) = \mathcal{N}(0, \sigma_j^2)$.

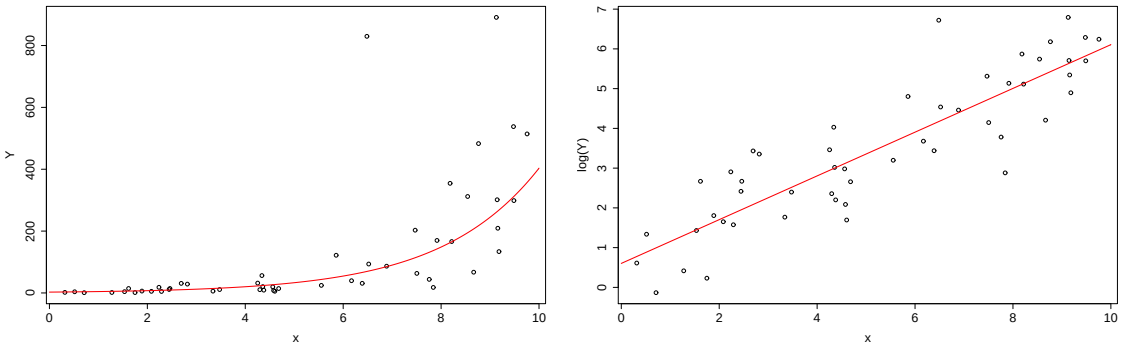

Figure 2.5: A transformation of the data can justify the use of a linear model. Here, a loglinear model is shown.

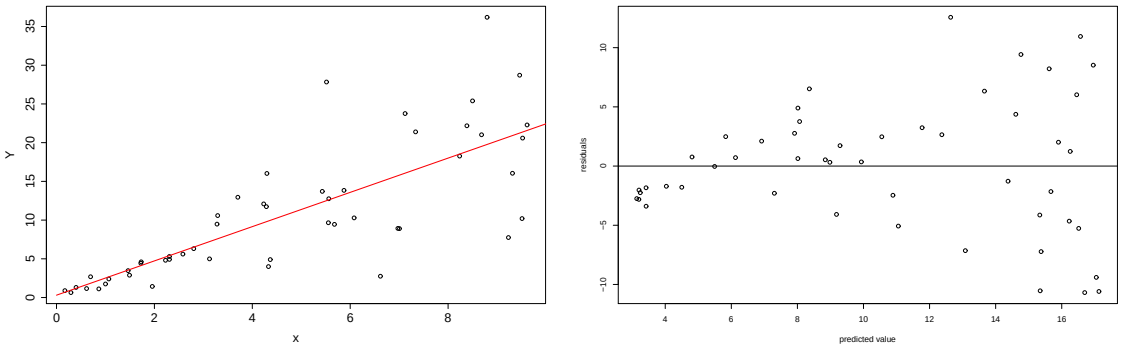

Figure 2.6: Example of a heteroskedastic data set.

Consequences:
The data are not used efficiently for estimation. More precise observations should get a higher weight.

Remedies:

1. Apply the model to transformed data. Use a variance-stabilizing transform, e.g., if $\sigma_j = c\mu_j$, i.e., if large data are more variable than small ones, we may use

$$Y_j^* = \log Y_j.$$

Then

$$\begin{aligned} \mathcal{L}(Y_j) = \mathcal{N}(\mu_j, \sigma_j^2) \quad &\Rightarrow \quad Y_j = \mu_j + c\mu_j V_j,\ V_1, \ldots, V_N \text{ iid } \mathcal{N}(0,1) \\ &\Rightarrow \quad Y_j^* = \log \mu_j + Z_j^*,\ Z_j^* = \log(1 + cV_j) \text{ iid.} \end{aligned}$$

2. In ordinary least squares $\hat{b}_1$ and $\hat{b}_2$ are obtained by minimizing

$$\sum_{j=1}^{N}(Y_j - b_1 - b_2 x_j)^2$$

w.r.t. b_1 and b_2.

Here, we may use weighted least squares instead (see Exercise):
Estimate σ_j^2 by $\hat{\sigma}_j^2$, e.g., by assuming that $\sigma_j^2 = c\mu_j^\alpha$ or by grouping the data w.r.t. the x_j-values. Then compute $\hat{b}_1$ and $\hat{b}_2$ by minimizing

$$\sum_{j=1}^{N}\left(\frac{Y_j - b_1 - b_2 x_j}{\hat{\sigma}_j}\right)^2$$

w.r.t. b_1 and b_2.

2.4.3 Outliers

Problem:
Some single points are not compatible with the model.

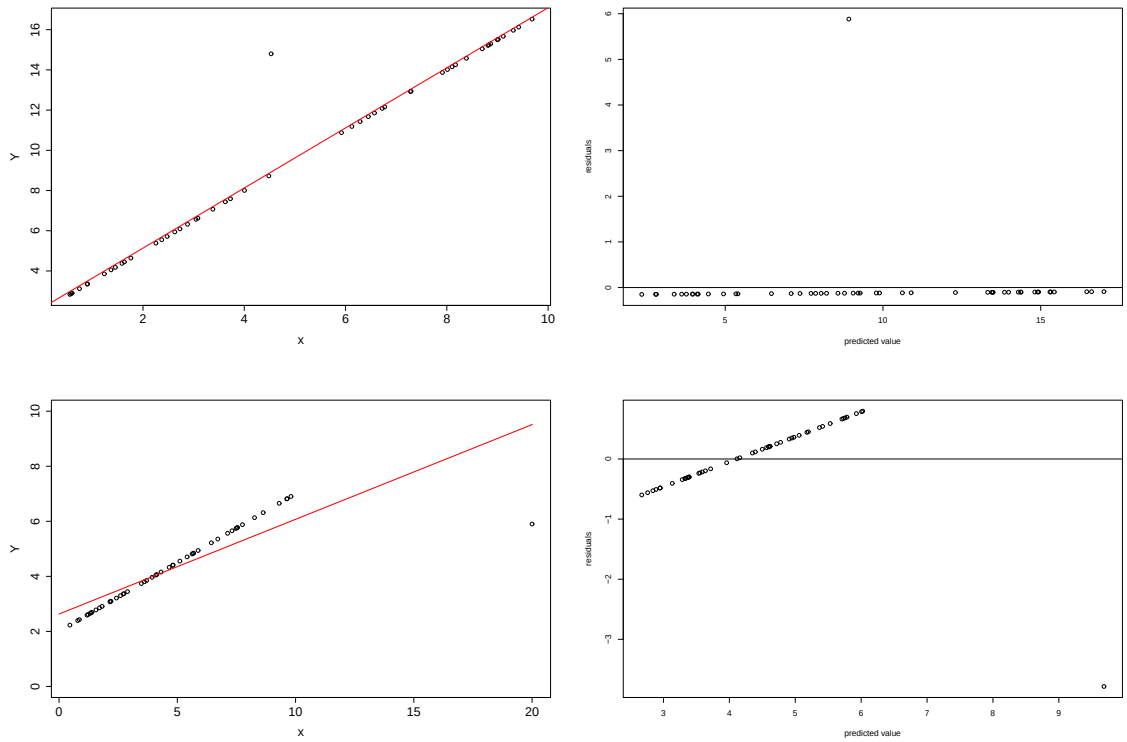

Figure 2.7: Examples of data sets with outliers.

Consequences:
The consequences depend on the location of the x_k-value of the outlier, e.g.

- Figure 2.7 left: x_j of the outlier in the center of the factor value range. In this case $\hat{b}_1$ will be slightly too large (small).

- Figure 2.7 right: x_j of the outlier far from the remaining factor values (leverage point). In this case $\hat{b}_2$ will be much too small (large).

Remedies:

1. Check the data for typing errors, transmission errors, reading errors, ...

2. Repeat the observation. Be careful: An extreme observation could be a hint that something special happened.

3. Exclude outliers from the further analysis of the data, but do not forget them! Reason: It is inconsistent and dangerous to use observations which cannot be explained by the model for fitting the model. On the other hand, these observations may contain important information.

4. Reject the model.

2.4.4 Nonnormal data

Problem:
$\mathcal{L}(Z_k) \neq \mathcal{N}(0, \sigma^2)$

Consequences:
Not so important if the difference is not too extreme. However, significance levels of, e.g., the F-test should not be taken too seriously (see Section 2.8).
Otherwise: Use methods of robust statistics.

2.4.5 Dependence of data

Problem:
$\text{cov}(Z_j, Z_k) \neq 0$ for $j \neq k$. Quite usual for time series (i.e., if the factor is time).

Consequences:
Confidence bands, significance levels etc. will be severely biased.

Remedies:

1. Estimate the correlation matrix $\Sigma = (corr(Y_j, Y_k))_{1 \leq k,j \leq N}$ and transform the data linearly such that the transformed data Y_k^* are uncorrelated (generalized least-squares).

2. Use methods from time series analysis (see Chapter 6).

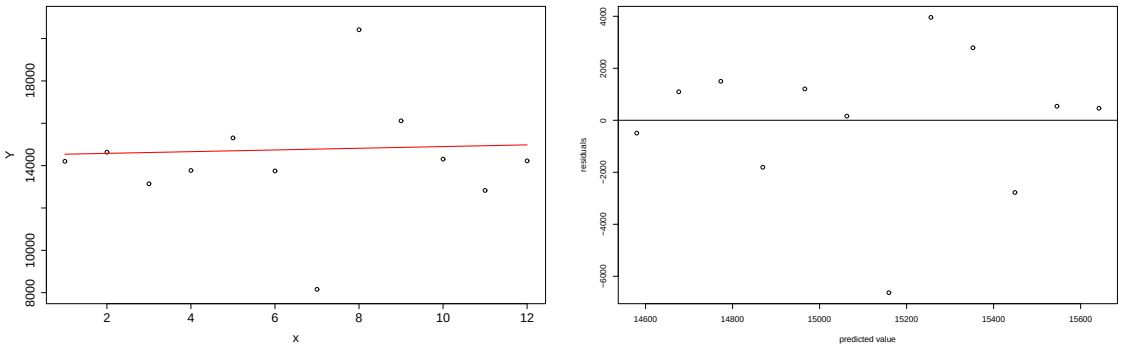

Figure 2.8: Car production per month influenced by a strike in July followed by extra shifts in the subsequent months.

2.5 Testing in Gaussian linear models

Residual plots yield a graphical tool for investigating the fit of a linear regression model. In the following, we will derive formal techniques based on significance tests. These can then be used to define methods for automatic model selection. The ideas are based on a generalization of the likelihood ratio test discussed for model (R1) in Section 2.1. Here, we only consider the geometric interpretation.

Theorem 2.5.1 (Fundamental Theorem of Linear Models)
Assume $Y_j = \mu_j + Z_j,\ j = 1, \ldots, N$, *where* $Z_1, \ldots, Z_N$ *are i.i.d.* $\mathcal{N}(0, \sigma^2)$*-distributed, i.e.* $Y = (Y_1, \ldots, Y_N)^T$ *is* $\mathcal{N}_N(\mu, \sigma^2 I_N)$*-distributed with* $\mu = (\mu_1, \ldots, \mu_N)^T$.

Let M *be a* d*-dimensional subspace of the sample space* $\mathbb{R}^N$ *and* M_0 *a* q*-dimensional subspace of* $M, 1 \leq q < d < N$. *Denote the squared distances from* Y *to* M *and* M_0 *by*

$$\widehat{D} = \min_{\mu \in M} ||Y - \mu||^2 \text{ and } \widehat{D}_0 = \min_{\mu \in M_0} ||Y - \mu||^2, \text{ respectively.}$$

a) If $E[Y] = \mu \in M$, *then* $\hat{s}^2 = \frac{1}{N-d}\widehat{D}$ *is an unbiased estimator for* σ^2, *and* $\frac{1}{\sigma^2}\widehat{D}$ *is* χ^2_{N-d}*-distributed.*

b) If $E[Y] \in M_0$, *then the ratio*

$$R = \frac{N-d}{d-q} \, \frac{\widehat{D}_0 - \widehat{D}}{\widehat{D}}$$

is $F_{d-q,N-d}$*-distributed.*

Proof:
Consider $e_1, \ldots, e_N$ an orthonormal basis of $\mathbb{R}^N$ such that $e_1, \ldots, e_q$ and $e_1, \ldots, e_d$ span the subspaces M_0 and M, respectively. O is the orthogonal $N \times N$-matrix whose j-th column is e_j, $j = 1, \ldots, N$. We define:

$$U = (U_1, \ldots, U_N)^T = O^T Y, \qquad \text{i.e.} \quad U_k = e_k^T Y$$

1) We first prove:
$U_k,\ k = 1, \ldots, N$, are independent, $\mathcal{N}(\nu_k, \sigma^2)$-distributed random variables with

$$\nu_k = 0,\ k = d+1, \ldots, N, \quad \text{if } \mathrm{E}Y \in M, \text{ and}$$

$$\nu_k = 0,\ k = q+1, \ldots, N, \quad \text{if } \mathrm{E}Y \in M_0.$$

Proof:

$U_1, \ldots, U_N$ are normally distributed as linear combinations of the normally distributed random variables $Y_1, \ldots, Y_N$. From the definition of U we have

$$\nu_k = \mathrm{E}[U_k] = \mathrm{E}\left[e_k^T\ Y\right] = e_k^T \mathrm{E}[Y] = 0 \text{ for all } k > d\ ,$$

if $\mathrm{E}[Y] \in M$. Analogously $\nu_k = 0$ for all $k > q$, if $\mathrm{E}[Y] \in M_0$. Let Σ_U be the covariance matrix of U. From Lemma 1.1.8

$$\Sigma_U = O^T(\sigma^2 I_N)O = \sigma^2 I_N,$$

as Y has $\sigma^2 I_N$ as covariance matrix. Hence, $U_1, \ldots, U_N$ are independent, and $\operatorname{var} U_k = \sigma^2,\ k = 1, \ldots, N$.

2) Now we write $\widehat{D}$ and R in terms of $U_1, \ldots, U_N$, so that one can immediately read off the distributional properties claimed in the theorem using the definitions of the χ^2- and the F-distribution.

Since O is orthogonal,

$$\begin{aligned}
\widehat{D} &= \min_{\mu \in M} ||Y - \mu||^2 = \min_{\mu \in M} ||O^T Y - O^T \mu||^2 \\
&= \min_{\mu \in M} \sum_{k=1}^{N} (U_k - e_k^T \mu)^2 \\
&= \min_{\mu \in M} \sum_{k=1}^{d} (U_k - e_k^T \mu)^2 + \sum_{k=d+1}^{N} U_k^2\ ,
\end{aligned}$$

as $e_k \perp M$ for $k > d$ by the definition of $e_1, \ldots, e_N$. The first summand vanishes because we can choose

$$\mu = \sum_{j=1}^{d} U_j e_j \in M$$

Thus,

$$\begin{aligned}
\widehat{D} &= \sum_{k=d+1}^{N} U_k^2 \\
\mathrm{E}[\widehat{D}] &= (N-d)\sigma^2, \text{ if } \mathrm{E}Y \in M,
\end{aligned}$$

and a) follows. If $\mathrm{E}[Y] \in M_0$, it can be shown analogously that

$$\widehat{D}_0 = \sum_{k=q+1}^{N} U_k^2,$$

and therefore

$$R = \frac{N-d}{d-q}\frac{\hat{D}_0 - \hat{D}}{\hat{D}} = \frac{N-d}{d-q} \cdot \frac{\sum_{k=q+1}^{d} U_k^2}{\sum_{k=d+1}^{N} U_k^2}.$$

We get b) from the definition of the F distribution.

□

Remark 2.5.2 (Construction of the test)
A typical question related to the model (LRG) is if the model is unnecessarily complicated, i.e. whether some factors, e.g. the last $d-q$ ones, are redundant. This question can be translated into the following decision problem:
In model (LRG) test the hypothesis $H_0 : b_{q+1} = \ldots = b_d = 0$ against the alternative $H_1 : b_k \neq 0$ for at least one $k > q$.

With the help of Theorem 2.5.1 we can construct a suitable test. Consider

$$\widehat{D}_0 = \min_{b_1,\ldots,b_q} \sum_{k=1}^{N} (Y_k - b_1 x_{k1} - \ldots - b_q x_{kq})^2 ,$$

$$\begin{aligned} \widehat{D} &= \min_{b_1,\ldots,b_d} \sum_{k=1}^{N} (Y_k - b_1 x_{k1} - \ldots - b_d x_{kd})^2 , \\ R &= \frac{N-d}{d-q}\,\frac{\widehat{D}_0 - \widehat{D}}{\widehat{D}} . \end{aligned}$$

Then, by Theorem 2.5.1 $P_{H_0}(R > f_\alpha) = \alpha$, where f_α is the $(1-\alpha)$-quantile of the $F_{d-q,N-d}$ distribution. Hence, we get a level α test if we reject the hypothesis if $R > f_\alpha$.

If we test for the significance of only one parameter, an equivalent alternative to the above F-test is a t-test based on the following result:

Proposition 2.5.3
For the model (LRG) and under the assumptions of Theorem 2.2.3, let b_ℓ be the true value of the ℓ-th parameter, and let $\sigma^2 C_{\ell\ell}$ be the variance of $\hat{b}_\ell$, i.e. the ℓ-th diagonal element of $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Let $\hat{s}^2 = \frac{1}{N-d}\widehat{D}$ denote the unbiased estimator of σ^2. Then,

$$\mathcal{L}\left(\frac{\hat{b}_\ell - b_\ell}{\hat{s}\sqrt{C_{\ell\ell}}}\right) = t_{N-d}.$$

2.6 Data-adaptive model selection

A crucial problem in regression analysis is to choose the right model if several or even a lot of different models are plausible. This includes the problem of selecting the right predictor variables as well as choosing which function of the predictor variables should determine the expectation of the response. We start from available *data*:

N independent sets of observations $Y_j, \xi_{j1}, \ldots, \xi_{jm}, j = 1, \ldots, N$.

Goal: Describe $\mathrm{E}Y_j$ as a suitable *regression function* of some $\xi_{j\ell}, 1 \le \ell \le m$.

We may choose the regression function from a general class of Gaussian linear models

$$
\begin{aligned}
&Y_j = \mu_j + Z_j, \quad j = 1, \ldots, N \\
&Z_1, \ldots, Z_N \text{ i.i.d. } \mathcal{N}(0, \sigma^2) \\
&\mu_j = \mathrm{E}Y_j = g(\xi_{ji_1}, \ldots, \xi_{ji_p}; b) = (\mathbf{X}b)_j \qquad \text{for some } 1 \le i_1 < \ldots < i_p \le m,\ 1 \le p \le m, \\
&g(u_1, \ldots, u_p; b) = \sum_{k=1}^{d} b_k f_k(u_1, \ldots, u_p),\ f_1, \ldots, f_d \text{ known}
\end{aligned}
$$

Typical in applications:

$$
\begin{aligned}
g(u_1, \ldots, u_p; b) &= b_1 + \sum_{k=1}^{p} b_{k+1} u_k \qquad p, i_1, \ldots, i_p = ? \\
g(u_1; b) &= b_1 + \sum_{k=1}^{d-1} b_{k+1} u_1^k \qquad d = ?
\end{aligned}
$$

For polynomial models, a perfect fit by interpolation is possible but not desirable. Too complicated models will also reproduce the structure of the residuals of the observed data (overfitting). Such a model will not perform well in predictions.

Wanted:

a) Model with good fit to the data, i.e. $\widehat{Z}_j = Y_j - \widehat{Y}_j$ "small" for all j.

b) Model which is as simple as possible, i.e. the number of parameters d and/or the number of independent variables p is small.

a) and b) are antagonistic! Therefore, as a program to do:

1) Find a measure of fit for quantifying requirement a)
2) Balance between a) and b)
3) Develop efficient strategies for finding good models with a feasible amount of computation

Why 3)? For fixed m, there are already $2^m - 1$ models of the simple form

$$
\mu_j = b_1 + \sum_{k=1}^{p} b_{k+1} \xi_{ji_k}, \quad 1 \le i_1 < \ldots < i_p \le m,\ p = 1, \ldots, m.
$$

In practice often $m \geq 20$.

2.6.1 The R^2-statistic

As $Y_j = \mu_j + Z_j$, we have two sources of variability in the data: the dependence of μ_j on j and the randomness of the residuals Z_j. The effect of these two aspects of variability can be measured in the following manner:
Let $\widehat{Z}_j = Y_j - \widehat{Y}_j$ denote the sample residuals as in the previous section. We define:

- the *residual sum of squares*

$$\mathrm{RSS} = \sum_{j=1}^{N} \widehat{Z}_j^2,$$

  a measure for the discrepancy between the data and the prediction by the model. For a correct model, it can also be interpreted as a measure for the variability in the data $Y_1, \ldots, Y_N$ due to random errors.

- the *sum of squares due to regression*

$$\mathrm{SS}_{reg} = \sum_{j=1}^{N} (\widehat{Y}_j - \overline{Y}_N)^2,$$

  a measure for the variability due to the dependence of $\mathrm{E}Y_j$ on j.

To get a scale invariant measure, we have to standardize. For that purpose, we consider

- the *total sum of squares*

$$\mathrm{TSS} = \sum_{j=1}^{N} (Y_j - \overline{Y}_N)^2,$$

  a measure for the total variability of the data Y_j. This is independent of the particular model which we are considering.

- the R^2 statistic or coefficient of determination

$$R^2 = \frac{SS_{reg}}{TSS}$$

  as a measure of how well a model describes the systematic variability in the data, i.e. the dependence of Y_j on the predictor variables.

To avoid the discussion of special cases we assume now always $f_1(u_1, \ldots, u_p) = 1$, i.e. there is an additive constant term in the regression function and the first column of the design matrix is $x^{(1)} = (1, \ldots, 1)^T$.

Proposition 2.6.1
Let (LRG) hold, and let $\widehat{Y} = \mathbf{X}\hat{b}$. *Then, under the assumptions of Theorem 2.2.3*

a) $TSS = SS_{reg} + RSS$, *i.e.* $R^2 = 1 - \frac{RSS}{TSS}$.

b) $R = \sqrt{R^2}$ *is the sample correlation of* Y *and* $\hat{Y}$, *i.e.*

$$R = \frac{\sum_{j=1}^{N}(Y_j - \overline{Y}_N)(\widehat{Y}_j - \overline{\hat{Y}}_N)}{\sqrt{\sum_{j=1}^{N}(Y_j - \overline{Y}_N)^2 \sum_{j=1}^{N}(\widehat{Y}_j - \overline{\hat{Y}}_N)^2}}.$$

Proof:
Note: $\widehat{Z}_j = Y_j - \widehat{Y}_j$, i.e. $Y = \widehat{Y} + \widehat{Z}$. From Proposition 2.3.3, $\widehat{Z} \perp x^{(1)}, \ldots, x^{(d)}$ and, therefore,

$$\widehat{Z}^T\widehat{Y} = 0 \text{ as } \widehat{Y} \in M_X, \quad \text{and} \quad \widehat{Z}^T x^{(1)} = \sum_{j=1}^{N} \widehat{Z}_j = 0. \tag{2.7}$$

a) We have

$$TSS = ||Y - \overline{Y}_N x^{(1)}||^2, \quad RSS = ||\widehat{Z}||^2, \text{ and } SS_{reg} = ||\widehat{Y} - \overline{Y}_N x^{(1)}||^2.$$

The assertion follows from Pythagoras' Theorem, as, by (2.7),

$$Y - \overline{Y}_N x^{(1)} = \widehat{Z} + (\widehat{Y} - \overline{Y}_N x^{(1)}), \qquad \widehat{Z} \perp \widehat{Y} - \overline{Y}_N x^{(1)}.$$

b) From (2.7),

$$\overline{\widehat{Y}_N} = \frac{1}{N}\sum_{j=1}^{N} \widehat{Y}_j = \frac{1}{N}\sum_{j=1}^{N}(Y_j - \widehat{Z}_j) = \overline{Y}_N.$$

Using (2.7) again,

$$\sum_{j=1}^{N}(Y_j - \overline{Y}_N)(\widehat{Y}_j - \overline{Y}_N) = \sum_{j=1}^{N}(\widehat{Y}_j + \widehat{Z}_j - \overline{Y}_N)(\widehat{Y}_j - \overline{Y}_N) = \sum_{j=1}^{N}(\widehat{Y}_j - \overline{Y}_N)^2 = SS_{reg}.$$

We conclude that the sample correlation of $(Y_j, \widehat{Y}_j), j = 1, \ldots, N$, is given by

$$\frac{SS_{reg}}{\sqrt{SS_{reg}\, TSS}} = \sqrt{\frac{SS_{reg}}{TSS}} = \sqrt{R^2}.$$

□

2.6.2 Model selection by maximizing (or minimizing) a criterion function

Idea:

- R^2 measures goodness-of-fit
  $\Rightarrow$ Maximize R^2 over a selection of models
- The fit becomes better if the models get more complicated.
  $\Rightarrow$ Incorporate a penalty for larger models.

Assume that we are given a hierarchy of models $\mathbf{M}_1 \subseteq \mathbf{M}_2 \subseteq \ldots \subseteq \mathbf{M}_d \subseteq \ldots$ where model $\mathbf{M}_d$ has d parameters. We may use the following approaches:

I) Adjusted R^2

Recall that $R^2 = 1 - \frac{RSS}{TSS}$.
By the *Fundamental Theorem of Linear Models*, Theorem 2.5.1:

$$\begin{array}{ll} \mathcal{L}\left(\frac{1}{\sigma^2}RSS(d)\right) = \chi^2_{N-d} & \text{if the model is correct, and} \\ \mathcal{L}\left(\frac{1}{\sigma^2}TSS\right) = \chi^2_{N-1} & \text{if } \mathrm{E}Y_j \equiv b_1. \end{array}$$

As $\mathrm{E}[X] = q$, if $\mathcal{L}(X) = \chi^2_q$, we adjust RSS and TSS such that they have the same mean σ^2 under the corresponding hypotheses, and get the *adjusted R^2 statistic*:

$$\overline{R}^2(d) := 1 - \frac{\frac{1}{N-d}RSS(d)}{\frac{1}{N-1}TSS} = 1 - (1 - R^2(d))\frac{N-1}{N-d}$$

Then $\frac{N-1}{N-d}$ is an increasing function of d while $1 - R^2(d)$ is decreasing.

Select the model $\mathbf{M}_{\widehat{d}}$ with

$$\widehat{d} = \arg\max_d \overline{R}^2(d).$$

II) Residual mean-square (rms)

A quantity that is equivalent to $\overline{R}^2$ is the residual mean square

$$\hat{s}^2_d = \frac{1}{N-d}\sum_{j=1}^{N}\widehat{Z}^2_j(d) = \frac{1}{N-d}RSS(d).$$

For increasing d, $\frac{1}{N-d}$ is increasing and $RSS(d)$ is decreasing. Hence, we select the model with

$$\widehat{d} = \arg\min_d \hat{s}^2_d.$$

This is equivalent to maximizing $\overline{R}^2$ as $\overline{R}^2(d) = 1 - \frac{\hat{s}^2_d}{\frac{1}{N-1}TSS}$.

I) and II) are still preferring too large models. Therefore, we consider another approach which uses an additive rather than a multiplicative penalty.

III) Mallow's C_p-statistic
Assume that there is a "largest" model $\overline{\mathbf{M}}$, i.e. $\mathbf{M} \subseteq \overline{\mathbf{M}}$ for all models considered. Denote the residual mean square w.r.t. $\overline{\mathbf{M}}$ by $\hat{s}^2_{\max}$.

If $\overline{\mathbf{M}}$ describes $\mathrm{E}Y_j$ correctly, then $\hat{s}^2_{\max}$ is an unbiased estimator of σ^2. There is a high chance that this is true to a very good approximation, if $\overline{\mathbf{M}}$ is large enough. We set

$$C_d := \frac{RSS(d)}{\hat{s}^2_{\max}} - (N - 2d) \quad \text{(in Mallow's paper, } d \text{ was called } p, \text{ therefore } C_p)$$

and we select the model with

$$C_{\widehat{d}} = \min_d C_d.$$

C_d penalizes large d more severely than $\overline{R}^2_d$ or $\hat{s}^2_d$:
For a large model close to $\overline{\mathbf{M}}$, we have $\overline{R}^2(d) \approx 1$ and $\hat{s}^2_d \approx \sigma^2$. Hence, the values do not change much when further increasing the number of parameters. In contrast, we have $C_d \approx d$ which depends on d.

2.6.3 Strategies for selecting good models

I) All-subset regression

Calculate a measure for the goodness of fit, e.g., Mallow's C_p for all models. Select the model with the best criterion value (not feasible if there are many variables and/or parameters).

II) Backward elimination

1. Fit the largest model $\overline{\mathbf{M}}$ containing all variables to the data.
Set: $\mathbf{M}^* = \overline{\mathbf{M}}$ (current model).

2. For each parameter b_k in model $\mathbf{M}^*$, test

$$H_0 : b_k = 0 \text{ against } H_1 : b_k \neq 0 (\text{except for the constant } b_1).$$

This results in $d-1$ test statistics $T_2, \ldots, T_d$. Define M by $T_M := \min_{k \geq 2} T_k$, i.e. b_M is the least significant parameter.

3. Select a bound c_α (depending on N, d, and the level α, e.g. $\alpha = 10\%$).

   a. If $T_M < c_\alpha$, set $b_M = 0$ in the old model $\mathbf{M}^*$ to obtain a new smaller model $\mathbf{M}^*$ with $d-1$ parameters, goto 2.

   b. If $T_M > c_\alpha$, adopt model $\mathbf{M}^*$. STOP!

Advantage: A model with all variables included is seen once.

Drawbacks:

1) For $\overline{\mathbf{M}}$, $\mathbf{X}^T\mathbf{X}$ is often nearly singular, causing computational problems with step 1.
2) Once a variable is removed, it is gone forever.
3) The "level" of the test is not a strict bound on the error probability in the sense of hypothesis testing, as $d-1$ tests are applied simultaneously.

III) Stepwise regression

Start with the simplest model: $\mathbf{M}^* = \mathbf{M}_0 : \mathrm{E}Y_j = b_1,\ j = 1, \ldots, N$. Then, include additional variables and build more complex models until a satisfactory fit is achieved.

The selection of variables to be included is based on "partial correlations" which, due to lack of time, is not explained here. We will see a similar concept in the time series part.

Example 2.6.2
We are studying the heat evolving during the hardening process of concrete depending on the cement components. Assume that for $N = 13$ samples the following information is given:

$$\begin{aligned} Y_k &= \text{heat evolving during the hardening process (in cal/g)} \\ \xi_{k1} &= \text{percentage of weight Tricalcium-Aluminate} \\ \xi_{k2} &= \text{percentage of weight Tricalcium-Silicate} \\ \xi_{k3} &= \text{percentage of weight Tetracalcium-Aluminium-Ferrite} \\ \xi_{k4} &= \text{percentage of weight Dicalcium-Silicate} \end{aligned}$$

Consider only models of the form

$$\begin{array}{lll} M_0 & EY_k = b_1 & \\ M_i & EY_k = b_1 + b_2\xi_{ki} & y = g(\xi_i) \\ M_{ij} & EY_k = b_1 + b_2\xi_{ki} + b_3\xi_{kj} & y = g(\xi_i, \xi_j) \\ M_{ijl} & EY_k = b_1 + b_2\xi_{ki} + b_3\xi_{kj} + b_4\xi_{kl} & y = g(\xi_i, \xi_j, \xi_l) \\ \overline{\mathbf{M}} = M_{1234} & EY_k = b_1 + \sum_{j=1}^{4} b_{j+1}\xi_{kj} & y = g(\xi_1, \xi_2, \xi_3, \xi_4) \end{array}$$

We look for the best model using the strategies introduced above.

1. R^2-statistic

d = # parameters	best model	R^2
2	$y = g(\xi_4)$	0.675
3	$y = g(\xi_1, \xi_2)/y = g(\xi_1, \xi_4)$	0.979/0.972
4	$y = g(\xi_1, \xi_2, \xi_4)$	0.98234
5	$y = g(\xi_1, \xi_2, \xi_3, \xi_4)$	0.98237

The increase in the quality from $d = 3$ to $d = 4$ is small, as

$$\widehat{\text{corr}}(\xi_1, \xi_3) = -0.824$$
$$\widehat{\text{corr}}(\xi_2, \xi_4) = -0.973$$

Consequently, if ξ_1 and ξ_2 or ξ_1 and ξ_4 are in the model, adding a third variable does not improve the fit much. The increase from $d = 4$ to $d = 5$ is extremely small, as

$$\xi_1 + \xi_2 + \xi_3 + \xi_4 \approx 1 (\in [0.95, 0.99])$$

as percentages of the major cement components.
Best models: $y = g(\xi_1, \xi_2)$ or $y = g(\xi_1, \xi_4)$

2. Residual mean-square $\hat{s}^2$

d						
2	**1**: 115.06	**2**: 82.39	**3**: 176.31	**4**: 80.35		
3	**12**: 5.79	**13**: 122.71	**14**: 7.48	**23**: 41.54	**24**: 86.89	**34**: 17.57
4	**123**: 5.35	**124**: 5.33	**134**: 5.65	**234**: 8.20		
5	**1234**: 5.98					

Best model: $y = g(\xi_1, \xi_2, \xi_4)$

3. Mallow's C_p

d						
1	443.2					
2	**1**: 202.5	**2**: 142.5	**3**: 315.2	**4**: 138.7		
3	**12**: 2.7	**13**: 198.1	**14**: 5.5	**23**: 62.4	**24**: 138.2	**34**: 22.4
4	**123**: 3.0	**124**: 3.0	**134**: 3.5	**234**: 7.3		
5	**1234**: 5.0					

Best model: $y = g(\xi_1, \xi_2)$

4. Backward Elimination

- Fit model $\overline{\mathbf{M}} = M_{1234}, y = g(\xi_1, \xi_2, \xi_3, \xi_4)$ with $d = 5$ and $N = 13$
- F-tests, $\alpha = 0.10$. $M = 3$, $T_3 = 0.018 < 3.46 = (1 - \alpha)$-quantile of $F_{1,8}$
  $\Longrightarrow$Remove ξ_3 and fit the new current model $M_{124}, y = g(\xi_1, \xi_2, \xi_4)$
- F-tests, $\alpha = 0.10$
  $M = 4, T_4 = 1.86 < 3.36 = (1 - \alpha)$-quantile of $F_{1,9}$.
  $\Longrightarrow$ Remove ξ_4 and fit the new current model $M_{12}, y = g(\xi_1, \xi_2)$
- F-tests, $\alpha = 0.001$
  $T_1 \approx 146, T_2 \approx 208 > 14.91 = (1 - \alpha)$-quantile of $F_{1,10}$
  $\Longrightarrow$ Stop! Retain model M_{12}.

2.7 Confidence bands for regression functions

Now we know the estimator and the test for the general linear regression model (LRG). In the next step, we are interested in confidence intervals.

An application would be to quantify the reliability of forecasts for future observations. Let us consider data $Y_1, \ldots, Y_N$ generated by model (LRG) and assume that we are adding a new observation Y_{N+1} with $\xi = (x_{N+1,1}, \ldots, x_{N+1,d})$ known. A consideration analogous to Section 2.3 yields

$$\widehat{Y}_{N+1} = m(\xi; \hat{b}) = \xi \hat{b}.$$

as a feasible prediction for Y_{N+1}. From Theorem 2.2.3 and Lemma 1.1.8, we have

$$\mathcal{L}(\widehat{Y}_{N+1}) = \mathcal{N}\big(\xi b, \sigma^2 \xi (\mathbf{X}^T \mathbf{X})^{-1} \xi^T\big),$$

and we get as an approximate $(1-\alpha)$-confidence interval for $\mathrm{E}[Y_{N+1}] = m(\xi; b) = \xi b$:

$$m(\xi; b) \in \widehat{Y}_{N+1} \pm q_{1-\frac{\alpha}{2}} \sqrt{\hat{s}^2 \xi (\mathbf{X}^T \mathbf{X})^{-1} \xi^T}$$

where we have replaced σ^2 by its unbiased estimator $\hat{s}^2$ and where q_β denotes the β-quantile of the standard normal distribution.

This is a confidence interval for one value of the function $m(\cdot; b)$, but we would like to have simultaneous confidence bands for all reasonable values of ξ. We restrict attention to the situation where we have only one predictor variable, i.e. $\mathrm{E}Y_j$ is a function of a real-valued variable x_j:

$$\mathrm{E}Y_j = m(x_j; b), \;\; b \in \mathbb{R}^d \quad \text{unknown parameter}$$

Examples:

$$\begin{aligned} m(x; b) &= b_1 \cos x + b_2 \sin x \;, \\ m(x; b) &= b_1 + b_2 x + \ldots + b_d x^{d-1}, \\ m(x; b) &= b_1 f_1(x) + \ldots + b_d f_d(x) \;\; \text{with known functions } f_1, \ldots, f_d. \end{aligned}$$

Definition 2.7.1
A γ-confidence band for the function $m(\cdot; b)$ defined on the interval (c, d) is given by two functions $G(x|Y)$ and $H(x|Y)$ that depend on the data vector Y in such a way that

$$P_{b,\sigma^2}\big\{G(x|Y) \le m(x; b) \le H(x|Y) \;\text{ for all }\; x \in (c, d)\big\} \ge \gamma \qquad \text{for all } \; b \in \mathbb{R}^d, \; \sigma^2 > 0.$$

We consider arbitrary linear combinations of the parameters $b_1, \ldots, b_d$. For fixed q, $1 \le q \le d$, define

$$L_q(b) = \left\{ \beta : \;\; \beta = \sum_{k=1}^{q} \alpha_k b_k, \;\; \alpha_1, \ldots, \alpha_q \in \mathbb{R} \right\},$$

the set of all linear combinations of the first q parameters. Let $\beta = \sum_{k=1}^{q} \alpha_k b_k \in L_q(b)$. Then an elementary calculation shows that

$$\hat{\beta} = \sum_{k=1}^{q} \alpha_k \hat{b}_k$$

is the least squares estimator of β. As a function of the ML estimator $\hat{b}_j$ $\hat{\beta}$ is also the ML estimator for β. Using Theorem 2.2.3 and Lemma 1.1.8, we get

$$\begin{aligned} \mathrm{E}[\hat{\beta}] &= \sum_{k=1}^{q} \alpha_k \mathrm{E}[\hat{b}_k] = \beta \text{ and} \\ \text{var } \hat{\beta} &= \mathrm{E}(\hat{\beta}-\beta)^2 = \sum_{j,k=1}^{q} \alpha_j \alpha_k \ \mathrm{cov}(\hat{b}_j, \hat{b}_k) \\ &= \sigma^2 \sum_{j,k=1}^{q} \alpha_j \alpha_k \, (\mathbf{X}^T\mathbf{X})^{-1}_{jk} = \sigma^2 a^T (\mathbf{X}^T\mathbf{X})^{-1} \, a \,, \end{aligned}$$

where $a = (\alpha_1, \ldots, \alpha_q, \ 0, \ldots, 0)^T \in \mathbb{R}^d$.

The unbiased estimator of var $\hat{\beta}$ is given by $\hat{s}^2_\beta = \hat{s}^2 a^T (X^T X)^{-1} \, a$, where $\hat{s}^2$ is the unbiased estimator of σ^2.

Theorem 2.7.2 (Scheffé's method)
Let $Y = (Y_1, \ldots, Y_N)^T$ satisfy the regression model (LRG). Assume that $\mathbf{X}^T\mathbf{X}$ is invertible. Let $\hat{b}$ denote the LS estimator of b, $\hat{s}^2$ the unbiased estimator of σ^2, and $\hat{s}^2_\beta = \hat{s}^2 a^T (X^T X)^{-1} \, a$. For any fixed $q \le d$ let f_γ be the γ-quantile of the $F_{q,N-d}$-distribution $(1 \le q \le d)$.
Then

$$P_{b,\sigma^2}\big\{\hat{\beta} - \sqrt{q f_\gamma \, \hat{s}^2_\beta} \le \beta \le \hat{\beta} + \sqrt{q f_\gamma \, \hat{s}^2_\beta} \text{ for all } \beta \in L_q(b)\big\} \ge \gamma$$

for all $b \in \mathbb{R}^d$, $\sigma^2 > 0$.

Proof:
(Sketch for $q = d$)
Show that

$$\mathcal{L}\left(\frac{(\hat{b}-b)^T (X^T X)(\hat{b}-b)}{d\hat{s}^2}\right) = F_{d,N-d}$$

Additionally, we use that for any positive definite matrix B we have

$$\sup_{\alpha \neq 0} \frac{(\alpha^T x)^2}{\alpha^T B \alpha} = x^T B^{-1} x.$$

Apply this with $x = \hat{b} - b$ and $B = (X^T X)^{-1}$ and assume that

$$\sup_{\alpha \neq 0} \frac{(\alpha^T(\hat{b}-b))^2}{\alpha^T (X^T X)^{-1} \alpha} = (\hat{b}-b)^T (X^T X)(\hat{b}-b) \le d\hat{s}^2 f_\gamma.$$

This implies

$$\frac{(\alpha^T(\hat{b}-b))^2}{\alpha^T (X^T X)^{-1} \alpha} \le d\hat{s}^2 f_\gamma \text{ for all } \alpha \neq 0.$$

Therefore,

$$(\alpha^T(\hat{b}-b))^2 \le d\hat{s}^2 f_\gamma \alpha^T (X^T X)^{-1} \alpha \text{ for all } \alpha \neq 0.$$

Taking square roots yields the desired form of the confidence region. □

As a special case for $\alpha = (1, 0, \ldots, 0)$, $\alpha = (0, 1, 0, \ldots, 0)$, $\ldots$, we get a γ-confidence region for the parameter vector b.

Theorem 2.7.3
Let the assumptions of Theorem 2.7.2 hold. The unbiased estimator of var $\hat{b}_k$ (compare Theorem 2.2.3) is given by

$$\hat{s}_k^2 = (\mathbf{X}^T\mathbf{X})^{-1}_{kk}\hat{s}^2.$$

A γ-confidence region for $(b_1, \ldots, b_q)^T$ is given by:

$$P_{b,\sigma^2}\{\hat{b}_k - \sqrt{qf_\gamma \hat{s}_k^2} \le b_k \le \hat{b}_k + \sqrt{qf_\gamma \hat{s}_k^2}\,,\ \ k = 1, \ldots, q\} \ge \gamma$$

for all $b \in \mathbb{R}^d$, $\sigma^2 > 0$.

Our original aim was to derive confidence bands for the regression function. This can be done by using Scheffé's method as the following example demonstrates.

Example: We consider the model (R1), and we want to fit a regression line of the form $m(x; b) = b_1 + b_2 x = \beta(x) \in L_2(b)$ to the data. With $\alpha_1 = 1$, $\alpha_2 = x$; $\beta = \beta(x) = b_1 + b_2 x$ is a linear combination of the parameters b_1, b_2. We denote $\hat{\beta} = \hat{\beta}(x) = \hat{b}_1 + \hat{b}_2 x$ the estimated value of the linear regression function at the point x. From Theorem 2.7.2, we obtain

$$P_{b,\sigma^2}\{\hat{\beta}(x) - \sqrt{2f_\gamma \hat{s}^2_{\beta(x)}} \le \beta(x) \le \hat{\beta}(x) + \sqrt{2f_\gamma \hat{s}^2_{\beta(x)}} \text{ for all } \ x \in \mathbb{R}\} \ge \gamma$$

for all $b \in \mathbb{R}^2$, $\sigma^2 > 0$.

The design matrix for the model (R1) is

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix},$$

so that

$$\mathbf{X}^T\mathbf{X} = N \begin{pmatrix} 1 & \overline{x}_N \\ \overline{x}_N & \overline{x_N^2} \end{pmatrix}$$

with $\overline{x}_N = \frac{1}{N}\sum_{j=1}^N x_j$, $\overline{x_N^2} = \frac{1}{N}\sum_{j=1}^N x_j^2$, hence

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{N}\{\overline{x_N^2} - (\overline{x}_N)^2\}^{-1} \begin{pmatrix} \overline{x_N^2} & -\overline{x}_N \\ -\overline{x}_N & 1 \end{pmatrix}.$$

Hence, the estimator for the variance of $\hat{\beta}(x)$ is

$$\begin{aligned} \hat{s}_\beta^2 &= \hat{s}^2_{\beta(x)} = \hat{s}^2(1\ x)(\mathbf{X}^T\mathbf{X})^{-1}\begin{pmatrix}1\\ x\end{pmatrix} \\ &= \frac{\hat{s}^2}{N}\left\{1 + \frac{(x - \overline{x}_N)^2}{\overline{x_N^2} - (\overline{x}_N)^2}\right\}. \end{aligned}$$

As γ- confidence band for the linear regression function $m(x;b) = b_1 + b_2 x$ we obtain

$$m(x;b) \in \left[\hat{\beta}(x) - \sqrt{2f_\gamma \frac{\hat{s}^2}{N}} \sqrt{1 + \frac{(x-\overline{x}_N)^2}{\overline{x_N^2} - (\overline{x}_N)^2}}, \hat{\beta}(x) + \sqrt{2f_\gamma \frac{\hat{s}^2}{N}} \sqrt{1 + \frac{(x-\overline{x}_N)^2}{\overline{x_N^2} - (\overline{x}_N)^2}} \right].$$

2.8 Least-squares for non-Gaussian data

For the Gaussian residuals, the LS estimators of the parameters $b_1, \ldots, b_d$ coincide with the ML estimators in the regression model (LRG). Hence, we know from mathematical statistics that these estimators are asymptotically efficient. If $Z_1, \ldots, Z_N$ are no longer Gaussian, but still i.i.d. with mean 0 and finite variance σ^2, the LS estimators retain some properties shown for the Gaussian case: they have asymptotically the same distribution, and they are the best possible linear and unbiased estimators.

In this section, we consider the following general linear regression model:

$$(LR) \qquad Y = \mathbf{X}b + Z, \;\; Z_1, \ldots, Z_N \;\text{ i.i.d. }\; \mathrm{E}Z_j = 0, \;\; \mathrm{var}Z_j = \sigma^2 < \infty$$

Our first goal is to show that the LS estimator $\hat{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$ is approximately normal for large sample size N. We use the Cramér-Wold device which allows proving asymptotic normality of a random vector by relying only on a one-dimensional central limit theorem.

Lemma 2.8.1 (Cramér-Wold device)
Let $B, B_N \in \mathbb{R}^d, N \geq 1$, be random vectors. Then,

$$B_N \xrightarrow[\mathcal{L}]{} B \qquad \textit{iff} \qquad \alpha^T B_N \xrightarrow[\mathcal{L}]{} \alpha^T B \quad \textit{for all } \alpha \in \mathbb{R}^d.$$

Proof:
The proof uses characteristic functions (Lecture Probability Theory). □

Furthermore, we use the fact that a random vector $V = (V_1, \ldots, V_d)^T$ has a multivariate normal distribution iff $\alpha^T V = \sum_{k=1}^{d} \alpha_k V_k$ is a one-dimensional normal random variable for each $\alpha = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{R}^d$.

Additionally, we need a general central limit theorem (CLT) for non-identically distributed independent real random variables. To avoid unnecessary moment conditions, we do not use the Lyapunov CLT, but the Lindeberg CLT instead.

Theorem 2.8.2 (Lindeberg central limit theorem)
Let $V_1, V_2, \ldots$ be independent real random variables with mean 0 and $\mathrm{var}V_k = \sigma_k^2 < \infty$. Let $v_N^2 = \sum_{k=1}^{N} \sigma_k^2 = var\left(\sum_{k=1}^{N} V_k\right)$. If the Lindeberg condition

$$\sum_{k=1}^{N} E\left[\frac{V_k^2}{v_N^2} 1_{(\delta,\infty)}\left(\frac{|V_k|}{v_N}\right)\right] \to 0 \quad (N \to \infty) \qquad \textit{for all } \delta > 0$$

is satisfied, then the sum of the V_k is asymptotically normal, i.e.

$$\frac{1}{v_N}\sum_{k=1}^{N} V_k \xrightarrow[\mathcal{L}]{} \mathcal{N}(0,1).$$

Corollary 2.8.3 (CLT for weighted sums)
$U_1, U_2, \ldots$, i.i.d. with $EU_j = 0$, var $U_j = 1$. Then

$$\frac{1}{v_N}\sum_{k=1}^{N} w_k U_k \xrightarrow[\mathcal{L}]{} \mathcal{N}(0,1)\ (N\to\infty) \quad \textit{with } v_N^2 = \sum_{k=1}^{N} w_k^2$$

if $\dfrac{\max_{1\le k\le N} w_k^2}{v_N^2} \to 0 \ \ (N\to\infty)$.

Proof:
Use Lindeberg's CLT with $V_k = w_k U_k$, $\sigma_k^2 = w_k^2$. Set $\gamma_N = \max_{1\le k\le N}|w_k|$ s.th.

$$\frac{\delta v_N}{|w_k|} \ge \frac{\delta v_N}{\gamma_N} \quad \text{for all} \quad k \Longrightarrow$$

$$\sum_{k=1}^{N} E\left\{\frac{V_k^2}{v_N^2} 1_{(\delta,\infty)}\left(\frac{|V_k|}{v_N}\right)\right\} \le \sum_{k=1}^{N} E\left\{\frac{w_k^2 U_k^2}{v_N^2}\ 1_{\left(\frac{\delta v_N}{\gamma_N},\infty\right)}(|U_k|)\right\}$$

$$= E\left\{U_1^2 1_{\left(\frac{\delta v_N}{\gamma_N},\infty\right)}(|U_1|)\right\} \longrightarrow 0, \text{ as } \frac{v_N}{\gamma_N} \longrightarrow \infty.$$

The limit relation follows from Lebesgue's theorem of dominated convergence and $1_{(c_N,\infty)}(u) \to 0$ for all u if $c_N \to \infty$. □

If Z_j is $\mathcal{N}(0,\sigma^2)$-distributed, i.e. (LRG) holds:

$$\mathcal{L}\Big((\mathbf{X}^T\mathbf{X})^{1/2}(\hat{b}-b)\Big) = \mathcal{N}_d(0,\sigma^2 I_d)$$

Recall: $\mathbf{X}^T\mathbf{X}$ is by definition symmetric, by assumption non-singular and, moreover, positive definite, as

$$\sum_{k,l=1}^{d} \alpha_k\alpha_l(\mathbf{X}^T\mathbf{X})_{kl} = \alpha^T(\mathbf{X}^T\mathbf{X})\alpha = (\alpha^T\mathbf{X}^T)(\mathbf{X}\alpha) = ||\mathbf{X}\alpha||^2 > 0$$

for all $\alpha \neq 0$.
Therefore, there exists a $d\times d$-matrix $\mathbf{Q} = (\mathbf{X}^T\mathbf{X})^{1/2}$ s.th. $\mathbf{Q}^2 = \mathbf{X}^T\mathbf{X}$.
Construction: $\mathbf{X}^T\mathbf{X}$ has eigenvalues $0 < \lambda_1 \le \ldots \le \lambda_d$, and for an orthogonal matrix $\mathbf{O}$:

$$\mathbf{X}^T\mathbf{X} = \mathbf{O}\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix}\mathbf{O}^T.$$

Then,

$$\mathbf{Q} = \mathbf{O}\begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_d} \end{pmatrix}\mathbf{O}^T.$$

Theorem 2.8.4
Assume model (LR) with $\mathbf{X}^T\mathbf{X}$ *invertible. Let the hat matrix* $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ *satisfy* $\max_{1\le j\le N} H_{jj} \longrightarrow 0$ *for* $N \to \infty$. *Then,*

$$(\mathbf{X}^T\mathbf{X})^{1/2}(\hat{b} - b) \xrightarrow[\mathcal{L}]{} \mathcal{N}_d(0, \sigma^2 I_d) \quad (N \to \infty).$$

Proof:

i) By Theorem 2.2.3 and $Y = \mathbf{X}b + Z$, we have, using again $\mathbf{Q} = (\mathbf{X}^T\mathbf{X})^{1/2}$,

$$\mathbf{Q}(\hat{b} - b) = \mathbf{Q}\Big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}b + Z) - b\Big) = \mathbf{Q}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Z = \mathbf{Q}^{-1}\mathbf{X}^T Z$$

As Z has mean 0 and covariance matrix $\sigma^2 I_N$, we get from Lemma 1.1.8

$$E\ (\mathbf{Q}^{-1}\mathbf{X}^T Z) = 0, \qquad \operatorname{cov}(\mathbf{Q}^{-1}\mathbf{X}^T Z) = \sigma^2\mathbf{Q}^{-1}(\mathbf{X}^T\mathbf{X})\mathbf{Q}^{-1} = \sigma^2 I_d.$$

It remains to show that $\mathbf{Q}(\hat{b} - b)$ is asymptotically normal.

ii) By Lemma 2.8.1 (Cramér-Wold device), it suffices to show

$$\alpha^T\mathbf{Q}^{-1}\mathbf{X}^T Z \xrightarrow[\mathcal{L}]{} \mathcal{N}(0, \sigma^2\alpha^T\alpha) = \mathcal{N}(0, \sigma^2||\alpha||^2) \quad \text{for all } \alpha \in \mathbb{R}^d.$$

Write: $\alpha^T\mathbf{Q}^{-1}\mathbf{X}^T Z = a^T Z$ with $a = \mathbf{X}\mathbf{Q}^{-1}\alpha$. Then,

$$a^T Z = \sum_{j=1}^N a_j Z_j = \sum_{j=1}^N \underbrace{(\sigma a_j)}_{w_j} \underbrace{\frac{Z_j}{\sigma}}_{U_j} \xrightarrow[\mathcal{L}]{} \mathcal{N}(0, \sigma^2||\alpha||^2)$$

by Corollary 2.8.3, if

$$\frac{\max_{1\le j\le N} a_j^2}{\sum_{j=1}^N a_j^2} \longrightarrow 0, \quad (*)$$

as $v_N^2 = \sum_{j=1}^N w_j^2 = \sigma^2 \sum_{j=1}^N a_j^2 = \sigma^2 a^T a = \sigma^2\alpha^T\mathbf{Q}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{Q}^{-1}\alpha = \sigma^2||\alpha||^2$ is independent of N.

iii) It remains to show $(*)$. Let $\xi_1^T, \ldots, \xi_N^T$ be the rows of $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} \xi_1^T \\ \vdots \\ \xi_N^T \end{pmatrix}, \ Y_j = \xi_j^T b + Z_j \ , \ \xi_j \in \mathbb{R}^d.$$

By definition of a:

$$a_j = \xi_j^T\mathbf{Q}^{-1}\alpha = \langle \mathbf{Q}^{-1}\xi_j, \ \alpha\rangle$$

The inequality of Cauchy-Schwarz implies

$$a_j^2 \leq ||\mathbf{Q}^{-1}\xi_j||^2||\alpha||^2 \overset{ii)}{=} ||\mathbf{Q}^{-1}\xi_j||^2||a||^2.$$

Hence,

$$\begin{aligned}\frac{\max a_j^2}{\sum_{j=1}^N a_j^2} &\leq \max_{1\leq j\leq N} ||\mathbf{Q}^{-1}\xi_j||^2 = \max_j \xi_j^T\mathbf{Q}^{-2}\xi_j = \max_j \xi_j^T(\mathbf{X}^T\mathbf{X})^{-1}\xi_j \\ &= \max_{1\leq j\leq N} H_{jj} \to 0 \quad (N\to\infty)\end{aligned}$$

by assumption.

□

Corollary 2.8.5
Under the conditions of Theorem 2.8.4

$$\sqrt{N}(\hat{b}-b) \xrightarrow[\mathcal{L}]{} \mathcal{N}_d(0,\Sigma) \qquad \textit{with } \Sigma = \lim_{N\to\infty} N\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Proof:
Multiply the limit relationship of the theorem by $\sqrt{N}(\mathbf{X}^T\mathbf{X})^{-1/2}$. □

We finish this section by showing that LS estimators still have some optimal properties in the case of non-Gaussian data. We consider only a restricted class of estimators which share two characteristics with LS estimators: they are unbiased, and they are linear functions of the data $Y_1,\ldots,Y_N$.

Definition 2.8.6

a) A *linear, unbiased* estimator $\tilde{b}$ of the parameter b in the model (LR) satisfies

$$\tilde{b} = AY \text{ for some } d\times N-\text{matrix } A, \quad \mathrm{E}[\tilde{b}] = b \text{ for all } b\in\mathbb{R}^d,\ \sigma^2>0\ .$$

b) $\tilde{b}$ is called the BLUE *(Best Linear Unbiased Estimator)* of b, if for each linear, unbiased estimator b^* of b

$$\mathrm{E}\left[||\tilde{b}-b||^2\right] \leq \mathrm{E}\left[||b^*-b||^2\right] \text{ for all } b\in\mathbb{R}^d,\ \sigma^2>0.$$

Theorem 2.8.7 (Gauss-Markov Theorem)
In the general linear regression model (LR), the least squares estimator $\hat{b}$ is the BLUE of the parameter vector b.

Proof:
We present the proof just for the case where $\mathbf{X}^T\mathbf{X}$ is invertible i.e. the columns of $\mathbf{X}$ are linearly independent. Consider $\tilde{b}$ an arbitrary linear estimator, i.e. an estimator of the form $\tilde{b}=AY$ with suitable $d\times N$-Matrix A. If $\tilde{b}$ is unbiased, we have

$$b = \mathrm{E}[\tilde{b}] = AE[Y] = A\mathbf{X}b\ ,$$

if b is the parameter-vector related to Y. Since $b \in \mathbb{R}^d$ is arbitrary, it follows

$$A\mathbf{X} = I_d \ .$$

Therefore,

$$\begin{aligned} \mathrm{E}\left[||\tilde{b} - b||^2\right] &= \mathrm{tr}\ \left\{\mathrm{E}\left[(\tilde{b} - b)(\tilde{b} - b)^T\right]\right\} \\ &= \mathrm{tr}\left\{A\mathrm{E}\left[\{(Y - \mathbf{X}b)(Y - \mathbf{X}b)^T\}\right] A^T\right\} \\ &= \mathrm{tr}\{A\ \mathrm{E}[ZZ^T]A^T\} \\ &= \sigma^2\,\mathrm{tr}\{AA^T\}\,, \end{aligned}$$

since $\mathrm{E}[ZZ^T] = \mathrm{cov}(Z) = \sigma^2 I_N$. The estimator $\tilde{b} = AY$ is the BLUE, if A solves the extremum problem

(E) minimize $\mathrm{tr}\{AA^T\}$ under the constraint $AX = I_d$.

Consider the least squares estimator

$$\hat{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = A_0 Y.$$

We prove in three steps that A_0 is a solution of (E) and therefore $\hat{b}$ is the BLUE of b.

(i) Consider M_X, the d-dimensional linear hull of the columns of $\mathbf{X}$ in $\mathbb{R}^N$. Additionally, let $A = B + C$ the unique decomposition of the $d \times N$-Matrix A in matrices B and C, whose rows are in M_X and orthogonal to M_X, respectively. In particular, the rows of C and the columns of B^T are orthogonal, so that $CB^T = 0_d$ (the $d \times d$-zero matrix). Then

$$\begin{aligned} \mathrm{tr}\{AA^T\} &= \mathrm{tr}\{(B + C)\ (B + C)^T\} = \mathrm{tr}\{BB^T + CC^T\} \\ &\geq \mathrm{tr}(BB^T) \end{aligned}$$

and, because $CX = 0$,

$$BX = AX = I_d.$$

If A is a solution of (E), it follows that B is a solution of (E).

It is enough to show that A_0 solves the extremum problem

(E') minimize $\mathrm{tr}\{AA^T\}$ under the constraints

a) $AX = I_d$,

b) $\{\text{rows of } A\} \subseteq M_X$

(ii) The rows of $A_0 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ are linear combinations of the rows of X^T and consequently lie in M_X. Additionally $A_0\mathbf{X} = I_d$, so A_0 satisfies the constraints of (E').

(iii) A_0 is the unique matrix that satisfies the constraints of (E') and therefore the solution of this extremum problem: Consider D an arbitrary matrix satisfying a) and b). Because the rows of D and A_0 lie in M_X, the rows of $D - A_0$ lie in M_X. And because $(D - A_0)\ \mathbf{X} = 0_d$, the rows of $D - A_0$ are orthogonal to M_X, so that we can conclude $D = A_0$. □

Chapter 3

Analysis of Variance (ANOVA)

In this chapter, we return to the starting point of the previous chapter. We have data

$$(x_1, y_1), \ldots, (x_N, y_N)$$

where the means μ_j of the y_j depend somehow on one or several explanatory variables. In contrast to regression, x_j which represents the conditions of the j^{th} experiment is no longer real or vector-valued, but highly discrete (assuming only a few different values) or even qualitative (i.e. may not be represented by a number in a natural way). Therefore, in the following, we identify x_j with its index j which is nothing else but the number of the corresponding experiment (or group of experiments under the same conditions).

3.1 The one-factor layout

First, we consider the so-called *one-factor layout analysis of variance* which is based on the following model:

$$(V1) \qquad Y_{jk} = \mu_j + Z_{jk}\ ,\ \ j = 1, \ldots, p,\ \ k = 1, \ldots, n_j,\ \ Z_{jk} \text{ i.i.d. } \mathcal{N}(0, \sigma^2).$$

The total sample of size $N = n_1 + \ldots + n_p$ is subdivided in p groups. In each group the explanatory variable has a fixed value. As in regression analysis, we assume that the value of the explanatory variables influences the sample mean but not the variance. Note that for the case $p = 2$ the model (V1) corresponds to the classical two sample problem where the two-sample t-test is applied.

Example: 21 male test persons are subdivided in 3 homogeneous groups w.r.t. some socio-economical parameters. The cholesterol level in the blood is measured:

Group A:	403	269	311	336	259					
Group B:	312	222	302	420	420	386	353	210	286	290
Group C:	403	244	353	235	319	260				

The question is whether the groups can be distinguished with respect to the mean cholesterol level in the blood. The group means are 315.6 for group A, 320.1 for group B, and 302.33

for group C.

To deal with such questions and to be able to generalize them to much more complicated situations, we proceed systematically and first isolate the overall mean from the effect caused by the different values of the explanatory factor:

$$\mu_j = \mu + \delta_j, \;\; j = 1, \ldots, p, \;\; \text{with} \sum_{i=j}^{p} n_j \delta_j = 0.$$

μ is called the overall mean, δ_i the factor effect. Then, the model of the one-factor layout analysis of variance can be written as

$$(V1) \qquad Y_{jk} = \mu + \delta_j + Z_{jk}, \;\; j = 1, \ldots, p, \;\; k = 1, \ldots, n_j \quad \text{with} \quad \sum_j n_j \delta_j = 0, \;\; Z_{jk} \text{ i.i.d. } \mathcal{N}(0, \sigma^2)$$

The question stated in the example whether the different values of the factor have an influence on the observation can then be formulated as a test problem:

Test in Model (V1) the hypothesis

$$H_0 : \mu_1 = \ldots = \mu_p, \quad \text{i.e.} \;\; \delta_1 = \ldots = \delta_p = 0,$$

against the alternative

$$H_1 : \delta_j \neq 0 \quad \text{for at least one } \; j.$$

First, we observe that the ML estimators (= LS estimators for the Gaussian model (V1)) for $\mu, \; \delta_1, \ldots, \delta_p$ are of the form:

$$\begin{aligned} \widehat{\mu} &= \frac{1}{N} \sum_{j=1}^{p} \sum_{k=1}^{n_j} Y_{jk} = \overline{Y}_{\bullet\bullet} \\ \hat{\delta}_j &= \frac{1}{n_j} \sum_{k=1}^{n_j} (Y_{jk} - \overline{Y}_{\bullet\bullet}) = \overline{Y}_{j\bullet} - \overline{Y}_{\bullet\bullet} \end{aligned}$$

Notation: In a multi-indexed sequence, e.g a_{ijk}, the use of a dot means that we sum over this index:

$$a_{i\bullet k} = \sum_j a_{ijk}, \;\; a_{i\bullet\bullet} = \sum_{j,k} a_{ijk} \;\; \text{etc.}$$

The overline refers to an average over this index. For example $\overline{a}_{i\bullet k}$ is $a_{i\bullet k}$ divided by the number of summands.

Remark: a) $\sum_{j=1}^{p} n_j \hat{\delta}_j = 0$

b) $\sum_{j=1}^{p} \hat{\delta}_j = 0, \quad \text{if } \; n_j = n, \;\; j = 1, \ldots, p.$

Define the sample residuals via

$$\widehat{Z}_{jk} = Y_{jk} - \widehat{\mu}_j = Y_{jk} - \widehat{\mu} - \hat{\delta}_j.$$

Then $\sum_{k=1}^{n_j} \hat{Z}_{jk} = 0$ for all j, and

$$\begin{aligned} Y_{jk} &= \widehat{\mu} + \hat{\delta}_j + \widehat{Z}_{jk} \\ &= \overline{Y}_{\bullet\bullet} + (\overline{Y}_{j\bullet} - \overline{Y}_{\bullet\bullet}) + (Y_{jk} - \overline{Y}_{j\bullet}) \end{aligned}$$

or in the vectorial form

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{pn_p} \end{pmatrix} = \widehat{\mu} \begin{pmatrix} 1 \\ \vdots \\ \cdot \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} \hat{\delta}_1 \\ \vdots \\ \hat{\delta}_1 \\ \hat{\delta}_2 \\ \vdots \\ \hat{\delta}_p \end{pmatrix} + \begin{pmatrix} \widehat{Z}_{11} \\ \vdots \\ \widehat{Z}_{1n_1} \\ \widehat{Z}_{21} \\ \vdots \\ \widehat{Z}_{pn_p} \end{pmatrix}$$

$$=: \qquad g \qquad + \qquad f \qquad + \qquad r.$$

The decomposition of the data vector Y into the vectors g (global mean), f (factor effects) and r (residuals) corresponds to a decomposition of $\mathbb{R}^N$ into three subspaces:

$$\begin{aligned} M_g &= \{c(1,\ldots,1)^T;\ c \in \mathbb{R}\}, \\ &\quad \dim M_g = 1,\ g \in M_g; \\ M_f &= \{(c_1,\ldots,c_1,\ c_2,\ldots,c_p)^T;\ c_j \in \mathbb{R},\ \sum_{j=1}^{p} n_j c_j = 0\}, \\ &\quad \dim M_f = p-1,\ f \in M_f; \\ M_r &= \{(c_{11},\ldots,c_{1n_1},\ c_{21},\ldots,c_{pn_p})^T;\ c_{jk} \in \mathbb{R},\ \sum_{k=1}^{n_j} c_{jk} = 0,\ j = 1,\ldots,p\}, \\ &\quad \dim M_r = N-p,\ r \in M_r. \end{aligned}$$

Proposition 3.1.1
The subspaces $M_g,\ M_f,\ M_r \subseteq \mathbb{R}^N$ are mutually orthogonal.

Proof:
Consider $\gamma \in M_g, \phi \in M_f$, and $\rho \in M_r$. Then

$$\gamma^T\phi = c\sum_{j=1}^{p} n_j c_j = 0, \quad \gamma^T\rho = c\sum_{j,k} c_{jk} = 0, \quad \phi^T\rho = \sum_{j=1}^{p} c_j \sum_{k=1}^{n_j} c_{jk} = 0.$$

□

From Pythagoras' theorem we get

$$\begin{aligned} \sum_{j,k} Y_{jk}^2 &= ||Y||^2 = ||g||^2 + ||f||^2 + ||r||^2 \\ &= N\overline{Y}_{\bullet\bullet}^2 + \sum_{j=1}^{p} n_j(\overline{Y}_{j\bullet} - \overline{Y}_{\bullet\bullet})^2 + \sum_{j,k}(Y_{jk} - \overline{Y}_{j\bullet})^2 \end{aligned}$$

or, if we move $N\overline{Y}_{\bullet\bullet}^2$ to the left-hand side,

$$\begin{aligned}\widehat{D}_G := \sum_{j,k}(Y_{jk} - \overline{Y}_{\bullet\bullet})^2 &= \sum_{j=1}^{p} n_j(\overline{Y}_{j\bullet} - \overline{Y}_{\bullet\bullet})^2 + \sum_{j,k}(Y_{jk} - \overline{Y}_{j\bullet})^2 \\ &=: \widehat{D}_B + \widehat{D}_I\end{aligned}$$

$\frac{1}{N}\widehat{D}_G$ is the sample variance for the overall sample $Y_{11}, \ldots, Y_{pn_p}$. We have decomposed this total variability in two parts. The first part $\frac{1}{N}\widehat{D}_B$ represents the variability of the data that is due to the difference in the means between the subsamples $Y_{i1}, \ldots, Y_{in_i}$, $i = 1, \ldots, p$, . The second part $\frac{1}{N}\widehat{D}_I$ represents the variability that exists within each subsample, i.e. the variability due to the randomness of the residuals Z_{jk} within the subsamples. This decomposition and its multi-factor generalizations gave its name to the Analysis of Variance:

total variability	=	variability between subsamples (due to the factor effects)	+	variability within subsamples (due to the random residuals)

If $\widehat{D}_G$ does not differ substantially from $\widehat{D}_I$, this hints at the absence of factors effects (i.e. $H_0 : \delta_1 = \ldots = \delta_p = 0$ holds). Based on this intuition, we want to develop a test for the above mentioned problem. As a test statistic, we use

$$R = \frac{N-p}{p-1}\,\frac{\widehat{D}_G - \widehat{D}_I}{\widehat{D}_I} = \frac{N-p}{p-1}\cdot\frac{\widehat{D}_B}{\widehat{D}_I}$$

The hypothesis H_0 is rejected if R is large enough.

Proposition 3.1.2
Let $n_j > 1$ for all $j = 1, \ldots, p$. Then $\hat{\sigma}^2 = \frac{1}{N-p}\widehat{D}_I$ is an unbiased estimator of σ^2. Furthermore, under the hypothesis $H_0 : \mu_1 = \ldots = \mu_p$, R is $F_{p-1,N-p}$-distributed.

Proof:
First, we observe that $\overline{Y}_{j\bullet}$ is the LS estimator of μ_j, hence

$$\widehat{D}_I = \min_{c_1,\ldots,c_p\in\mathbb{R}} \sum_{j=1}^{p}\sum_{k=1}^{n_j}(Y_{jk} - c_j)^2 = \min_{\nu\in M_f\oplus M_g} ||Y - \nu||^2.$$

It follows from Theorem 2.5.1 a) and dim $M_f \oplus M_g = p$, that $\frac{1}{\sigma^2}\widehat{D}_I$ is $\chi^2_{N-p}-$ distributed. So, $\hat{\sigma}^2 = \frac{1}{N-p}\widehat{D}_I$ is an unbiased estimator of σ^2.

Under the hypothesis $\mu_1 = \ldots = \mu_p = \mu$, $\mathrm{E}Y \in M_g \subseteq M_g \oplus M_f$, and additionally we have

$$\widehat{D}_G = \sum_{j,k}(Y_{jk} - \overline{Y}_{\bullet\bullet})^2 = \min_{\nu\in M_g} ||Y - \nu||^2.$$

The assertion follows from Theorem 2.5.1 b) and dim $M_g = 1$. □

Let $f_{1-\alpha}$ denote the $(1-\alpha)-$ quantile of the $F_{p-1,N-p}$-distribution. The test with acceptance region

$$\left\{y \in \mathbb{R}^N;\ \frac{N-p}{p-1} \cdot \frac{\widehat{D}_B}{\widehat{D}_I} \leq f_{1-\alpha}\right\}$$

is then a level α-test of the hypothesis $H_0 : \delta_1 = \ldots = \delta_p = 0$ (no factor effects) against the alternative $\delta_j \neq 0$ for at least one j in the model (V1).

Example (cont.): The essential quantities for the test are usually presented in the form of an ANOVA table ("Analysis of Variance") . For the one-factor layout analysis of variance we have the following table:

	Sums of squares	Degrees of freedom	Mean sums of squares	$F-$value
between groups	$\widehat{D}_B$	$p-1$	$\frac{1}{p-1}\widehat{D}_B = Q_B$	$\frac{Q_B}{Q_I} = R$
within groups	$\widehat{D}_I$	$N-p$	$\frac{1}{N-p}\widehat{D}_I = Q_I$	
total	$\widehat{D}_G$	$N-1$		

For the data of the example we obtain

between groups	1 202.5	2	601.2	0.126
within groups	85 750.5	18	4 763.9	
total	86 953.5	20		

Because the 0.95–quantile of the $F_{2,18}$-distribution has the value 3.95, we do not reject the hypothesis on the level 5%.

3.2 The two-factor layout

In the next step, we focus on situations where the data are influenced by two factors that assume p and q different values, respectively. In this *two-factor layout analysis of variance* we assume the following model:

$$(V2) \qquad Y_{ijk} = \mu_{ij} + Z_{ijk},\ i = 1, \ldots, p,\ j = 1, \ldots, q,\ k = 1, \ldots, m \quad Z_{ijk} \text{ are i.i.d. } \mathcal{N}(0, \sigma^2).$$

To simplify the notation, we limit ourselves to the case where for each pair (i, j) of values of the explanatory variables there is the same number m of i.i.d observations $Y_{ijk},\ k = 1, \ldots, m$. This situation is called a *balanced design*.

Once more we decompose the mean value μ_{ij} of these observations into the global mean of the overall sample, the average deviations from the global mean caused by the 1^{st} and the 2^{nd} factor and the contribution caused by the interaction between the value i of the 1^{st} factor and the value j of the 2^{nd} factor:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

where

μ	$=$	$\overline{\mu}_{\bullet\bullet}$	global mean
α_i	$=$	$\overline{\mu}_{i\bullet} - \mu$	contribution of the 1^{st} factor value i to the mean
β_j	$=$	$\overline{\mu}_{\bullet j} - \mu$	contribution of the 2^{nd} factor value j to the mean
γ_{ij}			interaction effect between the 1^{st} factor value i and the 2^{nd} factor value j.

It follows:

$$\sum_{i=1}^{p} \alpha_i = \sum_{j=1}^{q} \beta_j = 0, \;\; \sum_{i=1}^{p} \gamma_{il} = \sum_{j=1}^{q} \gamma_{lj} = 0 \;\; \text{for all } \; l.$$

The origins of analysis of variance lies in biometrics, where the second factor usually appears because one cannot hold the experimental conditions during the investigation of the 1^{st} factor permanently constant. For example, if we want to study the effect of different fertilizers on the crop yield, we need to collect the data from different fields that differ with respect to the soil, the local climate etc. The fertilizer is then the 1^{st} factor, the field the 2^{nd} factor. Since, e.g., an otherwise good fertilizer i can always be worse on the sour soil of field j we need to consider interaction terms.

We first limit ourselves to situations where the absence of interaction is plausible, i.e. we assume additivity of effects: the contribution of the 2^{nd} factor j to the mean is always independent of the corresponding value of the 1^{st} factor and vice versa. The model for the two-factor layout analysis of variance without interaction is the following:

(V2') $Y_{ijk} = \mu + \alpha_i + \beta_j + Z_{ijk}, \;\; i = 1, \ldots, p, \;\; j = 1, \ldots, q, \;\; k = 1, \ldots, m,$
with $\sum_{i=1}^{p} \alpha_i = 0, \;\; \sum_{j=1}^{q} \beta_j = 0,$
where the $N = pqm$ random variables Z_{ijk} are i.i.d. $\mathcal{N}(0, \sigma^2)$.

Analogous to the one-factor layout analysis of variance we decompose the data vector

$$Y = g + f_1 + f_2 + r \in \mathbb{R}^N,$$

where the components are defined as follows:

$$\begin{aligned} Y_{ijk} &= \overline{Y}_{\bullet\bullet\bullet} + (\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}) + (\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet}) + (Y_{ijk} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet}). \\ &= \widehat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{Z}_{ijk} \end{aligned}$$

The decomposition of the data vector is equivalent to the decomposition of $\mathbb{R}^N$ into the following 4 orthogonal subspaces:

$$\begin{array}{lcll} & & & \text{Dimension:} \\ g \in M_g & = & \{c(1,\ldots,1)^T;\ c \in \mathbb{R}\} & 1 \\ f_1 \in M_1 & = & \{(\underbrace{c_1,\ldots,c_1}_{mq},\underbrace{c_2,\ldots,c_2}_{mq},c_3,\ldots,c_p)^T;\ c_i \in \mathbb{R},\ \sum_{i=1}^{p} c_i = 0\} & p-1 \\ f_2 \in M_2 & = & \{(\underbrace{\underbrace{c_1',\ldots,c_1'}_{m},\underbrace{c_2',\ldots,c_2'}_{m},c_3',\ldots,c_3',c_q',\ldots,c_q'}_{p \text{ times}})^T;\ c_j' \in \mathbb{R},\ \sum_{j=1}^{q} c_j' = 0\} & q-1 \\ r \in M_r & = & \{(c_{111},\ldots,c_{11m},\ c_{121},\ldots,c_{12m},\ c_{131},\ldots,c_{pqm})^T;\ c_{ijk} \in \mathbb{R}, & \\ & & \sum_{i,k} c_{ijk} = 0,\ j = 1,\ldots,q,\ \sum_{j,k} c_{ijk} = 0\ ,\ i = 1,\ldots,p\} & N-p-q+1 \end{array}$$

The orthogonality of the subspaces $M_g,\ M_1,\ M_2$, and M_r is proven as in Proposition 3.1.1. Since they are linear combinations of normally distributed random variables, $g,\ f_1,\ f_2,\ r$ have multivariate Gaussian distributions, too.

Using again Pythagoras' theorem we get the following decomposition of variability:

$$\widehat{D}_G = ||Y-g||^2 = ||f_1||^2 + ||f_2||^2 + ||r||^2 = \widehat{D}_1 + \widehat{D}_2 + \widehat{D}_R$$

or in detail

$$\begin{array}{lcccccc} \widehat{D}_G & = & \sum\limits_{i,j,k} (Y_{ikj} - \overline{Y}_{\bullet\bullet\bullet})^2 & & & & \\ & = & qm \sum\limits_{i=1}^{p} (\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2 & + & pm \sum\limits_{j=1}^{q} (\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2 & + & \sum\limits_{i,j,k} \widehat{Z}_{ijk}^2 \\ & = & \widehat{D}_1 & + & \widehat{D}_2 & + & \widehat{D}_R. \end{array}$$

The interpretation is the following:

$$\begin{array}{lcll} \text{Total variance } \frac{1}{N}\widehat{D}_G = & & \text{variance } \frac{1}{N}\widehat{D}_1 & \text{due to the variability of the } 1^{st} \text{ factor} \\ & + & \text{variance } \frac{1}{N}\widehat{D}_2 & \text{due to the variability of the } 2^{nd} \text{ factor} \\ & + & \text{variance } \frac{1}{N}\widehat{D}_R & \text{due to randomness of the residuals } Z_{ijk}. \end{array}$$

The various hypotheses can be expressed in terms of these subspaces:

$$\begin{array}{rl} \mathrm{E}Y \in M_g & \text{if } \alpha_1 = \ldots = \alpha_p = \beta_1 = \ldots, \beta_q = 0 \\ \mathrm{E}Y \in M_g \oplus M_1 & \text{if } \beta_1 = \ldots = \beta_q = 0 \\ \mathrm{E}Y \in M_g \oplus M_2 & \text{if } \alpha_1 = \ldots = \alpha_p = 0 \\ \mathrm{E}Y \in M_g \oplus M_1 \oplus M_2 & \text{always} \end{array}$$

Assume we want to test

$$H_0^\alpha : \alpha_1 = \ldots = \alpha_p = 0 \text{ against } H_1^\alpha : \alpha_i \neq 0 \text{ for at least one } i.$$

In general, $E[Y] \in M_g \oplus M_1 \oplus M_2$, $\dim M_g \oplus M_1 \oplus M_2 = p + q - 1$ and

$$\hat{D}_R = \min_{\nu \in M_g \oplus M_1 \oplus M_2} ||Y - \nu||^2.$$

Under H_0^α, $E[Y] \in M_g \oplus M_2$, $\dim M_g \oplus M_2 = q$ and $\min_{\nu \in M_g \oplus M_2} ||Y - \nu||^2 = \hat{D}_R + \hat{D}_1$. Hence, Theorem 2.5.1 implies that under this hypothesis

$$R_1 = \frac{N - p - q + 1}{p - 1} \cdot \frac{\widehat{D}_1}{\widehat{D}_R}$$

is $F_{p-1,N-p-q+1}$-distributed.
The ANOVA-table for this decision problem and the analogous question if the second factor has an effect (testing $H_0^\beta : \beta_1 = \ldots = \beta_q = 0$) is represented as follows:

	sum of squares ($\widehat{D}$)	degrees of freedom	mean sum of squares	F-value (=R)
1^{st} factor	$\widehat{D}_1 = qm \sum_{i=1}^{p} (\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2$	$p - 1$	$\frac{1}{p-1}\widehat{D}_1 = Q_1$	$\frac{Q_1}{Q_R}$
2^{nd} factor	$\widehat{D}_2 = pm \sum_{j=1}^{q} (\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2$	$q - 1$	$\frac{1}{q-1}\widehat{D}_2 = Q_2$	$\frac{Q_2}{Q_R}$
residuals	$\widehat{D}_R = \sum_{i,j,k} (\overline{Y}_{ijk} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2$	$N - p - q + 1$	$\frac{1}{N-p-q+1}\widehat{D}_R = Q_R$	
total	$\widehat{D}_G = \sum_{i,j,k} (Y_{ikj} - \overline{Y}_{\bullet\bullet\bullet})^2$	$N - 1$		

Example 3.2.1
The tensile strength of 4 blades of a titanium base alloy is measured in the center (Z), at a corner (E) and on the longitudinal (L) and horizontal (H) edge. We have two factors: the number of the blade and the location where the stress is applied. For each combination of factors, one measurement is reported. This means $p = 4, q = 4$ and $m = 1$.

	blade			
	1	2	3	4
location E	137.1	142.2	128.0	136,6
location Z	140.1	139.4	116.8	136.5
location L	141.8	139.6	132.5	140.8
location H	136.1	140.8	132.2	129.0

The associated ANOVA-Table is given in the following form

location	66	3	22	1.05
blade	407	3	136	6.3
residuals	194	9	21	
total	667	15		

The 0.95 quantile of the $F_{3,9}$ distribution is 3.86. The data give no reason to doubt about the hypothesis that the tensile strength is constant over the entire area of the blade. The hypothesis that the all four blades have the same tensile strength can be rejected on the level 5% .

We have assumed that there is no interaction between both factors in the two–factor layout model (V2'). One can check again if this hypothesis is consistent with the data with a test. We return to our general model:

$$\begin{aligned}(\text{V2}) \quad & Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + Z_{ijk}, \;\; i = 1, \ldots, p, \;\; j = 1, \ldots, q, \;\; k = 1, \ldots, m, \\ & \text{with } \sum_{i=1}^{p} \alpha_i = 0, \;\; \sum_{j=1}^{q} \beta_j = 0, \;\; \sum_{i=1}^{p} \gamma_{i\ell} = 0 = \sum_{j=1}^{q} \gamma_{\ell j} \text{ for all } \ell, \\ & \text{the } Z_{ijk} \text{ are i.i.d. } \mathcal{N}(0, \sigma^2).\end{aligned}$$

In this model we want to test the hypothesis of no interaction, i.e. $\gamma_{ij} = 0, \;\; i = 1, \ldots, p, \;\; j = 1, \ldots, q$. Again, with the help of Theorem 2.5.1 we want to construct an F-test. Under model (V2), EY always lies in the pq-dimensional subspace $M_{1,2}$ of $\mathbb{R}^N$ ($N = pqm$) defined via

$$M_{1,2} = \{(\underbrace{c_{11}, \ldots, c_{11}}_{m}, \underbrace{c_{12}, \ldots, c_{12}}_{m}, \ldots, \underbrace{c_{pq}, \ldots, c_{pq}}_{m})^T; \;\; c_{ij} \in \mathbb{R}\}.$$

Since $\overline{Y}_{ij\bullet}$ is the LS–estimator of E$Y_{ijk}, \;\; k = 1, \ldots, m$, we obtain

$$\widehat{D}_{1,2} = \sum_{i,j,k} (Y_{ijk} - \overline{Y}_{ij\bullet})^2 = \min_{\nu \in M_{1,2}} ||Y - \nu||^2.$$

On the other hand, under the hypothesis of no interaction we have E$Y \in M_g \oplus M_1 \oplus M_2 =: M_0$, and

$$\min_{\nu \in M_0} ||Y - \nu||^2 = \widehat{D}_R \, .$$

Since dim $M_0 = p + q - 1$, it follows again from Theorem 2.5.1b that under this hypothesis

$$\begin{aligned} R &= \frac{N - pq}{pq - (p + q - 1)} \cdot \frac{\widehat{D}_R - \widehat{D}_{1,2}}{\widehat{D}_{1,2}} \\ &= \frac{N - pq}{(p-1)(q-1)} \cdot \frac{m}{\widehat{D}_{1,2}} \sum_{i,j} (\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2 \end{aligned}$$

is $F_{(p-1)(q-1),N-pq}$-distributed. The test can be used only in the case $m > 1$ since otherwise $\widehat{D}_{1,2} = 0$. For the case $m = 1$, other tests exist.

Chapter 4

Generalized linear models

4.1 Binary response variables and logistic regression

Linear regression works well if the responses Y_j and the explanatory variables x_{jk} can be well interpreted as real numbers. In the previous chapter on ANOVA, we have looked at the case where that is no longer true for the x_{jk}. In this chapter, we consider situations where the Y_j are discrete, e.g. assume only the two values 0 and 1.

Let $Z_1, \ldots, Z_N$ be independent Bernoulli variables with

$$\pi_j = P(Z_j = 1) = \mathrm{E}(Z_j),\ j = 1, \ldots, N.$$

Additionally, we are given the corresponding predictor variables $(x_{j1}, \ldots, x_{jp}) = \xi_j^T,\ j = 1, \ldots, N$.

Goal (as in linear regression): Describe $\mathrm{E}(Z_j)$ as a function of ξ_j.

However, linear regression (assuming $\mathrm{E}(Z_j) = \xi_j^T b,\ b \in \mathbb{R}^p$) does not work as $0 \leq \mathrm{E}(Z_j) \leq 1$ and linear functions are unbounded. Therefore, we extend the linear model by introducing a fixed nonlinear transformation into the relationship between $\mathrm{E}(Z_j)$ and $\xi_j^T b$:

Model: For some parameter vector $b \in \mathbb{R}^p$ and some given link function g

$$\pi_j = P(Z_j = 1) = g^{-1}(\xi_j^T b) \qquad \text{or} \qquad g(\pi_j) = \xi_j^T b\ ,\ \ j = 1, \ldots, N.$$

Common choices are

a) Probit model: $g(\pi) = \Phi^{-1}(\pi)$, where Φ is the distribution function of $\mathcal{N}(0, 1)$.

The idea of this approach is to introduce random variables

$$Y_i = \xi_i^T b + \varepsilon_i, i = 1, \ldots, N,\ \varepsilon_1, \ldots, \varepsilon_N \text{ i.i.d.}, \mathcal{L}(\varepsilon_i) = \mathcal{N}(0, 1),$$

If we set $Z_i = 1$ if and only if $Y_i > 0$ then $P(Z_i = 1) = \Phi(\xi_i^T b)$.
The motivation of the model is as follows:

Assume that the random variables Y_i refer to some measurements that follow a normal law, e.g. the blood pressure. A person is diagnosed to suffer from high blood pressure if the corresponding Y_i exceeds a certain threshold. Hence, $Z_i = 1$ can be interpreted as "person suffers from high blood pressure".

b) Logit model or logistic regression model:

$$g(\pi) = \log \frac{\pi}{1-\pi}\,, \text{ i.e. } g^{-1}(u) = \frac{1}{1+e^{-u}} = \frac{e^u}{1+e^u}$$

g^{-1} is called the logistic function, and it is the distribution function of the logistic distribution.
This model is motivated by using the same interpretation as above just with replacing the normal distribution by a logistic distribution. Additionally, it has the following interpretation:
Consider the odds

$$\text{Odds}(Z_i) = \frac{P(Z_i = 1)}{1 - P(Z_i = 1)} = \frac{P(Z_i = 1)}{P(Z_i = 0)} \in [0, \infty).$$

Then $\text{Logit}(Z_i) = \log(\text{Odds}(Z_i))$ is unbounded, so we may assume

$$\text{Logit}(Z_i) = \xi_i^T b.$$

We consider a slightly more general setup with repeated observations:

$$Z_{i1}, \ldots, Z_{in_i} \text{ i.i.d. Bernoulli variables with } \pi_i = P(Z_{ik} = 1),\ k = 1, \ldots, n_i,\ i = 1, \ldots, N.$$

We have $n_1 + \ldots + n_N$ observations, but we can combine them to N independent binomial random variables without loss of information:

$$Y_i = \sum_{k=1}^{n_i} Z_{ik} \text{ is } \mathcal{B}(n_i, \pi_i) \text{ - distributed},\ i = 1, \ldots, N.$$

This yields the following transformation of the model:
Model: $Y_1, \ldots, Y_N$ independent, $\mathcal{L}(Y_i) = \mathcal{B}(n_i, \pi_i),\ i = 1, \ldots, N,$ with

$$g(\pi_i) = \xi_i^T b,\ i = 1, \ldots, N.$$

The parameter b has to be estimated, e.g. by maximum likelihood (ML). For abbreviation, we write π_i in the likelihood function, but we keep in mind that $\pi_i = g^{-1}(\xi_i^T b)$ is a (known) function of the parameter b. The log-likelihood then is

$$\ell(b|Y_1, \ldots, Y_N) = \sum_{i=1}^{N} \log B(Y_i;\ n_i, \pi_i)$$

by independence of the Y_i, where

$$B(k;n,p) = \left(\begin{array}{c} n \\ k \end{array} \right) \; p^k(1-p)^{n-k}$$

denotes the weights of the binomial distribution. Therefore, we get

$$\ell(b|Y_1,\ldots,Y_N) = \sum_{i=1}^{N} \left\{ Y_i \log \frac{\pi_i}{1-\pi_i} + n_i \log(1-\pi_i) + \log \left(\begin{array}{c} n_i \\ Y_i \end{array} \right) \right\}$$

Note that the last term does not depend on the π_i's and therefore on b, i.e. it can be omitted in maximizing ℓ. The ML-estimator is

$$\hat{b} = \arg\max_b \ell(b|Y_1,\ldots,Y_N).$$

In practice, this estimator has to be computed numerically, e.g. by using Newton's method. In the following, we will discuss how to find suitable starting values.
If the model is saturated (i.e. $p = N$) such that without loss of generality $b = (\pi_1,\ldots,\pi_N)^T$, the ML estimator for b is given by $(b_{\max})_i = \frac{Y_i}{n_i}$. For the deviance (defined in Section 4.2) we get

$$\begin{aligned} D &= 2[\ell(b_{\max}|Y_1,\ldots,Y_N) - \ell(\hat{b}|Y_1,\ldots,Y_N)] \\ &= 2\sum_{i=1}^{N} Y_i \log \frac{Y_i}{n_i - Y_i} + n_i \log(1 - \frac{Y_i}{n_i}) - Y_i \log \frac{\hat{\pi}_i}{1-\hat{\pi}_i} - n_i \log(1-\hat{\pi}_i) \\ &= 2\sum_{i=1}^{N} Y_i \log \frac{Y_i}{n_i\hat{\pi}_i} + (n_i - Y_i) \log \frac{n_i - Y_i}{n_i - n_i\hat{\pi}_i} \end{aligned}$$

Because $\hat{b}$ maximizes $\ell(b)$, $\hat{b}$ minimizes

$$D(b) = 2[\ell(b_{\max}|Y_1,\ldots,Y_N) - l(b|Y_1,\ldots,Y_N)].$$

A Taylor expansion of $\log \frac{x}{t}$ about $x = t$ yields

$$s \log \frac{s}{t} = (s-t) + \frac{1}{2}\frac{(s-t)^2}{t} + \ldots.$$

Applying this to the equation for D tells us that we get an approximate estimate of b by minimizing

$$Q(b) = \sum_{i=1}^{N} \frac{(Y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)},$$

where again $\pi_i = g^{-1}(\xi_i^T b)$. The solution $\tilde{b}$ is a weighted nonlinear least-squares estimator where the weights depend on the unknown parameter. It is computationally advantageous to first determine $\tilde{b}$ and then use it as starting value for the numerical calculation of $\hat{b}$.

An even simpler nonlinear least-squares estimator with weights not depending on the parameter is defined as minimizer $\tilde{b}_{\text{mod}}$ of

$$Q_{\text{mod}}(b) = \sum_{i=1}^{N} \frac{(Y_i - n_i\pi_i)^2}{Y_i(1-\frac{Y_i}{n_i})}$$

where we use that by the law of large numbers $\pi_i \approx Y_i/n_i$ for large enough n_i. For logistic regression, $\tilde{b}_{\text{mod}}$ can be approximated by $\hat{b}_{\text{mod}}$ minimizing

$$\hat{Q}_{\text{mod}}(b) = \sum_{i=1}^{N} w_i (U_i - \xi_i^T b)^2$$

with $U_i = g\left(\frac{Y_i}{n_i}\right)$ and weights $w_i = \{Y_i(1 - \frac{Y_i}{n_i})\}^{-1}$. Here we use that if $Y_i \approx n_i\pi_i$ then $U_i = g\left(\frac{Y_i}{n_i}\right) \approx g(\pi_i) = \xi_i^T b$. Now, calculating $\hat{b}_{\text{mod}}$ can be done explicitly as we just have to solve a weighted linear least-squares problem

$$\text{minimize } (U - \mathbf{X}b)^T W (U - \mathbf{X}b) \text{ w.r.t } b$$

with diagonal matrix W with diagonal elements $w_1, \ldots, w_N$ and $\mathbf{X} = (x_{il})_{i=1,\ldots,N,\ l=1,\ldots,p}$ as usual. The solution is

$$\hat{b}_{\text{mod}} = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W U,$$

and provides a quick and good starting value for calculating the ML-estimator $\hat{b}$ numerically.

4.2 Checking the adequacy of a model: the deviance

Assume there is a maximal model with parameter $b_{\max} \in \mathbb{R}^N$, i.e. the number of parameters equals the number of observations.
We consider the log likelihood ratio (LR) statistic for testing a given model with $b \in \mathbb{R}^p$ against the maximal model $b_{\max} \in \mathbb{R}^N$:

$$\log \lambda = \log \frac{L(\hat{b}|Y_1, \ldots, Y_N)}{L(\hat{b}_{\max}|Y_1, \ldots, Y_N)} = \ell(\hat{b}|Y_1, \ldots, Y_N) - \ell(\hat{b}_{\max}|Y_1, \ldots, Y_N).$$

We define the (scaled) *deviance* as

$$D = -2 \log \lambda = 2\big(\ell(\hat{b}_{\max}|Y_1, \ldots, Y_N) - \ell(\hat{b}|Y_1, \ldots, Y_N)\big).$$

In practice, this transformation of the LR statistic is often used as its asymptotic distribution is known.
For computing the asymptotic distribution of D, we use the following decomposition

$$\begin{aligned} D &= 2\big\{\ell(\hat{b}_{\max}|Y_1, \ldots, Y_N) - \ell(b_{\max}|Y_1, \ldots, Y_N)\big\} \\ &\quad +2\big\{\ell(b_{\max}|Y_1, \ldots, Y_N) - \ell(b|Y_1, \ldots, Y_N)\big\} \\ &\quad -2\big\{\ell(\hat{b}|Y_1, \ldots, Y_N) - \ell(b|Y_1, \ldots, Y_N)\big\}, \end{aligned}$$

where the second term is obviously always ≥ 0, but close to 0 if the fit of the model with parameter b is almost as good as the fit of the maximal model. From general results on LR statistics (discussed in the lecture 'Mathematical Statistics'), the first and the last term are asymptotically χ^2_N and χ^2_p, distributed, respectively, i.e. their difference is roughly χ^2_{N-p}-distributed. This yields $E[D] \approx N - p$.

Examples

(i) linear regression, $\mathcal{L}(Y_j) = \mathcal{N}(\mu_j, \sigma^2)$ with known σ^2, $\mu_j = (\mathbf{X}b)_j$, $j = 1, \ldots, N$
maximal model: $\hat{\mu}_{\max,j} = Y_j$

$$\begin{aligned} \Longrightarrow \qquad 2\ell(\hat{b}) &= -N\log(2\pi\sigma^2) - \frac{1}{\sigma^2}\sum_{j=1}^{N}(Y_j - \widehat{Y}_j)^2 \text{ with } \widehat{Y} = \mathbf{X}\hat{b}, \\ 2\ell(\hat{b}_{\max}) &= -N\log(2\pi\sigma^2) \text{ as, in this model, } \widehat{Y}_j = \hat{\mu}_j = Y_j. \end{aligned}$$

Together:

$$D = \frac{1}{\sigma^2}\sum_{j=1}^{N}(Y_j - \widehat{Y}_j)^2 = \frac{RSS}{\sigma^2}$$

From Theorem 2.5.1 we get $\mathcal{L}(D) = \chi^2_{N-p}$.

(ii) logistic regression, $\mathcal{L}(Y_j) = \mathcal{B}(n_j, \pi_j),\ j = 1, \ldots, N,$

$b_{\max} = (\pi_1, \ldots, \pi_N)^T,\ \hat{\pi}_{\max,j} = \frac{Y_j}{n_j}$

$$D = 2\sum_{j=1}^{N}\left(Y_j \log\frac{Y_j}{n_j\hat{\pi}_j} + (n_j - Y_j)\log\frac{n_j - Y_j}{n_j(1-\hat{\pi}_j)}\right) \qquad \text{with } \hat{\pi}_j \equiv g^{-1}(\xi_j^T\hat{b}) = \frac{1}{1+e^{-\xi_j^T\hat{b}}}.$$

Remark 4.2.1
Usually, we have $D = \sum_{j=1}^{N} d_j$ with appropriate d_j depending on Y_j and ξ_j. Then,

$$r_j := \operatorname{sgn}(d_j)\sqrt{|d_j|},\ j = 1, \ldots, N$$

are asymptotically normally distributed if the model is good. For diagnostic purposes, $r_1, \ldots, r_N$ play the same role as the sample residuals $\widehat{Z}_j$ in linear regression.

Example 4.2.2
Some insects are exposed to gaseous carbon disulphide at various concentrations for five hours. The number of insects killed is reported.

Dose x_i	Number of insects n_i	Number killed y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

We consider the logistic model

$$\pi_i = \frac{\exp(b_1 + b_2 x_i)}{1 + \exp(b_1 + b_2 x_i)},$$

so

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = b_1 + b_2 x_i.$$

The log-likelihood is

$$l = \sum_{i=1}^{N} \left[y_i(b_1 + b_2 x_i) - n_i \log[1 + \exp(b_1 + b_2 x_i)] + \log \binom{n_i}{y_i} \right].$$

The maximum likelihood estimator is obtained using Newton's method. As starting value, we choose $b_1^{(0)} = b_2^{(0)} = 0$. The solution is $\hat{b}_1 = -60.72$ and $\hat{b}_2 = 34.27$. With this, we get $D = 11.23$. The 0.95-quantile of the χ^2_6-distribution is 12.59. Hence, the logistic regression model is not rejected by a test based on the χ^2-distribution. However, the fit is not particularly good, either.

4.3 The generalized linear model (GLIM)

The techniques from the previous two sections can also be applied in a more general context.

Definition 4.3.1
An exponential family of distributions is characterized by a common form for the probability density

$$p_\vartheta(x) = s(x)t(\vartheta)\ e^{a(x)c(\vartheta)} = \exp\{a(x)c(\vartheta) + d(\vartheta) + f(x)\},$$

$\vartheta \in \Theta \subseteq \mathbb{R}$, or for the probability weights

$$P_\vartheta(X = x_k) = \exp\{a(x_k)c(\vartheta) + d(\vartheta) + f(x_k)\}.$$

If $c(\vartheta) = \vartheta$, the distribution is said to be in canonical form, and the corresponding ϑ is called the natural parameter. Other parameters of the distribution in addition to ϑ are called nuisance parameters.

Examples:
$P(\lambda)$, $\mathrm{Exp}(\lambda)$, $\mathcal{B}(n,p)$ (with nuisance parameter n), $\mathcal{N}(\mu,\sigma^2)$ (with nuisance parameter σ^2).

Exponential families share some convenient technical properties of the Gaussian family, and the theory can be developped along similar lines. Combining it with the idea of modifying linear regression by nonlinear link functions, we get the *Generalized Linear Regression Model*:

(GLIM) $Y_1, \ldots, Y_N$ are independent having distributions belonging to the same exponential family in canonical form.
Let $\xi_i^T = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, N$, be the corresponding predictor variables. For some monotone link function $g \in C^1$, we have

$$\mu_i = \mathrm{E}Y_i = g^{-1}(\xi_i^T b),\ i = 1, \ldots, N,$$

for some $b \in \mathbb{R}^p$.

Examples:
a) $g(u) = u,\ \mathcal{L}(Y_i) = \mathcal{N}(\mu_i, \sigma^2)$.
Then, with $\mathbf{X} = (x_{ik})_{i=1,\dots,N,\ k=1,\dots,p}$, (GLIM) reduces to the linear regression model $Y = \mathbf{X}b + Z$ with i.i.d. $\mathcal{N}(0, \sigma^2)$-variables Z_i.

b) logistic regression: $\vartheta = \pi,\ g(\pi) = \log \frac{\pi}{1-\pi}$

We consider maximum likelihood (ML) estimators for the parameter vector b of a GLIM. Due to the independence of the data, the likelihood function factorizes:

$$L(b|Y_1, \dots, Y_N) = \prod_{j=1}^{N} p_{\vartheta_j}(Y_j)$$

where the GLIM parameter b and the parameters of the exponential family ϑ_j, in the case of densities, are related by

$$g^{-1}(\xi_j^T b) = \mathrm{E}_{\vartheta_j}(Y_j) = \int y p_{\vartheta_j}(y) dy.$$

Considering the log likelihood and the special form of the densities (or weights) of the exponential family, we finally get

$$\ell(b) = \ell(b|Y_1, \dots, Y_N) = \log L(b|Y_1, \dots, Y_N) = \sum_{j=1}^{N} \log p_{\vartheta_j}(Y_j) = \sum_{j=1}^{N} \big\{a(Y_j)\vartheta_j + d(\vartheta_j)\big\} + \text{const},$$

where the constant does not depend on the parameters. So the ML estimator is given by

$$\hat{b} = \arg\max_b \ell(b|Y_1, \dots, Y_N) = \arg\max_b \sum_{j=1}^{N} \big\{a(Y_j)\vartheta_j + d(\vartheta_j)\big\}.$$

Assuming that the log likelihood is twice continuously differentiable w.r.t. b, we define the $p \times p$ *Fisher information* matrix I by

$$I_{km} = \mathrm{E}\left[\frac{\partial \ell(b)}{\partial b_k} \frac{\partial \ell(b)}{\partial b_m}\right] = -\mathrm{E}\left[\frac{\partial^2 \ell(b)}{\partial b_k \partial b_m}\right], \qquad 1 \le k, m \le p,$$

where the identity follows from integration by parts. Similar to proving the asymptotic normality of ML estimators in more general situations (lecture 'Mathematical Statistics') we get

Theorem 4.3.2
Assume (GLIM) where the ML-estimator $\hat{b}$ of b is uniquely defined with probability 1 and I is nonsingular. Then,

a) $\sqrt{N}(\hat{b} - b) \xrightarrow[\mathcal{L}]{} \mathcal{N}_p(0, \Sigma)$ *with* $\Sigma = \lim NI^{-1}$

b) $(\hat{b} - b)^T I(\hat{b} - b) \xrightarrow[\mathcal{L}]{} \chi^2_p$ *(Wald statistic).*

Chapter 5

Time Series - Preliminaries

5.1 Examples of Time Series

A time series in discrete time is a sequence of observations that is ordered in time. Each observation is interpreted as a realisation of a random variables. Hence, we are looking at a sequence of random variables that are ordered in time (stochastic process). In many cases, independence of the random variables cannot be assumed.
Time series are recorded in various fields of application including

- Finance: stock prices, market risk, ...
- Economics: annual growth, price indices, unemployment rates...
- Environmental sciences: temperature, rainfall, erosion,...
- Social sciences: annual birth rate,...
- Medicine: brain activity, heart rates, ...

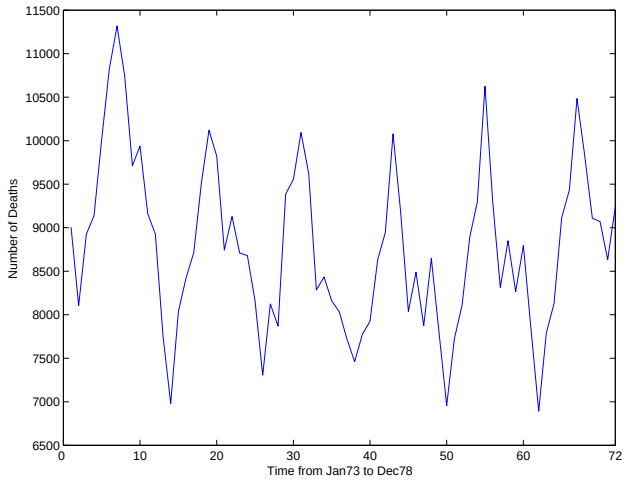

Figure 5.1: Accidental Deaths USA Jan 1973 to Dec 1978

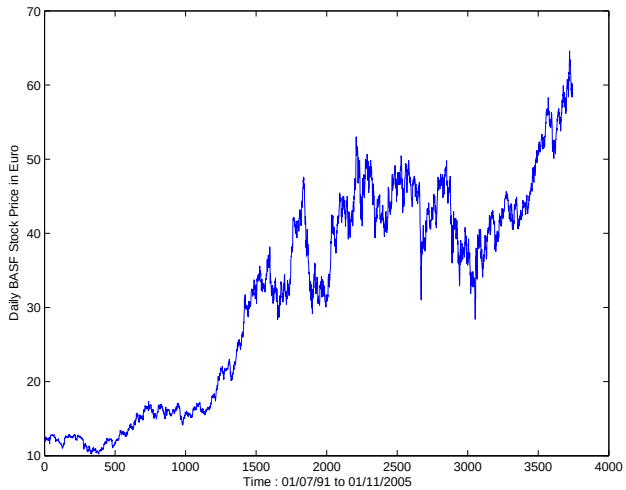

Figure 5.2: BASF Daily Stock Value

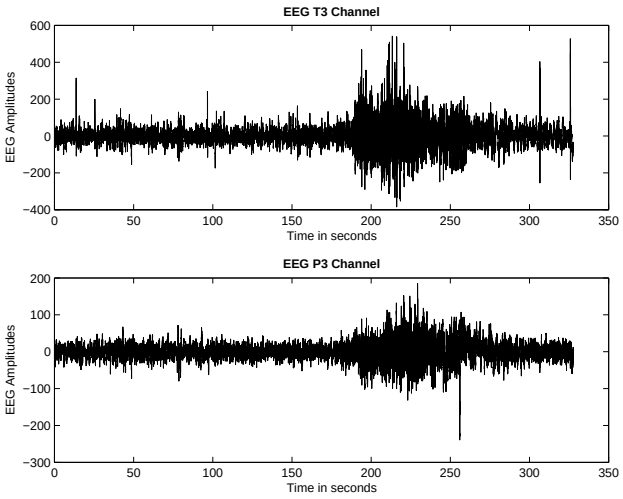

Figure 5.3: Brain Activity

5.2 Elementary theory of Hilbert spaces

Definition 5.2.1
A complex or real vector space $\mathcal{H}$ is an inner product space if there exists a map

$$< ., . >: \mathcal{H} \times \mathcal{H} \longrightarrow \mathbb{C}(\mathbb{R})$$

that satisfies the following properties:

1. $< x, y >= \overline{< y, x >}$ for all $x, y \in \mathcal{H}$
2. $< x + y, z >=< x, z > + < y, z >$ for all $x, y, z \in \mathcal{H}$
3. $< x, \alpha y >= \overline{\alpha} < x, y >$ for all $x, y \in \mathcal{H}, \alpha \in \mathbb{C}$
4. $< x, x > \geq 0$ for all $x \in \mathcal{H}$
5. $< x, x >= 0$ if and only if $x = 0$

A norm of x on $\mathcal{H}$ equipped with $< ., . >$ is

$$\|x\| = \sqrt{< x, x >}$$

For example, if $\mathcal{H} = \mathbb{R}^n$, based on

$$< x, y >= \sum_{i=1}^{n} x_i y_i,$$

the Euclidean norm is defined as

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Lemma 5.2.2
For all $x, y \in \mathcal{H}$ we have the following inequalities:

1. *Cauchy-Schwarz Inequality*
$$| < x, y > | \leq \|x\|\|y\|$$
2. *Triangular Inequality*
$$\|x + y\| \leq \|x\| + \|y\|$$

Note that the Cauchy-Schwarz inequality implies that the inner product is a continuous function on $\mathcal{H} \times \mathcal{H}$.

Definition 5.2.3
A sequence $\{x_n; n = 1, 2, \ldots\}$ of elements of an inner product space $\mathcal{H}$ is said to be a Cauchy sequence, if
$$\|x_n - x_m\| \longrightarrow 0 \text{ as } n, m \longrightarrow \infty.$$

Definition 5.2.4 (Hilbert Space)
A Hilbert space $\mathcal{H}$ is an inner product space which is complete, i.e. every Cauchy sequence converges in norm to some $x \in \mathcal{H}$, i.e.
$$\|x_n - x\| \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

Theorem 5.2.5 (Projection)
If $\mathcal{M}$ is closed subspace of a Hilbert space $\mathcal{H}$ and $x \in \mathcal{H}$, then

1. *there exists a unique element $\hat{x} \in \mathcal{M}$ such that*
$$\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|.$$
2. *$\hat{x} \in \mathcal{M}$ and $\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$ if and only if $\hat{x} \in \mathcal{M}$ and $(x - \hat{x}) \in \mathcal{M}^{\perp}$, where $y \in \mathcal{M}^{\perp}$ if and only if $< x, y >= 0$ for all $x \in \mathcal{M}$*

Definition 5.2.6

- The ***closed span*** $\overline{sp}\{x_t, t \in T\}$ of any subset of a Hilbert space $\mathcal{H}$ is the smallest closed space of $\mathcal{H}$ which contains all $x_t, t \in T$.
- A set $\{e_t, t \in T\}$ is called orthonormal if for every $s, t \in T$
$$< e_s, e_t >= \delta_{st} = \begin{cases} 1 & \text{if } t = s \\ 0 & \text{otherwise.} \end{cases}$$
- A Hilbert space $\mathcal{H}$ with orthonormal basis $\{e_t, t \in T\}$ is separable if T is finite or countably infinite.

Theorem 5.2.7
If $\mathcal{H}$ is a separable Hilbert space with orthonormal basis $\{e_t, t = 1, 2, \ldots\}$, then

1. $x = \sum_{i=1}^{\infty} < x, e_i > e_i$ *for all* $x \in \mathcal{H}$, *i.e.*

$$\|x - \sum_{i=1}^{n} < x, e_i > e_i\| \longrightarrow 0 \text{ as } n \longrightarrow \infty$$

2. $< x, y >= \sum_{i=1}^{\infty} < x, e_i >< e_i, y >$ *for all* $x, y \in \mathcal{H}$ *(Parseval's identity) In particular,*

$$\|x\|^2 = \sum_{i=1}^{\infty} | < x, e_i > |^2.$$

Example
We consider the space $\mathcal{L}_2$ of complex random variables X on $(\Omega, \mathcal{B}, P)$, with $\mathbb{E}|X|^2 < \infty$ and

$$< X, Y >= \mathbb{E}[X\overline{Y}].$$

Obviously, $< X, X >= 0 \Leftrightarrow \mathbb{E}|X|^2 = 0$ does not necessarily imply $X = 0$. Hence, $\mathcal{L}_2$ is not an inner product space.
Let $\mathcal{L}^2_{\mathbb{C}}(\Omega, \mathcal{A}, F)$ denote the space of complex measurable functions h such that

$$\int_{\Omega} |h(\omega)|^2 F(d\omega) < \infty,$$

where F is a finite, non degenerated measure on $(\Omega, \mathcal{A})$.
We consider the space $L^2_{\mathbb{C}}(\Omega, \mathcal{A}, F)$ of equivalence classes of $\mathcal{L}^2_{\mathbb{C}}(\Omega, \mathcal{A}, F)$ where two functions $f, g \in \mathcal{L}^2_{\mathbb{C}}$ are equivalent if and only if

$$\int_{\Omega} |(f - g)(\omega)|^2 F(d\omega) = 0.$$

Then $L^2_{\mathbb{C}}(\Omega, \mathcal{A}, F)$ equipped with

$$< h, g >= \int_{\Omega} h(\omega)\overline{g(\omega)} F(d\omega)$$

is a Hilbert space.
In particular, consider $L^2_{\mathbb{C}}([-\pi, \pi], \mathcal{B}, U)$ where $\mathcal{B}$ is the Borel σ-algebra on $[-\pi, \pi]$ and U the uniform probability measure on $([-\pi, \pi], \mathcal{B})$, such that

$$< h, g >= \frac{1}{2\pi} \int_{-\pi}^{\pi} h(\omega)\overline{g(\omega)} d\omega.$$

The set $\{e_n = e^{in\omega} : n \in \mathbb{Z}\}$ is an orthonormal basis of $L^2_{\mathbb{C}}([-\pi, \pi], \mathcal{B}, U)$. Hence, for every $f \in L^2_{\mathbb{C}}([-\pi, \pi], \mathcal{B}, U)$,

$$f = \sum_{j=-\infty}^{\infty} < f, e_j > e_j$$

where the $< f, e_j >$ are the Fourier coefficients of f.
In the following, we will assign the scaling factor $(2\pi)^{-1}$ to the scalar product and write $L^2_{\mathbb{C}}([-\pi, \pi], \mathcal{B}, L)$ instead of $L^2_{\mathbb{C}}([-\pi, \pi], \mathcal{B}, U)$, where L denotes the Lebesgue measure.

Chapter 6

Linear Models for Stationary Time Series

6.1 Stationary Stochastic Processes

Definition 6.1.1
A stochastic process is a family of random variables $\{X_t : t \in \mathbb{T}\}$ defined on the same probability space $(\Omega, \mathcal{A}, P)$.

Definition 6.1.2
A real valued stochastic process $\{X_t : t \in \mathbb{Z}\}$ is said to be strictly stationary if

$$\mathcal{L}(X_{t+t_1}, \ldots, X_{t+t_n}) = \mathcal{L}(X_{t_1}, \ldots, X_{t_n})$$

for all $n \geq 1, t_1, \ldots, t_n, t \in \mathbb{Z}$.

Remark 6.1.3

1. In a strictly stationary stochastic process, X_t has the same distribution for every $t \in \mathbb{Z}$.
2. Strictly stationary means that the distribution of the stochastic process is invariant under a time shift of the origin.

In practice we have only a finite number of observations (data) from the stochastic process $\{X_t : t \in \mathbb{Z}\}$: $x_1, \ldots, x_n$. Therefore, it is impossible to study the whole distribution of $\{X_t : t \in \mathbb{Z}\}$ from the data. Usually, one will rely on the following popular notion.

Definition 6.1.4
A real valued stochastic process $\{X_t : t \in \mathbb{Z}\}$ is said to be (weakly) stationary if

1. $\mathrm{E}[X_t^2] < \infty \quad$ for all $t \in \mathbb{Z}$
2. $\mathrm{E}[X_t] = \mu \quad$ for all $t \in \mathbb{Z}$
3. $\mathrm{cov}(X_s, X_{s+t}) = r_t \quad$ for all $s, t \in \mathbb{Z}$.

$\{r_t, t \in \mathbb{Z}\}$ is called the autocovariance function (acf) and $\{\rho_t = \frac{r_t}{r_0}, t \in \mathbb{Z}\}$ is the autocorrelation function.

We will usually assume that $\mu = 0$.

Proposition 6.1.5 (Properties of the acf)

1. $r_0 \geq 0$.
2. $|r_t| \leq r_0 \quad$ *for all* $t \in \mathbb{Z}$.
3. $r_t = r_{-t} \quad$ *for all* $t \in \mathbb{Z}$.

Proof:
$r_0 = \text{var}X_t$ for all $t \in \mathbb{Z}$.

2. With $\langle X_s, X_t \rangle := \text{E}(X_s\, X_t) (= \text{cov}(X_s, X_t)$ if $\mu = 0), L^2(\Omega, \mathcal{A}, \mathcal{P})$ is an inner product space.
Using the Cauchy-Schwarz inequality $|\langle x, y \rangle| \leq ||x|| \cdot ||y||$, one derives

$$\begin{aligned} |r_t| &= |\text{E}(X_s - \text{E}X_s)(X_{s+t} - \text{E}X_{s+t})| \\ &\leq \sqrt{\text{E}(X_s - \text{E}X_s)^2\, \text{E}(X_{s+t} - \text{E}X_{s+t})^2} = \text{var}X_s = r_0. \end{aligned}$$

3. $r_t = \text{cov}(X_s, X_{s+t}) = \text{cov}(X_{s-t}, X_s) = \text{cov}(X_s, X_{s-t}) = r_{-t}$

Remark 6.1.6

- $\{X_t, t \in \mathbb{Z}\}$ strictly stationary $\not\Rightarrow$ $\{X_t, t \in \mathbb{Z}\}$ (weakly) stationary.
Indeed, if we consider $\{X_t, t \in \mathbb{Z}\}$ i.i.d. Cauchy, strict stationarity is fulfilled. However, $\text{E}X_t$ does not exist and it follows that $\{X_t, t \in \mathbb{Z}\}$ cannot be weakly stationary.
- $\{X_t, t \in \mathbb{Z}\}$ strictly stationary and $\text{E}X_t^2 < \infty \Rightarrow \{X_t, t \in \mathbb{Z}\}$ (weakly) stationary.
- $\{X_t, t \in \mathbb{Z}\}$ sequence of independent random variables

$$\mathcal{L}(X_t) = \begin{cases} Exp(1) & \text{if } t = 2k+1 \\ \mathcal{N}(1,1) & \text{if } t = 2k \end{cases}, k \in \mathbb{Z}.$$

  $\{X_t, t \in \mathbb{Z}\}$ is weakly stationary but not strictly stationary.
- Consider X_t a weakly stationary time series. Then
  - $Y_t = t\mu + X_t$ is non stationary,
  - $Z_t = Y_t - Y_{t-1}$ is stationary.

Definition 6.1.7
A stochastic process $\{X_t, t \in \mathbb{Z}\}$ is said to be Gaussian if $(X_{t_1}, \ldots, X_{t_n})$ has an n-variate Gaussian distribution for all $n \geq 1, t_1, \ldots, t_n \in \mathbb{Z}$

In this case the distribution is fully described by the mean and the covariance matrix.

Proposition 6.1.8
If $\{X_t, t \in \mathbb{Z}\}$ *is a weakly stationary Gaussian process with mean* μ *and acf* $\{r_t, t \in \mathbb{Z}\}$ *then*

1. $\mathcal{L}(X_{\tau+1}, \ldots, X_{\tau+n}) = \mathcal{N}_n((\mu, \ldots, \mu)^T, R_n)$ *for all* $\tau \in \mathbb{Z}, n \geq 1$ *where* $R_n = (r_{t-s})_{1 \leq s,t \leq n}$ *is the autocovariance matrix.*

2. $\{X_t, t \in \mathbb{Z}\}$ *is strictly stationary.*

Proof:

1. For all $\tau \in \mathbb{Z}, s, t \in \{1, \ldots, n\}, \operatorname{cov}(X_{\tau+t}, X_{\tau+s}) = \operatorname{cov}(X_{t-s}, X_0) = r_{t-s}$

2. Follows from 1.

Proposition 6.1.9
Consider $\{r_t, t \in \mathbb{Z}\}$ *the acf of a weakly stationary stochastic process and* $R_n, n \geq 1$, *its autocovariance matrix.*

1. $\{r_t, t \in \mathbb{Z}\}$ *is positive semidefinite, i.e. for all* $n \geq 1, z_1, \ldots, z_n \in \mathbb{C}, \sum_{s,t=1}^{n} \overline{z}_t r_{t-s} z_s \geq 0$

2. R_n *is a positive semidefinite, symmetric Toeplitz*[1] *matrix*

Proof: 1.

$$\begin{aligned}\sum_{s,t=1}^{n} z_s \bar{z}_t r_{t-s} &= \mathrm{E}\left[\sum_{s,t=1}^{n} z_s \bar{z}_t \,(X_t - \mu)(X_s - \mu)\right] = \mathrm{E}\left[\left(\sum_{t=1}^{n} \bar{z}_t (X_t - \mu)\right)\left(\sum_{s=1}^{n} z_s (X_s - \mu)\right)\right] \\ &= \mathrm{E}\left[\overline{\sum_{t=1}^{n} z_t (X_t - \mu)} \sum_{s=1}^{n} z_s (X_s - \mu)\right] = \mathrm{E}\left|\sum_{t=1}^{n} z_t (X_t - \mu)\right|^2 \geq 0.\end{aligned}$$

2. Because $(R_n)_{st} = r_{t-s} = r_{s-t} = (R_n)_{ts}$, R_n is symmetric. By definition,

$$R_n = \begin{pmatrix} r_0 & r_1 & r_2 & \ldots & r_{n-1} \\ r_1 & r_0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & r_2 \\ \vdots & \ddots & \ddots & \ddots & r_1 \\ r_{n-1} & \ldots & \ldots & r_1 & r_0 \end{pmatrix},$$

i.e. R_n is a symmetric Toeplitz-Matrix. Finally, R_n is positive semidefinite which follows from 1.

□

Theorem 6.1.10
A sequence $\{r_t, t \in \mathbb{Z}\}$ *is an acf of a weakly stationary stochastic process if and only if it is symmetric and positive semidefinite.*

Proof:

- $\Rightarrow$ Follows from Proposition 6.1.9.
- $\Leftarrow$ see Brockwell and Davis, Theorem 1.5.1., p.27.

□

[1] A matrix in which each diagonal descending from left to right is constant.

Additionally to the above theorem we have Herglotz's Theorem which provides a criterion for positive semidefiniteness.

Theorem 6.1.11 (Herglotz)
A real valued symmetric sequence $\{r_t, t \in \mathbb{Z}\}$ *is positive semidefinite if and only if there exists a right continuous, non decreasing, bounded function* F *on* $[-\pi, \pi]$, *with* $F(-\pi) = 0$ *such that*

$$r_t = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{it\omega} F(d\omega) \quad for \quad t \in \mathbb{Z},$$

i.e. $\{r_t, t \in \mathbb{Z}\}$ *are the Fourier-Stieltjes coefficients of* F.

Proof:
See Brockwell and Davis, Theorem 4.3.1, p. 117-119

Remark 6.1.12
F defines a finite measure $\tilde{F}$ on $([-\pi, \pi], \mathcal{B})$ via

$$\int_{-\pi}^{\lambda} \tilde{F}(d\omega) = \int_{-\pi}^{\pi} 1_{[-\pi,\lambda]}(\omega)\tilde{F}(d\omega) = \tilde{F}([-\pi, \lambda]) = F(\lambda) - F(-\pi) = F(\lambda).$$

In the following, we will often identify the function F and the measure $\tilde{F}$.

Definition 6.1.13
If $\{r_t, t \in \mathbb{Z}\}$ is the acf of the stationary stochastic process $\{X_t, t \in \mathbb{Z}\}$, the finite measure defined above is also called the spectral measure of r_t. If

$$F(\lambda) = F(\lambda) - F(-\pi) = \int_{-\pi}^{\lambda} f(\omega)d\omega, -\pi \leq \lambda \leq \pi,$$

i.e.

$$r_t = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{it\omega} f(\omega)d\omega \quad \text{for} \quad t \in \mathbb{Z},$$

then f is called a spectral density of r_t.

Remark 6.1.14
A spectral density has the following properties:

- $f(\omega) \geq 0$ for all $\omega \in [-\pi, \pi]$
- $f(\omega) = f(-\omega)$ for all $\omega \in [-\pi, \pi]$
- f is integrable
- $F(\pi) = \int_{-\pi}^{\pi} f(\omega)d\omega = 2\pi r_0$

- If $\sum_{t=-\infty}^{\infty} |r_t| < \infty$, then there exists a spectral density given by

$$f(\omega) = \sum_{t=-\infty}^{\infty} r_t e^{-it\omega}.$$

Indeed, by Fourier theory one has

$$f(\omega) = \sum_{t=-\infty}^{\infty} < f, e^{it\cdot} > e^{it\omega} = \sum_{t=-\infty}^{\infty} r_{-t} e^{it\omega} = \sum_{t=-\infty}^{\infty} r_t e^{-it\omega}.$$

6.2 Linear Processes

Definition 6.2.1
A sequence $\{\varepsilon_t, t \in \mathbb{Z}\}$ of random variables is called

- *strict white noise* if ε_t i.i.d., $\mathrm{E}[\varepsilon_t] = 0$ and $\operatorname{var} \varepsilon_t = \sigma_\varepsilon^2 < \infty$.
- *white noise* if the ε_t are uncorrelated, $\mathrm{E}[\varepsilon_t] = 0$ and $\operatorname{var} \varepsilon_t = \sigma_\varepsilon^2 < \infty$.

Definition 6.2.2
A stationary process $\{X_t, t \in \mathbb{Z}\}$ of the form

$$X_t = \sum_{k=-\infty}^{\infty} b_k \varepsilon_{t-k}, t \in \mathbb{Z}$$

with $\sum_{k=-\infty}^{\infty} b_k^2 < \infty$ is called

1. a *linear process* if ε_t is strict white noise.
2. a *generalized linear process* if ε_t is white noise.

Remark 6.2.3
Is $\sum_{k=-\infty}^{\infty} b_k \varepsilon_{t-k}$ well defined?
Define

$$X_{t,n} = \sum_{k=-n}^{n} b_k \varepsilon_{t-k},$$

for $m \geq n$

$$X_{t,m} = \sum_{k=-m}^{-n-1} b_k \varepsilon_{t-k} + \sum_{k=n+1}^{m} b_k \varepsilon_{t-k} + X_{t,n}.$$

Then,

$$\|X_{t,n} - X_{t,m}\|^2 = \left\langle \sum_{|k|=n+1}^{m} b_k \varepsilon_{t-k}, \sum_{|j|=n+1}^{m} b_j \varepsilon_{t-j} \right\rangle = \sigma_\varepsilon^2 \sum_{|k|=n+1}^{m} b_k^2$$

which implies that $X_{t,n}$ is a Cauchy sequence in the Hilbert space $L^2(\Omega, \mathcal{A}, P)$. Hence, there exists an $X_t \in L^2(\Omega, \mathcal{A}, P)$ with

$$\|X_{t,n} - X_t\|^2 = \mathrm{E}(X_{t,n} - X_t)^2 \underset{n\to\infty}{\to} 0,$$

i.e. $X_{t,n} \xrightarrow{L_2} X_t (n \to \infty)$.

Theorem 6.2.4
Consider $\{X_t, t \in \mathbb{Z}\}$ a generalized linear process. Then the acf is given by

$$r_t = \sigma_\varepsilon^2 \sum_{j=-\infty}^{\infty} b_j b_{j+t}, \text{ for all } t \in \mathbb{Z}$$

and the spectral density by

$$f(\omega) = \sigma_\varepsilon^2 |b(\omega)|^2$$

with $b(\omega) = \sum\limits_{j=-\infty}^{\infty} b_j e^{-ij\omega}$.

Proof:
First, we note that $E(X_s) = 0$. Hence, by using the continuity of the inner product,

$$\begin{aligned} r_t &= < X_s, X_{s+t} >= \lim_{n\to\infty} < X_{s,n}, X_{s+t,n} > \\ &= \lim_{n\to\infty} \left\langle \sum_{|j|\le n} b_j \varepsilon_{s-j}, \sum_{|k|\le n} b_k \varepsilon_{s+t-k} \right\rangle = \lim_{n\to\infty} \sum_{|j|,|k|\le n} b_k b_j < \varepsilon_{s-j}, \varepsilon_{s+t-k} > \\ &= \lim_{n\to\infty} \sum_{-n\le j, j+t\le n} b_j b_{j+t} \sigma_\varepsilon^2 = \sigma_\varepsilon^2 \sum_{j=-\infty}^{\infty} b_j b_{j+t}, \end{aligned}$$

since $< \varepsilon_{s-j}, \varepsilon_{s+t-k} >= \begin{cases} \sigma_\varepsilon^2 & k = j+t \\ 0 & \text{otherwise.} \end{cases}$
Since r_t does not depend on s this also shows the weak stationarity of X_t.
For the second part of the theorem, it suffices to prove that $r_t = \frac{1}{2\pi} \int e^{it\omega} f(\omega)\, d\omega$. First, we need to prove that $b(\omega)$ is well defined.
Consider $b_n(\omega) = \sum\limits_{j=-n}^{n} b_j e^{-ij\omega}$. For $m > n$

$$\| b_n(\cdot) - b_m(\cdot) \|^2 = \left\langle \sum_{|j|=n+1}^{m} b_j e^{-ij\omega}, \sum_{|k|=n+1}^{m} b_k e^{-ik\omega} \right\rangle = \sum_{|j|=n+1}^{m} b_j^2.$$

In particular $||b_n(\cdot) - b_{n+1}(\cdot)||^2 = b_{n+1}^2$. By Brockwell & Davis, Proposition 2.10.1, $b_n(\omega)$ converges to $b(\omega) := \sum_{j=-\infty}^{\infty} b_j e^{-ij\omega}$ for almost all $\omega \in \Omega$.

Finally,

$$\begin{aligned}
\frac{1}{2\pi}\int e^{it\omega} f(\omega)\, d\omega &= \frac{\sigma_\varepsilon^2}{2\pi}\int e^{it\omega}|b(\omega)|^2\, d\omega \\
&= \frac{\sigma_\varepsilon^2}{2\pi}\lim_{n\to\infty}\int e^{it\omega}|b_n(\omega)|^2\, d\omega \\
&= \frac{\sigma_\varepsilon^2}{2\pi}\lim_{n\to\infty}\int e^{it\omega}\sum_{|j|,|k|\le n} b_j b_k e^{-ij\omega}e^{ik\omega}\, d\omega \\
&= \sigma_\varepsilon^2 \lim_{n\to\infty}\sum_{|j|,|k|\le n} b_j b_k \langle e^{-ij\omega}, e^{-i(k+t)\omega}\rangle \\
&= \sigma_\varepsilon^2 \lim_{n\to\infty}\sum_{-n\le k,k+t\le n} b_k b_{k+t} = r_t
\end{aligned}$$

□

6.3 Estimators for the mean and the autocovariances

In this section, the general assumptions are: $\{X_t, t\in\mathbb{Z}\}$ stationary with $\mathrm{E}[X_t]=\mu$, $\mathrm{cov}(X_t,X_s)=r_{t-s}$ and $f(\omega)$ a spectral density.

Estimator for the mean

Motivation: for X_t i.i.d. random variables with $\mathrm{E}|X_1|<\infty$ by the strong law of large numbers (SLLN)

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu.$$

Does $\bar{X}_n$ remain a reasonable estimator of μ if the X_ts are dependent?

Proposition 6.3.1

1. *$E[\bar{X}_n]=\mu$, i.e. $\bar{X}_n$ is an unbiased estimator of μ.*
2. *$var\bar{X}_n = \frac{1}{n}\sum\limits_{t=-(n-1)}^{n-1}(1-\frac{|t|}{n})r_t$*
3. *If $\sum\limits_{t=-\infty}^{\infty}|r_t|<\infty$, then $nvar\bar{X}_n \longrightarrow\limits_{n\to\infty}\sum\limits_{t=-\infty}^{\infty} r_t = f(0)$.*

Proof:

1. $\mathrm{E}\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n}\mathrm{E}X_t = \mu$
2. $$\begin{aligned}
\mathrm{var}\bar{X}_n &= \frac{1}{n^2}\mathrm{var}(X_1+\ldots+X_n) = \frac{1}{n^2}\sum_{s,t=1}^{n}\mathrm{cov}(X_t,X_s) \\
&= \frac{1}{n^2}\sum_{l=-(n-1)}^{n-1}\sum_{1\le s,t\le n, s-t=l} r_l = \frac{1}{n}\sum_{t=-(n-1)}^{n-1}(1-\frac{|t|}{n})r_t
\end{aligned}$$

3.

$$n\text{var}\bar{X}_n = \sum_{t=-(n-1)}^{n-1}(1-\frac{|t|}{n})r_t \leq \sum_{t=-(n-1)}^{n-1}(1-\frac{|t|}{n})|r_t| =: g_n.$$

Obviously, $g_n \leq g_{n+1} \leq \sum_{t=-\infty}^{\infty}|r_t| < \infty$. Hence, g_n is convergent such that $\sum_{t=-(n-1)}^{n-1}(1-\frac{|t|}{n})r_t$ is also convergent. Finally, using the dominated convergence theorem for series it follows

$$\lim_{n\to\infty} n\text{var}\bar{X}_n = \sum_{t=-\infty}^{\infty}\lim_{n\to\infty}\left(1-\frac{|t|}{n}\right)r_t = \sum_{t=-\infty}^{\infty} r_t$$

Since $\sum_{t=-\infty}^{\infty}|r_t| < \infty$, a spectral density exists and $\sum_{t=-\infty}^{\infty} r_t = f(0)$ (see Remark 6.1.14).

□

Remark 6.3.2

- By 1., $\bar{X}_n$ is unbiased. If $\sum_{t=-\infty}^{\infty}|r_t| < \infty$ 3. implies that $\text{var}\bar{X}_n$ converges to zero. In this case, $\bar{X}_n$ is a consistent estimator of μ.
- If we assume that $f(\omega)$ is continuous in $(-\delta,\delta)$ for a $\delta > 0$ then one can also show that $n\text{var}\bar{X}_n \underset{n\to\infty}{\longrightarrow} f(0)$. The assumption is weaker than the one used above but the result requires a more involved proof.

Applications

1. If $\{X_t, t \in \mathbb{Z}\}$ is a stationary Gaussian process, then

$$\mathcal{L}\left(\sqrt{n}(\bar{X}_n - \mu)\right) = \mathcal{N}\left(0, \sum_{|t|<n}(1-\frac{|t|}{n})r_t\right).$$

2. If $\{X_t, t \in \mathbb{Z}\}$ is stationary and linear but not Gaussian, the asymptotic distribution of $\bar{X}_n$ is still computable as presented in the following proposition.

Proposition 6.3.3 (Central Limit Theorem for Linear Processes)
Let $X_t = \mu + \sum_{k=-\infty}^{\infty} b_k\varepsilon_{t-k}$ *be a linear process with mean* μ. *If* $\sum_{k=-\infty}^{\infty} b_k \neq 0$ *and* $\sum_{k=-\infty}^{\infty}|b_k| < \infty$ *then*

$$\sqrt{n}(\bar{X}_n - \mu) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, f(0))$$

with $f(0) = \sigma_\varepsilon^2\left(\sum_{k=-\infty}^{\infty} b_k\right)^2$.

Estimators for autocovariances

The goal in this section is to estimate $r_t, t = 0, 1, \ldots, n-1$ given the data $X_1, \ldots, X_n$ with unknown mean $\mu = \mathrm{E}X_t$.

Consider $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. with $\mathrm{var}X_i < \infty, \mathrm{var}Y_i < \infty$. A standard estimator for $\mathrm{cov}(X_i, Y_i)$ is the sample covariance

$$\hat{\mathrm{cov}}(X_i, Y_i) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)(Y_j - \bar{Y}_n).$$

For the time series data $X_1, \ldots, X_n$ design the pairs $(X_1, X_{t+1}), \ldots, (X_{n-t}, X_n)$ and estimate r_t by:

$$\tilde{r}_t = \frac{1}{n-t}\sum_{k=1}^{n-t}(X_k - \bar{X}_n)(X_{t+k} - \bar{X}_n) \text{ and } \tilde{r}_t = \tilde{r}_{-t} \text{ for } t < 0.$$

or

$$\hat{r}_t = \frac{1}{n}\sum_{k=1}^{n-t}(X_k - \bar{X}_n)(X_{t+k} - \bar{X}_n) \text{ and } \hat{r}_t = \hat{r}_{-t} \text{ for } t < 0.$$

Which of these estimators performs better?

1. There is no general proof that one is superior.
2. If μ was known, then $\mathrm{E}[\tilde{r}_t] = r_t$ but $\hat{r}_t$ is not unbiased. However, $\mathrm{E}\hat{r}_t \underset{n\to\infty}{\longrightarrow} r_t$.
3. $\mathrm{mse}(\tilde{r}_t)$ and $\mathrm{mse}(\hat{r}_t)$ are of comparable size.
4. The decisive argument for $\hat{r}_t$: $\hat{r}_0, \ldots, \hat{r}_{n-1}$ is always positive semi-definite, but $\tilde{r}_0, \ldots, \tilde{r}_{n-1}$ is not.

Indeed,

$$\hat{R}_n = \begin{pmatrix} \hat{r}_0 & \hat{r}_1 & \hat{r}_2 & \ldots & \hat{r}_{n-1} \\ \hat{r}_1 & \hat{r}_0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \hat{r}_2 \\ \vdots & \ddots & \ddots & \ddots & \hat{r}_1 \\ \hat{r}_{n-1} & \ldots & \ldots & \hat{r}_1 & \hat{r}_0 \end{pmatrix}$$

can be rewritten as $\hat{R}_n = \frac{1}{n}TT^T$ with an $n \times 2n$-matrix

$$T = \begin{pmatrix} 0 & \ldots & 0 & 0 & Y_1 & Y_2 & \ldots & Y_n \\ 0 & \ldots & 0 & Y_1 & Y_2 & \ldots & Y_n & 0 \\ \vdots & & Y_1 & Y_2 & \ldots & Y_n & 0 & \vdots \\ \vdots & & & & & & & \vdots \\ 0 & Y_1 & Y_2 & \ldots & Y_n & 0 & \ldots & 0 \end{pmatrix},$$

with $Y_i = X_i - \overline{X}_n$. For $z \in \mathbb{R}^n$ we get

$$z^T\hat{R}_n z = \frac{1}{n}(T^T z)^T(T^T z) \geq 0.$$

Theorem 6.3.4 (Asymptotic normality of the empiric autocovariance function)
Let

$$X_t = \mu + \sum_{k=-\infty}^{\infty} b_k \varepsilon_{t-k}$$

a linear process with $E[\varepsilon_t^4] < \infty$ *and* $\sum_{k=-\infty}^{\infty} |b_k| < \infty$. *Define* $\hat{c}_t = \sqrt{n}(\hat{r}_t - r_t)$. *Then for all* $m \geq 1, t_1, \ldots, t_m \geq 0$*:*

$$(\hat{c}_{t_1}, \ldots, \hat{c}_{t_m}) \xrightarrow{\mathcal{L}} \mathcal{N}_m(0, \Sigma)$$

with

$$\Sigma_{t,s} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f^2(\omega) \left\{ e^{i(t-s)\omega} + e^{i(t+s)\omega} \right\} d\omega + \frac{\kappa_\varepsilon}{\sigma_\varepsilon^2} r_t r_s, \text{ with } \kappa_\varepsilon = E[\varepsilon_t^4] - 3\sigma_\varepsilon^4$$

Proof:
Brockwell, Davis, Chapter 7.

Remark 6.3.5

1. The asymptotic results given in this section are useful for the construction of confidence intervals or test statistics, e.g. for

$$H_0 : \mu = 0 \text{ against } H_1 : \mu \neq 0$$

2. If $\mathcal{L}(\varepsilon_t) = \mathcal{N}(0, \sigma_\varepsilon^2)$, then $\kappa_\varepsilon = 0$.
3. $\int_{-\pi}^{\pi} f^2(\omega) \left\{ e^{i(t-s)\omega} + e^{i(t+s)\omega} \right\} d\omega$ exists because $f(\omega) = \sigma_\varepsilon^2 |b(\omega)|^2$ is bounded.

Finally, we are interested in estimating the spectral density.

Definition 6.3.6
Consider $\mathbb{C}^n$ with scalar product $\langle u, v \rangle = \sum_{i=1}^{n} u_i \bar{v}_i$.
The frequencies $\omega_j = \frac{2\pi j}{n}$, $-\pi < \omega_j \leq \pi$, are called Fourier frequencies. Let

$$F_n = \{ j \in \mathbb{Z} : -\pi < \omega_j \leq \pi \}.$$

Then the vectors

$$e_j = \frac{1}{\sqrt{n}} (e^{i\omega_j}, e^{i2\omega_j}, \ldots, e^{in\omega_j})^T, \; j \in F_n,$$

form an orthonormal basis of $\mathbb{C}^n$.
Hence, for any $x \in \mathbb{C}^n$

$$x = \sum_{j \in F_n} a_j e_j,$$

where $a_j = \langle x, e_j \rangle = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t e^{-it\omega_j}$.

Definition 6.3.7

1. The discrete Fourier transform of $x \in \mathbb{C}^n$ is the sequence $\{a_j : j \in F_n\}$.

2.
$$I_n(\omega_j) := |a_j|^2 = |\langle x, e_j\rangle|^2 = \frac{1}{n}\left|\sum_{t=1}^{n} x_t e^{-it\omega_j}\right|^2$$
is called the periodogram of x at ω_j.

Note that by Parseval's identity
$$||x||^2 = \sum_{j\in F_n} |\langle x, e_j\rangle|^2 = \sum_{j\in F_n} I_n(\omega_j).$$

Proposition 6.3.8
Let $\{X_t : t \in \mathbb{Z}\}$ be a stationary time series. If $\omega_k = \frac{2\pi k}{n} \neq 0$ then
$$I_n(\omega_k) = \sum_{t=-(n-1)}^{n-1} \hat{r}_t e^{-it\omega_k}. \tag{6.1}$$

Proof:

$$\begin{aligned} I_n(\omega_k) &= \frac{1}{n}\sum_{t=1}^{n} X_t e^{-it\omega_k} \sum_{s=1}^{n} \bar{X}_s e^{is\omega_k} \\ &= \frac{1}{n}\sum_{t=1}^{n}(X_t - \bar{X_n})e^{-it\omega_k} \sum_{s=1}^{n}(\bar{X}_s - \overline{\bar{X}_n})e^{is\omega_k} = \frac{1}{n}\sum_{t,s=1}^{n}(X_t - \bar{X_n})(\bar{X}_s - \overline{\bar{X}_n})e^{i(s-t)\omega_k} \\ &= \sum_{t=-(n-1)}^{n-1} \hat{r}_t e^{-it\omega_k}. \end{aligned}$$

□

Remark 6.3.9

1. Note the resemblance between (6.1) and the formula $f(\omega_k) = \sum_{s=-\infty}^{\infty} r_s e^{-is\omega_k}$ for the spectral density. Hence, $\hat{f}(\omega_k) = I_n(\omega_k)$ is a natural estimator for $f(\omega_k)$. Furthermore, the periodogram can be extended to any value $\omega \in [-\pi, \pi]$. We have
$$E[\hat{f}(\omega)] = E[I_n(\omega)] \longrightarrow_{n\to\infty} f(\omega),$$
i.e. $\hat{f}(\omega)$ is asymptotically unbiased.

2. $I_n(\omega)$ is in general not consistent. Indeed, there is a $c > 0$ such that $\mathrm{var}(I_n(\omega)) \geq c$ for all n.

6.4 Autoregressive Processes

Definition 6.4.1
A stochastic process $\{X_t, t \in \mathbb{Z}\}$ is called autoregressive of order $p \geq 1$ (AR(p)) if
$$X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t, \quad t \in \mathbb{Z}, \alpha_p \neq 0, \tag{6.2}$$
$\{\varepsilon_t, t \in \mathbb{Z}\}$ white noise.

Question:

- Does a stationary solution of (6.2) exist? If yes is it unique?
- What happens if the starting distribution is predefined, e.g. $X_0 = x_0, \ldots, X_{1-p} = x_{1-p}$?

Theorem 6.4.2
Consider $\{X_t, t \in \mathbb{Z}\}$ *an AR(1)-process, i.e.*

$$X_t = \alpha X_{t-1} + \varepsilon_t \text{ with } |\alpha| < 1.$$

1. *The unique stationary solution is given by* $\eta_t = \sum_{k=0}^{\infty} \alpha^k \varepsilon_{t-k}$.
2. *With some predefined distribution for* X_0, *we get*

$$X_t = \eta_t + \xi_t, t \geq 1,$$

where $\xi_t = \alpha^t \xi_0$ *and* $\xi_0 = X_0 - \eta_0$.

Proof:

1. For any solution η_t we have

$$\begin{aligned} \eta_t &= \alpha\eta_{t-1} + \varepsilon_t \\ &= \alpha^2\eta_{t-2} + \alpha\varepsilon_{t-1} + \varepsilon_t \\ &= \varepsilon_t + \alpha\varepsilon_{t-1} + \ldots + \alpha^k\varepsilon_{t-k} + \alpha^{k+1}\eta_{t-k-1} \end{aligned} \tag{6.3}$$

   Now assume that η_t is stationary. Then $\|\eta_t\|^2 = \mathrm{E}[\eta_t^2]$ is constant. Hence,

$$\|\eta_t - \sum_{j=0}^{k} \alpha^j \varepsilon_{t-j}\|^2 = \alpha^{2k+2}\|\eta_{t-k-1}\|^2 \underset{k\to\infty}{\longrightarrow} 0.$$

   Since $\sum_{j=0}^{k} \alpha^j \varepsilon_{t-j}$ is a Cauchy sequence(w.r.t. k), that converges in L_2 one can conclude $\eta_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}$.

2. Obviously, $X_0 = \eta_0 + (X_0 - \eta_0) = \eta_0 + \xi_0$. By induction, we get $X_t = \alpha(\eta_{t-1} + \xi_{t-1}) + \varepsilon_t = \eta_t + \xi_t$.

□

Application:
Simulation of an $AR(1)$-process:

i) Fix X_0' and generate ε_t, $t \geq 1$, i.i.d. (e.g. $\mathcal{N}(0, \sigma_\varepsilon^2)$).

ii) Compute $X_t' = \alpha X_{t-1}' + \varepsilon_t$, $t \geq 1$, recursively.

iii) Discard the first T ($= 200$ or 300) values and set

$$X_t = X'_{t+T}$$

for $t \geq 1$. $X_1, X_2, \ldots$ is asymptotically stationary.

Remark 6.4.3

1. Why does the simulation work?

$$X_t = \eta_t + \xi_t = \eta_t + \alpha^t(X_0 - \eta_0) \to \eta_t \text{ for } t \to \infty.$$

Hence, X_t will forget its starting distribution exponentially fast.

2. If $|\alpha| > 1$

$$X_{t-1} = \frac{1}{\alpha}X_t - \frac{\varepsilon_t}{\alpha} = \frac{1}{\alpha^{k+1}}X_{t+k} - (\frac{1}{\alpha^k}\varepsilon_{t+k} + \ldots + \frac{1}{\alpha}\varepsilon_t)$$

therefore one can derive a unique stationary solution as previously which is given by

$$\eta_t = -\sum_{j=1}^{\infty}\left(\frac{1}{\alpha}\right)^j \varepsilon_{t+j}.$$

Note that in this case X_t is correlated with the ε_s for $s > t$ (future). Hence, this case is often considered unnatural.

3. For $|\alpha| = 1$, no stationary solution exists.

Proposition 6.4.4
If $\{X_t, t \in \mathbb{Z}\}$ is a stationary AR(1)-process with $|\alpha| < 1$ then

1. $EX_t = 0$,
2. $varX_t = \frac{\sigma_\varepsilon^2}{1-\alpha^2}$,
3. $r_t = \sigma_\varepsilon^2 \frac{\alpha^{|t|}}{1-\alpha^2}$

Now, let us focus on the case of higher order autoregressive processes, i.e. $p \geq 1$.

Definition 6.4.5

1. Define the shift operator U on $\mathcal{H} = \overline{sp}\{X_t, t \in \mathbb{Z}\}$ by

$$U(X_t) := X_{t+1}.$$

Then $U^k(X_t) = X_{t+k}, \quad U^{-1}(X_t) = X_{t-1}$.

2. $A(z) = 1 - \alpha_1 z - \ldots - \alpha_p z^p, z \in \mathbb{C}$ is called the generating polynomial of the AR(p)-process $X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t$.

For $p = 1$, we have $A(z) = 1 - \alpha z$. Furthermore,

$$A(U^{-1})X_t = \varepsilon_t.$$

Given $A(z)$, when does a stationary solution of (6.2) exist? The idea is to find a

$$B(z) = \sum_{k=0}^{\infty} \psi_k z^k$$

with $\sum_{k=0}^{\infty} |\psi_k| < \infty$ such that

$$B(U^{-1})A(U^{-1}) = U^0 = I$$

and hence $X_t = B(U^{-1})\varepsilon_t$.

Remark 6.4.6
$B(z) = \sum_{k=0}^{\infty} \psi_k z^k$ with $\sum_{k=0}^{\infty} |\psi_k| < \infty$ does not always exist!

Definition 6.4.7
An AR(p)-process is said to be causal if there exists a sequence of constants $\{\psi_k, k = 0, 1, \ldots\}$ such that

$$\sum_{k=0}^{\infty} |\psi_k| < \infty$$

and

$$X_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}$$

with $\{\varepsilon_t\}$ white noise.

Remark 6.4.8
A causal AR(p)-process is stationary since in that case $X_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}$ with $\sum_{k=0}^{\infty} \psi_k^2 \leq (\sum_{k=0}^{\infty} |\psi_k|)^2 < \infty$ is a generalized linear process.

Theorem 6.4.9

1. *For $p \geq 1$, there exists a causal solution of $A(U^{-1})X_t = \varepsilon_t$ if and only if $A(z)$ has no zeroes in $\{z \in \mathbb{C} : |z| \leq 1\}$. This solution is unique and stationary. The coefficients $\{\psi_k\}$ are determined by the relation*

$$\psi(z) = \sum_{k=0}^{\infty} \psi_k z^k = \frac{1}{A(z)}.$$

   Furthermore, there exist $0 < \rho < 1$ and $c \geq 0$ such that $|\psi_k| \leq c\rho^k$ for all k.

2. *Under the condition of 1., the solution of $A(U^{-1})X_t = \varepsilon_t$ with arbitrary initial distribution is asymptotically stationary and the non stationary part vanishes with an exponential rate.*

Proof:

1. Brockwell, Davis, Theorem 3.1.1.
2. Analogous to 6.4.3.

□

Theorem 6.4.10
Consider $\{X_t, t \in \mathbb{Z}\}$ *a causal (stationary) AR(p)-process with* $\{\varepsilon_t, t \in \mathbb{Z}\}$ *white noise and* $A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^k$. $\{X_t, t \in \mathbb{Z}\}$ *has a spectral density*

$$f(\omega) = \frac{\sigma_\varepsilon^2}{|A(e^{-i\omega})|^2}.$$

Proof:
For a generalized linear process $f(\omega) = \sigma_\varepsilon^2 |b(\omega)|^2$. Since $b(\omega) = \sum_{k=0}^{\infty} \psi_k e^{-i\omega k} = \frac{1}{A(e^{-i\omega})}$ the result follows. □

Parameter Estimation

In this section we work with the general assumption that ε_t is independent of X_s for $s < t$. This is clearly true for causal AR-processes.

Proposition 6.4.11 (Yule-Walker equations)
Consider $\{X_t, t \in \mathbb{Z}\}$ *a causal AR(p)-process with* $r_t = cov(X_t, X_0)$. *Then*

$$\sum_{k=1}^{p} \alpha_k r_{t-k} = r_t, \text{ for } t \geq 1$$

and

$$\sum_{k=1}^{p} \alpha_k r_k = r_0 - \sigma_\varepsilon^2.$$

With $\alpha = (\alpha_1, \ldots, \alpha_p)^T, R_p = (r_{t-k})_{1 \leq t,k \leq p}, r(p) = (r_1, \ldots, r_p)^T$ *we have*

$$R_p \alpha = r(p) \text{ (matrix representation).}$$

Proof:
Clearly, $\mathrm{E}X_t = 0$. Hence,

$$\begin{aligned} r_t &= \operatorname{cov}(X_t, X_0) = \mathrm{E}[X_t X_0] = \mathrm{E}\left[\left(\sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t\right) X_0\right] \\ &= \sum_{k=1}^{p} \alpha_k \mathrm{E}[X_{t-k} X_0] + \mathrm{E}[\varepsilon_t X_0] = \sum_{k=1}^{p} \alpha_k r_{t-k} + \operatorname{cov}(\varepsilon_t, X_0). \end{aligned}$$

Now,

1. if $t \geq 1, \operatorname{cov}(\varepsilon_t, X_0) = \mathrm{E}[\varepsilon_t X_0] = 0$ by the independence assumption.
2. if $t = 0, \operatorname{cov}(\varepsilon_t, X_0) = \mathrm{E}[\varepsilon_0 X_0] = \mathrm{E}\left[\varepsilon_0 \left(\sum_{k=1}^{p} \alpha_k X_{0-k} + \varepsilon_0\right)\right] = \mathrm{E}[\varepsilon_0^2] = \sigma_\varepsilon^2$ also by the independence assumption.

□

Lemma 6.4.12
Let $\{r_t, t \in \mathbb{Z}\}$ be the acf of a stationary process. If $r_0 > 0$ and $r_t \to 0$ as $t \to \infty$, then R_n is positive definite for every n.

Proof:
Brockwell, Davis, Proposition 5.1.1 □

From the above lemma, we can derive

$$\alpha = R_p^{-1} r(p) \text{ and } \sigma_\varepsilon^2 = r_0 - \alpha^T r(p).$$

Definition 6.4.13 (Yule-Walker estimators)
Consider $X_1, \ldots, X_n$ observations from a causal AR(p)-process and $\hat{r}_0, \ldots, \hat{r}_{n-1}$ the sample autocovariances. Then

$$\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)^T = \hat{R}_p^{-1} \hat{r}(p)$$

with $\hat{r}(p) = (\hat{r}_1, \ldots, \hat{r}_p)^T$ and

$$\hat{\sigma}_\varepsilon^2 = \hat{r}_0 - \hat{\alpha}^T \hat{r}(p)$$

are the Yule-Walker estimators of $\alpha_1, \ldots, \alpha_p, \sigma_\varepsilon^2$.

Remark 6.4.14
Since $\hat{r}_t \xrightarrow{p} r_t$ for $n \to \infty$ and $t = 0, 1, 2, \ldots, p$, application of Slutsky's Lemma yields

$$\hat{\alpha} = \hat{R}_p^{-1} \hat{r}(p) \xrightarrow{p} R_p^{-1} r(p) = \alpha,$$

$$\hat{\sigma}_\varepsilon^2 \xrightarrow{p} \sigma_\varepsilon^2 \quad (n \longrightarrow \infty).$$

Hence the Yule-Walker estimators are consistent. It can be shown (see Thm. 6.1.10.) that the sequence
$\ldots, 0, \hat{r}_{-p}, \ldots, \hat{r}_0, \ldots, \hat{r}_p, 0, \ldots$ is the acf of a linear process. Hence, $\hat{R}_p$ is invertible by Lemma 6.4.12.

Alternative:
Due to the similarity of the AR model to linear regression, the parameters of an AR-process can also be estimated by least squares:

$$X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t, \quad t = p+1, \ldots, n$$

and the least squares estimators are defined by

$$\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_p)^T = \arg\min Q(\alpha),$$

$$\tilde{\sigma}_\varepsilon^2 = \frac{1}{n} Q(\tilde{\alpha}) = \frac{1}{n} \min Q(\alpha)$$

with $Q(\alpha) = \sum_{t=p+1}^{n} (X_t - \sum_{k=1}^{p} \alpha_k X_{t-k})^2$.

Theorem 6.4.15
Consider a causal AR(p)-process with strict white noise $\{\varepsilon_t\}$. Then,

1. $\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \sigma_\varepsilon^2 R_p^{-1})$ *as* $n \to \infty$
2. $\sqrt{n}(\tilde{\alpha} - \alpha) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \sigma_\varepsilon^2 R_p^{-1})$ *as* $n \to \infty$

Proof:
(Sketch)

2. As in regression we define
   $Y = (X_{p+1}, \ldots, X_n)^T$, $\alpha = (\alpha_1, \ldots, \alpha_p)^T$, $\varepsilon = (\varepsilon_{p+1}, \ldots, \varepsilon_n)^T$ and

$$X = \begin{pmatrix} X_p & \ldots & X_1 \\ X_{p+1} & \ldots & X_2 \\ \vdots & & \vdots \\ X_{n-1} & \ldots & X_{n-p} \end{pmatrix}.$$

   We then derive

$$\begin{aligned} \tilde{\alpha} - \alpha &= (X^T X)^{-1} X^T Y - \alpha \\ &= (X^T X)^{-1} X^T (X\alpha + \varepsilon) - \alpha \\ &= (X^T X)^{-1} X^T \varepsilon \end{aligned}$$

   For the asymptotic behaviour of $\sqrt{n}(\tilde{\alpha} - \alpha)$ we need to prove, for example using some version of the Central Limit Theorem, that $\frac{1}{\sqrt{n}} X^T \varepsilon \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \sigma_\varepsilon^2 R_p)$ and $n(X^T X)^{-1} \xrightarrow{p} R_p^{-1}$ and use the Slutsky lemma to conclude.

1. Show that $\tilde{\alpha} - \hat{\alpha} \xrightarrow{p} 0$.

□

Proposition 6.4.16
Consider $\{X_t, t \in \mathbb{Z}\}$ a causal linear process

$$X_t = \sum_{k=0}^{\infty} b_k \varepsilon_{t-k}$$

with $|b_k| \leq c\rho^k$ for $0 < \rho < 1$. Then, X_t has an AR(∞)-representation

$$X_t = \sum_{k=1}^{\infty} \alpha_k X_{t-k} + \varepsilon_t.$$

The coefficients α_k are given by

$$A(z) = 1 - \sum_{k=1}^{\infty} \alpha_k z^k = \frac{1}{B(z)}$$

with $B(z) = \sum_{k=0}^{\infty} b_k z^k, |z| \leq 1, z \in \mathbb{C}$. The α_k decrease with an exponential rate.

Linear Forecasting

Goal: Given $X_1, \ldots, X_n$ observations from an AR-process. We would like to predict X_{n+s} for some $s \geq 1$.
Here, we will consider only linear prediction

$$\hat{X}_{n+s} = \sum_{t=1}^{n} \beta_t(s) X_t.$$

To do this, we need to minimize

$$\mathrm{E}(X_{n+s} - \hat{X}_{n+s})^2 = \|X_{n+s} - \hat{X}_{n+s}\|^2$$

with respect to $\beta_1, \ldots, \beta_n$ (mean squared error prediction).
For this, we compute the orthogonal projection of X_{n+s} on $\overline{sp}\{X_1, \ldots, X_n\}$

$$X^*_{n+s} = P_{\overline{sp}\{X_1,\ldots,X_n\}} X_{n+s}$$

which yields the following result for an AR(p)-process.

Proposition 6.4.17
Let $\{X_t, t \in \mathbb{Z}\}$ be an AR(p)-process, i.e. $X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t$. Given the data $X_1, \ldots, X_n, n \geq p$ we have

$$X^*_{n+1} = P_{\overline{sp}\{X_1,\ldots,X_n\}} X_{n+1} = \sum_{k=1}^{p} \alpha_k X_{n+1-k}, \quad n > p,$$

since $P_{\overline{sp}\{X_1,\ldots,X_n\}} \varepsilon_{n+1} = 0.$

$$\begin{aligned} X^*_{n+2} &= P_{\overline{sp}\{X_1,\ldots,X_n\}} \left(\alpha_1 X_{n+1} + \sum_{k=2}^{p} \alpha_k X_{n+2-k} + \varepsilon_{n+2} \right) \\ &= \alpha_1 X^*_{n+1} + \sum_{k=2}^{p} \alpha_k X_{n+2-k}, \\ &\vdots \\ X^*_{n+p} &= \alpha_1 X^*_{n+p-1} + \ldots + \alpha_{p-1} X^*_{n+1} + \alpha_p X_n. \end{aligned}$$

for $s > p$:

$$X^*_{n+s} = \sum_{k=1}^{p} \alpha_k X^*_{n+s-k}.$$

Remark 6.4.18
Consider now $\{X_t, t \in \mathbb{Z}\}$ a stationary process with acf $\{r_t, t \in \mathbb{Z}\}$. For $p \geq 1$ let

$$X_{p+1}^* = P_{\overline{sp}\{X_1,\ldots,X_p\}} X_{p+1} = \sum_{k=1}^{p} \alpha_k(p) X_{p+1-k}$$

be the best linear prediction of X_{p+1} given the last p observations and

$$\sigma_\varepsilon^2(p) = \mathrm{E}(X_{p+1} - X_{p+1}^*)^2.$$

By the definition of X_{p+1}^*,

$$X_{p+1} = X_{p+1}^* + \varepsilon_{p+1}(p)$$

with $\varepsilon_{p+1}(p) \perp \overline{sp}\{X_1, \ldots, X_p\}$. Analogous to the proof of the Yule-Walker equations we obtain the generalized Yule-Walker equations

$$\sum_{k=1}^{p} \alpha_k(p) r_{t-k} = r_t, \quad t = 1, 2, \ldots, p$$

$$\sum_{k=1}^{p} \alpha_k(p) r_k = r_0 - \sigma_\varepsilon^2(p).$$

The solutions $\alpha_1(p), \ldots, \alpha_p(p)$ of the generalized Yule-Walker equations provide the coefficients of the best linear prediction for X_t given $X_{t-1}, \ldots, X_{t-p}$. The AR(p)-process with coefficients $\alpha_1(p), \ldots, \alpha_p(p)$ is called the autoregressive approximation of order p of $\{X_t, t \in \mathbb{Z}\}$.

6.5 Order Selection for Autoregressive Processes

We want to fit an AR(p)-process to the data $X_1, \ldots, X_n$. If p was known, we could simply estimate the model parameters, e.g. by the Yule-Walker equations. However, in general, p is unknown.
Hence, we need to find a way of estimating p and derive the goodness of this estimation. As in regression, a high value of p will lead to an accurate description of the data while from a modelling point of view, a small p is desirable. Indeed, for the autoregressive processes the approximation gets better as p increases since $\ldots \subset AR(p) \subset AR(p+1) \subset \ldots$.

Partial Autocorrelation

Definition 6.5.1
Consider $\{X_t : t \in \mathbb{Z}\}$ a stationary process.

1. $\rho_t = \frac{r_t}{r_0} = corr(X_t, X_0), t \geq 0$ is the autocorrelation function of $\{X_t, t \in \mathbb{Z}\}$.
2. The partial autocorrelation function (pacf) $\pi_t, t = 0, 1, 2, \ldots$ is defined as follows: $\pi_0 = 1, \pi_1 = \rho_1$ and for $t \geq 2$

   $$\pi_t = corr(X_t - P_{\overline{sp}\{X_1,\ldots,X_{t-1}\}} X_t,\ X_0 - P_{\overline{sp}\{X_1,\ldots,X_{t-1}\}} X_0).$$

   It measures the linear dependence of X_0 and X_t after the connecting influence of $X_1, \ldots, X_{t-1}$ is deleted.

Proposition 6.5.2
Consider $\{X_t : t \in \mathbb{Z}\}$ *a causal (hence, stationary) AR(p)-process with coefficients* $\alpha_1, \ldots, \alpha_p$. *Then,*

1. $\pi_t = 0$ *for* $t > p$
2. $\pi_p = \alpha_p$.

Proof:

1. $X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t$ with $\varepsilon_t \perp X_s$ for $t > s$ by causality.
   Hence for $t > p$:

$$\begin{aligned} P_{\overline{sp}\{X_1,\ldots,X_{t-1}\}}(X_t) &= \sum_{k=1}^{p} \alpha_k P_{\overline{sp}\{X_1,\ldots,X_{t-1}\}}(X_{t-k}) + P_{\overline{sp}\{X_1,\ldots,X_{t-1}\}}(\varepsilon_t) \\ &= \sum_{k=1}^{p} \alpha_k X_{t-k}. \end{aligned}$$

   This implies

$$\begin{aligned} \pi_t &= \operatorname{corr}(X_t - P_{\overline{sp}\{X_1,\ldots,X_{t-1}\}}(X_t), X_0 - P_{\overline{sp}\{X_1,\ldots,X_{t-1}\}}(X_0)) \\ &= \operatorname{corr}(\varepsilon_t,\ X_0 - P_{\overline{sp}\{X_1,\ldots,X_{t-1}\}}(X_0)) = 0, \end{aligned}$$

   since $X_0 - P_{\overline{sp}\{X_1,\ldots,X_{t-1}\}}(X_0) \in sp\{\varepsilon_s : s < t\}$.

2. For $t = p$: $X_p = \sum_{k=1}^{p-1} \alpha_k X_{p-k} + \alpha_p X_0 + \varepsilon_p$ yields

$$P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_p) = \sum_{k=1}^{p-1} \alpha_k X_{p-k} + \alpha_p P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_0).$$

   and therefore

$$X_p - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_p) = \alpha_p(X_0 - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_0)) + \varepsilon_p,$$

   hence

$$\operatorname{cov}(X_p - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_p), X_0 - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_0)) = \alpha_p \operatorname{var}(X_0 - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_0)).$$

   It remains to prove that

$$\operatorname{var}(X_p - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_p)) = \operatorname{var}(X_0 - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_0)).$$

   Since $\langle X_p, X_s \rangle = \langle X_{p-s}, X_0 \rangle$, $s = 1, \ldots, p-1$, the geometric relations of X_p and X_0 to $\overline{sp}\{X_1, \ldots, X_{p-1}\}$ are identical. Therefore,

$$||X_p - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_p)||^2 = ||X_0 - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_0)||^2,$$

   which yields

$$\pi_p = \frac{\operatorname{cov}(X_p - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_p), X_0 - P_{\overline{sp}\{X_1,\ldots,X_{p-1}\}}(X_0))}{\operatorname{var}(X_0 - P_{\overline{sp}\{X_1,\ldots,X_{t-1}\}}(X_0))} = \alpha_p. \quad \square$$

Proposition 6.5.3
Let $\{X_t : t \in \mathbb{Z}\}$ *be a stationary process with acf* $\{r_t : t \in \mathbb{Z}\}$. *For* $p \geq 1$ *let* $\alpha(p) = (\alpha_1(p), \ldots, \alpha_p(p))^T$ *be the solution of the generalized Yule-Walker equations of order* p, $R_p\alpha(p) = r(p)$. *Then,* $\pi_p = \alpha_p(p)$. *If* $\{X_t : t \in \mathbb{Z}\}$ *is an AR(*p_0*)-process then* $\pi_{p_0} = \alpha_{p_0}(p_0) = \alpha_{p_0}$.

Hence, we estimate π_p by $\hat{\pi}_p = \hat{\alpha}_p(p)$ with $\hat{\alpha}(p) = \hat{R}_p^{-1}\hat{r}(p)$, i.e. the Yule-Walker estimate under

$$H_0 : \{X_t : t \in \mathbb{Z}\} \text{ is an AR}(p)\text{-process.}$$

The idea is to find the largest p such that $\pi_p \neq 0$.
In general, however, $\hat{\pi}_p \neq 0$, even if $\pi_p = 0$. Therefore, we need to test the significance of $\hat{\pi}_p \neq 0$.

Procedure for model choice:
For $p = 1, 2, 3, \ldots$, test the hypothesis

$$H_0 : \{X_t, t \in \mathbb{Z}\} \text{ is an AR}(p)\text{-process, i.e. } \pi_{p+1} = 0$$

against

$$H_1 : \pi_{p+1} \neq 0$$

at the level α. We use that

$$\hat{Q} = \frac{\sqrt{n-p-1}\hat{\pi}_{p+1}}{\sqrt{1-\hat{\pi}_{p+1}^2}} \xrightarrow{\mathcal{L}} t_{n-p-1}$$

under H_0. Reject the hypothesis if $|\hat{Q}| > C_{1-\frac{\alpha}{2}} = 1 - \frac{\alpha}{2}$-quantile of the t_{n-p-1} distribution.
Alternatively, we may use the following test:
Under H_0: $\sqrt{n}\hat{\alpha}_{p+1} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ in analogy to the CLT for $\hat{\alpha}$. Hence, $\sqrt{n}\hat{\pi}_{p+1}$ is approximately $\mathcal{N}(0,1)$-distributed for large n. This means that we reject H_0 at level α if

$$\sqrt{n}|\hat{\pi}_{p+1}| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Remark 6.5.4 (Graphical method)

- Plot $\hat{\pi}_t$ against $t > 0$.
- Choose p such that $\hat{\pi}_t \approx 0$ for all $t > p$.
- Note: $\hat{\pi}_{p+1}, \hat{\pi}_{p+2}, \ldots$ are approximatively $\mathcal{N}(0, \frac{1}{n})$ distributed for large n.

Automatic Model Selection Criterion: Final Prediction Error (FPE)

Definition 6.5.5
Define

$$FPE(p) := \frac{n+p}{n-p}\hat{\sigma}_\varepsilon^2(p)$$

with $\hat{\sigma}_\varepsilon^2(p) = \hat{r}_0 - \hat{\alpha}(p)^T\hat{r}(p)$ and $\hat{\alpha}(p) = \hat{R}_p^{-1}\hat{r}(p)$. We then choose as estimate of p

$$\hat{p} = \arg\min_{p\geq 1} FPE(p).$$

Motivation: Assume that $\{X_t, t \in \mathbb{Z}\}$ is an AR(p)-process. Then the best linear prediction of X_t given $X_s, s < t$, is

$$\hat{X}_t = \hat{X}_t(\alpha) = \sum_{k=1}^{p} \alpha_k X_{t-k}$$

with

$$\mathrm{E}[(X_t - \hat{X}_t)^2] = \mathrm{E}[\varepsilon_t^2] = \sigma_\varepsilon^2$$

However, we need to estimate $\alpha_1, \ldots, \alpha_p$ therefore

$$\hat{\hat{X}}_t = \hat{X}_t(\hat{\alpha}(p)) = \sum_{k=1}^{p} \hat{\alpha}_k X_{t-k}$$

where $\hat{\alpha}$ is estimated by the Yule-Walker equations. Then

$$\hat{\sigma}_\varepsilon^2(p) = \mathrm{E}[(X_t - \hat{\hat{X}}_t)^2] \geq \mathrm{E}[(X_t - \hat{X}_t)^2]$$

as the estimation of $\hat{\alpha}(p)$ produces an additional error.
For a fixed value of p, the consistency of $\hat{\sigma}_\varepsilon^2(p)$ yields

$$FPE(p) := \frac{n+p}{n-p}\hat{\sigma}_\varepsilon^2(p) \underset{n\to\infty}{\longrightarrow} \sigma_\varepsilon^2.$$

Furthermore,

- if p increases, then $\frac{n+p}{n-p}$ also increases
- if p increases, then $\hat{\sigma}_\varepsilon^2$ decreases.

Therefore, $\frac{n+p}{n-p}$ is a penalty term for large values of p and will prevent the choice of large p. Finally, minimizing the $FPE(p)$ is equivalent to choosing p such that $\mathrm{E}\left[(X_t - \sum_{k=1}^{p} \hat{\alpha}_k X_{t-k})^2\right]$ is as small as possible.

Akaike Information Criterion (AIC)

This method can be used in a more general context than only for AR(p)-processes. Let $M_1 \subset M_2 \subset \ldots \subset M_p$ be a sequence of models where model M_i is known up to i parameters.

Example 6.5.6

$M_q = \{(X_1, \ldots, X_n)^T$ is part of a Gaussian AR(p)-process with $p = q-1, EX_t = 0, \theta = (\alpha, \sigma_\varepsilon^2)^T \in \Theta_q\}$ and

$$\Theta_q = \{(\alpha, \sigma_\varepsilon^2)^T : \alpha_1, \ldots, \alpha_p \text{ fulfill the causality criterion, } 0 < \sigma_\varepsilon^2 < \infty\}.$$

Assume that the distribution P_θ of $(X_1, \ldots, X_n)$ under M_q has the density p_θ. Consider the log-likelihood function

$$l_q(\theta) = \log p_\theta(X_1, \ldots, X_n)$$

and let $\hat{\theta}_q$ be the ML-estimator for θ in model M_q. Increasing q will increase the likelihood $l_q(\hat{\theta}_q)$. Hence, we consider a penalized likelihood.

Definition 6.5.7
Define

$$AIC(q) = -2l_q(\hat{\theta}_q) + 2q$$

and choose $\hat{q} = \arg\min_{q \geq 1} AIC(q)$.

For the Gaussian AR(p)-process we assume that $X_1, \ldots, X_p$ are fixed and consider the log-likelihood function of $(X_{p+1}, \ldots, X_n)$:

$$L_q(\theta | X_{p+1}, \ldots, X_n) = (2\pi\sigma_\varepsilon^2)^{-\frac{n-p}{2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} Q_p(\alpha)\right)$$

with $Q_p(\alpha) = \sum_{t=p+1}^{n} \left(X_t - \sum_{k=1}^{p} \alpha_k X_{t-k}\right)^2$. Then

$$l_q(\theta | X_{p+1}, \ldots, X_n) = -\frac{n-p}{2} \log 2\pi\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2} Q_p(\alpha)$$

For fixed σ_ε^2 maximizing $l_q(\theta)$ is equivalent to minimizing $Q_p(\alpha)$. Hence, the ML-estimator of α is the LS-estimator $\tilde{\alpha} = \text{argmin}_\alpha Q(\alpha)$. One can show that $\tilde{\alpha}$ and the Yule-Walker estimator $\hat{\alpha}(p)$ coincide up to terms of order $\frac{p}{n}$ (see also Theorem 6.4.15) and that $\hat{\sigma}_\varepsilon^2 = \hat{r}_0 - \hat{\alpha}(p)^T \hat{r}(p) \approx \frac{1}{n} Q_p(\hat{\alpha}(p))$. Then

$$\begin{aligned} l_q(\hat{\theta}_q) &\approx l_q(\hat{\alpha}(p), \hat{\sigma}_\varepsilon^2) \\ &= -\frac{n-p}{2} \log 2\pi\hat{\sigma}_\varepsilon^2 - \frac{1}{2\hat{\sigma}_\varepsilon^2} Q_p(\hat{\alpha}(p)) \\ &\approx -\frac{n-p}{2} \log 2\pi - \frac{n-p}{2} \log \hat{\sigma}_\varepsilon^2 - \frac{n}{2} \end{aligned}$$

which implies

$$-2l_q(\hat{\theta}_q) + 2p \approx n \log \hat{\sigma}_\varepsilon^2 + 2p + const(n).$$

Ignoring the constant term we may define the AIC for Gaussian AR(p)-processes:

Definition 6.5.8

$$AIC(p) := n \log \hat{\sigma}_\varepsilon^2 + 2p$$

and we choose

$$\hat{p}_{AIC} = \arg\min_{p \geq 1} AIC(p).$$

For p increasing, $2p$ increases and $\log \hat{\sigma}_\varepsilon^2$ decreases.

Proposition 6.5.9
For Gaussian AR(p)-processes, FPE and AIC are asymptotically equivalent as $n \longrightarrow \infty$.

Proof:

$$\begin{aligned} n \log FPE(p) &= n \log \hat{\sigma}_\varepsilon^2(p) + n \log \frac{1 + \frac{p}{n}}{1 - \frac{p}{n}} \\ &= n \log \hat{\sigma}_\varepsilon^2(p) + 2p + \frac{2p^3}{3n^2} + \dots \\ &\quad \text{(applying a Taylor expansion of } \log(1+x) - \log(1-x)) \\ &= AIC(p) + O\left(\frac{p^3}{n^2}\right) \end{aligned}$$

□

Remark 6.5.10
A drawback of AIC is that if $\{X_t, t \in \mathbb{Z}\}$ is an AR(p_0)-process then

1. $P(\hat{p}_{AIC} \geq p_0) \longrightarrow 1$ as $n \to \infty$
2. $P(\hat{p}_{AIC} = p_0) \approx 0.8 \not\longrightarrow 1$ as $n \to \infty$

Hence, AIC is not consistent and tends to overestimate p.

As alternative, use e.g. the Schwarz-Kashyap-Rissanen criterion which is equivalent to Akaike's BIC:

Definition 6.5.11

$$S(p) := n \log \hat{\sigma}_\varepsilon^2(p) + p \log n$$

$$\hat{p}_s := \arg\min_{p \geq 1} S(p)$$

Remark 6.5.12

1. $\hat{p}_s \xrightarrow{p} p_0$ as $n \to \infty$ if $\{X_t : t \in \mathbb{Z}\}$ is an AR(p_0)-process.
2. $p \log n > 2p$ for $n \geq 8$ which makes the penalty term stronger for large n.

6.6 Moving Average and ARMA Processes

Definition 6.6.1
A stochastic process $\{X_t, t \in \mathbb{Z}\}$ is said to be a Moving Average of order $q \geq 0$ (MA(q)) if

$$X_t = \sum_{k=1}^{q} \nu_k \varepsilon_{t-k} + \varepsilon_t$$

with $\{\varepsilon_t, t \in \mathbb{Z}\}$ white noise.

Definition 6.6.2
A stochastic process $\{X_t, t \in \mathbb{Z}\}$ is said to be an autoregressive Moving Average of order (p, q) (ARMA(p, q)), $p \geq 0, q \geq 0$ if

$$X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \sum_{k=1}^{q} \nu_k \varepsilon_{t-k} + \varepsilon_t$$

with $\{\varepsilon_t, t \in \mathbb{Z}\}$ white noise.

Remark 6.6.3
By convention $ARMA(0, q) = MA(q)$ for $q \geq 0$, $ARMA(p, 0) = AR(p)$ for $p \geq 1$ and $ARMA(0, 0) = MA(0) =$ white noise.

Theorem 6.6.4
Consider $\{X_t, t \in \mathbb{Z}\}$ an ARMA(p, q)-process such that the polynomials

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^k$$

and

$$D(z) = 1 + \sum_{k=1}^{q} \nu_k z^k$$

have no common root. Then $\{X_t : t \in \mathbb{Z}\}$ has a causal stationary solution if and only if $A(z) \neq 0$ for all $z \in \{z \in \mathbb{C} : |z| \leq 1\}$. Moreover, the coefficients of the $MA(\infty)$-representation of $\{X_t, t \in \mathbb{Z}\}$ are given by

$$B(z) = \sum_{j=0}^{\infty} Y_j z^j = \frac{D(z)}{A(z)}.$$

Remark 6.6.5
An ARMA(p, q)-process can be written as

$$A(U^{-1})X_t = D(U^{-1})\varepsilon_t$$

where A is a polynomial of degree p and D is a polynomial of degree q. Then, the ARMA$(p+m, q+m)$

$$\Theta(U^{-1})A(U^{-1})X_t = \Theta(U^{-1})D(U^{-1})\varepsilon_t$$

where Θ is a polynomial of degree m is another representation of the same process. Indeed,

$$X_t = \frac{\Theta(U^{-1})D(U^{-1})}{\Theta(U^{-1})A(U^{-1})}\varepsilon_t.$$

If we assume that A and D have no common roots, then the ARMA polynomial representation is unique.

Corollary 6.6.6
If $\{X_t, t \in \mathbb{Z}\}$ is a stationary ARMA(p, q)-process, then a spectral density of $\{X_t, t \in \mathbb{Z}\}$ is given by

$$f_X(\omega) = \frac{\sigma_\varepsilon^2 |D(e^{-i\omega})|^2}{|A(e^{-i\omega})|^2}.$$

Application of the Theorem and Corollary
Consider an ARMA$(1,1)$-process, i.e.

$$X_t = \alpha X_{t-1} + \nu\varepsilon_{t-1} + \varepsilon_t, |\alpha| < 1, |\nu| < 1.$$

It follows,

$$A(z) = 1 - \alpha z,\ D(z) = 1 + \nu z.$$

$A(z) = 0$ implies $z_A = \frac{1}{\alpha}$ and $D(z) = 0$ implies $z_D = -\frac{1}{\nu}$. If $z_A \neq z_D$ then $A(z)$ and $D(z)$ have no common zero and $|\alpha| < 1$ implies $|z_A| > 1$. In this case

$$\begin{aligned} B(z) &= \frac{1+\nu z}{1-\alpha z} = (1+\nu z)(1+\sum_{k=1}^{\infty}\alpha^k z^k) \\ &= 1 + \sum_{k=1}^{\infty}\alpha^{k-1}(\alpha+\nu)z^k \end{aligned} \tag{6.4}$$

and a spectral density is derived as

$$f_X(\omega) = \sigma_\varepsilon^2 \left| 1 + \sum_{k=1}^{\infty}\alpha^{k-1}(\alpha+\nu)e^{-ik\omega} \right|^2.$$

Fitting ARMA Processes

Example: $\{X_t, t \in \mathbb{Z}\}$ MA(1), i.e.

$$X_t = \nu\varepsilon_{t-1} + \varepsilon_t$$

with $\{\varepsilon_t, t \in \mathbb{Z}\}$ white noise.
Goal: Given the observations $X_1, \ldots, X_n$ we want to estimate ν and σ_ε^2.

1. $r_1 = \mathrm{E}[X_1X_0] = \mathrm{E}[(\varepsilon_1 + \nu\varepsilon_0)(\varepsilon_0 + \nu\varepsilon_{-1})] = \nu\sigma_\varepsilon^2$ and $r_0 = \mathrm{E}[X_1^2] = (1+\nu^2)\sigma_\varepsilon^2$. Hence,

$$\rho_1 = \frac{r_1}{r_0} = \frac{\nu}{1+\nu^2} \Leftrightarrow \rho_1\nu^2 - \nu + \rho_1 = 0$$

and we derive

$$\nu_{1/2} = \frac{1 \pm \sqrt{1 - 4\rho_1^2}}{2\rho_1}.$$

The idea here is to replace ρ_1 by $\hat{\rho}_1 = \frac{\hat{r}_1}{\hat{r}_0}$ and derive an estimate of ν.
We will not discuss here how to select the estimate from the two possible solutions. This approach is not efficient, anyway. For instance, for Gaussian MA(1)-processes there are other estimates $\tilde{\nu}$ for which we have

$$\frac{\text{mse}(\tilde{\nu})}{\text{mse}(\hat{\nu})} \underset{n\to\infty}{\longrightarrow} const < 1$$

2. As alternative we consider LS-estimation
This approach is complicated by the fact that only the X_t, and not the ε_t, are observed. Under some mild assumptions one can make use of the AR(∞)-representation of X_t: For $|\nu| < 1$ X_t has an $AR(\infty)$-representation, i.e.

$$X_t = \sum_{k=1}^{\infty} \alpha_k X_{t-k} + \varepsilon_t$$

with the coefficients given by

$$1 - \sum_{k=1}^{\infty} \alpha_k z^k \;=\; A(z) = \frac{1}{D(z)} = \frac{1}{1+\nu z} = 1 + \sum_{k=1}^{\infty} (-\nu)^k z^k$$

Hence,

$$\hat{\nu} = \arg\min_{\nu} \sum_{t=-\infty}^{\infty} \left(X_t + \sum_{k=1}^{\infty} (-\nu)^k X_{t-k} \right)^2.$$

Since only $X_1, \ldots, X_n$ are observable, we reduce the estimate to

$$\hat{\nu} = \arg\min_{\nu} \sum_{t=2}^{n} \left(X_t + \sum_{k=1}^{t-1} (-\nu)^k X_{t-k} \right)^2.$$

This can be solved by setting

$$\begin{aligned} e_0(\nu) &= 0, \\ e_t(\nu) &= X_t - \nu e_{t-1}(\nu) \text{ for } t \geq 1, \end{aligned}$$

such that

$$\hat{\nu} = \arg\min_{\nu} \sum_{t=2}^{n} e_t^2(\nu).$$

Remark 6.6.7
For Gaussian processes it is known that the LS-estimator and the Maximum-Likelihood estimator are asymptotically equivalent. Moreover, the ML-estimator is efficient. For practical application, one can make use of a nonlinear optimization routine to solve the problem numerically.

For estimation in general ARMA-processes, we will make use of the following result.

Proposition 6.6.8
Let $\{X,t \in \mathbb{Z}\}$ be an MA(q)-process with $r_t = cov(X_t, X_0)$. Then $r_t = 0$ for $t > q$.

Proof:

$$\begin{aligned} r_t &= \operatorname{cov}(X_t, X_0) = \mathrm{E}[X_t X_0] \\ &= E\left[\varepsilon_t\varepsilon_0 + \varepsilon_t \sum_{k=1}^{q} \nu_k \varepsilon_{-k} + \varepsilon_0 \sum_{k=1}^{q} \nu_k \varepsilon_{t-k} + \sum_{j,k=1}^{q} \nu_k \nu_j \varepsilon_{t-j}\varepsilon_{-k}\right] \\ &= 0, \text{ for } t - q > 0. \end{aligned}$$

□

If we want to fit an ARMA(p, q)-process to given data $X_1, \ldots, X_n$ we need to

- Estimate p and q by $\hat{p}$ and $\hat{q}$.
- Estimate the parameters $\alpha_1, \ldots, \alpha_{\hat{p}}, \nu_1, \ldots, \nu_{\hat{q}}, \sigma_\varepsilon^2$.

Estimation procedure:

1. Investigate plots of the empirical acf and pacf. Do they give evidence that we have an AR(p)- or an MA(q)-process? In this case, the order may be determined from the plots.

2. Use an objective model selection criterion, e.g. the generalized AIC

   $$AIC(p, q) = -2 \log L(\hat{\alpha}, \hat{\nu}, \hat{\sigma}_\varepsilon^2) + 2(p + q)$$

   or

   $$AIC(p, q) = n \log \hat{\sigma}_\varepsilon^2 + 2(p + q).$$

   Find $\operatorname{argmin}_{p,q} AIC(p, q)$. To compute AIC, we need estimates of α, ν and σ_ε^2.
   For given values of p and q these may be obtained by:

   - a generalization of the Yule-Walker estimates (not efficient),
   - LS-estimation, (asymptotically efficient for Gaussian processes),
   - ML-estimation (efficient).

   The idea is to use an easy method such as Yule-Walker estimation in the AIC to get $\hat{p}$ and $\hat{q}$. Then $\hat{\alpha}(\hat{p})$, $\hat{\nu}(\hat{q})$, and $\hat{\sigma}_\varepsilon^2$ can be computed by ML.

3. Model validation: Do the empirical residuals behave like white noise? If yes: accept the model. If no: refine the model.

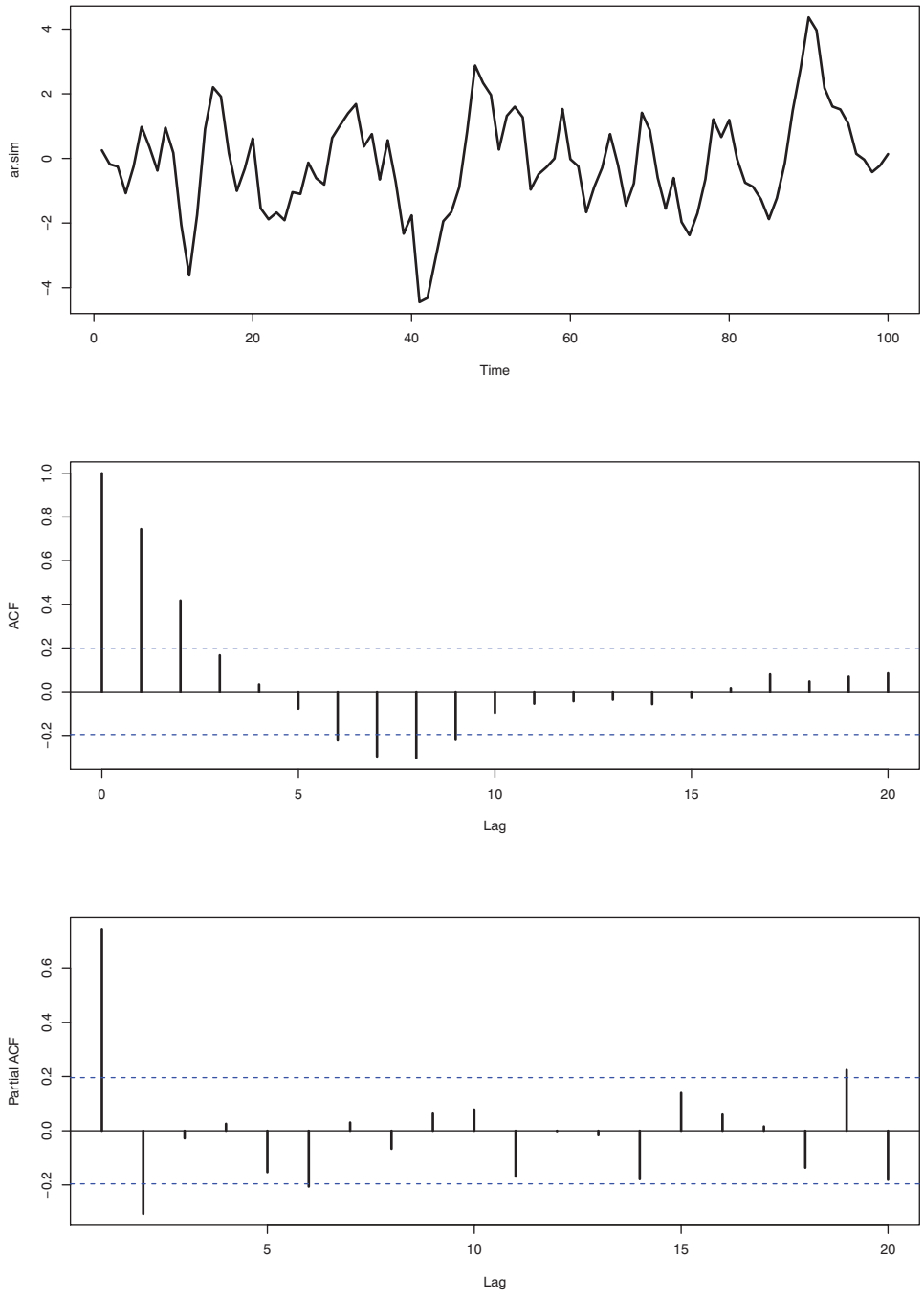

Figure 6.1: Realization of an AR(2)-process with $\alpha_1 = 0.9$, $\alpha_2 = -0.2$ and estimates of its acf and pacf. The ε_t are iid N(0,1)-distributed.

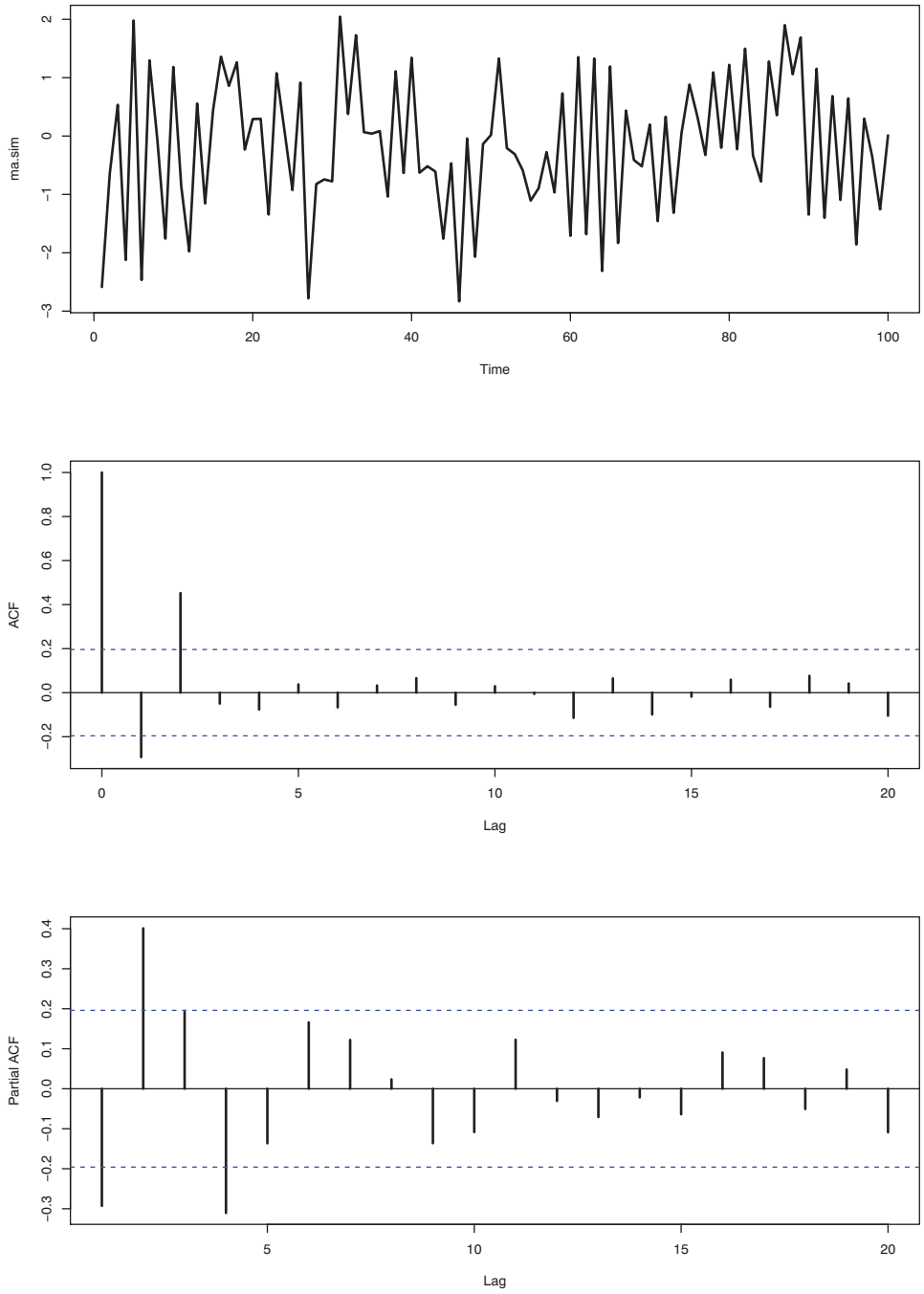

Figure 6.2: Realization of an MA(2)-process with $\nu_1 = -0.3$, $\nu_2 = 0.9$ and estimates of its acf and pacf. The ε_t are iid N(0,1)-distributed.

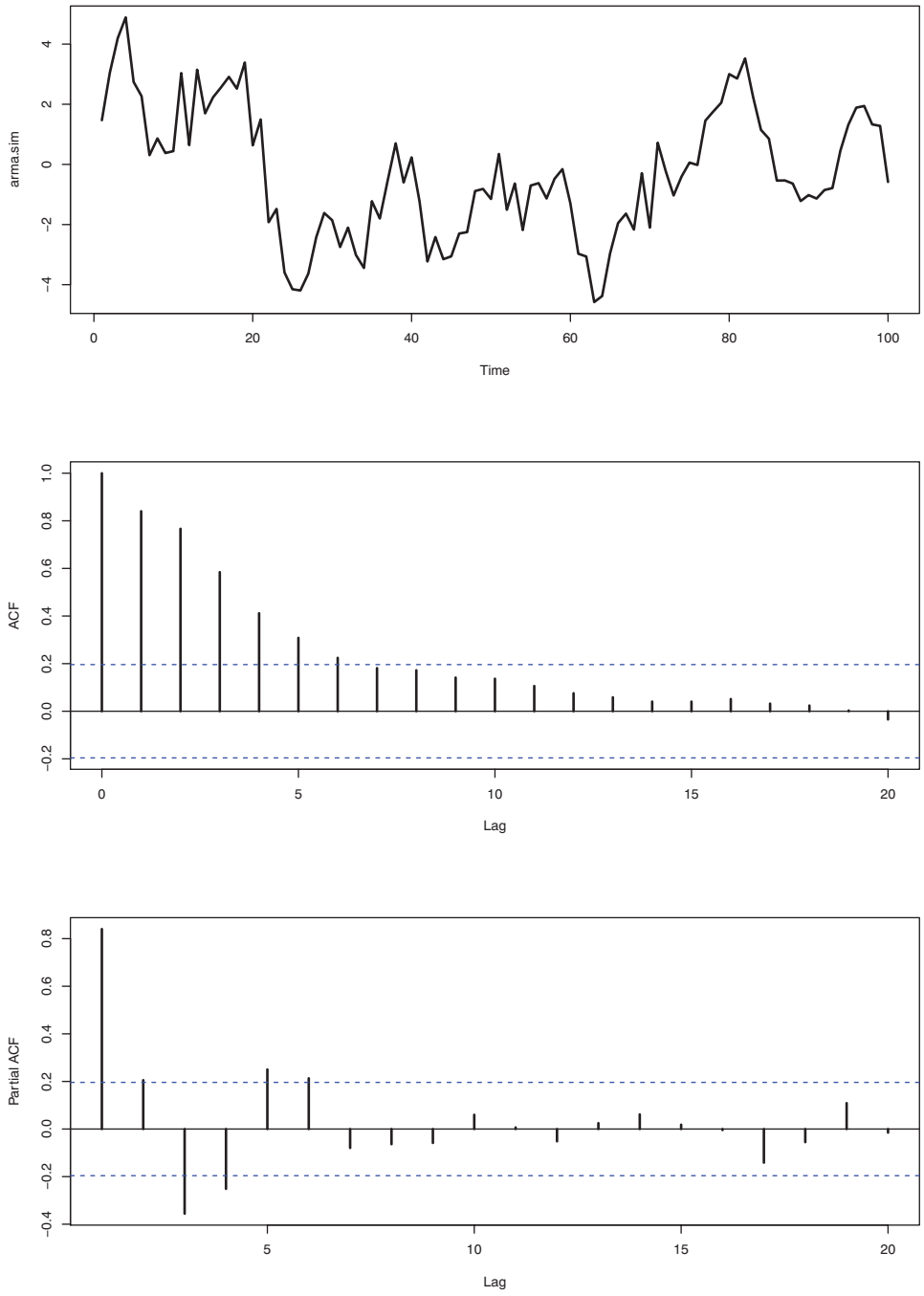

Figure 6.3: Realization of an ARMA$(2,2)$-process with $\alpha_1 = 0.9$, $\alpha_2 = -0.2$, $\nu_1 = -0.3$, $\nu_2 = 0.9$ and estimates of its acf and pacf. The ε_t are iid N(0,1)-distributed.

Chapter 7

Time Series with Trend and Seasonal Components

Causal, stationary and invertible ARMA processes are particularly suitable if

1. the data $X_1, \ldots, X_n$ do not exhibit any clear departure from stationarity.

2. the sample acf or pacf decreases (exponentially) fast.

If 1. or 2. are violated, we can transform the data until the transformed data satisfy these conditions.
Here, we are going to consider models of the following form:

$$X_t = m_t + s_t + \varepsilon_t,$$

where m_t is a trend component, s_t is a seasonal component, and ε_t is a stationary process.

7.1 ARIMA Processes

Consider the following example: Let $X_t = \mu + at + R_t$ be a process with linear trend $\mu + at$ and stationary rest $\{R_t : t \in \mathbb{Z}\}$. Then

$$(\nabla X)_t = X_t - X_{t-1} = \mu + at + R_t - \mu - a(t-1) - R_{t-1} = a + R_t - R_{t-1}$$

is a stationary process with $E(\nabla X)_t = a$. ∇ is called the difference operator.
For $X_t = at^2 + bt + c + R_t$ we have to apply ∇ twice to get a stationary process.
More generally, we have the following proposition.

Proposition 7.1.1
Consider a stochastic process $\{X_t\}$ given by

$$X_t = P_d(t) + R_t,$$

with a stationary process $\{R_t\}$ and a polynomial $P_d(t)$ of degree $d \geq 1$, $P_d(t) = a_d t^d + \ldots + a_1 t + a_0, a_d \neq 0$. Then, $\{(\nabla^d X)_t, t \in \mathbb{Z}\}$ is a stationary process with $E(\nabla^d X)_t = d!\, a_d$.

Proof:

$$\begin{aligned}(\nabla X)_t &= \underbrace{P_d(t) - P_d(t-1)}_{:=P_{d-1}(t)} + R_t - R_{t-1} \\ &= P_{d-1}(t) + (\nabla R)_t\end{aligned}$$

$$\begin{aligned}P_{d-1}(t) &= a_d t^d + \ldots - a_d(t-1)^d - \ldots \\ &= d a_d t^{d-1} + \ldots,\end{aligned}$$

Since $(t-1)^d = t^d - dt^{d-1} + \ldots$. $P_{d-1}(t)$ is a polynomial of order $d-1$ with leading coefficient da_d. d-fold application of ∇ yields

$$(\nabla^d X)_t = d \cdot (d-1) \cdot (d-2) \ldots \cdot 1 \cdot a_d + (\nabla^d R)_t.$$

Using the linearity of the expectation, it is easy to show that $\{(\nabla^d X)_t, t \in \mathbb{Z}\}$ is a stationary process with expectation $\mathrm{E}(\nabla^d X)_t = d\,!\, a_d$. □

Now we can try to model $(\nabla X)_t$ by an ARMA-process: Using the shift operator,

$$U : X_t \longmapsto X_{t+1}$$

we rewrite

$$(\nabla X)_t = X_t - X_{t-1} = (1 - U^{-1}) X_t,$$

hence,

$$\nabla = 1 - U^{-1}.$$

Definition 7.1.2
Consider $p, d, q \geq 0$. $\{X_t, t \in \mathbb{Z}\}$ is called an autoregressive integrated moving average-process of order (p, d, q) (ARIMA(p, d, q)) if

$$(1 - U^{-1})^d X_t = (\nabla^d X)_t$$

is a stationary ARMA(p, q)-process.

With $\tilde{A}(z) = A(z)(1-z)^d$, an ARIMA-process is characterized by the equation

$$\tilde{A}(U^{-1}) X_t = A(U^{-1})(1 - U^{-1})^d X_t = D(U^{-1})\varepsilon_t.$$

Note that $\tilde{A}$ has a d-fold zero in $z = 1$, hence in the unit circle.

Remark 7.1.3

- For an ARIMA$(p, 1, q)$-process $\{X_t, t \in \mathbb{Z}\}$ the difference process $\{Y_t, t \in \mathbb{Z}\}$ with $Y_t = X_t - X_{t-1}$ is a stationary ARMA(p, q)-process. Since

  $$X_t = X_0 + \sum_{j=1}^{t} Y_j,$$

  X_t can be interpreted as a sum of ARMA-processes which explains the term "integrated".

- The range of applicability of ARIMA-processes is wider than just time series with polynomial trend. There exist extensions such as fractional ARIMA-processes (for non-integer d), vectorial ARIMA-processes or seasonal ARIMA-processes (SARIMA, see later).
- The acf of ARIMA-processes is typically positive and slowly decreasing.

Time series whose acf is decreasing but oscillates have a periodic component whose period can be read from the acf. For such processes we consider a generalization of ARIMA-processes.

7.2 Seasonal ARIMA processes

Proposition 7.2.1
Consider a stochastic process $\{X_t, t \in \mathbb{Z}\}$ *with*

$$X_t = Z_t + R_t$$

where $\{R_t, t \in \mathbb{Z}\}$ *is a stationary process and* $\{Z_t, t \in \mathbb{Z}\}$ *a periodic process with period* s*, i.e.* $Z_t = Z_{t+s}$ *for all* $t \in \mathbb{Z}$*. Let* $\nabla_s = 1 - U^{-s}$ *such that* $\nabla_s X_t = X_t - X_{t-s}$*. Then,* $\{(\nabla_s X)_t, t \in \mathbb{Z}\}$ *is a stationary process with* $E(\nabla_s X)_t = 0$*.*

Definition 7.2.2
A process $\{X_t, t \in \mathbb{Z}\}$ is called a seasonal ARIMA-process with period s (SARIMA$(p, d, q) \times (P, D, Q)_s$) if

$$Y_t = (1 - U^{-1})^d (1 - U^{-s})^D X_t$$

is a stationary ARMA-process of the form

$$A(U^{-1})F(U^{-s})Y_t = B(U^{-1})G(U^{-s})\varepsilon_t$$

with $\{\varepsilon_t, t \in \mathbb{Z}\}$ white noise,

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^k, \quad F(z) = 1 - \sum_{k=1}^{P} \rho_k z^k,$$

$$B(z) = 1 + \sum_{k=1}^{q} \nu_k z^k, \quad G(z) = 1 + \sum_{k=1}^{Q} \gamma_k z^k.$$

Example Consider X_t a SARIMA$(1, d, 0) \times (1, D, 0)_{12}$-process (monthly data). Then,

$$A(z) = 1 - \alpha z, F(z) = 1 - \rho z, B(z) = 1 \text{ and } G(z) = 1.$$

From the stationary ARMA-process Y_t

$$\begin{aligned} A(U^{-1})F(U^{-s})Y_t &= \varepsilon_t \\ Y_t - \rho Y_{t-s} - \alpha Y_{t-1} + \alpha\rho Y_{t-s-1} &= \varepsilon_t \end{aligned}$$

we derive the following representation

$$Y_t = \alpha Y_{t-1} + \rho\, Y_{t-12} - \alpha\rho\, Y_{t-13} + \varepsilon_t.$$

Hence, Y_t depends on Y_{t-1}, Y_{t-12} (value of the series for the same month one year earlier) and Y_{t-13}.

Remark 7.2.3
Y_t is a special representation of an ARMA$(p + Ps, q + Qs)$ process.

Motivation for Definition 7.2.2:
Assume that monthly data X_t are observed and recorded as follows:

		1	2	3	...	12	Month
Year	1	X_1	X_2	X_3	...	X_{12}	
	2	X_{13}	X_{14}	...	...	X_{24}	
	3	X_{25}	X_{26}	...	...	X_{36}	
	4	X_{37}	...	...	...	X_{48}	
	$\vdots$	$\vdots$					

The columns of this matrix are also time series and represent the data for a fixed month over several years. For each month, we can define a time series

$$Z_t^{[m]} = X_{m+12t} \quad m = 1, \ldots, 12, \quad t \in \mathbb{Z}$$

We will assume that the monthly data $\{Z_t^{[m]}, t \in \mathbb{Z}\}$ can be modelled by an ARMA(P, Q)-process, i.e.

$$Z_t^{[m]} = \sum_{k=1}^{P} \rho_k \, Z_{t-k}^{[m]} + V_t^{[m]} + \sum_{k=1}^{Q} \gamma_k \, V_{t-k}^{[m]},$$

where for fixed m $\{V_t^{[m]}, t \in \mathbb{Z}\}$ is white noise. This means that the temporal dependence from year to year does not depend on the considered month.
Define a new time series via

$$V_{m+12t} := V_t^{[m]} \quad \text{for} \quad m = 1, \ldots, 12 \quad \text{and} \quad t \in \mathbb{Z}.$$

$\{V_t, t \in \mathbb{Z}\}$ need not be white noise. Indeed, $V_t^{[m]}$ for constant t and consecutive m might be strongly correlated.
Now

$$X_t = \sum_{k=1}^{P} \rho_k X_{t-12k} + \sum_{k=1}^{Q} \gamma_k V_{t-12k} + V_t$$

or, using the shift operator,

$$F(U^{-12})X_t = G(U^{-12})V_t, \tag{7.1}$$

which yields a year-to-year model.
Finally, we include the dependence from month to month by assuming that $\{V_t, t \in \mathbb{Z}\}$ follows an ARMA(p, q)-model, i.e.

$$A(U^{-1})V_t = B(U^{-1})\varepsilon_t \tag{7.2}$$

with $\{\varepsilon_t\}$ white noise. Combining (7.1) and (7.2) and replacing the fixed period $s = 12$ by an arbitrary s we get

$$A(U^{-1})F(U^{-s})X_t = B(U^{-1})G(U^{-s})\varepsilon_t,$$

which is a seasonal ARMA$(p, q) \times (P, Q)_s$-model.

Note that V_t, V_{t+12}, V_{t+24} are no longer uncorrelated in this model. However, due to the exponential decay of the acf in an ARMA-process we can hope that they are only weakly correlated.
If X_t has both a trend and a periodic component we get a general ARIMA-model as in Definition 7.2.2.

7.3 Fitting SARIMA Models to Data

Given $X_1, \ldots, X_n$, we want to fit a SARIMA-model to the data. This is done in the following steps:

1. Plot the data to check if there are obvious trends, outliers,... in the data. Is there any obvious departure from stationarity?

2. If necessary, remove the heteroskedasticity (time-dependence of the variance), e.g. via a logarithmic or Box-Cox transformation. Consider for example
$$X_t = \mu_t + V_t$$
with $\mathrm{E}V_t = 0, \mathrm{var}V_t = \sigma^2\mu_t^2$. We then define
$$\begin{aligned} X_t' &= \log X_t = \log \mu_t(1 + \frac{V_t}{\mu_t}) \\ &= \log \mu_t + \log(1 + \frac{V_t}{\mu_t}) \\ &\approx \log \mu_t + \frac{V_t}{\mu_t} = \log \mu_t + V_t' \end{aligned}$$
since often $V_t << \mu_t$. Obviously $\mathrm{E}V_t' = 0$ and $\mathrm{var}V_t' = \sigma^2$.

3. Select s, d, and D such that
$$Y_t = (1 - U^{-1})^d(1 - U^{-s})^D X_t$$
looks stationary. Usually, we have $0 \leq d \leq 2$ (no, linear or quadratic trend) and $0 \leq D \leq 1$ (period or no period).

4. Plot the sample acf and pacf of the (transformed) data Y_t: Plot $\hat{r}_t$ and $\hat{\pi}_t$ for $t = ks, k \geq 0$. Choose P and Q such that the plots look like the acf and pacf of an ARMA(P, Q)-process.

5. Choose p and q such that $\hat{r}_1, \ldots, \hat{r}_{s-1}$ and $\hat{\pi}_1, \ldots, \hat{\pi}_{s-1}$ look like the acf and pacf of an ARMA(p, q)-process.

6. Estimate the parameters $\alpha_k, \nu_k, \rho_k, \gamma_k, \sigma_\varepsilon^2$ by maximum-likelihood.

7. If the steps 4. to 6. yield several plausible models choose one of them e.g. by AIC.

8. Check the selected model by an examination of the fitted residuals.

Remark 7.3.1
The steps 4. to 7. may be replaced by a direct application of AIC in the class SARIMA$(p, 0, q) \times (P, 0, Q)_s$ for the data Y_t.

Examination of the Fitted Residuals

The residuals $\{\varepsilon_t\}$ are by assumption white noise. Therefore, we need to answer the following question: Do the sample residuals really behave like white noise?
Assume for example that $\{X_t, t \in \mathbb{Z}\}$ follows an ARMA(p,q)-model. Then the true residuals satisfy

$$\varepsilon_t(p,q) = X_t - \sum_{k=1}^{p} \alpha_k X_{t-k} - \sum_{k=1}^{q} \nu_k \varepsilon_{t-k}(p,q)$$

and the sample residuals are defined as

$$e_t(p,q) = X_t - \sum_{k=1}^{p} \hat{\alpha}_k X_{t-k} - \sum_{k=1}^{q} \hat{\nu}_k e_{t-k}(p,q)$$

where $X_t = 0,\ e_t = 0$ for $t \leq 0$.
The e_t should at least asymptotically behave like white noise.

1. Approach:
Plot $e_t(p,q)$ against time. They should look uncorrelated with mean 0 and variance σ_ε^2.

2. Approach:
Investigate the acf $\{\hat{r}_{k,e}\}$ of the sample residuals. If $\{e_t\}$ is white noise then the $\hat{r}_{k,e}$ are asymptotically independent $\mathcal{N}(0, \sigma_k^2)$-variables with $\sigma_k^2 = \frac{1}{n}$ for large k and $\sigma_k^2 < \frac{1}{n}$ for small k.

3. Approach:
Perform a Portmanteau or Box-Pierce test (formulated for ARMA-models here):
Test $H_0 : \{e_t\}$ is white noise against $H_1 : \{e_t\}$ is not.
Choose $\{h_n\}$ such that $h_n \to \infty$ and $h_n = O(\sqrt{n})$, e.g., $h_n = 2\sqrt{n}$. Then

$$Q = n \sum_{k=1}^{h_n} \hat{r}_{k,e}^2 \sim_{\mathcal{L}} \chi^2_{h_n-(p+q)}$$

under H_0.
Alternatively, use the Ljung-Box test with

$$Q = n(n+2) \sum_{k=1}^{h_n} \frac{1}{n-k} \hat{r}_{k,e}^2 \sim_{\mathcal{L}} \chi^2_{h_n-(p+q)}$$

under H_0 which has a better finite sample size behaviour.
The problem with these tests is that they reject unsuitable models but they are not able to choose the best among several suitable models.

7.4 Linear Forecasting of Nonstationary Time Series

Given the data $X_1, \ldots, X_n$ we want to predict $X_{n+1}, X_{n+2}, \ldots$.

Box-Jenkins Approach

A SARIMA model is fitted to the data and used for the prediction of the future values of the series.
For AR-processes we already know the procedure: If $X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t$ then the best linear prediction of $X_{n+1}, X_{n+2}, \ldots$, i.e., the linear prediction minimizing $E[(X_{n+i} - \hat{X}_{n+i})^2]$, is given by

$$\hat{X}_{n+1} = \sum_{k=1}^{p} \alpha_k X_{n+1-k}$$

$$\hat{X}_{n+2} = \alpha_1 \hat{X}_{n+1} + \sum_{k=2}^{p} \alpha_k X_{n+2-k}, \ldots$$

If $q = Q = 0$, i.e. in models without MA-components, this approach also works for SARIMA-processes:

1. Transform X_t to $Y_t = (1 - U^{-1})^d (1 - U^{-s})^D X_t$.
2. Since $A(U^{-1})F(U^{-s})Y_t = \varepsilon_t$, Y_t is a special AR$(p + sP)$-process. Hence, predictions for Y_t are obtained as above.
3. Backtransformation of the $\hat{Y}_t$ yields $\hat{X}_t$

Note: The sample size has to be large enough since the application of the difference operator reduces the number of Y-data compared to the X-data.
For the prediction of general SARIMA-models we need to know how to predict ARMA(p, q)-processes

$$Y_t = \sum_{k=1}^{p} \alpha_k Y_{t-k} + \varepsilon_t + \sum_{k=1}^{q} \nu_k \varepsilon_{t-k}.$$

Since ε_t is orthogonal to Y_s for $s < t$ and since $\varepsilon_s \in \overline{sp}\{Y_j : j < t\}$ for $s < t$ the best linear prediction for Y_{n+1} given $Y_t, t \leq n$ is

$$\tilde{Y}_{n+1} = \sum_{k=1}^{p} \alpha_k Y_{n+1-k} + \sum_{k=1}^{q} \nu_k \varepsilon_{n+1-k}.$$

We cannot observe the ε_t and have only finitely many observations $Y_1, \ldots, Y_n$. Hence, the ε_t are replaced by the residuals

$$e_t = Y_t - \sum_{k=1}^{p} \alpha_k Y_{t-k} - \sum_{k=1}^{q} \nu_k e_{t-k}, \quad t = 1, \ldots, n,$$

with $e_t = 0$ and $Y_t = 0$ for $t \leq 0$. Then

$$\hat{Y}_{n+1} = \sum_{k=1}^{p} \alpha_k Y_{n+1-k} + \sum_{k=1}^{q} \nu_k e_{n+1-k}.$$

This approach is useful if many data are available and only short term predictions are required. In other cases one uses ad-hoc methods since either the data are not sufficient to fit a complex model or the model might change over time.

Exponential Smoothing

Given $X_0, X_1, \ldots, X_n$, we first assume that $\{X_t, t \in \mathbb{Z}\}$ has neither a seasonal component nor a deterministic trend.

1. Predict X_{n+1} as a weighted sum of the past observations
$$\hat{X}_{n+1|n} = c_0 X_n + c_1 X_{n-1} + \ldots$$
where $|c_k|$ is small for large k, i.e. we put more weight on the recent observations.
2. Choose $c_k = \delta(1-\delta)^k, k \geq 0$ for a $\delta \in (0,1)$. Then
$$\hat{X}_{n+1|n} = \delta X_n + \delta(1-\delta)X_{n-1} + \delta(1-\delta)^2 X_{n-2} + \ldots$$

This prediction is optimal in the sense that
$$\mathrm{E}(X_{n+1} - \hat{X}_{n+1|n})^2$$
is minimal when $\{X_t, t \in \mathbb{Z}\}$ is an AR(∞)-process of the form
$$X_t = \sum_{k=1}^{\infty} \delta(1-\delta)^{k-1} X_{t-k} + \varepsilon_t.$$
In this case
$$X_t = \delta X_{t-1} + (1-\delta)\sum_{k=1}^{\infty} \delta(1-\delta)^{k-1} X_{t-k-1} + \varepsilon_t$$
and
$$(1-\delta)X_{t-1} = (1-\delta)\sum_{k=1}^{\infty} \delta(1-\delta)^{k-1} X_{t-k-1} + (1-\delta)\varepsilon_{t-1}.$$
Hence,
$$X_t = \delta X_{t-1} + [(1-\delta)X_{t-1} - (1-\delta)\varepsilon_{t-1} + \varepsilon_t] = X_{t-1} - (1-\delta)\varepsilon_{t-1} + \varepsilon_t.$$
Therefore,
$$(1-U^{-1})X_t = X_t - X_{t-1} = \varepsilon_t - (1-\delta)\varepsilon_{t-1} \equiv \text{ ARMA}(0,1)$$
and $\{X_t, t \in \mathbb{Z}\}$ is an ARIMA$(0,1,1)$-process.

Remark 7.4.1

- If we add an observation X_n to the sequence $X_1, \ldots, X_{n-1}$ then
$$\begin{aligned} \hat{X}_{n+1|n} &= \sum_{k=1}^{\infty} \delta(1-\delta)^{k-1} X_{n+1-k} \\ &= \delta X_n + (1-\delta)\sum_{k=1}^{\infty} \delta(1-\delta)^{k-1} X_{n-k} \\ &= \delta X_n + (1-\delta)\hat{X}_{n|n-1}. \end{aligned}$$
Setting $\eta_n = X_n - \hat{X}_{n|n-1}$ we get
$$\hat{X}_{n+1|n} = \delta\eta_n + \hat{X}_{n|n-1}.$$

- Choice of δ:
  As long as δ is not too small, its choice is often not critical. δ can be chosen using a LS-approach as follows:

$$\begin{aligned}
\hat{X}_{2|1} &= X_1 \\
\eta_2 &= X_2 - \hat{X}_{2|1} \\
\hat{X}_{3|2} &= \delta\eta_2 + \hat{X}_{2|1} \\
&\dots \\
\eta_n &= X_n - \hat{X}_{n|n-1}
\end{aligned}$$

  Then minimize $Q(\delta) = \sum_{k=2}^{n} \eta_k^2$ w.r.t $0 < \delta < 1$ (minimization of the one-step prediction errors).

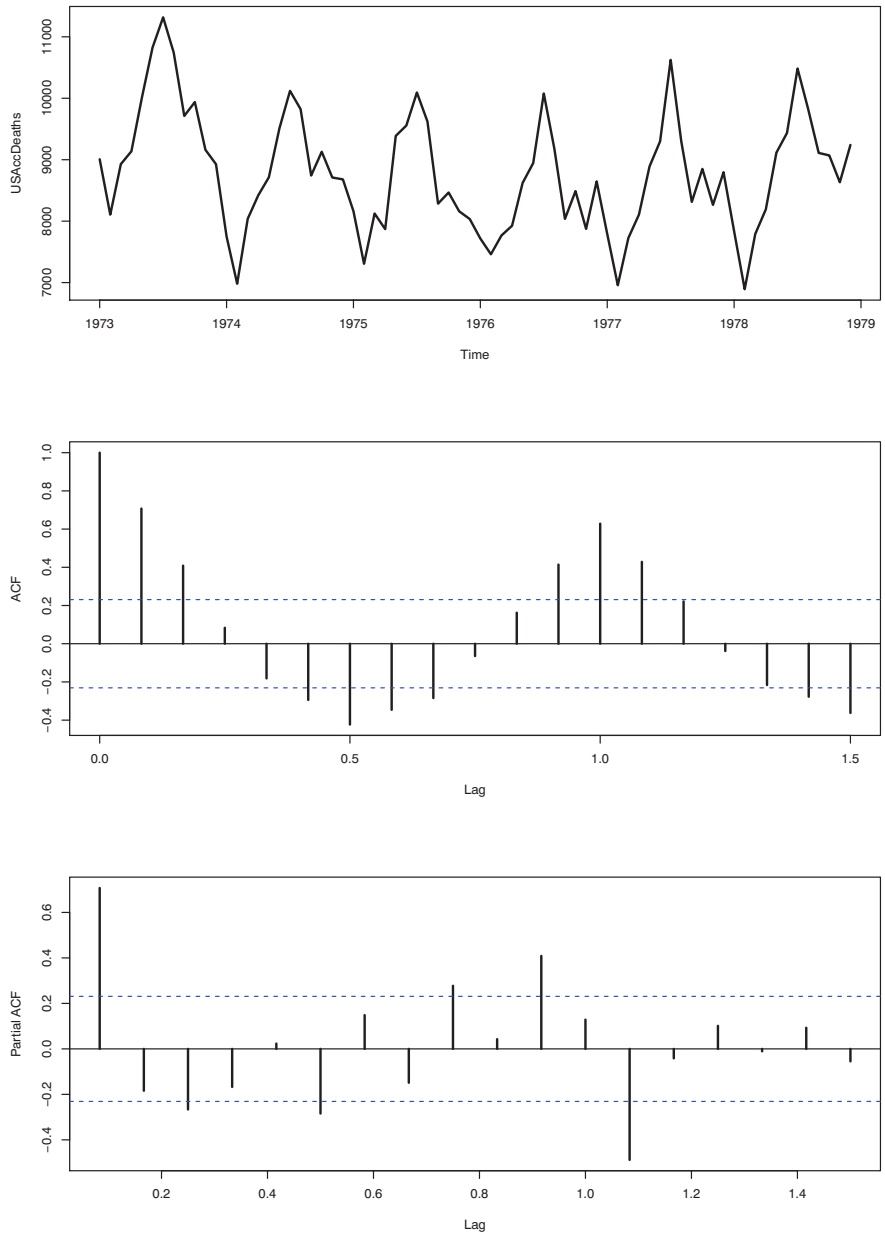

Figure 7.1: A time series giving the monthly totals of accidental deaths in the USA.