University of Groningen

Faculty of Economics

Linear Models

Sytse Knypstra

2006

version 060127

In econometrics linear models are traditionally very important; analysis of variance (ANOVA) and linear regression analysis both being special cases. The word 'linear' means that the vector of expected values of a certain vector of n observations $\boldsymbol{Y}$ lies in a linear subspace of dimension smaller than n.

In our treatment of the subject the geometrical interpretation is stressed. In many textbooks a more algebraic approach is followed.

In a linear model we assume that the outcomes of n random variables are observed, which are arranged in a vector $\boldsymbol{Y}$ of dimension n. Further it is assumed that the components of the vector $\boldsymbol{Y}$ are uncorrelated with identical variance. The expected value of one component depends linearly on one or more factors, whose values are contained in a matrix $\boldsymbol{X}$.

More specifically, $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X\beta}$, where $\boldsymbol{X}$ is a matrix of n rows and p columns, containing only known real numbers, and $\boldsymbol{\beta}$ is a vector of p unknown coefficients.

On the next page six examples of linear models are specified.

1. **One sample problem**
   $\mathcal{E}(Y_i) = \mu$, all expectations are identical; the value of μ is unknown and we want to estimate μ or test hypothesis about μ.

2. **Two samples problem**
   $\mathcal{E}(Y_1) = \ldots = \mathcal{E}(Y_{n_1}) = \mu_1$ and $\mathcal{E}(Y_{n_1+1}) = \ldots = \mathcal{E}(Y_n) = \mu_2$.
   We want to estimate μ_1 and μ_2 and test the hypothesis that $\mu_1 = \mu_2$.

3. **One factor ANOVA**
   The first n_1 expectations are μ_1; the next n_2 expectations are μ_2, ..., the last n_k expectations are μ_k. We want to estimate $\mu_1, \ldots, \mu_k$ and test the hypothesis that all μ's are identical.

4. **Two factor ANOVA**
   $\mathcal{E}(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$, where the sum of the α's, the sum of the β's and the sum of the γ's, the latter taken over i as well as over j, is equal to 0. We want to estimate the parameters μ, the α's, the β's and the γ's. And we might want to test the hypothesis that all α's are zero.

5. **Simple linear regression**
   $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i$, where the x_i's are known real numbers. We want to estimate β_0 and β_1 and test the hypothesis that $\beta_1 = 0$.

6. **Multiple linear regression**
   $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$, where the x's are again known real numbers. We want to estimate the β's and test hypothesis regarding the β's.

All six examples have in common that the vector of expectations can be written as $\boldsymbol{X\beta}$, where $\boldsymbol{X}$ is a matrix of known numbers and $\boldsymbol{\beta}$ a vector of parameters with unknown values. We are interested in the value of $\boldsymbol{\beta}$ and possibly σ^2: we want to **estimate** (functions of) $\boldsymbol{\beta}$ and σ^2 and we want to test hypothesis about (functions of) $\boldsymbol{\beta}$ and σ^2.
The estimators and tests can be derived for the general linear model and then be applied to special cases such as the **six examples** shown.

Example 1. Does increasing the amount of calcium in our diet reduce blood pressure? An experiment gave one group of 10 black men a calcium supplement for 12 weeks. The control group of 11 black men received a placebo that appeared identical. The experiment was double-blind. The decrease of blood pressure after the experiment was measured.

The vector $\boldsymbol{Y}$ consists of 21 observations, the first $n_1 = 10$ belonging to the 'calcium' group, the remainimg $n_2 = 11$ to the placebo group. We assume that all 21 observations are independently distributed, with identical variance σ^2; the first 10 observations with expectation μ_1, the remaining 11 with expectation μ_2. Therefore we can write:

$$\mathcal{E}\begin{pmatrix} Y_{1\,1} \\ \vdots \\ Y_{1\,10} \\ Y_{2\,1} \\ \vdots \\ Y_{2\,11} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \boldsymbol{X\beta}.$$

We want to estimate μ_1 and μ_2 or test the hypothesis that both μ's are identical.

Assumptions for the linear model are:

1. $\boldsymbol{Y}$ is a vector of n observations,
2. the components of $\boldsymbol{Y}$ are uncorrelated,
3. the components of $\boldsymbol{Y}$ have equal variance σ^2,
4. $\mathcal{E}(\boldsymbol{Y})$ lies in the linear subspace, spanned by the column vectors of the matrix $\boldsymbol{X}$:
   $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X\beta}$, where $\boldsymbol{X}$ is an $n \times p$-matrix with known elements and $\boldsymbol{\beta}$ is a vector of p unknown parameters.

Or, equivalently:

a. $\boldsymbol{Y}$ is a vector of n observations,
b. $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is an $n \times p$-matrix with known elements and $\boldsymbol{\beta}$ is a vector of p unknown parameters,
c. the components of $\boldsymbol{\varepsilon}$ have expectation 0,
d. the components of $\boldsymbol{\varepsilon}$ are uncorrelated,
e. the components of $\boldsymbol{\varepsilon}$ have equal variance σ^2.

Under these assumptions it is possible to derive the so-called Least Squares estimator for $\boldsymbol{\beta}$. Under the additional assumption:

5. $\boldsymbol{Y}$ (and therefore each component of $\boldsymbol{Y}$) is normally distributed

f. $\boldsymbol{\varepsilon}$ (and therefore each component of $\boldsymbol{\varepsilon}$) is normally distributed

it is also possible to derive the maximum likelihood estimator, likelihood ratio tests and confidence intervals for (linear combinations of) components of $\boldsymbol{\beta}$.

The assumptions **(1) – (5)** can be summarised as:

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}).$$

The assumptions **(a) – (f)** can be summarised as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \text{and} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Under the first four **assumptions** estimators for the unknown parameters can be derived on the basis of the so-called least squares (LS) criterion. Despite the rather mild assumptions they have nice properties. If the assumption of normality **(5)** of **(f)** is added, the maximum likelihood (ML) estimator for $\boldsymbol{\beta}$ appears to coincide with the least squares estimator. Under the assumption of normality the distribution of the ML-estimator can be derived.

The least squares criterion requires us to find the point $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ (in the linear subspace spanned by the column vectors of $\boldsymbol{X}$) for which the distance $\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|$ is minimal. Or, equivalently:

$$\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

should be minimal.

The norm used here is the Euclidean norm; the sum of squares of the components of the difference vector. This explains the term 'least squares'.

Below the situation is shown geometrically. The vector $\boldsymbol{Y}$ is considered to be a point in R^n. The vector of expectations $\boldsymbol{X\beta}$ is bound to lie in the linear subspace $\mathcal{L}$ of R^n, spanned by the column vectors of $\boldsymbol{X}$. The projection theorem now states that the point $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ in $\mathcal{L}$ which is closest to $\boldsymbol{Y}$ is the projection of $\boldsymbol{Y}$ onto $\mathcal{L}$. Therefore $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ should be perpendicular to $\mathcal{L}$, and thus perpendicular to the spanning vectors of $\mathcal{L}$: the column vectors of $\boldsymbol{X}$. This means that $\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ or, equivalently,

$$\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{2.1}$$

should hold. These equations are also called the normal equations.

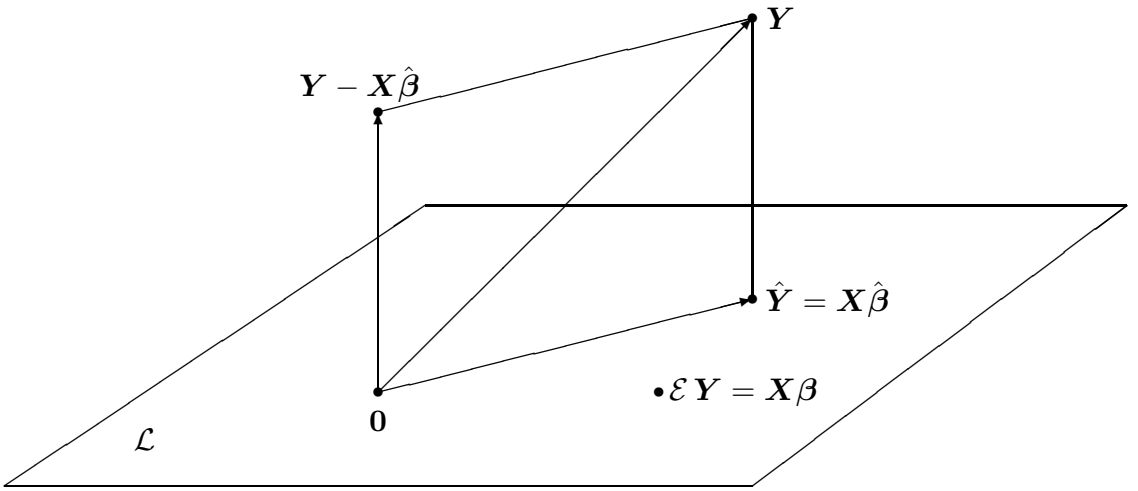

Figure 2.1 Geometrical representation.

If $\boldsymbol{X}$ has rank p = the number of columns, then $\boldsymbol{X}^T\boldsymbol{X}$ also has rank p and the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular. In this case the normal equations have only one solution:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{2.2}$$

The projection of $\boldsymbol{Y}$ onto $\mathcal{L}$ is then $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y}$, where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the projection matrix. In **exercise 3** it is left to the reader to prove that $\boldsymbol{P}$ is symmetric and idempotent and has rank p. We can write the vector $\boldsymbol{Y}$ as the sum of two terms: its projection $\boldsymbol{P}\boldsymbol{Y} = \hat{\boldsymbol{Y}}$ on $\mathcal{L}$ and the residual vector $\boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$. These two vectors are orthogonal. In **exercise 3** the reader is also asked to prove that $\boldsymbol{I} - \boldsymbol{P}$ is a projection matrix: it projects all vectors $\boldsymbol{Y}$ onto the orthogonal complement of $\mathcal{L}$ in R^n. The rank of $\boldsymbol{I} - \boldsymbol{P}$ is $n - p$. Now Pythagoras' theorem holds:

$$\|\boldsymbol{Y}\|^2 = \|\hat{\boldsymbol{Y}}\|^2 + \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2.$$

If the rank of $\boldsymbol{X}$ is smaller than p, then $\boldsymbol{X}^T\boldsymbol{X}$ is singular and the **normal equations** have more than one solution; in that case different linear combinations of the column vectors of $\boldsymbol{X}$ produce the same result $\hat{\boldsymbol{Y}}$; moreover different values of the parameters lead to the same model for $\mathcal{E}(\boldsymbol{Y})$. In that case we say that the vector of parameters $\boldsymbol{\beta}$ is not identifiable. In the sequel we always assume that $\boldsymbol{X}$ has rank p.

Example 2. Continuation of **example 1**.
In this case

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 10 & 0 \\ 0 & 11 \end{pmatrix},$$

hence

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ &= \begin{pmatrix} \frac{1}{10}\sum_{j=1}^{10} Y_{1j} \\ \frac{1}{11}\sum_{j=1}^{11} Y_{2j} \end{pmatrix} \\ &= \begin{pmatrix} \bar{Y}_{1.} \\ \bar{Y}_{2.} \end{pmatrix}. \end{aligned}$$

A dot in a subscript means that the average was taken over the corresponding subscript.
The least squares estimators are in this case the two sample averages.

The least squares estimator $\hat{\boldsymbol{\beta}}$ has the nice property that it is unbiased and that it is the best linear unbiased estimator in the sense that it has the 'smallest' covariance matrix of all linear unbiased estimators of $\boldsymbol{\beta}$.
'Linear' means here: it is a linear function of $\boldsymbol{Y}$ and we say that one (covariance) matrix is smaller that another if their difference matrix is positive (semi-)definite.

Theorem 1. The Gauss-Markov theorem.
We assume that the assumptions **(1) – (4)** *or* **(a) – (f)** *for a linear model hold. Moreover we assume that* $\boldsymbol{X}$ *has rank* p*, hence that* $\boldsymbol{X}^T\boldsymbol{X}$ *is non-singular. Then*

i. $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ *is an unbiased estimator for* $\boldsymbol{\beta}$*,*
ii. $var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$*,*
iii. *The difference of the covariance matrix of an arbitrary linear unbiased estimator for* $\boldsymbol{\beta}$ *and the covariance matrix of* $\hat{\boldsymbol{\beta}}$ *is positive semi-definite.*

Proof.

i. See **exercise 6**.
ii.
$$\begin{aligned} \operatorname{var}(\hat{\boldsymbol{\beta}}) &= \operatorname{var}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[\operatorname{var}(\boldsymbol{Y})]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned}$$

iii. We compare the covariance matrices of $\hat{\boldsymbol{\beta}}$ and of an arbitrary linear unbiased estimator $\boldsymbol{AY}$.
Define the matrix $\boldsymbol{D} = \boldsymbol{A} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.
Because $\boldsymbol{AY}$ is an unbiased estimator for $\boldsymbol{\beta}$, we find that
$\boldsymbol{DX\beta} = \mathcal{E}(\boldsymbol{DY}) = \mathcal{E}(\boldsymbol{AY}) - \mathcal{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = \boldsymbol{\beta} - \boldsymbol{\beta} = \boldsymbol{0}$ for all $\boldsymbol{\beta}$, hence $\boldsymbol{DX} = \boldsymbol{0}$. Moreover:
$$\begin{aligned} &\operatorname{var}(\boldsymbol{AY}) = \boldsymbol{A}[\operatorname{var}(\boldsymbol{Y})]\boldsymbol{A}^T \\ &= [(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D}](\sigma^2\boldsymbol{I})[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} + \boldsymbol{D}^T] \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} + \sigma^2\boldsymbol{DD}^T. \end{aligned}$$
The difference matrix $\sigma^2\boldsymbol{DD}^T$ is positive semi-definite.

Example 3. The estimator from **example 2** is an unbiased estimator for $\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$. Moreover, this is the 'best' linear unbiased estimator.

In order to estimate σ^2 we base ourselves on the sum of squares of the residuals $(\boldsymbol{Y}-\hat{\boldsymbol{Y}})^T(\boldsymbol{Y}-\hat{\boldsymbol{Y}}) = \|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2$. Under the same assumptions as for the **Gauss-Markov theorem** with rank $\boldsymbol{X} = p$, we find that (see **exercise 4**) $S^2 = \frac{1}{n-p}\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2$ is an unbiased estimator for σ^2.
In the proof of the Gauss-Markov theorem we found that $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and because $S^2 = \frac{1}{n-p}\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2$ is an unbiased estimator for σ^2, an unbiased estimator for the covariance matrix of $\hat{\boldsymbol{\beta}}$ is: $S^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

If we want to estimate a linear parameter function $\boldsymbol{c}^T\boldsymbol{\beta}$, then it follows from the **Gauss-Markov theorem** that $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator and that its variance equals $\sigma^2\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}$. So if we want to estimate for example β_i, we take: $\hat{\beta}_i$. This estimator has variance: $\sigma^2\times$ the (i,i)-th element of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Example 4. Continuation of **example 1 - 3**.
An unbiased estimator for σ^2 is the 'pooled variance' $S^2 = \frac{1}{19}\left[\sum_{j=1}^{10}(Y_{1j}-\bar{Y}_{1.})^2 + \sum_{j=1}^{11}(Y_{2j}-\bar{Y}_{2.})^2\right]$.
The variance of the estimator $\hat{\mu}_1$ is $\frac{\sigma^2}{10}$.

As an alternative to the least squares criterion we can also apply the maximum likelihood criterion for deriving estimators. To this end we have to assume a (normal) distribution for the observations.

Assuming normality: $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$,
and rank $\boldsymbol{X} = p$, the log-likelihood is:

$$\ln L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{Y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2}{2\sigma^2}. \tag{2.3}$$

Given the value of σ^2, this function attains its maximum if $\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2$ is at its minimum; therefore the ML-criterion gives the same estimator for $\boldsymbol{\beta}$ as the LS-criterion.

Equating the derivative with respect to σ^2 to 0 the ML-estimator for σ^2 is derived:

$$\hat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{n-p}{n}S^2. \tag{2.4}$$

The optimality properties of the maximum likelihood estimators are stronger (for example UMVU) than the optimality properties found in the **Gauss-Markov theorem**. Even if no assumptions are made about the type of distribution of the observations, we are still able to derive the **expectation and the variance of the least squares estimators**.
But thanks to the assumption of normality we are able to derive the exact **distribution of the estimators** for $\boldsymbol{\beta}$ and σ^2.

If $\boldsymbol{Y}$ has a normal distribution then its density function f is:

$$f(\boldsymbol{Y};\boldsymbol{\beta},\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\boldsymbol{X\beta})\right]$$
$$= a(\boldsymbol{\beta},\sigma^2)\exp\left[\sum_{i=1}^{p}\tfrac{\beta_i}{\sigma^2}Z_i - \tfrac{1}{2\sigma^2}\boldsymbol{Y}^T\boldsymbol{Y}\right]$$

where $\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{Y}$. This density function belongs to the exponential family. If there are no restrictions for $\boldsymbol{\beta}$ and σ^2, then the statistic $t(\boldsymbol{Y}) = (Z_1,\ldots,Z_p,\boldsymbol{Y}^T\boldsymbol{Y})$ is complete and minimal sufficient, hence according to the Lehmann-Scheffé theorem an unbiased estimator for (a function of) $\boldsymbol{\beta}$ and/or σ^2 is UMVU (uniformly minimal variance unbiased). Without the assumption of normality the LS estimator for $\boldsymbol{\beta}$ would only be UMVU in the class of linear estimators.

If $\boldsymbol{X}$ has rank p and $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta},\sigma^2\boldsymbol{I})$, then $\hat{\boldsymbol{\beta}}$ is also normally distributed because it is a linear transformation of $\boldsymbol{Y}$:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}). \tag{2.5}$$

Now we derive the distribution of (a constant times) $\hat{\sigma}^2$ by rewriting:

$$\frac{n}{\sigma^2}\hat{\sigma}^2 = \frac{1}{\sigma^2}[(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}]^T[(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}]$$
$$= \frac{1}{\sigma}\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\frac{1}{\sigma}\boldsymbol{Y}.$$

$\frac{1}{\sigma}\boldsymbol{Y} \sim \mathcal{N}_n(\frac{1}{\sigma}\boldsymbol{X\beta},\boldsymbol{I})$ and $\boldsymbol{I}-\boldsymbol{P}$ is idempotent with rank $n-p$ (**exercise 3**), hence:

$$\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} \sim \chi'^2(n-p;\lambda)$$

with $\lambda = \frac{1}{\sigma}\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X}\boldsymbol{\beta}\frac{1}{\sigma} = 0$.

Thus

$$\frac{n}{\sigma^2}\hat{\sigma}^2 = \frac{n-p}{\sigma^2}S^2 \sim \chi^2_{n-p}.$$

If we define: $\boldsymbol{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, then $\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{0}$. Now $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent.

Example 5. Continuation of **example 1 – 4**. If the observations Y_{ij} are normally distributed, then $\hat{\mu}_1$ is the UMVU-estimator for μ_1.
Moreover $\hat{\mu}_1$ has a $\mathcal{N}(\mu_1, \frac{\sigma^2}{10})$-distribution and

$$\frac{1}{\sigma^2}\left[\sum_{j=1}^{10}(Y_{1j}-\bar{Y}_{1.})^2+\sum_{j=1}^{11}(Y_{2j}-\bar{Y}_{2.})^2\right]$$

has a χ^2_{19}-distribution.
This variable and $\hat{\mu}_1$ are independent.

Under the **assumptions (1) – (5) or (a) – (f)** the vector of expectations $\mathcal{E}(\boldsymbol{Y})$ lies in a linear subspace $\mathcal{L}$ of R^n. We will derive tests for two kinds of null hypotheses.

In the **first kind** the null hypothesis has the form: the vector $\mathcal{E}(\boldsymbol{Y})$ lies in a linear subspace $\mathcal{L}_0$ of the space $\mathcal{L}$. The alternative is that the expectation does not lie in $\mathcal{L}_0$, but in its relative complement $\mathcal{L} \setminus \mathcal{L}_0$.

In the **second kind** the null hypothesis states that a linear combination of the coefficients $\boldsymbol{\beta}$ equals a constant k. With this null hypothesis also one-sided alternatives can be formulated. Moreover, confidence intervals can be constructed for $\boldsymbol{c}^T\boldsymbol{\beta}$.

All tests and confidence intervals are derived under the **assumptions (1) – (5) or (a) – (f)** for linear models, so the assumption of a normal distribution for the observations is included.

Under the assumption that the n observations are normally distributed, i.e. $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, we derive by geometrical arguments likelihood ratio tests for the null hypothesis that the vector of expectations $\mathcal{E}(\boldsymbol{Y})$ lies in a certain linear subspace $\mathcal{L}_0$ with dimension r against the alternative hypothesis that $\mathcal{E}(\boldsymbol{Y})$ does not lie in $\mathcal{L}_0$ but in a linear subspace $\mathcal{L}$ (with dimension $q > r$) which contains $\mathcal{L}_0$ as a subspace (see the **figure**; here the expectation is drawn in $\mathcal{L}_0$).

Moreover we derive confidence intervals and likelihood ratio tests for linear combinations of the parameters $\boldsymbol{\beta}$.

Maximum likelihood estimators under the null hypothesis will be denoted by a dot and without the restrictie of the null hypothesis by a hat.

The projection of $\boldsymbol{Y}$ on $\mathcal{L}$ will be called $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and the projection of $\boldsymbol{Y}$ (and hence also of $\hat{\boldsymbol{Y}}$) on $\mathcal{L}_0$ will be called $\dot{\boldsymbol{Y}} = \boldsymbol{X}_0\dot{\boldsymbol{\beta}}$.

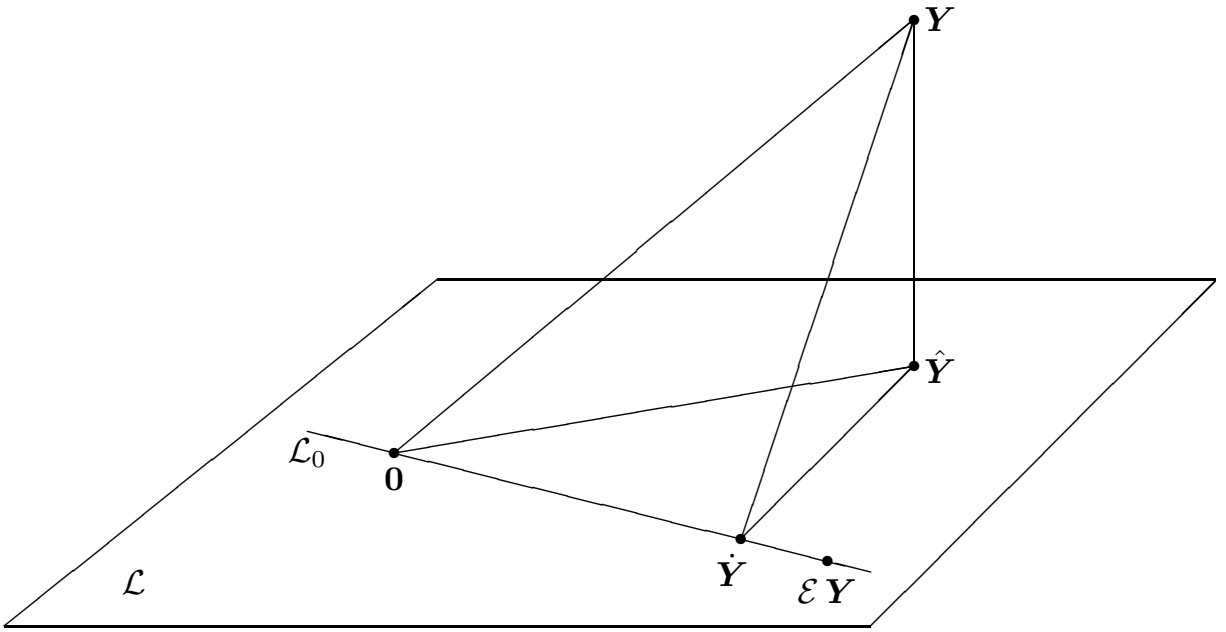

Figure 3.1 Geometrical representation of a linear hypothesis test.

The (generalised) likelihood-ratio test of level α for the testing problem:
$H_0 : \mathcal{E}(\boldsymbol{Y}) \in \mathcal{L}_0$ against $H_a : \mathcal{E}(\boldsymbol{Y}) \in \mathcal{L} \setminus \mathcal{L}_0$
rejects H_0 if

$$\frac{f(\boldsymbol{Y}; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)}{f(\boldsymbol{Y}; \dot{\boldsymbol{\beta}}, \dot{\sigma}^2)} > k$$

where k is chosen such that the test has level α.

This condition corresponds with (see **formula (2.3)**):

$$\frac{(2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}{2\hat{\sigma}^2}\right]}{(2\pi\dot{\sigma}^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{Y} - \dot{\boldsymbol{Y}}\|^2}{2\dot{\sigma}^2}\right]} > k$$

or (see **formula (2.4)**):

$$\left(\frac{\dot{\sigma}^2}{\hat{\sigma}^2}\right)^{\frac{n}{2}} = \left(\frac{\|\boldsymbol{Y} - \dot{\boldsymbol{Y}}\|^2}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}\right)^{\frac{n}{2}} > k.$$

This is equivalent to $\dfrac{\|\boldsymbol{Y} - \dot{\boldsymbol{Y}}\|^2}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2} > k^{\frac{2}{n}}$

and to

$$\frac{\|\boldsymbol{Y} - \dot{\boldsymbol{Y}}\|^2 - \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2} = \frac{\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|^2}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2} > k^{\frac{2}{n}} - 1.$$

The equality of the numerators follows from the application of Pythagoras' theorem on the triangle $(\boldsymbol{Y}, \hat{\boldsymbol{Y}}, \dot{\boldsymbol{Y}})$.

The statistics $\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2$ and $\|\boldsymbol{Y} - \dot{\boldsymbol{Y}}\|^2$ are called the residual sum of squares under the general model and under the null hypothesis.

From the **Fisher-Cochran theorem** it follows that under the null hypothesis: $\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|^2/\sigma^2 \sim \chi^2_{q-r}$ and $\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/\sigma^2 \sim \chi^2_{n-q}$; these quantities are independent.

Therefore under H_0 the quotient

$$F = \frac{\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|^2/(q-r)}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-q)}$$

has an $F_{q-r,n-q}$-distribution.
This statistic is used as a test statistic.

Example 6. Continuation of **example 1 - 5**.
For testing the the null hypothesis $\mu_1 = \mu_2$ against $\mu_1 \neq \mu_2$ we compute the projections of $\boldsymbol{Y}$ on $\mathcal{L}$ and on $\mathcal{L}_0$. In **example 2** we already computed $\hat{\boldsymbol{\beta}}$ and thus $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Under H_0 is $\mu_1 = \mu_2$ $(= \mu)$, hence $\mathcal{E}(\boldsymbol{Y}) = (\mu, \ldots, \mu, \mu, \ldots, \mu)^T = \boldsymbol{X}_0\boldsymbol{\beta}$. Here $\boldsymbol{X}_0$ is the 'matrix' with only one column: $(1, \ldots, 1)^T$ and $\boldsymbol{\beta}$ a vector of only one element, μ. Under H_0 we therefore have: $\dot{\boldsymbol{\beta}} = \hat{\mu} = (\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^T\boldsymbol{Y} = \frac{1}{21}\left[\sum_{j=1}^{10} Y_{1j} + \sum_{j=1}^{11} Y_{2j}\right] = \bar{Y}_{..}$.
Summarising:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{Y}_{1.} \\ \vdots \\ \bar{Y}_{1.} \\ \bar{Y}_{2.} \\ \vdots \\ \bar{Y}_{2.} \end{pmatrix} \quad \text{and} \quad \dot{\boldsymbol{Y}} = \boldsymbol{X}_0\dot{\boldsymbol{\beta}} = \begin{pmatrix} \bar{Y}_{..} \\ \vdots \\ \bar{Y}_{..} \\ \bar{Y}_{..} \\ \vdots \\ \bar{Y}_{..} \end{pmatrix}.$$

For the dimensions of the subspaces we find: $n = 21$, $q = 2$, $r = 1$. The test statistic is

$$\begin{aligned} F &= \frac{\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|^2/(q-r)}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-q)} \\ &= \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\hat{Y}_{ij} - \dot{Y}_{ij})^2/1}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - \hat{Y}_{ij})^2/19} \\ &= \frac{\sum_{i=1}^{2} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2/1}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2/19}. \end{aligned}$$

The numerator of this statistic represents the variation *between* the group means, weighted with the group size.
The denominator is the pooled estimator of the common variance σ^2 *within* the groups.

In order to calculate the test statistic's outcome we only need the squared distances $\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2$ and $\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|^2$ and the dimensions of the linear (sub)spaces in which the various vectors are supposed to lie: $n-q$ for the vector $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ and $q-r$ for the vector $\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}$.

The test statistic depends on the rate $\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|$ to $\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|$. From **figure 3.1** it becomes clear that we tend to reject the null hypothesis when $\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|$ is large compared to $\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|$ and accept it when this rate is small. The results are often displayed in an ANOVA table.

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F
Hypothesis or Treatment	$q-r$	$\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|^2$	$\dfrac{\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|^2}{q-r}$	$\dfrac{\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|^2/(q-r)}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-q)}$
Error	$n-q$	$\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2$	$\dfrac{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}{n-q}$	
Total	$n-r$	$\|\boldsymbol{Y} - \dot{\boldsymbol{Y}}\|^2$		

Table 3.1 ANOVA table

Assuming that $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ we can derive a **test** for testing $H_0 : \boldsymbol{c}^T\boldsymbol{\beta} = k$ against:
$H_a : \boldsymbol{c}^T\boldsymbol{\beta} > k$, $H_a : \boldsymbol{c}^T\boldsymbol{\beta} < k$ or $H_a : \boldsymbol{c}^T\boldsymbol{\beta} \neq k$.
We can also derive a confidence interval for $\boldsymbol{c}^T\boldsymbol{\beta}$.

Let $\boldsymbol{c}^T\hat{\boldsymbol{\beta}} = \boldsymbol{a}^T\boldsymbol{Y}$ be an unbiased linear estimator for $\boldsymbol{c}^T\boldsymbol{\beta}$. Then the linear combination $\boldsymbol{a}^T\boldsymbol{Y}$ has a normal distribution with expectation $\boldsymbol{c}^T\boldsymbol{\beta}$ and variance $\boldsymbol{a}^T\boldsymbol{a}\sigma^2$. The variable $(\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{c}^T\boldsymbol{\beta})/\sqrt{\boldsymbol{a}^T\boldsymbol{a}\sigma^2}$ then has a standard normal distribution.
It can be shown that this variable is independent from $\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/\sigma^2$, which has a χ^2_{n-q}-distribution. Hence the variable

$$T = \frac{(\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{c}^T\boldsymbol{\beta})/\sqrt{\boldsymbol{a}^T\boldsymbol{a}}}{\sqrt{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-q)}}$$

has a t-distribution with $n - q$ degrees of freedom.
We can use T als a test statistic for testing the null hypothesis $\boldsymbol{c}^T\boldsymbol{\beta} = k$ against the alternatives $\boldsymbol{c}^T\boldsymbol{\beta} > k$, $\boldsymbol{c}^T\boldsymbol{\beta} < k$ or $\boldsymbol{c}^T\boldsymbol{\beta} \neq k$.

When $\boldsymbol{c}^T\boldsymbol{\beta}$ is not specified we can use T as a pivotal quantity for constructing a confidence interval for $\boldsymbol{c}^T\boldsymbol{\beta}$.

Example 7. (Continuation of example 6.)
An interesting linear combination of the parameters is $\mu_1 - \mu_2$ (the vector $\boldsymbol{c}$ is now $(1\ -1)^T$).

An unbiased linear estimator for $\mu_1 - \mu_2$ is $\bar{Y}_{1.} - \bar{Y}_{2.} = \boldsymbol{a}^T\boldsymbol{Y}$ with $\boldsymbol{a}^T = (\frac{1}{10}, \ldots, \frac{1}{10}, -\frac{1}{11}, \ldots, -\frac{1}{11})$. This estimator has a normal distribution with variance $(\frac{1}{10} + \frac{1}{11})\sigma^2$.

For testing the null hypothesis $\mu_1 - \mu_2 = 0$ against the one-sided alternative $\mu_1 - \mu_2 > 0$ we construct the test statistic

$$\begin{aligned} T &= \frac{(\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{c}^T\boldsymbol{\beta})/\sqrt{\boldsymbol{a}^T\boldsymbol{a}}}{\sqrt{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-q)}} \\ &= \frac{(\bar{Y}_{1.} - \bar{Y}_{2.})/\sqrt{\frac{1}{10} + \frac{1}{11}}}{\sqrt{\sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2/19}} \\ &= \frac{\bar{Y}_{1.} - \bar{Y}_{2.}}{S\sqrt{\frac{1}{10} + \frac{1}{11}}}, \end{aligned}$$

where S^2 = the pooled variance = $\dfrac{9S_1^2 + 10S_2^2}{19}$.

This is the classical test statistic for the two samples problem (equal variances assumed). This statistic has under H_0 a t_{19}-distribution.

A 95%-confidence interval for $\mu_1 - \mu_2$ has the limits:

$$\bar{Y}_{1.} - \bar{Y}_{2.} \pm 2.093S.$$

Comparing the expectations of k populations is called one way ANOVA. An obvious hypothesis to test would be that all k populations have identical expectations. The corresponding likelihood ratio test **was already derived** for a more general case. Here we have observations $Y_{i1}, \ldots, Y_{in_i}$ from population i. All Y_{ij} are assumed independent and normally distributed with expectations $\mathcal{E}(Y_{ij}) = \mu_i$ and equal variances σ^2. More concise: $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ with

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}.$$

Matrix $\boldsymbol{X}$ and hence also $\boldsymbol{X}^T\boldsymbol{X}$ have rank k.
The ML-estimator for the vector of μ_i's is $(\hat{\mu}_1, \ldots, \hat{\mu}_k)^T = \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = (\bar{Y}_{1.}, \ldots, \bar{Y}_{k.})^T$.

For testing H_0: $\mu_1 = \ldots = \mu_k$ we need the dimensions of the linear subspaces $\mathcal{L}$ and $\mathcal{L}_0$ and the projections of $\boldsymbol{Y}$ onto these spaces.

The vector of observations $\boldsymbol{Y}$ lies in a space with dimension $n = n_1 + \ldots + n_k$;
the vector $\mathcal{E}(\boldsymbol{Y})$ lies under H_a in the space $\mathcal{L}$, spanned by the column vectors of $\boldsymbol{X}$, with dimension $q = k$,
and under the null hypothesis in the space $\mathcal{L}_0$, spanned by the vector $(1, \ldots, 1)^T$, with dimension $r = 1$.

The projections of $\boldsymbol{Y}$ are
$\hat{\boldsymbol{Y}} = (\bar{Y}_{1.}, \ldots, \bar{Y}_{1.}, \bar{Y}_{2.}, \ldots, \bar{Y}_{2.}, \ldots, \bar{Y}_{k.}, \ldots, \bar{Y}_{k.})^T$ and
$\dot{\boldsymbol{Y}} = (\bar{Y}_{..}, \ldots, \bar{Y}_{..}, \bar{Y}_{..}, \ldots, \bar{Y}_{..}, \ldots, \bar{Y}_{..}, \ldots, \bar{Y}_{..})^T$. A part of the ANOVA table is shown **below**.

Source of variation	Degrees of freedom *(df)*	Sum of squares *(SS)*	Mean squares *(MS)*
Groups	$k-1$	$\sum n_i(\bar{Y}_{i.}-\bar{Y}_{..})^2$	$\sum n_i(\bar{Y}_{i.}-\bar{Y}_{..})^2/(k-1)$
Error	$n-k$	$\sum\sum(Y_{ij}-\bar{Y}_{i.})^2$	$\sum\sum(Y_{ij}-\bar{Y}_{i.})^2/(n-k)$
Total	$n-1$	$\sum\sum(Y_{ij}-\bar{Y}_{..})^2$	

Table 4.1 ANOVA table for one way ANOVA

The test statistic

$$F = \frac{\|\hat{\boldsymbol{Y}}-\dot{\boldsymbol{Y}}\|^2/(q-r)}{\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2/(n-q)} = \frac{\sum\sum(\hat{Y}_{ij}-\dot{Y}_{ij})^2/(k-1)}{\sum\sum(Y_{ij}-\hat{Y}_{ij})^2/(n-k)} = \frac{\sum n_i(\bar{Y}_{i.}-\bar{Y}_{..})^2/(k-1)}{\sum\sum(Y_{ij}-\bar{Y}_{i.})^2/(n-k)}$$

has under H_0 a $F_{k-1,n-k}$-distribution.
We tend to reject the null hypothesis when the variation *between* the groups (numerator) is large, compared to the variation *within* the groups (denominator).

Instead of the k parameters $\mu_1,\ldots,\mu_k$ sometimes the $k+1$ parameters $\mu,\tau_1,\ldots,\tau_k$ are chosen with the restriction $\sum n_i\tau_i = 0$. Here $\mu = \sum n_i\mu_i/n$ is the over-all expectation for all populations and $\tau_i = \mu_i - \mu$ the deviation from μ in population i.

The maximum likelihood estimators for the new parameters can be expressed in terms of the old ones: $\hat{\mu} = \sum n_i\hat{\mu}_i/n$ en $\hat{\tau}_i = \hat{\mu}_i - \hat{\mu}$.

The test for the corresponding null hypothesis H_0 : $\tau_1 = \ldots = \tau_k = 0$ is not affected by a change in parametrisation.

Usually we will not be satisfied by the conclusion that expectations in the k populations differ. We want to know: which (group of) population(s) differs from which other (group)?

Pairwise comparisons of k populations involves $k(k-1)/2$ tests; the number of all possible linear contrasts of the form $\sum a_i\mu_i$ (with $\sum a_i = 0$) is even infinite. Moreover the tests are not independent. If each test is applied with level of significance α, then the probability that at least one null hypothesis is falsely rejected cannot be exactly determined but is certainly larger than α.

Two solutions are presented in the form of simultaneous confidence intervals (equivalent to simultaneous tests): according to Bonferroni and according to Scheffé.

The **Bonferroni** inequality tells us that for m events $A_1, \ldots, A_m$ (which need not be disjoint):

$$\mathrm{P}(A_1 \cap A_2 \cap \ldots, \cap A_m) \geq 1 - \sum_{i=1}^{n} \mathrm{P}(A_i^c).$$

If $A_1, \ldots, A_m$ are confidence intervals, each with confidence level $1 - \frac{\alpha}{m}$, then the probability of their intersection

$A_1 \cap A_2 \cap \ldots, \cap A_m$ is at least $1 - \alpha$ so that the confidence level of the combined intervals is also at least $1 - \alpha$.

The method of **Scheffé** even allows for an unlimited number of contrasts leaving the total confidence at least at level $1 - \alpha$. Scheffé proved that

$$\mathrm{P}\left(\sum a_i\bar{Y}_{i.} - S\sqrt{b} \leq \sum a_i\mu_i \leq \sum a_i\bar{Y}_{i.} + S\sqrt{b} \quad \text{for each contrast } \boldsymbol{a}\right) \geq \alpha$$

where $b = (k-1)f_\alpha \sum \frac{a_i^2}{n_i}$.

This defines simultaneous confidence intervals for all possible contrasts $\sum a_i\mu_i$ with total confidence level at least $1 - \alpha$.

Example 8. In the table below we find the yields of a certain process, determined by four different analysts (from [Lindgren], p.522-524).

Three simultaneous confidence intervals were constructed for $\mu_4 - \mu_3$, $\mu_2 - \mu_1$ and $\frac{1}{3}(\mu_1 + \mu_2 + \mu_4) - \mu_3$ (with a confidence level of at least 95%) according the method of Scheffé. The quantity S is the square root of the pooled variance; its outcome is $\sqrt{11.47}$. For the computation of b for the three contrasts we use $k-1 = 3$ and $f_\alpha = 3.34$. For example the third contrast is $\sum \frac{a_i^2}{n_i} = \frac{1/9}{6} + \frac{1/9}{5} + \frac{1/9}{4} + \frac{1}{3} = 0.402$, hence $b = 4.026$. The remaining values of b are 5.845 and 3.674.

For each contrast $\sum a_i\mu_i$ we find the corresponding interval $\sum a_i Y_i \pm S\sqrt{b}$. This gives successively: 5 ± 8.19, 5.93 ± 6.49 and 4.91 ± 6.79.

If we are only interested in the six possible *differences* $\mu_i - \mu_j$, we can apply Bonferroni's method with $m = 6$. We construct $(100 - \frac{5}{6})\%$-confidence intervals for each of the six differences $\mu_i - \mu_j$. Their formulas are given by $\bar{Y}_{i.} - \bar{Y}_{j.} \pm t^* S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, where t^* is the $(100 - \frac{5}{6})\%$- point of the t-distribution with $n - k = 18 - 4$ degrees of freedom.

Analyst	Yield	n_i	$\bar{Y}_i$	$\sum(Y_{ij} - \bar{Y}_i)^2$	$n_i(Y_i - \bar{Y})^2$
1	8, 5, −1, 6, 5, 3	6	4.33	47.33	2.67
2	7, 12, 5, 3, 10	5	7.4	53.2	28.8
3	4, −2, 1	3	1	18	48
4	1, 6, 10, 7	4	6	42	4

In a two way ANOVA we have $k = rc$ populations which can be arranged according to *two* factors in a table with r rows and c columns. From each population we have s observations. This is a so-called balanced experiment: the number of observations per cell is constant. The theory that follows does not hold for unbalanced experiments.

We distinguish between the case that we have more than one observation per cell and the case that we have only **one observation per cell**.

More than one observation per cell.

Suppose we have rc populations which we can arrange in a table with r rows and c columns. From each population we have s observations, where $s > 1$. We will derive tests for

- the null hypothesis that for each row there is no difference between the expectations within the row (no column effect),
- the null hypothesis that for each column there is no difference between the expectations within the column (no row effect),
- the null hypothesis that the expectation in a cell is the sum of a general level plus a roweffect plus a columneffect (no interaction).

The observations Y_{ijk} $(i = 1,\ldots,r;\ j = 1,\ldots,c;\ k = 1,\ldots,s)$ are independent and normally distributed with expectations $\mathcal{E}(Y_{ijk}) = \mu_{ij}$ and equal variance σ^2. In contrast with one factor ANOVA we now write the various populations with two instead of one index and the number of observations per sample is constant. Therefore the ML-estimator for μ_{ij} is again the average of the sample from the i,j-th population:
$\hat{\mu}_{ij} = \frac{1}{s}\sum_k Y_{ijk} = \bar{Y}_{ij.}$.

A more common parametrisation is with the $1 + r + c + rc$ parameters $\mu, \theta_1,\ldots,\theta_r, \phi_1,\ldots,\phi_c, \xi_{11},\ldots,\xi_{rc}$ and the $1+1+c+r-1$ restrictions: $\sum_i \theta_i = 0$, $\sum_j \phi_j = 0$, $\sum_i \xi_{ij} = 0$, $\sum_j \xi_{ij} = 0$. The connection between both parametrisations is given by

$$\mu_{ij} = \mu + \theta_i + \phi_j + \xi_{ij}.$$

By summing this equality over i and j we find:

$$\sum_i \sum_j \mu_{ij} = rc\mu$$

$$\sum_j \mu_{ij} = c\mu + c\theta_i$$

$$\sum_i \mu_{ij} = r\mu + r\phi_j.$$

Now we can express the new parameters μ, θ_i, ϕ_j and ξ_{ij} in the μ_{ij}'s. The corresponding maximum likelihood estimates are then:

$$\hat{\mu} = \frac{1}{rc}\sum\sum \hat{\mu}_{ij} = \bar{Y}_{...} \tag{4.1}$$

$$\hat{\theta}_i = \frac{1}{c}\sum_j \hat{\mu}_{ij} - \hat{\mu} = \bar{Y}_{i..} - \bar{Y}_{...} \tag{4.2}$$

$$\hat{\phi}_j = \frac{1}{r}\sum_i \hat{\mu}_{ij} - \hat{\mu} = \bar{Y}_{.j.} - \bar{Y}_{...} \tag{4.3}$$

$$\hat{\xi}_{ij} = \hat{\mu}_{ij} - \hat{\theta}_i - \hat{\phi}_j - \hat{\mu} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...} \tag{4.4}$$

An interesting null hypothesis could be 'no interaction' (all $\xi_{ij} = 0$) against the alternative that not all ξ_{ij}'s are equal to 0. Then under the null hypothesis $\mathcal{E}(Y_{ijk}) = \mu + \theta_i + \phi_j$ holds (in total $1 + r + c$ parameters) with the two restrictions $\sum_i \theta_i = 0$ and $\sum_j \phi_j = 0$. Under H_0 the vector $\mathcal{E}(\boldsymbol{Y})$ lies again in a linear subspace $\mathcal{L}_0$ of $\mathcal{L}$. However, in order to derive its projection $\dot{\boldsymbol{Y}}$ either additional theory is needed for least squares estimation under linear restrictions or the problem should be restated with exactly $r + c - 1$ new parameters. Anyway, the outcome appears to be $\dot{Y}_{ijk} = \hat{\mu} + \hat{\theta}_i + \hat{\phi}_j$, with $\hat{\mu}$, $\hat{\theta}_i$ and $\hat{\phi}_j$ from **formula (4.1)** through **(4.3)**. $\boldsymbol{Y}$ lies in a space with dimension rcs, the spaces $\mathcal{L}$ and $\mathcal{L}_0$ have dimensions rc and $r + c - 1$. The test statistic F has under H_0 an $F_{rc-r-c+1,rcs-rc} = F_{(r-1)(c-1),rc(s-1)}$-distribution:

$$
\begin{aligned}
F &= \frac{\|\hat{\boldsymbol{Y}} - \dot{\boldsymbol{Y}}\|^2/(r-1)(c-1)}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/rc(s-1)} \\
&= \frac{\sum\sum\sum(\hat{Y}_{ijk} - \dot{Y}_{ijk})^2/(r-1)(c-1)}{\sum\sum\sum(Y_{ijk} - \hat{Y}_{ijk})^2/rc(s-1)} \\
&= \frac{\sum\sum s(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2/(r-1)(c-1)}{\sum\sum\sum(Y_{ijk} - \bar{Y}_{ij.})^2/rc(s-1)}.
\end{aligned}
$$

If we do not reject the null hypothesis of no interaction (we henceforth assume that the ξ_{ij}'s are equal to 0), it may be interesting to test the null hypothesis that there is no row effect (H_{0A} : $\theta_i = 0$ for all i) or the hypothesis that there is no column effect ($H_{0B} : \phi_j = 0$ for all j).
The linear subspaces corresponding with these hypotheses are shown in **figure 4.1**, in which the original vector $\boldsymbol{Y}$ of observations is left out. The figure is restricted to the space $\mathcal{L}$ with dimension rc. In this space we find the point $\hat{\boldsymbol{Y}}$.

Within the space $\mathcal{L}$ we find the linear subspace $\mathcal{L}_0$ with dimension $r + c - 1$ and in its interior the linear subspaces $\mathcal{L}_A$, which contains the vector $\mathcal{E}(\boldsymbol{Y})$ under the null hypothesis $H_{0A} : \mathcal{E}(Y_{ijk}) = \mu + \phi_j$, and $\mathcal{L}_B$, which contains the vector $\mathcal{E}(\boldsymbol{Y})$ under the null hypothesis $H_{0B} : \mathcal{E}(Y_{ijk}) = \mu + \theta_i$. Their intersection is the linear subspace $\mathcal{L}_{AB}$, given by $\mathcal{E}(Y_{ijk}) = \mu$, so spanned by the vector with all components equal to 1.

In **figure 4.1** and **table 4.2** the projections of $\boldsymbol{Y}$ on the various linear subspaces can be found. The projection of $\hat{\boldsymbol{Y}}$ (and of $\boldsymbol{Y}$) on $\mathcal{L}_0$ is now called $\hat{\boldsymbol{Y}}_0$ instead of $\dot{\boldsymbol{Y}}$.

Description	$\mathcal{E}(Y_{ijk}) =$	space	dimension	vector	ijk-th comp.
		R^n	$n = rcs$	$\boldsymbol{Y}$	Y_{ijk}
interaction	$\mu + \theta_i + \phi_j + \xi_{ij}$	$\mathcal{L}$	rc	$\hat{\boldsymbol{Y}}$	$\bar{Y}_{ij.}$
no interaction	$\mu + \theta_i + \phi_j$	$\mathcal{L}_0$	$r + c - 1$	$\hat{\boldsymbol{Y}}_0$	$\bar{Y}_{i..} + \bar{Y}_{.j.} - \bar{Y}_{...}$
column effect only	$\mu + \phi_j$	$\mathcal{L}_A$	c	$\hat{\boldsymbol{Y}}_A$	$\bar{Y}_{.j.}$
row effect only	$\mu + \theta_i$	$\mathcal{L}_B$	r	$\hat{\boldsymbol{Y}}_B$	$\bar{Y}_{i..}$
no effects	μ	$\mathcal{L}_{AB}$	1	$\hat{\boldsymbol{Y}}_{AB}$	$\bar{Y}_{...}$

Table 4.2 Vectors, linear spaces and their dimensions

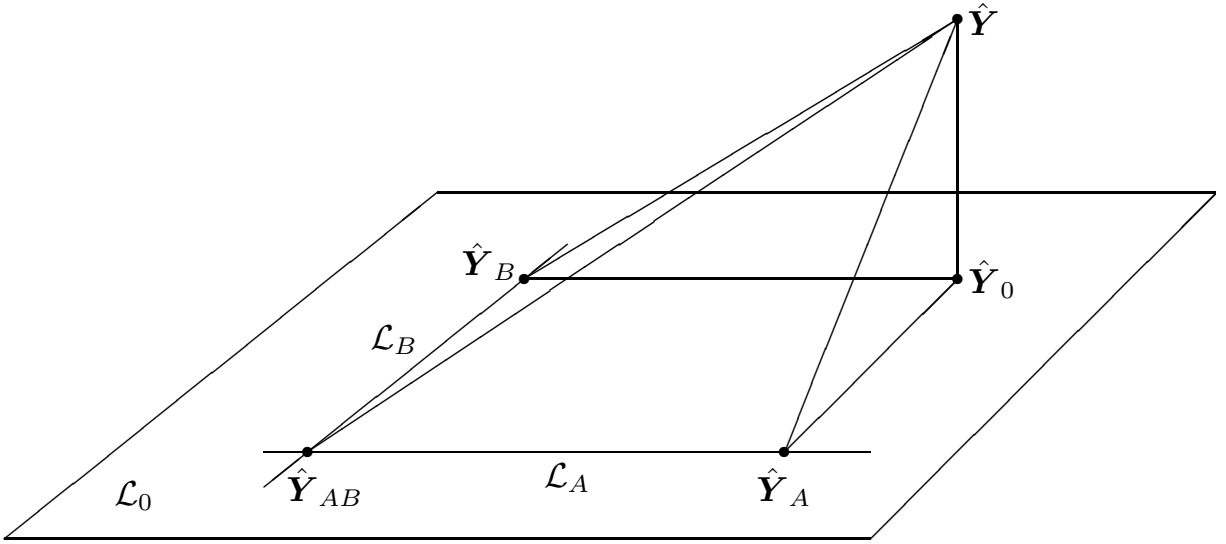

Figure 4.1 two factors, s observations per cell.

We already saw that the component ijk of the vector $\hat{\boldsymbol{Y}}_0$ is equal to $\bar{Y}_{i..} + \bar{Y}_{.j.} - \bar{Y}_{...}$. The space $\mathcal{L}_B$ is spanned by the r vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_r$, in which $\boldsymbol{e}_i$ is filled with ones in the places ijk $(j = 1, \ldots, c; k = 1, \ldots, s)$ and zeroes in the remaining places. From this we deduce that the projection $\hat{\boldsymbol{Y}}_B$ is a vector with the value $\bar{Y}_{i..}$ as its ijk-th component. In the same way we derive that $\hat{\boldsymbol{Y}}_A$ is a vector whose ijk-th component equals $\bar{Y}_{.j.}$. All components of the vector $\hat{\boldsymbol{Y}}_{AB}$ are equal to $\bar{Y}_{...}$.

The orthogonal complement of $\mathcal{L}_{AB}$ within $\mathcal{L}_A$ is orthogonal to the orthogonal complement of $\mathcal{L}_{AB}$ within $\mathcal{L}_B$ and the vector $\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_0$ is orthogonal to $\mathcal{L}_0$. Therefore Pythagoras' theorem holds as follows:

$$\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{AB}\|^2 = \|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_0\|^2 + \|\hat{\boldsymbol{Y}}_0 - \hat{\boldsymbol{Y}}_A\|^2 + \|\hat{\boldsymbol{Y}}_0 - \hat{\boldsymbol{Y}}_B\|^2.$$

By means of the **Fisher-Cochran theorem** it can be derived that the terms on the right hand side are independent and, after division by σ^2, all have a χ^2-distribution with $rc - r - c + 1 = (r-1)(c-1)$, $r-1$ and $c-1$ degrees of freedom respectively.

Moreover, all these terms are independent of $SSE = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2$, where SSE/σ^2 has a χ^2-distribution with $rc(s-1)$ degrees of freedom, as we saw earlier.
Because $H_0 : \mathcal{E}(\boldsymbol{Y}) \in \mathcal{L}_{AB}$ also fits into the theory of linear models, **we know** that

$$\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{AB}\|^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 + \|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{AB}\|^2.$$

In total we get: $SSA + SSB + SSI + SSE = SST$, where

$$SSA = \|\hat{\boldsymbol{Y}}_0 - \hat{\boldsymbol{Y}}_A\|^2 = \sum cs(\bar{Y}_{i..} - \bar{Y}_{...})^2,$$
$$SSB = \|\hat{\boldsymbol{Y}}_0 - \hat{\boldsymbol{Y}}_B\|^2 = \sum rs(\bar{Y}_{.j.} - \bar{Y}_{...})^2,$$
$$SSI = \|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_0\|^2 = \sum\sum s(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2,$$
$$SSE = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 = \sum\sum\sum(Y_{ijk} - \bar{Y}_{ij.})^2,$$
$$SST = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{AB}\|^2 = \sum\sum\sum(Y_{ijk} - \bar{Y}_{...})^2.$$

For testing the hypotheses H_0 (no interaction), H_{0A} (no row effect) and H_{0B} (no column effect) we again construct an ANOVA-table (see **table 4.3**).

All test statistics in the last column have an F-distribution under the corresponding null hypothesis with the obvious numbers of degrees of freedom.

Although the distinct sums of squares are independent, the various test statistics are not because the same denominator plays a role in all statistics.

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F
row effect	$r-1$	SSA	$\frac{SSA}{r-1}$	$\frac{SSA/(r-1)}{SSE/rc(s-1)}$
column effect	$c-1$	SSB	$\frac{SSB}{c-1}$	$\frac{SSB/(c-1)}{SSE/rc(s-1)}$
Interaction	$(r-1)(c-1)$	SSI	$\frac{SSI}{(r-1)(c-1)}$	$\frac{SSI/(r-1)(c-1)}{SSE/rc(s-1)}$
Error	$rc(s-1)$	SSE	$\frac{SSE}{rc(s-1)}$	
Total	$rcs-1$	SST		

Table 4.3 ANOVA-table, two factors, s observations per cell

Suppose we have rc populations which we can arrange in a table with r rows and c columns. From each population we now only have one observation. We want to derive tests for

- the null hypothesis that for each row there is no difference between the expectations within the row (no column effect),
- the null hypothesis that for each column there is no difference between the expectations within the column (no row effect),

Therefore, in this situation the null hypothesis of 'no interaction' cannot be tested.
Now we have in total rc observations $Y_{11}, \ldots, Y_{rc}$. If we are modeling: $\mathcal{E}(Y_{ij}) = \mu + \theta_i + \phi_j + \xi_{ij}$ (with the well-known restrictions), then the space $\mathcal{L}$ coincides with R^n and therefore $\hat{Y}_{ij} = Y_{ij}$ and $SSE = 0$. Then it is impossible to carry out tests of hypotheses for the various parameters.

A workable model will have to contain less parameters. A possible model is: $\mathcal{E}(Y_{ij}) = \mu + \theta_i + \phi_j$ with the restrictions $\sum \theta_i = 0$ and $\sum \phi_j = 0$.
In that case the space $\mathcal{L}$ is a linear subspace of R^n with dimension $r + c - 1$.
The ML-estimators for the parameters now become:

$$\hat{\mu} = \bar{Y}_{..}, \quad \hat{\theta}_i = \bar{Y}_{i.} - \bar{Y}_{..}, \quad \hat{\phi}_j = \bar{Y}_{.j} - \bar{Y}_{..}.$$

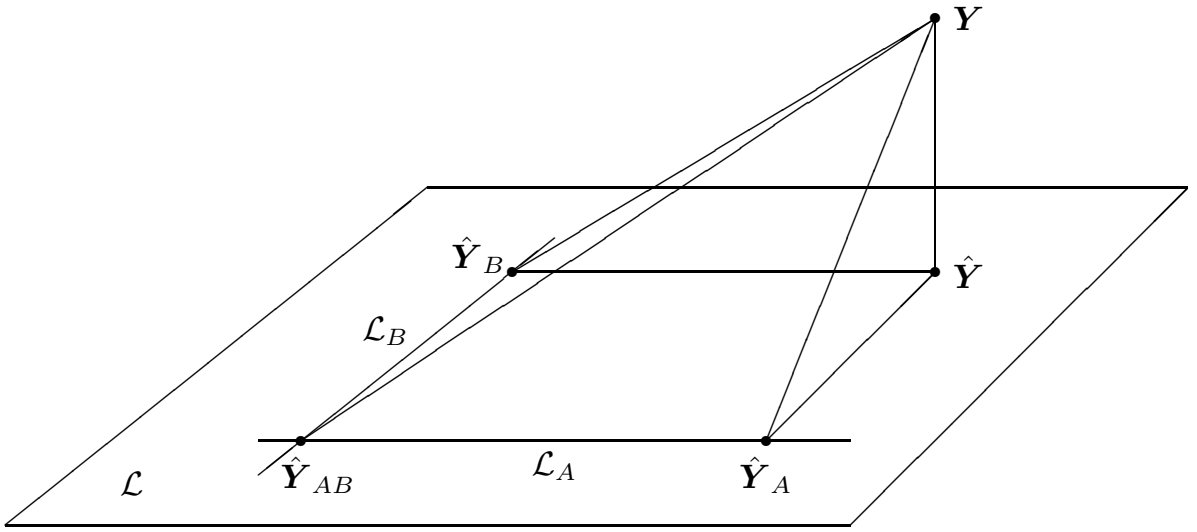

Figure 4.2 two factors, 1 observation per cell.

Description	$\mathcal{E}(Y_{ij}) =$	space	dimension	vector	ij-th component
		R^n	$n = rc$	$\boldsymbol{Y}$	Y_{ij}
row- and column effect	$\mu + \theta_i + \phi_j$	$\mathcal{L}$	$r + c - 1$	$\hat{\boldsymbol{Y}}$	$\bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}$
column effect only	$\mu + \phi_j$	$\mathcal{L}_A$	c	$\hat{\boldsymbol{Y}}_A$	$\bar{Y}_{.j}$
row effect only	$\mu + \theta_i$	$\mathcal{L}_B$	r	$\hat{\boldsymbol{Y}}_B$	$\bar{Y}_{i.}$
no effects	μ	$\mathcal{L}_{AB}$	1	$\hat{\boldsymbol{Y}}_{AB}$	$\bar{Y}_{..}$

Table 4.4 Vectors, spaces and their dimensions.

Two interesting null hypotheses are: 'no row effect', ($H_{0A} : \theta_i = 0$ for all i) and 'no column effect' ($H_{0B} : \phi_j = 0$ voor all j). These null hypotheses and their intersection define the linear subspaces $\mathcal{L}_A$, $\mathcal{L}_B$ and $\mathcal{L}_{AB}$ within $\mathcal{L}$ with dimensions c, r and 1, respectively. The complement of $\mathcal{L}_{AB}$ within $\mathcal{L}_A$ is orthogonal to the complement of $\mathcal{L}_{AB}$ within $\mathcal{L}_B$. In **figure 4.2** and **table 4.4** the various spaces and the corresponding projections are shown.
Application of the Pythagoras' theorem leads to

$$\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{AB}\|^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 + \|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_A\|^2 + \|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_B\|^2.$$

This results the ANOVA-**table 4.5**, in which
$SSA = \|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_A\|^2 = \sum c(\bar{Y}_{i.} - \bar{Y}_{..})^2$,
$SSB = \|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_B\|^2 = \sum r(\bar{Y}_{.j} - \bar{Y}_{..})^2$,
$SSE = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 = \sum\sum(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$,
$SST = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{AB}\|^2 = \sum\sum(Y_{ij} - \bar{Y}_{..})^2$.

With the help of the **Fisher-Cochran theorem** it can be shown that the test statistics in the last column of **table 4.5** have under the corresponding null hypothesis an F-distribution with the obvious number of degrees of freedom.

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F
Row effect	$r-1$	SSA	$\frac{SSA}{r-1}$	$\frac{SSA/(r-1)}{SSE/(r-1)(c-1)}$
Column effect	$c-1$	SSB	$\frac{SSB}{c-1}$	$\frac{SSB/(c-1)}{SSE/(r-1)(c-1)}$
Error	$(r-1)(c-1)$	SSE	$\frac{SSE}{(r-1)(c-1)}$	
Total	$rc-1$	SST		

Table 4.5 Two way ANOVA-table, one observation per cell.

In a simple linear regression model the expectation of Y_i is assumed to depend linearly on a variable x_i, so: $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i$. Then for the whole vector of observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ we find:

$$\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} \text{ with } \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

From this we conclude that

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix},$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix},$$

$$\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}.$$

Thus the least squares estimator for $\boldsymbol{\beta}$ becomes:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{Y} - \dfrac{s_{xy}}{s_{xx}}\bar{x} \\ \dfrac{s_{xy}}{s_{xx}} \end{pmatrix},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$,
$s_{xx} = \sum(x_i - \bar{x})^2$, $s_{xy} = \sum(x_i - \bar{x})(Y_i - \bar{Y})$.

Inverting $\boldsymbol{X}^T\boldsymbol{X}$ becomes easier if we rewrite the model in terms of the parameters $\gamma_0 = \beta_0 + \beta_1\bar{x}$ and $\gamma_1 = \beta_1$, hence $\mathcal{E}(Y_i) = \gamma_0 + \gamma_1(x_i - \bar{x})$. The two columns of the corresponding $\boldsymbol{X}$-matrix are then orthogonal, making $\boldsymbol{X}^T\boldsymbol{X}$ a diagonal matrix. Eventually the least squares estimators $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are easily transformed back to $\hat{\beta}_0$ and $\hat{\beta}_1$.

An important quantity that is related to these estimators, is the sample correlation coefficient:
$r = \dfrac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \hat{\beta}_1\sqrt{\dfrac{s_{xx}}{s_{yy}}}$, where $s_{yy} = \sum(Y_i - \bar{Y})^2$.
This quantity is a measure for the strength of the linear relationship between the two variables.

Under the additional assumptions of normality and constant variance σ^2 of the observations distributions of estimators and test statistics can be derived.
The distribution of $\hat{\boldsymbol{\beta}}$ is given in **formula (2.5)**.
Testing whether Y depends linearly on x, implies formulating the hypotheses $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$. Based on the theory of **section 3** the corresponding likelihood ratio test is derived.

The vector of observations $\boldsymbol{Y}$ lies in an n-dimensional space; under $H_0 \cup H_1$ its expectation lies in a two-dimensional ($q = 2$) space $\mathcal{L}$, spanned by the two columns of $\boldsymbol{X}$. The projection $\hat{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ on $\mathcal{L}$ has as its i-th component: $\hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{Y} + \hat{\beta}_1 (x_i - \bar{x})$. Under H_0 the vector of expectations lies in a one-dimensional ($r = 1$) space $\mathcal{L}_0$, spanned by the vector $(1, \ldots, 1)^T$. The projection of $\boldsymbol{Y}$ on $\mathcal{L}_0$ is $\dot{\boldsymbol{Y}} = (\bar{Y}, \ldots, \bar{Y})^T$.

Source of variation	Degrees of freedom *(df)*	Sums of squares *(SS)*	Mean squares *(MS)*	F
Regression	$q - r = 1$	$\frac{s_{xy}^2}{s_{xx}}$	$\frac{s_{xy}^2}{s_{xx}} / 1$	$\frac{(n-2)s_{xy}^2}{s_{xx}s_{yy} - s_{xy}^2}$
Error	$n - q = n - 2$	$s_{yy} - \frac{s_{xy}^2}{s_{xx}}$	$\left(s_{yy} - \frac{s_{xy}^2}{s_{xx}}\right) / (n-2)$	
Total	$n - r = n - 1$	s_{yy}		

Table 4.6 ANOVA table for simple linear regression.

The test statistic F has under the null hypothesis an F-distribution with 1, and $n-2$ degrees of freedom respectively.

For testing the null hypothesis $H_0 : \beta_1 = 0$ we can not only use this F-test, but also a t-test. In the latter case we can also formulate one-sided alternatives.
This t-test is derived in the same **section 3**:

$$\boldsymbol{c}^T\hat{\boldsymbol{\beta}} = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \hat{\beta}_1$$
$$= \frac{s_{xy}}{s_{xx}} = \frac{1}{s_{xx}} \sum (x_i - \bar{x})(Y_i - \bar{Y}) = \sum \frac{x_i - \bar{x}}{s_{xx}} Y_i$$

is an unbiased estimator for β_1 and a linear combination $\boldsymbol{a}^T\boldsymbol{Y}$ of the components of $\boldsymbol{Y}$. The i-th component of $\boldsymbol{a}$ is $\frac{x_i - \bar{x}}{s_{xx}}$.
The quantity

$$T = \frac{(\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{c}^T\boldsymbol{\beta})/\sqrt{\boldsymbol{a}^T\boldsymbol{a}}}{\sqrt{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-q)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\left(s_{xx}s_{yy} - s_{xy}^2\right)/(n-2)}}$$

then has a t-distribution with $n-2$ degrees of freedom.
For testing $H_0 : \beta_1 = \beta_1^{(0)}$ against a one-sided or two-sided alternative this statistic T is used (with $\beta_1 = \beta_1^{(0)}$) as a test statistic. We see that for testing $H_0 : \beta_1 = 0$ against a two-sided alternative the square of the test statistic T equals the test statistic F which we derived earlier.

Similarly a test can be derived for testing the H_0: 'the estimated expected reaction $\beta_0 + \beta_1 x^*$ at the value x^* equals a certain value k' against a one-sided or two-sided alternative. We will not derive this test here, but instead construct a confidence interval for $\beta_0 + \beta_1 x^*$. We are again dealing with a linear combination

$$\begin{aligned}\boldsymbol{c}^T\hat{\boldsymbol{\beta}} &= \begin{pmatrix} 1 & x^* \end{pmatrix}\begin{pmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{pmatrix} = \hat{\beta}_0 + \hat{\beta}_1 x^* \\ &= \bar{Y} + (x^* - \bar{x})\frac{s_{xy}}{s_{xx}} \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{x^* - \bar{x}}{s_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i.\end{aligned}$$

This is an unbiased estimator for $\beta_0 + \beta_1 x^*$ and moreover a linear combination $\boldsymbol{a}^T\boldsymbol{Y}$ of the components of $\boldsymbol{Y}$. The i-th component of $\boldsymbol{a}$ is

$$\frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{s_{xx}}.$$

The quantity T is used as a pivotal quantity for deriving a $(1-\alpha)$-confidence interval.

From tables we find a number $t_{\alpha/2}$ such that $\mathrm{P}(-t_{\alpha/2} < T < t_{\alpha/2}) = \alpha$

$$\text{or } \mathrm{P}\left(-t_{\alpha/2} < \frac{(\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{c}^T\boldsymbol{\beta})/\sqrt{\boldsymbol{a}^T\boldsymbol{a}}}{\sqrt{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-2)}} < t_{\alpha/2}\right) = \alpha$$

$$\text{or } \mathrm{P}\left(\boldsymbol{a}^T\boldsymbol{Y} - t_{\alpha/2}d < \boldsymbol{c}^T\boldsymbol{\beta} < \boldsymbol{a}^T\boldsymbol{Y} + t_{\alpha/2}d\right) = \alpha.$$

$$\text{with } d = \sqrt{\frac{(\boldsymbol{a}^T\boldsymbol{a})\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}{n-2}},$$

$$\begin{aligned}\boldsymbol{a}^T\boldsymbol{a} &= \sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{s_{xx}}\right)^2 \\ &= \sum_{i=1}^{n}\left(\frac{1}{n^2} + \frac{(x^* - \bar{x})^2(x_i - \bar{x})^2}{s_{xx}^2}\right) \\ &= \frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}},\end{aligned}$$

$$\text{and } \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 = \left(s_{yy} - \frac{s_{xy}^2}{s_{xx}}\right).$$

Hence a $(1-\alpha)$-confidence interval for $\boldsymbol{c}^T\boldsymbol{\beta}$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2}d.$$

In a multiple linear regression model the expectation of Y_i is assumed to depend linearly on a number of variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$:

$$\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki},$$

where the x's are well-known real numbers.
For the whole vector of observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ we find: $\mathcal{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$

with $\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \ldots & x_{kn} \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$.

The least squares estimator for $\boldsymbol{\beta}$ is given by **formula (2.2)**. This is also the maximum likelihood estimator when the observations are normally distributed with equal variance.

Under these assumptions we can also derive a test for the null hypothesis that none of the 'explanatory variables' $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ contributes to the prediction of $\boldsymbol{Y}$ against the alternative that at least one of the $\boldsymbol{x}$-variables does contribute.

Formally: $H_0 : \beta_1 = \ldots = \beta_k = 0$, where H_a states that at least one of the β's is different from 0.
Again we can construct an ANOVA-table. The dimensions of the various spaces are: $\dim(R^n) = n$, $\dim(\mathcal{L}) = q = k + 1$ and $\dim(\mathcal{L}_0) = r = 1$.
The numbers of degrees of freedom of the F-test statistic are therefore $(k+1) - 1 = k$ and $n - (k+1)$, respectively.

It is also possible to test whether certain $\boldsymbol{x}$-variables do contribute to the explanation of $\boldsymbol{Y}$. For example $H_0 : \beta_1 = \ldots = \beta_p = 0$ (whereas $\beta_{p+1}, \ldots, \beta_k$ may differ from 0).
In the corresponding ANOVA-table we find the following dimensions of the various spaces: $\dim(R^n) = n$, $\dim(\mathcal{L}) = q = k + 1$ and $\dim(\mathcal{L}_0) = r = k - p + 1$, hence the numers of degrees of freedom of the F-test statistic are: p and $n - k - 1$, respectively.

A special case of multiple linear regression is the so-called polynomial regression. This seems to imply a contradiction, but in fact it does not.

In polynomial regression powers of a measured variable are taken as explanatory variables, for example: $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$.

The word 'linear' in 'linear regression' applies to the way in which the parameters are incorporated into the model, it reflects the fact that we can write $\mathcal{E}(\boldsymbol{Y})$ as the product of a matrix of known real numbers $\boldsymbol{X}$ and a vector of unknown parameters $\boldsymbol{\beta}$. When certain columns of $\boldsymbol{X}$ were obtained by raising elements of other columns to a power this does not alter the linearity of the model.

Often the model is not fixed before carrying out the experiment or survey. For each subject or experimental unit a large number of variables is measured that are possibly related to $\boldsymbol{Y}$. The problem is then to select from these variables (and their powers and/or other functions) a limited set of variables that sufficiently explain the variation in $\boldsymbol{Y}$. This selection of variables is not easy and the predictive power of the ultimate model is hard to establish. See for example [Steerneman] and [Moore & McCabe], 2nd ed, section 9.2.

Similar to simple and multiple linear regression and various kinds of ANOVA being special cases of the general linear model, this linear model in its turn is, together with e.g. logistic regression and log-linear models, a special case of the theory of generalised linear models. For this theory a syllabus **Generalised Linear Models** is available.

In this appendix we assume that the random vectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are k-dimensional.

If $\boldsymbol{Y} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{I})$, then the quadratic form $U = \boldsymbol{Y}^T\boldsymbol{Y} = \sum_{i=1}^k Y_i^2$ has a non-central chi-squared distribution with k degrees of freedom and non-centrality parameter $\delta = \boldsymbol{\mu}^T\boldsymbol{\mu}$. We write: $U \sim \chi_k^2(\delta)$.
If $\delta = 0$ then U has a central chi-squared distribution.

Fisher-Cochran theorem.

Assume that $\boldsymbol{Y} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{I})$, and that the positive semidefinite matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$ with respective ranks $r_1, \ldots, r_m$ are such that $\boldsymbol{A}_1 + \cdots + \boldsymbol{A}_m = \boldsymbol{I}$.
Then each of the following statements implies the other:

i. *The quadratic forms $Q_j = \boldsymbol{Y}^T\boldsymbol{A}_j\boldsymbol{Y}$ $(j = 1, \ldots, m)$ are independent $\chi_{r_j}^2(\boldsymbol{\mu}\boldsymbol{A}_j^T\boldsymbol{\mu})$ random variables.*

ii. $r_1 + \cdots + r_m = k$.

A proof of this theorem can be found in [Azzalini], p.294.

Corollary

If $\boldsymbol{Y} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{I})$ and $\boldsymbol{A}$ is a symmetric idempotent matrix of order k, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi_r^2(\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu})$, where $r = \operatorname{rank}(\boldsymbol{A}) = \operatorname{tr}(\boldsymbol{A})$.

Proof. Apply the Fisher-Cochran theorem to the case $m = 2$ with $\boldsymbol{A}_1 = \boldsymbol{A}$ and $\boldsymbol{A}_2 = \boldsymbol{I} - \boldsymbol{A}$, which satisfy the required condition since $\operatorname{rank}(\boldsymbol{A}_1) + \operatorname{rank}(\boldsymbol{A}_2) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{I} - \boldsymbol{A}) = \operatorname{tr}(\boldsymbol{I}) = k$.

Remark.

In the case where $\operatorname{var}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$ the Fisher-Cochran theorem can be applied to the vector $\boldsymbol{Z} = \frac{1}{\sigma}\boldsymbol{Y} \sim \mathcal{N}_k(\frac{1}{\sigma}\boldsymbol{\mu}, \boldsymbol{I})$. If the other conditions are fulfilled, then $\boldsymbol{Z}\boldsymbol{A}_j\boldsymbol{Z} \sim \chi_{r_j}^2(\frac{1}{\sigma^2}\boldsymbol{\mu}^T\boldsymbol{A}_j\boldsymbol{\mu})$.

1. For the cases 1, 3, 5 and 6 of **section 1.2** find the matrix $\boldsymbol{X}$ and the vector $\boldsymbol{\beta}$. (Case 2 was already analysed in **example 1** and case 4 is easier to analyse using corner point restrictions.) Always first find the vector $\boldsymbol{\beta}$ of unknown parameters and then the matrix $\boldsymbol{X}$.

2. For the cases 1, 3 and 5 of **section 1.2** find the least squares estimators for the parameters. Also find the distribution of the estimators when the observations are normally distributed. Case 5 is easier to solve if the equivalent model $\mathcal{E}(Y_i) = \gamma + \beta(x_i - \bar{x})$ is taken as a starting point. Why?

3. Start from the assumptions for de **Gauss-Markov theorem** and show that the matrices $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{I} - \boldsymbol{P}$ are symmetric and idempotent and that they have rank p and $n - p$ respectively (hint: $\text{rank}(\boldsymbol{P}) = \text{tr}(\boldsymbol{P})$ and $\text{tr}(\boldsymbol{AB}) = \text{tr}(\boldsymbol{BA})$).

4. Start from the assumptions for the **Gauss-Markov theorem** and prove that $S^2 = \frac{1}{n-p}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 = \frac{1}{n-p}\|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}\|^2$ is an unbiased estimator for σ^2 (hint: $\mathcal{E}[\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}]$ $= \mathcal{E}[\text{tr}(\boldsymbol{Y} - \mathcal{E}(\boldsymbol{Y}))^T(\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{Y} - \mathcal{E}(\boldsymbol{Y}))]$).

5. Show that without the restrictions the parameters in case 4 of **section 1.2** are not identifiable.

6. When $\boldsymbol{X}^T\boldsymbol{X}$ is identifiable, show that the least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$, (first statement of the **Gauss-Markov theorem)** and that $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{c}^T\boldsymbol{\beta}$.

7. Construct an ANOVA table for the following cases of **section 1.2**: 2: $H_0 : \mu_1 = \mu_2$; 3: $H_0 : \mu_1 = \ldots = \mu_k$; 1: $H_0 : \mu = 0$; 5: $H_0 : \beta = 0$.

8. In the case of a one-way ANOVA sometimes it is said that in the test statistic the variation *between* the groups (the numerator) is compared with the variation *within* the groups (the denominator). Can the numerator and denominator be considered as varia*n*ces?

9. Show that in the one-way ANOVA case with $k = 2$ the test statistic is equal to the square of the test statistic in the two-sample problem with equal variances.

10. Find the least squares estimator in the simple linear regression case along the 'detour' (via y_0 and y_1).

11. Find the least squares estimator for the case of one explanatory variable, where the regression line is forced to pass through the origin: $\mathcal{E}(Y_i) = \beta x_i$.

12. Under the assumption of normality and constant variance of the observations you found least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. If you want to observe a new observation Y_{n+1} at the value x_{n+1}, what can be said about the distribution of Y_{n+1}? Derive a 95% prediction interval for Y_{n+1}.

AZZALINI, A. [1996] Statistical inference: based on the likelihood; Chapman & Hall, London.

LINDGREN, B.W. [1993] Statistical Theory, 4th ed.; Chapman & Hall, New York.

MOORE, D.S., MCCABE, G.P. [1998] Introduction to the Practice of Statistics, 3rd ed.; Freeman, New York.

SILVEY, S.D. [1970] Statistical Inference; Penguin Books, Harmondsworth.

STEERNEMAN, A.G.M. [1987] On the Choice of Variables in Discriminant and Regression Analysis; CWI Amsterdam.