TMA4267 - Linear statistical models

Trond Skaret Johansen

Spring semester 2024

Contents

Introduction

This is a brief summary of the course TMA4267 about linear statistical models. It includes the main content from the lecture held by TODO: ... recorded in, where some examples etc... are excluded.

The purpose of the notes is to give a good overview of the syllabus. I intend to add summaries of the lectures as I review them. I hope to include insights from projects / exercises where it is appropriate.

Topics

The first chapter begins by introducing multivariate distributions and how to compute expectations. We then move on to multivariate moments and transformations. Principal component analysis (PCA) is described, before we explain the need for charactestic functions and not just moment generating functions when working with multivariate distributions.

In the second chapter we introduce the multivariate normal distribution. We deal with estimation in the multivariate normal distribution and give the theory of quadratic forms and idempotent matrices.

The third chapter tackles multiple linear regression. We introduce the model and its assumptions and estimate the parameters. Properties of the estimators, fitted values and residuals are established, and then put to use in performing inference about the coefficients. We perform t-tests, do ANOVA, compute the coefficient of determination and perform general F-tests. Finally, we look at some way of transforming the data.

In the fourth chapter, we analyse the model and the selection of models. We perform multiple hypothesis testing and present examples. Two methods for controlling the FWER are investigated.

In the fifth and final chapter, we do more ANOVA and investigate design of experiment. Our focus is two-level factorial design. We discuss generators of design, resolution of design and the method of blocking.

Course progress

- First reading
  - ☑ Lecture 1-25
- Gjennomgang
  - ☑ Lecture 1-2
  - ☑ Lecture 3-4
  - ☑ Lecture 5-6
  - ☑ Lecture 7-8
  - ☑ Lecture 9-10
  - ☑ Lecture 11-12
  - ☑ Lecture 13-14
  - ☐ Lecture 15-16
  - ☑ Lecture 17-18
  - ☑ Lecture 19-20
  - ☑ Lecture 21-22
- Exams
  - ☑ May 2023
  - ☑ June 2019
  - ☑ May 2018
  - ☑ May 2017
  - ☑ June 2016
  - ☐ May 2015
  - ☐ May 2014
  - ☐ August 2014

1 Multivariate Distribution and its Generalisations

1.1 Matrix algebra

The main tool of the course is matrix algebra. We therefore repeat the most important convepts. The matrix $\boldsymbol{A}$ is said to be *symmetric* if $\boldsymbol{A}^\top = \boldsymbol{A}$. It is *orthogonal* if $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I}$, i.e. if the columns are orthogonal. The elements of the pair $(\lambda, \boldsymbol{\gamma})$ are called *eigenvalue* and *eigenvector* respectivelly if they satisfy $\boldsymbol{A}\boldsymbol{\gamma} = \lambda\boldsymbol{\gamma}$ and $\boldsymbol{\gamma} \neq \boldsymbol{0}$. λ can be found as the solution of $\det(\boldsymbol{A} - \lambda\boldsymbol{I})$. Recall that the *trace*, $\operatorname{tr}(\boldsymbol{A})$, of a matrix is the sum of the diagonal. We also have the formulas:

$$\det \boldsymbol{A} = \prod_{i=1}^{p} \lambda_i \quad \operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{p} \lambda_i.$$

For a symmetric matrix we may find the *Jordan decomposition*. Let $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_p\}$ and $\Gamma = (\boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma_p})$. Then we have:

$$\boldsymbol{A} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top.$$

For a symmetrix matrix $\boldsymbol{A}$ and a vector $\boldsymbol{x}$ we define the *quadratic form*:

$$Q(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{p}\sum_{j=1}^{p} x_i A_{ij} x_j.$$

Theorem 1. *Transforming* $\boldsymbol{y} = \boldsymbol{\Gamma}^\top\boldsymbol{x}$ *we obtain*

$$Q(\boldsymbol{x}) = \sum_{i=1}^{p} \lambda_i y_i^2.$$

A matrix is said to be *positive definite* if $Q(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \neq 0$ and positive semi definite if $Q(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \neq 0$. We write $A > 0$ and $A \geq 0$ respectivelly.

Theorem 2. *The symmetric matrix* A *is positive definite iff* $\lambda_i > 0$ *for all* i.

Proof. Using the transform of the previous theorem we find

$$\lambda_1 y_1^2 + \cdots + \lambda_p y_p^2 > 0 \quad \forall \boldsymbol{y} \in \mathbb{R}^p \Leftrightarrow \lambda_i > 0 \quad \forall i.$$

□

From this we obtain two more usefull results.

1. If $A > 0$ the inverse exists and the determinant is > 0
2. If $A > 0$ there exists a unique positive definite square root with decomposition:

$$A^{1/2} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Gamma}^\top.$$

1.2 Random vectors and distributions

A *random vector* $\boldsymbol{X}$ is a vector where each component is a *random variable*. Similarly, we define a *random matrix* as a matrix with random variables as component. As in the univariate case, we define the *cumulative distribution function* (CDF) by:

$$F(x) = \mathbb{P}\left[\boldsymbol{X} \leq \boldsymbol{x}\right] = \mathbb{P}\left[X_1 \leq x_1, \ldots, X_p \leq x_p\right].$$

A random vector is said to be *absolutelly continuous* if there exists a *probability density function* (PDF) f such that:

$$F(\boldsymbol{x}) = \int_{-\infty}^{x_p} \cdots \int_{-\infty}^{x_1} f(u_1, \ldots, u_p) du_1 \ldots du_p.$$

Then, we may compute the probability of the event $\boldsymbol{X} \in D$ by:

$$\mathbb{P}\left[\boldsymbol{X} \in D\right] = \int_D f(\boldsymbol{x}) d\boldsymbol{x} \quad \forall D \subseteq \mathbb{R}^p.$$

The random cector is said to be *discrete* if it is consentraded on a countable (finite or infinite) set of points. Then integral becomes a sum.

In the absolutelly continuous case, we may write the density as:

$$f(\boldsymbol{x}) = f\left(x_1, \ldots, x_p\right) = \frac{\partial^p F\left(x_1, \ldots, x_p\right)}{\partial x_1 \cdots \partial x_p}.$$

Let $\boldsymbol{X}_A, \boldsymbol{X}_B$ be two random vectors st. $\boldsymbol{X} = (\boldsymbol{X}_A, \boldsymbol{X}_B)^\top$ has cdf F. Then we may find the *marginal distribution*:

$$F_A(x_1, \ldots, x_k) = F(x_1, \ldots, x_k, \infty, \ldots, \infty).$$

In absolutelly continuous case we find the marginal density:

$$f_A(x_1, \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_p) du_p \ldots du_{k+1}$$

The *conditional distribution* of $\boldsymbol{X}_B$ given $\boldsymbol{X}_A = \boldsymbol{x}_A$ is:

$$f_{\boldsymbol{X}_B | \boldsymbol{X}_A = \boldsymbol{x}_A}(\boldsymbol{x}_B) = \frac{f(x_1, \ldots, x_p)}{f_A(x_1, \ldots, x_k)}$$

We say that the random vectors $\boldsymbol{X}_A, \boldsymbol{X}_B$ are *independent* if

$$F(x_1, \ldots, x_p) = F_A(x_1, \ldots, x_k) F_B(x_{k+1}, \ldots, x_p) \quad \forall x_1, \ldots x_p.$$

In the continuous case we have independence iff $f = f_A \cdot f_B$. In this case $f_{\boldsymbol{x}_B | \boldsymbol{X}_A = \boldsymbol{x}_A} = f_B(\boldsymbol{x}_B)$. Similar definition for independence when $\boldsymbol{X}$ has N components and not just 2.

1.3 Multivariate expecations and moments

We define the *expectation* of the random vector $\boldsymbol{X}$ as

$$\mathbb{E}[\boldsymbol{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_p])^\top.$$

Here, each $\mathbb{E}[X_i] = \int_{\mathbb{R}^p} x_i f(\boldsymbol{x}) d\boldsymbol{x}$. It is easy to show:

1. For constants a, b we have $\mathbb{E}[a\boldsymbol{X} + b\boldsymbol{Y}] = a\mathbb{E}[\boldsymbol{X}] + b\mathbb{E}[\boldsymbol{Y}]$.
2. For (shape compatible) matrices $\boldsymbol{A}, \boldsymbol{B}$ we have $\mathbb{E}[\boldsymbol{AXB}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{X}]\boldsymbol{B}$.
3. Let $\boldsymbol{X}, \boldsymbol{Y}$ be *independent* random matrices whose product is defined. Then $\mathbb{E}[\boldsymbol{XY}] = \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Y}]$.

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ and $\mathbb{E}[\boldsymbol{X}] =: \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$. We then define the *covariance matrix* of $\boldsymbol{X}$ as:

$$\mathrm{Var}[\boldsymbol{X}] = \mathrm{Cov}[\boldsymbol{X}] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{X_1X_1} & \cdots & \sigma_{X_1X_p} \\ \vdots & \ddots & \vdots \\ \sigma_{X_pX_1} & \cdots & \sigma_{X_pX_p} \end{pmatrix} = \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\right].$$

This matrix is *symmetric*. Note also that $\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}[X_i, X_j]$. We can also show:

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\top\right] - \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

The correlation matrix (with ones on the diagonal as $\rho_{X_iX_i} = 1$) is given by

$$\boldsymbol{\rho} = \begin{pmatrix} \rho_{X_1X_1} & \cdots & \rho_{X_1X_p} \\ \vdots & \ddots & \vdots \\ \rho_{X_pX_1} & \cdots & \rho_{X_pX_p} \end{pmatrix}, \quad \rho_{X_iX_j} = \frac{\sigma_{X_iX_j}}{\sqrt{\sigma_{X_i}}\sqrt{\sigma_{X_j}}}.$$

For two random vectors $\boldsymbol{X}, \boldsymbol{Y}$ we define their covariance matrix by

$$\boldsymbol{\Sigma}_{\boldsymbol{XY}} = \mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}] = \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}})^\top\right] = (\mathrm{Cov}[X_i, X_j])_{\substack{i=1,\ldots,p \\ j=1,\ldots,q}}$$

Proposition 1. *The covariance matrix $\boldsymbol{\Sigma}$ is positive semi-definite.*

Proof. Using the formula for the variance of a linear combination we obtain:

$$\boldsymbol{y}^\top\boldsymbol{\Sigma}\boldsymbol{y} = \begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1n} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{n1} & \Sigma_{n2} & \cdots & \Sigma_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^n \sum_{j=1}^n y_i \Sigma_{ij} y_j$$

$$= \sum_{i=1}^n y_i^2 \mathrm{Var}(X_i) + 2\sum_{i<j} y_i y_j \mathrm{Cov}(X_i, X_j) = \mathrm{Var}\left(\sum_{i=1}^n y_i X_i\right) \geq 0.$$

Which completes the proof. □

Remark. *We usually require that the covariance matrix is positive definite, since if it is only positive semi-definite there are nontrivial linear combinations with 0 variance. Indeed, if* $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ *we obtain using the formula for variance of linear combinations that:*

$$\text{Var}\left[X_1 - X_2\right] = 1^2\text{Var}\left[X_1\right] + (-1)^2\text{Var}\left[X_2\right] + 2(1)(-1)\text{Cov}\left[X_1, X_2\right] = 0.$$

This never happens for positive definite matrices since we for any nontrivial linear combination have:

$$\text{Cov}\left[\boldsymbol{c}^\top \boldsymbol{X}\right] = \boldsymbol{c}^\top \boldsymbol{\Sigma} \boldsymbol{c} > 0.$$

We have many more properties of covariance matrices:

1. $\boldsymbol{\Sigma_{XY}} = \boldsymbol{\Sigma_{YX}^\top}$.
2. If $\boldsymbol{X} \sim (\boldsymbol{\mu_X}, \boldsymbol{\Sigma_{XX}}), \boldsymbol{Y} \sim (\boldsymbol{\mu_Y}, \boldsymbol{\Sigma_{YY}})$ then $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})^\top$ has
$$\boldsymbol{\Sigma_{ZZ}} = \begin{pmatrix} \boldsymbol{\Sigma_{XX}} & \boldsymbol{\Sigma_{XY}} \\ \boldsymbol{\Sigma_{YX}} & \boldsymbol{\Sigma_{YY}} \end{pmatrix}.$$
3. Independence of $\boldsymbol{X}, \boldsymbol{Y}$ implies $\text{Cov}\left[\boldsymbol{X}, \boldsymbol{Y}\right] = \boldsymbol{0}$ (NB: the converse not true).
4. $\text{Var}\left[\boldsymbol{AX} + \boldsymbol{b}\right] = \boldsymbol{A}\text{Var}\left[\boldsymbol{X}\right]\boldsymbol{A}^\top$.
5. $\text{Cov}\left[\boldsymbol{X} + \boldsymbol{Y}, \boldsymbol{Z}\right] = \text{Cov}\left[\boldsymbol{X}, \boldsymbol{Z}\right] + \text{Cov}\left[\boldsymbol{Y}, \boldsymbol{Z}\right]$.
6. $\text{Var}\left[\boldsymbol{X} + \boldsymbol{Y}\right] = \text{Var}\left[\boldsymbol{X}\right] + \text{Cov}\left[\boldsymbol{X}, \boldsymbol{Y}\right] + \text{Cov}\left[\boldsymbol{Y}, \boldsymbol{Z}\right] + \text{Var}\left[\boldsymbol{Y}\right]$.
7. $\text{Cov}\left[\boldsymbol{AX}, \boldsymbol{BY}\right] = \boldsymbol{A}\text{Cov}\left[\boldsymbol{X}, \boldsymbol{Y}\right]\boldsymbol{B}^\top$.

1.4 Transformations

1.4.1 Mahalanobis transformation

Our first transformation is the *Mahalanobis transformation*. We recall that we can get 0 mean and unit variance in the univariate case by the transformation $Y = \frac{X-\mu}{\sigma}$. In the multivariate case, suppose $X = (X_1, \ldots, X_p)^\top \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ non-singular. Then using the unique positive definite square root $\boldsymbol{\Sigma}^{1/2}$ og $\boldsymbol{\Sigma}$, we have the transformation:

$$\boxed{\boldsymbol{Y} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim (0, \boldsymbol{I})} \tag{1}$$

Proof. Compute the expected value and variance. □

1.4.2 Principal component analysis (PCA)

In the following we suppose that we observe realisations of some random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The goal of *principal component analysis* is to reduce dimensionality by removing some components and keeping components with large variance and hence more information.

The idea is to first perform the transform $\boldsymbol{X} \mapsto \boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{b} = (Y_1, \ldots, Y_p)^\top$ in such a way that:

1. $\mathbb{E}\left[\boldsymbol{Y}\right] = \boldsymbol{0}$,

2. $\mathrm{Cov}\,[Y_i, Y_j] = 0$ for $i \neq j$, i.e. $\boldsymbol{\Sigma_Y}$ is diagonal,

3. $\mathrm{Var}\,[Y_1] \geq \mathrm{Var}\,[Y_2] \geq \cdots \geq \mathrm{Var}\,[Y_p]$.

Using again the Jordan decomposition we may find such a transform. Arange $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top$ such that the eigenvalues on the diagonal of $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$ are ordered decreasingly. A transformation satisfying the requirements is then $\boldsymbol{Y} = \boldsymbol{\Gamma}^{1/2}(\boldsymbol{X} - \boldsymbol{\mu})$. Indeed we may compute $\mathrm{Var}\,[\boldsymbol{Y}] = \boldsymbol{\Lambda}$. We call $Y_1, \ldots, Y_p$ the *principal components*.

For the interpretation, recall that for the trace of a matrix we have: $\mathrm{tr}\,(\boldsymbol{ABC}) = \mathrm{tr}\,(\boldsymbol{BCA}) = \mathrm{tr}\,(\boldsymbol{CAB})$. Using this we compute:

$$\text{“total variance”} = \sum_{i=1}^{p} \mathrm{Var}\,[X_i] = \mathrm{tr}\,(\boldsymbol{\Sigma}) = \mathrm{tr}\left(\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top\right) = \mathrm{tr}\left(\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\right) = \mathrm{tr}\,(\boldsymbol{\Lambda}) = \sum_{i=1}^{n} \lambda_i.$$

Since we care about variability in our data, we can use this interpretation to extract the features explaining as much cariance as possible. For example we can keep the m first principal components. Say we compute:

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} = 0.8.$$

Then we may say that “80% of variability is explained by the components”, and we have succesfully reduced dimensionality.

We can also do *empirical PCA* (also called *sample PCA*). Suppose that we have n observations of independent identically distributed random vectors $\boldsymbol{X}_i = \left(X_{i1}, \ldots, X_{ip}\right)^\top \sim N(\mu, \Sigma)$. Gather them in a $(n \times p)$ “data matrix”:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \boldsymbol{X}_2^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \cdots & X_{2p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}.$$

Estimate the unknown $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ by $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$ and $\boldsymbol{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top$. Then we may transform the matrix $\boldsymbol{X}$ into $\boldsymbol{Y} = (\boldsymbol{X} - \boldsymbol{1}_n\bar{\boldsymbol{X}})\boldsymbol{G}$ where $\boldsymbol{S} = \boldsymbol{GLG}$ is the Jordan decomposition of $\boldsymbol{S}$ with eigenvalues ordered decreasingly. Finally remove lowest components.

Remark. *Intuitivelly, PCA can be though of as finding a p-dimensional ellipsoid that fits the data. The axis of the ellipsoid represent the principal components.*

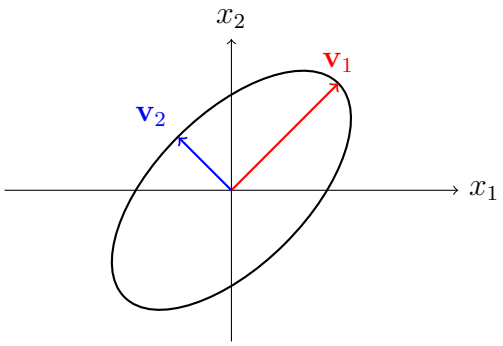

Figure 1: Principal Component Analysis (PCA) Ellipsoid. The red arrow indicates the first principal component (eigenvectors) $\mathbf{v}_1$ and the blue arrow indicates the second principal component $\mathbf{v}_2$.

1.4.3 General transformations

Suppose we transform the random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ into $\boldsymbol{Y} = g(\boldsymbol{X})$. If the function g is one-to-one and has differentiable inverse u, then

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = |\det \boldsymbol{J}| f_{\boldsymbol{X}}(u(\boldsymbol{y}))$$

where

$$\boldsymbol{J} = \left(\frac{\partial u_i(y)}{\partial y_j} \right)_{i,j=1,\ldots,p}$$

is the *Jacobian matrix*. An important special case is that of linear transformations $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$. If $\boldsymbol{A}$ is non-singular the inverse transform is $\boldsymbol{X} = \boldsymbol{A}^{-1}(\boldsymbol{Y} - \boldsymbol{b})$ and $\boldsymbol{J} = \boldsymbol{A}^{-1}$, so we obtain:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = |\det \boldsymbol{A}^{-1}| f_{\boldsymbol{X}}(\boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b})).$$

1.5 Characteristic function

In the univariate case we often studied the *moment generating function*:

$$M_X(t) = \mathbb{E}\left[e^{tX}\right].$$

Important properties include:

1. $M_X = M_Y \Rightarrow X \stackrel{d}{=} Y$,
2. $\mathbb{E}\left[X^k\right] = M_X^{(k)}(0)$,
3. Independence of X, Y implies $M_{X+Y} = M_X M_Y$.

However, it does noe exist for for instance the student-t distribution. It only exists when all moments exist. We shall now define the *characteristic function*, which **always** exists:

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right].$$

It has similar properties:

1. $\varphi_X = \varphi_Y \Rightarrow X \stackrel{d}{=} Y$,
2. $\mathbb{E}\left[X^k\right] = i^{-k}\varphi^{(k)}(0)$,
3. Independence of X, Y implies $\varphi_{X+Y} = \varphi_X \varphi_Y$.

In the p-variate case, the functions are functions of $\boldsymbol{t} = (t_1, \ldots, t_p)^\top$. Let as usual $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ be our random vector. We define:

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[e^{\boldsymbol{t}^\top \boldsymbol{X}}\right],$$
$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[e^{i\boldsymbol{t}^\top \boldsymbol{X}}\right].$$

We list some important properties.

1. If $\varphi_{\boldsymbol{X}}(t)$ is absolutelly integrable, the (Fourier) inversion formula holds:

$$f_{\boldsymbol{X}} = \frac{1}{(2\pi)^p} \int_{\mathbb{R}} e^{-i\boldsymbol{t}^\top \boldsymbol{x}} \varphi_{\boldsymbol{X}}(t) dt.$$

2. Denote $t_{(1)} = (t_1, \ldots, 0)^\top, \ldots, t_{(p)} = (0, \ldots, t_p)^\top$. Then

$$\varphi_{\boldsymbol{X}_k}(t_k) = \mathbb{E}\left[e^{it_1 X_1}\right] = \mathbb{E}\left[e^{i\boldsymbol{t}_{(1)}^\top \boldsymbol{X}}\right] = \varphi_{\boldsymbol{X}}(t_{(k)}).$$

3. Let $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}$ and $\boldsymbol{t} = \begin{pmatrix} \boldsymbol{t}_1 \\ \boldsymbol{t}_2 \end{pmatrix}$ be vectors with appropriate dimensions. Then:

$$\boxed{\boldsymbol{X}_1, \boldsymbol{X}_2 \text{ are independent } \Leftrightarrow \varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}_1}(\boldsymbol{t}_1)\varphi_{\boldsymbol{X}_2}(\boldsymbol{t}_2)} \tag{2}$$

4. Let $\boldsymbol{X}, \boldsymbol{Y}$ be independent p-variate random vectors. Then:

$$\varphi_{\boldsymbol{X}+\boldsymbol{Y}}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}}(\boldsymbol{t})\varphi_{\boldsymbol{Y}}(\boldsymbol{t}).$$

We end the section with a theorem linking the distributions of 1D random variables to the distribution of the p-variate random vector $\boldsymbol{X}$:

Theorem 3. Cramer-Wold *theorem states that the distribution of $\boldsymbol{X} \in \mathbb{R}^p$ is completelly determined by the set of all 1D distributions of $\boldsymbol{t}^\top \boldsymbol{X}, \boldsymbol{t} \in \mathbb{R}^p$.*

Proof. Suppose that for all $\boldsymbol{t} \in \mathbb{R}^p$ we have $\boldsymbol{t}^\top \boldsymbol{X} \overset{d}{=} \boldsymbol{t}^\top \boldsymbol{Y}$. Then:

$$\boldsymbol{t}^\top \boldsymbol{X} \overset{d}{=} \boldsymbol{t}^\top \boldsymbol{Y} \Rightarrow \varphi_{\boldsymbol{t}^\top \boldsymbol{X}}(u) = \varphi_{\boldsymbol{t}^\top \boldsymbol{Y}}(u).$$

Taking $u = 1$ gives $\mathbb{E}\left[e^{i\boldsymbol{t}^\top \boldsymbol{X}}\right] = \mathbb{E}\left[e^{i\boldsymbol{t}^\top \boldsymbol{Y}}\right]$ for all $\boldsymbol{t}$, so $\varphi_{\boldsymbol{X}} = \varphi_{\boldsymbol{Y}}$ and hence $\boldsymbol{X} \overset{d}{=} \boldsymbol{Y}$. □

2 The Multivariate Normal Distribution

2.1 From univariate to multivariate

2.1.1 Univariate case

We begin by recalling important properties and results of the univariate normal distribution. Let $-\infty < \mu < \infty, \sigma \geq 0$. We say $X \sim N(\mu, \sigma^2)$ if it has density:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

The case $\sigma = 0$ is considered normal degenerate and we have $X \overset{\text{a.s.}}{=} \mu$. One can show $\mathbb{E}[X] = \mu, \operatorname{Var}[X] = \sigma^2$ and that the moment generating and characteristic functions are:

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}},$$
$$\varphi_X(t) = e^{i\mu t - \frac{\sigma^2 t^2}{2}}.$$

We say that $Z \sim N(0,1)$ is standard normal. We can transform as follows:

$$X \sim N(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1),$$
$$Z \sim N(0,1) \Rightarrow \sigma Z + \mu \sim N(\mu, \sigma^2).$$

A final fundamental property is that linear combinations of independent normal random variables are also normal.

2.1.2 Multivariate case

We choose to present one of several equivalent definitions. Here we first define a special case and then generalise it.

Definition 1. A p-variate random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$ is a *standard normal vector* if its components are independent and each $Z_i \sim N(0,1)$. We imediatelly obtain $f_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod f_{Z_i}(z_i)$. We use the notation $Z \sim N(0, \boldsymbol{I})$.

Definition 2. We say that $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is *multivariate normal* if there exists a $(p \times q)$ matrix $\boldsymbol{A}$ and $\boldsymbol{\mu} \in \mathbb{R}^p$ such that

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}, \quad \boldsymbol{Z} \sim (0, \boldsymbol{I}_q)$$

Theorem 4. *For any vector $\boldsymbol{b} \in \mathbb{R}^p$ we have that $\boldsymbol{b}^\top \boldsymbol{X}$ is a univariate normal random variable.*

Proof. It is a linear combination of independent normal random variables:

$$\boldsymbol{b}^\top \boldsymbol{X} = \boldsymbol{b}^\top(\boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}) = \boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{b}^\top \boldsymbol{\mu} \sim N.$$

□

From the definition we find $\mathbb{E}[X] = \boldsymbol{\mu}, \operatorname{Var}[X] = \boldsymbol{A}\boldsymbol{A}^\top =: \boldsymbol{\Sigma}$. We write $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It turns out that the statement from the theorem would give an equivalent definition!

Theorem 5. *Let* $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *If* $\boldsymbol{\Sigma} > 0$, *then the probability density function (PDF) of* $\boldsymbol{X}$ *is:*

$$\boxed{f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2} \det \boldsymbol{\Sigma}^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}} \tag{3}$$

Proof. Recall that $\boldsymbol{X} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z} + \boldsymbol{\mu}$. Using this, the result follows from the transformation formula and using the pdf of the standard normal $\boldsymbol{Z}$. □

Theorem 6.

$$\boxed{\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} \sim N(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)} \tag{4}$$

Proof. Compute expected value and covariance. Normality follows from definition as we may write $\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{B}\boldsymbol{Z} + \boldsymbol{c}) + \boldsymbol{b} = \tilde{\boldsymbol{B}}\boldsymbol{Z} + \tilde{\boldsymbol{b}}$. □

The following corollary tells us that we may *standardize* in the multivariate case as well:

Corollary 1.

$$\boxed{\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N(0, \boldsymbol{I})} \tag{5}$$

Corollary 2. *Any subvector* $\boldsymbol{X}^*$ *of* $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *is also normal.*

Proof. Take $\boldsymbol{A}$ to be a "component removing" matrix and use Theorem 6. □

Theorem 7. *The characteristic function of* $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *is:*

$$\boxed{\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = e^{i\boldsymbol{t}^\top \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}}} \tag{6}$$

Proof. For $Z \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$, we may use independence to obtain:

$$\varphi_{\boldsymbol{Z}}(\boldsymbol{t}) = \prod_{i=1}^{p} \varphi_{Z_i}(t_i) = e^{-\frac{1}{2}\boldsymbol{t}^\top \boldsymbol{t}}.$$

Then write $\boldsymbol{X} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z} + \boldsymbol{\mu}$ and rewrite:

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[e^{i\boldsymbol{t}^\top \boldsymbol{X}}\right] = e^{i\boldsymbol{t}^\top} \mathbb{E}\left[e^{i\boldsymbol{t}^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}}\right] = e^{i\boldsymbol{t}^\top} \varphi_{\boldsymbol{Z}}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{t}) = e^{i\boldsymbol{t}^\top} e^{-\frac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}}.$$

Where we have used symmetry of $\boldsymbol{\Sigma}^{1/2}$ a few times. □

The following is among the most important results in the course:

Theorem 8. *Suppose* $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Then*

$$\boxed{\boldsymbol{A}\boldsymbol{X}, \boldsymbol{B}\boldsymbol{X} \text{ are independent } \Leftrightarrow \operatorname{Cov}[\boldsymbol{A}\boldsymbol{X}, \boldsymbol{B}\boldsymbol{X}] = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^\top = \boldsymbol{0}} \tag{7}$$

Proof. The rightward implication is true for any random vectors. To prove the leftward for normal $\boldsymbol{X}$ we show that we may factor the characteristic function

$$\varphi_{\begin{pmatrix}\boldsymbol{AX}\\ \boldsymbol{BX}\end{pmatrix}}(\boldsymbol{t}) = \varphi_{\boldsymbol{AX}}(\boldsymbol{t}_1)\varphi_{\boldsymbol{BX}}(\boldsymbol{t}_2).$$

□

Corollary 3. *Let $X = (X_1, \ldots, X_p)^\top \sim N\boldsymbol{\mu}, \boldsymbol{\Sigma}$. Then X_i, X_j independent iff $\operatorname{Cov}[X_i, X_j] = 0$.*

The next theorem tells us that we may "remove the dependent part" of a component from another:

Theorem 9. *Let $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be p-variate and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)^\top$ with $\boldsymbol{X}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Define $\boldsymbol{X}_2' = \boldsymbol{X}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1$ Then:*

1. *$X_2' \sim N(\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$.*
2. *$\boldsymbol{X}_1, \boldsymbol{X}_2'$ are independent.*

Theorem 10. *With notation as in the previous theorem, the conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$ is:*

$$\boldsymbol{X}_2|_{\boldsymbol{X}_1=\boldsymbol{x}_1} \sim N(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}).$$

Theorem 11. *If $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is non-singular. Then*

$$\boxed{U = (\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi_p^2}$$

Proof. By definition of the χ^2-distribution we have $U = \boldsymbol{Z}^\top\boldsymbol{Z} = Z_1^2 + \cdots + Z_p^2 \sim \chi_p^2$ for $\boldsymbol{Z} \sim N(0, \boldsymbol{I}_p)$. By the *Mahalanobis transformation* we have $\boldsymbol{Z} = \boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})$ from which the result follows. □

2.2 Estimation of the multivariate normal distribution

2.2.1 Univariate case

From the univatiate case we recall that if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ are independent, then the MLE estimators are:

$$\widehat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

But since we found that $\mathbb{E}[\sigma^2] = \frac{n-1}{n}\sigma^2$ is biased, we use instead the estimator

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We further proved that

1. $\bar{X} \sim N(\mu, \sigma^2/n)$,
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$,
3. $\bar{X}, S^2$ are independent (for the normal distribution),
4. $\sqrt{n}\frac{\bar{X}-\mu}{S} \sim t_{n-1}$ (student t distribution).

Our next goal is to obtain the result for the multivariate case.

2.2.2 Multivatiate case

In this case, we have p-variate independent randum vectors $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where we denote $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$. These make up the columns of the $(n \times p)$ *data matrix* or *feature matrix* $\boldsymbol{X}$ given as:

$$\boldsymbol{X} = \underbrace{\begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}}_{\text{features}\rightarrow} \overset{\text{samples}}{\downarrow} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = (\boldsymbol{X}_1 \cdots \boldsymbol{X}_n)^\top.$$

We want to estimate $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. Again we denote:

$$\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i = \frac{1}{n}\boldsymbol{X}^\top \mathbf{1},$$

$$\boldsymbol{S}^2 = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) = \frac{1}{n}\boldsymbol{X}^\top \left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\boldsymbol{X}.$$

The matrix $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is called the *centering matrix* because its action is to remove the mean of a vector:

$$\boldsymbol{C}\boldsymbol{y} = \begin{pmatrix} 1-\frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}.$$

We note that

1. $\boldsymbol{C}$ is symmetric
2. $\boldsymbol{C}$ is idempotent

For the estimators, we may prove:

Proposition 2. $\bar{\boldsymbol{X}} \sim N(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$

Proof. The proof is by factoring the characteristic function. We can do this since the $\boldsymbol{X}_i$'s are independent.

$$\varphi_{\bar{\boldsymbol{X}}}(\boldsymbol{t}) = \mathbb{E}\left[e^{i\boldsymbol{t}^\top \bar{\boldsymbol{X}}}\right] = \mathbb{E}\left[e^{i\left(\frac{\boldsymbol{t}}{n}\right)^\top (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)}\right] = \varphi_{\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n}\left(\frac{\boldsymbol{t}}{n}\right) = \varphi_{\boldsymbol{X}_1}\left(\frac{\boldsymbol{t}}{n}\right) \cdots \varphi_{x_n}\left(\frac{\boldsymbol{t}}{n}\right)$$

$$= \prod_{i=1}^n e^{i\left(\frac{\boldsymbol{t}}{n}\right)^\top \boldsymbol{\mu} - \frac{1}{2}\left(\frac{\boldsymbol{t}}{n}\right)^\top \boldsymbol{\Sigma}\left(\frac{\boldsymbol{t}}{n}\right)^\top} = e^{\sum_{j=1}^n i\left(\frac{\boldsymbol{t}}{n}\right)^\top \boldsymbol{\mu} - \frac{1}{2}\left(\frac{\boldsymbol{t}}{n}\right)^\top \boldsymbol{\Sigma}\left(\frac{\boldsymbol{t}}{n}\right)^\top} = e^{i\boldsymbol{t}^\top \mu - \frac{1}{2}\boldsymbol{t}^\top \frac{\boldsymbol{\Sigma}}{n}\boldsymbol{t}}.$$

□

Proposition 3. $\mathbb{E}[\boldsymbol{S}] = \frac{n-1}{n}\boldsymbol{\Sigma}$.

Proof. This is a more straigh forward computation. □

Hence we obtain an unbiased estimator as

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top.$$

2.2.3 Quadratic forms

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ be a $(p \times 1)$ random vector and $\boldsymbol{A} = (a_i j) \in \mathbb{R}^{p\times p}$ a matrix. This gives a *quadratic form*:

$$\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} = \sum_{i,j} X_i a_{ij} X_j.$$

Theorem 12. *Suppose* $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *then the* trace formula *holds:*

$$\boxed{\mathbb{E}\left[\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}\right] = \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}} \tag{8}$$

Proof. Using that $\operatorname{Cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j]$ we obtain the result:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}\right] &= \sum_{i,j} a_{ij}\mathbb{E}[X_i X_j] = \sum_{i,j} a_{ij}(\Sigma_{ij} - \mu_i\mu_j) \\ &= \sum_{i,j} a_{ij}\Sigma_{ij} - \sum_{ij} a_{ij}\mu_i\mu_j = \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) - \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}.\end{aligned}$$

□

2.2.4 Idempotent matrices

Recall that the (square) matrix $\boldsymbol{A}$ is said to be idempotent of $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. We have the following:

Proposition 4. *Let* $\boldsymbol{A}$ *be idempotent.*

1. $\boldsymbol{I} - \boldsymbol{A}$ *is also idempotent.*
2. $\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{A}) = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{A} = \boldsymbol{0}$.
3. *The only non-singular idempotent matrix is* $\boldsymbol{I}$.
4. *All eigenvalues of idempotent matrices are* 0 *or* 1.

Proof. We have:

1. $(\boldsymbol{I} - \boldsymbol{A})^2 = \boldsymbol{I}^2 - \boldsymbol{A}\boldsymbol{I} - \boldsymbol{I}\boldsymbol{A} + \boldsymbol{A}^2 = 0$.
2. Obvious
3. Suppose $\boldsymbol{A}$ is non-singular. Then $\boldsymbol{I} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$.
4. $\lambda\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{A}\boldsymbol{x} = \lambda^2\boldsymbol{x}$.

□

Let $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top \in \mathbb{R}^{p\times p}$ be symmetric and idempotent. Since only non-zero eigenvalues are 1, we have $\boldsymbol{\Lambda} = \boldsymbol{I}_r$. Also $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{\Lambda}) = r$.

The following theorem will be used extensivelly in the sequel.

Theorem 13. *$\boldsymbol{Z} \sim N(0, \boldsymbol{I}_p)$ and let $\boldsymbol{R}, \boldsymbol{S}$ be symmetric and idempotent of rank r, s respectivelly. Suppose also $\boldsymbol{RS} = 0$. Then*

1. *$\boldsymbol{Z}^\top \boldsymbol{R}\boldsymbol{Z} \sim \chi^2_r$*
2. *$\boldsymbol{Z}^\top \boldsymbol{R}\boldsymbol{Z}$ and $\boldsymbol{Z}^\top \boldsymbol{S}\boldsymbol{Z}$ are independent*
3. *$\frac{\boldsymbol{Z}^\top \boldsymbol{R}\boldsymbol{Z}/r}{\boldsymbol{Z}^\top \boldsymbol{S}\boldsymbol{Z}/s} \sim F_{r,s}$.*

Proof. We have:

1. Use spectral decomposition $\boldsymbol{R} = \boldsymbol{P}\boldsymbol{I}_r\boldsymbol{P}^\top$ to find:

$$\boldsymbol{Z}^\top \boldsymbol{R}\boldsymbol{Z} = (\boldsymbol{P}^\top \boldsymbol{Z})^\top \boldsymbol{I}_r \boldsymbol{P}^\top \boldsymbol{Z} = \boldsymbol{Y}^\top \boldsymbol{Y}.$$

   Note that $\boldsymbol{Y} = \boldsymbol{P}^\top \boldsymbol{Z}$ is r-variate normal and compute its expectation and variance to be 0 and $\boldsymbol{I}_r$. Then finally conclude that $\boldsymbol{Y}^\top \boldsymbol{Y}^\top \boldsymbol{Y} = Y_1^2 + \cdots + Y_r^2 \sim \chi^2_r$.

2. Compute $\operatorname{Cov}[\boldsymbol{R}\boldsymbol{Z}, \boldsymbol{S}\boldsymbol{Z}] = \boldsymbol{R}\boldsymbol{S} = \boldsymbol{0}$. Hence $\boldsymbol{R}\boldsymbol{Z}, \boldsymbol{S}\boldsymbol{Z}$. Then the measurable function $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{x}$ gives independence of $\boldsymbol{Z}^\top \boldsymbol{R}\boldsymbol{Z}$ and $\boldsymbol{Z}^\top \boldsymbol{S}\boldsymbol{Z}$.

3. By definition of Fisher distribution.

□

3 Multiple Linear Regression

3.1 Model and assumptions

We assume that we have $i = 1, \ldots, n$ observations of a *response variable* Y_i depending on k *explanatory variables* x_{ij} through a linear model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i.$$

It can be written on matrix form as:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{\boldsymbol{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}.$$

The matrix $\boldsymbol{X}$ is reffered to as the *design matrix*. The ε's are *errors* and the β's the *parameters*. We further assume:

1. $\boldsymbol{X}$ is of cull column rank.
2. $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$.
3. Homostochastic: $\operatorname{Var}[\varepsilon_i] = \sigma^2 \quad \forall i$.
4. If $\boldsymbol{X}$ is random, then 2 and 3 are conditioned on $\boldsymbol{X}$.
5. Normality of errors: $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$.

From the fift assumption it follows that when $\boldsymbol{X}$ is non-random we have

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

We denote the estimators of $\boldsymbol{\beta}, \boldsymbol{\sigma}^2$ by $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2$. From these we obtain *fitted values*:

$$\widehat{Y}_i = \widehat{\beta}_0 + \cdots + \widehat{\beta}_k x_{ik} = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}.$$

We define *residuals* by:

$$\begin{aligned} \widehat{\varepsilon}_i &= Y_i - \widehat{Y}_i, \\ \widehat{\boldsymbol{\varepsilon}} &= \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}. \end{aligned}$$

3.2 Parameter estimation

When estimating the above parameters, there are two approaches. We may use the *least squares estimator* (LSE):

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 = \underset{\boldsymbol{\beta} \in \mathbb{R}^{k+1}}{\arg\min} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

or we may use the *maximum likelihood estimator* (MLE):

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}{\arg\max}\, L(\beta), \quad L(\beta) = \prod_{i=1}^{n} f(Y_i).$$

It turns out that with our assumptions the result is the same. For the LSE, differentiating the sum of squares and equating to zero yields:

$$\boxed{\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}} \tag{9}$$

Having found this, we denote the *fitted values* by

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top}_{\boldsymbol{H}}\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}.$$

The matrix $\boldsymbol{H}$ is called the *prediction matrix* or *hat matrix*, and is of special interest:

Proposition 5. *For the hat matrix we have:*

1. $\boldsymbol{H}$ *is symmetric.*
2. $\boldsymbol{H}$ *is idempotent.*
3. $\operatorname{rank}(\boldsymbol{H}) = p$.
4. *Residuals can be expressed* $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ *with* $\boldsymbol{I} - \boldsymbol{H}$ *symmetric, idempotent of rank* $n - p$.

Our next goal is to estimate σ^2. More computation shows that the MLE is given by

$$\widehat{\boldsymbol{\sigma}}^2 = \frac{1}{n}\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}},$$

but this is skewed as

$$\mathbb{E}\left[\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}\right] = \sigma^2(n-p).$$

Hence our unbiased estimator is:

$$\boxed{\widehat{\boldsymbol{\sigma}}^2 = \frac{1}{n-p}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).} \tag{10}$$

3.2.1 Properties of the the estimators, fitted values and residuals

We begin by remarking that $\boldsymbol{X}\boldsymbol{\beta}$ is a linear combination of colums of $\boldsymbol{X}$ and hence lies in $\operatorname{col}(X)$. The same is true for $\widehat{\boldsymbol{Y}}$.

Proposition 6. *We have:*

1. $\boldsymbol{\varepsilon} \perp \widehat{\boldsymbol{Y}}$ *and* $\widehat{\boldsymbol{\varepsilon}} \perp \operatorname{col}(\boldsymbol{X})$.
2. $\sum_{i=1}^n \widehat{\varepsilon}_i = 0$ *and* $\sum_{i=1}^n Y_i = \sum_{i=1}^n \widehat{Y}_i$.
3. $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$.
4. $\widehat{\boldsymbol{\varepsilon}} \sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))$.

5. $\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.

6. $\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2$ *are independent.*

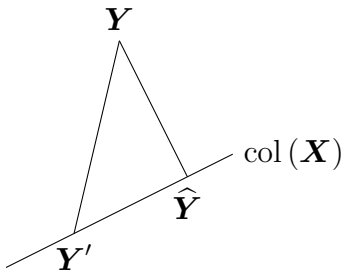

Figure 2: $\widehat{\boldsymbol{Y}}$ is the projection onto the column space of $\boldsymbol{X}$.

Proof. TODO: lecture 10 □

3.2.2 Inference about β_j

Our next goal is to make confidence intervals and to perform t-tests. Recall first that the random variable T has (by definition) the *Student's t-distribution* with m degrees of freedom if it can be written as:

$$T = \frac{Z}{\sqrt{V/m}}$$

where $Z \sim N(0,1), V \sim \chi^2_m$ are independent. We may then find values for $t_{\alpha,m}$ s.t. $\mathbb{P}\left[T \geq t_{\alpha,m}\right] = \alpha$ in tables.

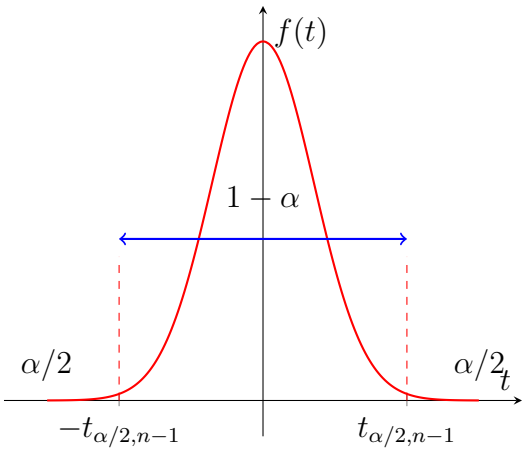

Figure 3: Two sided inference with the student-t distribution.

We have seen that $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$. Denote $(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = (e_{ij})_{i,j=1,\ldots,p}$. We then have $\widehat{\beta}_j \sim N(\beta_j, \sigma^2 e_{jj})$, so $\frac{\widehat{\beta}_j - \beta_j}{\sqrt{e_{jj}}\sigma} \sim N(0,1)$. Since the variance is unknown, consider the statistic:

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{e_{jj}\widehat{\sigma}}} = \frac{(\widehat{\beta}_j - \beta_j)/\sigma\sqrt{e_{jj}}}{\sqrt{\frac{(n-p)\widehat{\sigma}^2}{\sigma^2}/(n-p)}}.$$

Recalling the properties of the estimators, we know that $\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2$ are independent, that the numerator is $N(0,1)$-distributed and that $V = \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$. Hence, we may conclude:

$$\boxed{T = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{e_{jj}\widehat{\sigma}}} \sim t_{n-p}} \tag{11}$$

With this, we may constrict *confidence interval* by rewriting the inequalities in the expression:

$$\mathbb{P}\left[-t_{\frac{\alpha}{2},n-p} \leq \frac{\widehat{\beta}_j - \beta_j}{\sqrt{e_{jj}\widehat{\sigma}}} \leq t_{\frac{\alpha}{2},n-p}\right] = 1-\alpha.$$

We may also perform *hypothesis testing*. Consider the following test at significance level α:

$$\mathrm{H}_0 : \beta_j = 0, \qquad \mathrm{H}_1 : \beta_j \neq 0.$$

Under H_0 we have $T = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{e_{jj}\widehat{\sigma}}} \sim t_{n-p}$. The *critical region* under two-sided alternative is:

$$|T| \geq t_{\frac{\alpha}{2},n-p} \Rightarrow \mathrm{H}_0 \text{ is rejected.}$$

TODO: R printout with explanation of culumns

TODO: Does R do one or two sided hypothesis test ???

3.3 Some notes on independence

We sometimes use that plugging independente random variables through some functions result in new independent random variables. The theorem below tells us when this is okay.

Theorem 14. *Suppose X, Y are independent random variables and that f, g are two measurable functions. Then $f(X), g(Y)$ are also independent.*

A *measurable function* is a function s.t. the preimages of Borel sets are measurable in the given probability space. In particular, continous functions are measurable.

3.4 Analysis of variance (ANOVA)

The following theorem forms the basis on our discussion of ANOVA.

Theorem 15. *Assuming the necesarry assumptions, we have the* ANOVA decomposition*:*

$$\boxed{\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}_{\text{SSE}}}$$

Proof. We first split the sum as:

$$\begin{aligned}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 &= \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i + \widehat{Y}_i - \bar{Y}\right)^2 \\ &= \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}^{n}\left(Y_i - \widehat{Y}\right)\left(\widehat{Y}_i - \bar{Y}\right).\end{aligned}$$

Using again the properties of the estimators, we find that the last sum is 0:

$$\sum_{i=1}^{n}\left(Y_i-\widehat{Y}\right)\left(\widehat{Y}_i-\bar{Y}\right)=\underbrace{\sum_{i=1}^{n}\overbrace{(Y_i-\widehat{Y}_i)}^{\varepsilon_i}\widehat{Y}_i}_{=\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{Y}}=0}-\bar{Y}\underbrace{\sum_{i=1}^{n}\left(Y_i-\widehat{Y}\right)}_{=0\text{ by property 2}}.$$

□

The 3 sums are called *total sum of squares*, *regression sum of squares* and *error sum of squares* respectivelly. This decomposition motivates the following definition.

Definition 3. The part of the total variation due to the model is called the *coefficient of determination* or the *R2-score*:

$$\boxed{R^2=\frac{\text{SSR}}{\text{SST}}\overset{\text{thm}}{=}1-\frac{\text{SSE}}{\text{SST}}} \tag{12}$$

The R2-score is a measure of goodness-of-fit as it tells us how much of the variation in the data can be explained by the model. One may also prove another representation:

$$R^2=\frac{\left(\sum_{i=1}^{n}(Y_i-\bar{Y})(\widehat{Y}_i-\bar{Y})\right)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2\sum_{i=1}^{n}(\widehat{Y}_i-\bar{Y})}.$$

This is the square of the empirical correlation between $\boldsymbol{Y},\widehat{\boldsymbol{Y}}$.

3.4.1 Fictional model

The following discussion will examine what happens when an explanatory variable is explained by the other explanatory variables. We introduce a *fictional model* using x_{ij} as response for some fixed feature j. Wlog use feature k. The model is:

$$x_{i,k}=\alpha_0+\alpha_1x_{i,1}+\cdots+\alpha_{k-1}x_{i,k-1}+\delta_i.$$

As usual, we assume $\boldsymbol{\delta}=(\delta_1,\ldots,\delta_n)^\top\sim N(0,\sigma^2\boldsymbol{I})$. We may then estimate $\boldsymbol{\alpha}=(\alpha_0,\ldots,\alpha_{k-1})^\top$ and σ^2 by the usual $\widehat{\boldsymbol{\alpha}},\widehat{\sigma}^2$ to obtain fitted $\widehat{x}_{ik}$. We find the squared empirical correlation between $x_{i,k}$ and $\widehat{x_{i,k}}$ as:

$$R_k^2=\frac{\sum_{i=1}^{n}(\widehat{x}_{ik}-\bar{x_k})^2}{\sum_{i=1}^{n}(x_{ik}-\bar{x_k})^2}=\frac{\left(\sum_{i=1}^{n}(x_{ik}-\bar{x_k})(\widehat{x}_{ik}-\bar{x_k})\right)^2}{\sum_{i=1}^{n}(x_{ik}-\bar{x_k})^2\sum_{i=1}^{n}(\widehat{x}_{ik}-\bar{x_k})^2}.$$

We call it the coefficient of determination for x_{ik} as response. Repeating the procedure for the remaining $x_{ij},j=1,\ldots,k-1$ we obtain $R_1^2,\ldots,R_k^2$. It turns out that:

$$\text{Var}\left[\widehat{\beta}_j\right]=\frac{\sigma^2}{(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)}.$$

So the more a variable is explained by the other, the higher the variance of the estimator.

3.4.2 Further expressions for the sums of squares

Recall that $\boldsymbol{C}, \boldsymbol{H}$ are both symmetric an idempotent. For the total sum of squares, using that the centering matrix $\boldsymbol{C}$ is idempotent, we obain:

$$\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = (\boldsymbol{CY})^\top (\boldsymbol{CY}) = \boldsymbol{Y}^\top \boldsymbol{CY} = \boldsymbol{Y}^\top \left(\boldsymbol{I} - \frac{1}{n} \boldsymbol{1}\boldsymbol{1}^\top \right) \boldsymbol{Y}.$$

For the residual sum of squares we also need the fact that $\boldsymbol{H}\boldsymbol{x}_i = \boldsymbol{x}_i$ for all columns of $\boldsymbol{X}$. This follows readily as $\boldsymbol{HX} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{X}$. From this we have $\boldsymbol{H}\boldsymbol{1} = \boldsymbol{1}$ as this is the first column of $\boldsymbol{X}$.

$$\begin{aligned} \text{SSR} &= \sum_{i=1}^{n} (\widehat{Y}_i - \bar{Y})^2 = (\boldsymbol{CHY})^\top (\boldsymbol{CHY}) = \boldsymbol{Y}^\top \boldsymbol{HCHY} \\ &= \boldsymbol{Y}^\top \left(\boldsymbol{H} - \frac{1}{n} \boldsymbol{H}\boldsymbol{1}\boldsymbol{1}^\top \boldsymbol{H}^\top \right) \boldsymbol{Y} = \boldsymbol{Y}^\top \left(\boldsymbol{H} - \frac{1}{n} \boldsymbol{1}\boldsymbol{1}^\top \right) \boldsymbol{Y}. \end{aligned}$$

About the matrix $\boldsymbol{H} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top$, we note that it is symmetric, idempotent and of rank $p-1$:

$$\text{rank} \left(\boldsymbol{H} - \frac{1}{n} \boldsymbol{1}\boldsymbol{1}^\top \right) = \text{tr} \left(\boldsymbol{H} - \frac{1}{n} \boldsymbol{1}\boldsymbol{1}^\top \right) = \text{tr}\left(\boldsymbol{H}\right) - \frac{1}{n} \text{tr}\left(\boldsymbol{1}\boldsymbol{1}^\top\right) = p - 1.$$

Finally, for the error sum of squares we obtain using that $\boldsymbol{I} - \boldsymbol{H}$ is symmetric and idempotent:

$$\text{SSE} = \sum_{i=1}^{n} (\widehat{Y}_i - Y_i)^2 = \widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}} = ((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y})^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \cdots = \boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$$

3.5 F-test

We first consider the hypothesis test:

$$\text{H}_0 : \beta_i = 0 \quad \forall i \in \{1, \ldots, k\}, \qquad \text{H}_1 : \beta_i \neq 0 \text{ for at least one } i \in \{1, \ldots, k\}.$$

As usual we denote the significance level by α. Our test statistic F is:

$$\boxed{F = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} \sim F_{p-1,n-p}} \tag{13}$$

And as usual the test is $F \geq f_{\alpha,p-1,n-p} \Rightarrow \text{H}_0$ is rejected.

Proof. TODO: Uke 8 □

3.6 General F-test

We set up a much more general problem. Let $A \in \mathbb{R}^{r \times p}$, $r < p$, $\text{rank}(A) = r$, $\boldsymbol{d} \in \mathbb{R}^d$. We test the hypothesis:

$$\text{H}_0 : A\boldsymbol{\beta} = \boldsymbol{d}, \qquad \text{H}_1 : A\boldsymbol{\beta} \neq \boldsymbol{d}.$$

Some special cases of this general setup are.

- $r = 1, d = 0, A = (0, \ldots, 1, \ldots, 0)$ with 1 at index i, gives the test

$$\mathrm{H}_0 : \beta_i = 0, \qquad \mathrm{H}_1 : \beta_i \neq 0.$$

- $r = 1, d = 0, A = (0, \ldots, 1, \ldots, -1, \ldots, 0)$ with 1 at index i and -1 at index j, gives the test

$$\mathrm{H}_0 : \beta_i = \beta_j, \qquad \mathrm{H}_1 : \beta_i \neq \beta_j.$$

- $r = k, d = \mathbf{0} \in \mathbb{R}^k, A = (\mathbf{0}, \mathrm{diag}(1)) \in \mathbb{R}^{k \times p}$, gives the test

$$\mathrm{H}_0 : \beta_i = 0 \quad \forall i \in \{1, \ldots, k\}, \qquad \mathrm{H}_1 : \beta_i \neq 0 \text{ for some } i \in \{1, \ldots, k\}.$$

  This is the F-test of the previous section.

Let $\mathcal{B}$ be the space of $\boldsymbol{\beta}$ satisfying H_0. The restricted problem is:

$$\widehat{\boldsymbol{\beta}}^R = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Using lagrange multipliers and a bag of tricks, we obtain:

$$\widehat{\boldsymbol{\beta}}^R = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{A}^\top (\boldsymbol{A}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{A}^\top)^{-1} (\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{d}).$$

Denoting $\Delta = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^R$, we find:

$$\mathrm{SSE}^R = \mathrm{SSE} + \Delta^\top \boldsymbol{X}^\top \boldsymbol{X} \Delta$$

We claim that the under H_0, we have

$$\boxed{F = \frac{\left(\mathrm{SSE}^R - \mathrm{SSE}\right)/r}{\mathrm{SSE}/(n-p)} \sim F_{r,n-p}} \tag{14}$$

Proof. TODO: Uke 8 □

3.7 Transformations of data

Some models common in applications are:

1. $Y = \beta_0 + \beta_1 \frac{1}{x} + \varepsilon$
2. $Y = \frac{1}{\beta_0 x + \beta_1 + \varepsilon}$
3. $Y = \frac{x}{\beta_0 x + \beta_1 + x\varepsilon}$

These can also be analysed using linear regression after undergoing some *transformations*, namelly:

1. $\tilde{x} = \frac{1}{x} \rightsquigarrow Y = \beta_0 + \beta_1 \tilde{x} + \varepsilon$
2. $\tilde{Y} = \frac{1}{Y} \rightsquigarrow \tilde{Y} = \beta_0 + \beta_1 x + \varepsilon$
3. $\tilde{x} = \frac{1}{x}, \tilde{Y} = \frac{1}{Y} \rightsquigarrow \tilde{Y} = \beta_0 + \beta_1 \tilde{x} + \varepsilon$

We do not always know what form the model should have, which motivates the following section.

3.7.1 Box-Cox transformation

The *Box-Cox transformation* is a power transform of the form:

$$\boxed{u_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \ln y, & \text{if } \lambda = 0. \end{cases}} \tag{15}$$

We immediatelly see that this requires $\boldsymbol{Y} > 0$ the response $\boldsymbol{Y}$. To remedy negative $\boldsymbol{Y}$, simply shift them. To apply the transform, suppose that $\tilde{Y} = u_\lambda(Y)$ is normal and choose λ which maximises the log-likelihood function. This turns out to not have analytic solution. To find the optimal λ, we therefore perform a *grid search*. In practice, we use the R function `boxcox`.

3.7.2 Variance stabilising transformation

Suppose $\mu_i = \mathbb{E}[Y_i]$ and that $\text{Var}[Y_i]$ depends on μ_i by say $\text{Var}[Y_i] = h(\mu_i)$. Our goal is to transform by $\tilde{Y} = g(Y)$ so that the variance depends less on μ. First order taylor expansion of $g(Y)$ around μ gives:

$$g(y) \approx g(\mu) + g'(\mu)(y - \mu).$$

We find:

$$\begin{aligned} \mathbb{E}[g(Y)] &\approx g(\mu) + g'(\mu) \overbrace{(\mathbb{E}[Y] - \mu)}^{=0} = g(\mu) \\ \text{Var}[g(Y)] &\approx \text{Var}[g(\mu) + g'(\mu)(Y - \mu)] = g'(\mu)^2 \text{Var}[Y] \end{aligned}$$

Choose $g(y) = \int_c^y \frac{1}{\sqrt{h(\mu)}} d\mu$. Then $g'(y) = \frac{1}{\sqrt{h(y)}}$ so $\text{Var}[g(Y)] = 1$.

4 Model Analysis, Selection and Multiple Hypothesis Testing

4.1 Model analysis

Given a linear model, we can to some examination on the basis of visual analysis of residuals

TODO: Example QQ plots ?

TODO: not homostochastic, not independent ... transformed residuals free of trouble... TODO: long discussion of standardized / studentizes residuals ...

4.2 Model selection

Some challenges when selecting model include:

1. Unnecessary explanatory variables $\rightsquigarrow$ overfitting
2. Missed relevant explanatory variables $\rightsquigarrow$ underfitting
3. Quality of predictor $\not\Longleftrightarrow$ Quality of estimator

TODO: example from 2017 exam

Underfitting leads to biased estimation

Overfitting leads to greater variance in the estimators

4.2.1 Which model to choose

Suppose k covariates. Then we have 2^k possible models from maximal:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

to minimal:

$$Y_i = \beta_0.$$

We want to arrive at a compromise between simplisity and goodness of fit. It is clear that the usual R2-score will not decrease by introducing additional variables, so we need to intruduce penalty for extra variables. Let $\widehat{\sigma}^2$ be obtained from maximal model and SSE, p from the model of consideration.

1. *Adjusted coefficient of determination*:

$$\boxed{R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-p-1}}$$

2. *Mallow's Cp parameter* ("Complexity parameter"):

$$C_p = \frac{\text{SSE}}{\widehat{\sigma}^2} - n + 2p.$$

3. *Akaibe information criterion*:

$$\text{AIC} = \frac{\text{SSE}/n}{\widehat{\sigma}^2} + 2\frac{p}{n}.$$

4. *Bayesian information criterion*

$$\text{BIC} = \frac{\text{SSE}/n}{\widehat{\sigma}^2} + \ln n\frac{p}{n}.$$

Note that for the 3 last options we want to minimise the statistic. In practice selection is done either with software or in to steps:

1. For each $j = 1, \ldots, k$ choose the model with j covariates of maximal R2-score. This gives k models, so far not penalised.
2. Consider these k best models and choose "'best of the best" using one of the criteria.

4.3 Multiple hypothesis testing

As before we consider the test:

$$\text{H}_0 : \beta_1 = \cdots = \beta_k = 0.$$

If this is now rejected, what about the individual tests β_j? We can test individually using t-tests the hypothesis:

$$\text{H}_0 : \beta_j = 0.$$

But if we have many parameters, $\mathbb{P}$ [at least one type I error] is large. In the worst case, for m independent hypothesis we can compute this probability to be $1 - (1 - \alpha)^m$, which at significance $\alpha = 0.05$ and $m = 5, 20$ gives a probability $> 0.22, > 0.64$ respectivelly. We will now consider methods to resolve this issue.

4.3.1 p-value is a random variable

We recall that given an observation t for the test statistiv T we can compute the p-value, the probability of an equal of more extreme observation:

$$p(t) = \mathbb{P}_{\text{H}_0}(T \geq t).$$

Since t is the observet value of a random variable, it is clear that the p-value is also random. The rejection criteria for H_0 is:

$$\text{H}_0 \text{ is rejected} \Leftrightarrow p(t) \leq \alpha.$$

An equivalent expression for the rejection criteria is:

$$\text{H}_0 \text{ is rejected} \Leftrightarrow t \geq t' \text{ with } \mathbb{P}_{\text{H}_0}(T \geq t') = \alpha.$$

Hence we have

$$\mathbb{P}_{\text{H}_0}(p(t) \leq \alpha) = \alpha,$$

from which we conclude that under true H_0, the p-value is uniformly distributed on $[0, 1]$.

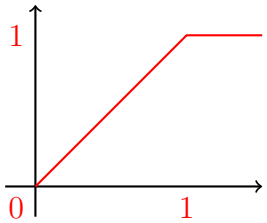

Figure 4: Distribuon of the p-value.

4.3.2 Testing m hypothesis

We begin by finding the m p-values $(p_1, \ldots, p_m)$ corresponding to each hypothesis. We then need to choose the *local significance level* α_{loc} and follow the rule that if p-value$\leq \alpha_{\text{loc}}$ we reject the corresponding H_0. We need to choose α_{loc} such that the probability of at least one type I error is α. The outcomes of our hypothesis test can be summarised by the table: Note that we know m,

Table 1: Multiple testing set-up

	H0 not rejected	H0 rejected	total
H_0 true	U	V	m_0
H_0 false	T	S	$m - m_0$
total	$m - R$	R	m

the number of hypothesis, and the number R of rejected hypothesis. Our goal is to control the *familywise error rate (FWER)*. This is the probability of at least one false positive finding:

$$\text{FWER} = \mathbb{P}\left[V \geq 1\right] = 1 - \mathbb{P}\left[V = 0\right].$$

For independent hypothesis we find

$$\begin{aligned}\text{FWER} &= 1 - \mathbb{P}\left[p_1 > \alpha_{\text{loc}}, \ldots, p_m > \alpha_{\text{loc}}\right] \\ &= 1 - \mathbb{P}\left[p_1 > \alpha_{\text{loc}}\right] \ldots \mathbb{P}\left[p_m > \alpha_{\text{loc}}\right] \\ &= 1 - (1 - \alpha_{\text{loc}})^m.\end{aligned}$$

If we for each hypothesis define the event:

$$R_j = \{\text{the j-th } H_0 \text{ is rejected but true}\} = \{p_j \leq \alpha_{\text{loc}}\}.$$

Then $\bar{R}_j = \{p_j > \alpha_{\text{loc}}\}$ is the complementary event. We may write:

$$\text{FWER} = \mathbb{P}\left[R_1 \cup \cdots \cup R_m\right] = 1 - \mathbb{P}\left[\bar{R}_1 \cap \cdots \cap \bar{R}_m\right].$$

We present two methods for choosing α_{loc}

1. The *Bonferrony method* uses subadditivity to bound the $FWER$ by:

$$\alpha = \text{FWER} = \mathbb{P}\left[R_1 \cup \cdots \cup R_m\right] \leq \sum_{j=1}^{m} \mathbb{P}\left[R_j\right] = m\alpha_{\text{loc}}.$$

Hence the Bonferrony method is to let:

$$\boxed{\alpha_{\text{loc}} = \frac{1}{m}\alpha} \tag{16}$$

We make the following remarks:

(a) This gives strong control (it works under any combination of true and false hypothesis).

(b) To get equality we need disjoint events / perfectly negativelly assosiated hypothesis.

(c) It is conservative: modelling dependencies may give lover requirement.

2. For the *Šidák method* we make the assumption that we have *independent tests*. Then we have seen that $\text{FWER} = 1 - (1 - \alpha_{\text{loc}})^m$. Hence the method is:

$$\boxed{\alpha_{\text{loc}} = 1 - (1 - \alpha)^{1/m}} \tag{17}$$

Again, we have strong control. The estimate is also exact when the tests are independent. It is conservative for positivelly dependent tests and liberal for negatively dependent tests. It is possible to show that the Šidák correction is always greater that the Bonferrony correction, so it is slighly less conservative; but we need to assume independence for exactness.

5 ANOVA and Design of Experiment

5.1 Analysis of varance (ANOVA)

From wikipedia; "in its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means. In other words, the ANOVA is used to test the difference between two or more means." We will investigate this.

Suppose we have p treatments/strategies, each called a *level*, for maximising some response (health/income etc.). Suppose further that we have n_i samples of each level $i = 1, \dots, p$:

1^{st} level	$Y_{1,1}, \dots, Y_{n_1,1}$
2^{nd} level	$Y_{2,1}, \dots, Y_{n_2,2}$
$\dots$	
p^{th} level	$Y_{p,1}, \dots, Y_{n_p,p}$

We model

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \quad 1 \le j \le p, \quad 1 \le i \le n_j.$$

We furhter assume that all $\varepsilon \sim N(0, \sigma^2)$ are independent. Our goal is to compare the values of μ. For instance we may want to test

$$\text{H}_0 : \mu_1 = \cdots = \mu_p \quad \text{VS} \quad \text{H}_1 : \text{not all equal.}$$

We want to rewrite the problem so that we can use the tools we have developed for linear regression. It is then more convenient to use the following model. Let $\mu = \frac{1}{N}\sum_{j=1}^{p} n_j \mu_j$ be the overall average, where $N = n_1 + \cdots + n_p$. Denote the deviation from the overall average by $\alpha_j = \mu_j - \mu$. Our model then becomes:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}.$$

Note further that by definition of μ, we have $\sum_{j=1}^{p} n_j \alpha_j = 0$. In other words, one of the deviations can be expressed by the rest. In the following we simplify by assuming equal samples $n_1 = \cdots = n_p$. Then we can express $\alpha_p = -\alpha_1 - \cdots - \alpha_{p-1}$. Some more rewriting left to the reader gives the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where:

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{p-1} \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} \mathbf{1}_n & \mathbf{1}_n & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{1}_n & \mathbf{0} & \mathbf{1}_n & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_n & -\mathbf{1}_n & -\mathbf{1}_n & \dots & -\mathbf{1}_n \end{pmatrix}.$$

Our hypothesis test is equivalent to

$$\text{H}_0 : \alpha_1 = \cdots = \alpha_{p-1} = 0 \quad \text{VS} \quad \text{H}_1 : \text{not all 0.}$$

Hence, we may use the F-test that we developed in Section 3.5. Using the general F-test we may test many other as well.

5.1.1 Two-way ANOVA

We may also have 2 levels both present at the same time. Our model is then:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Adding replicates at the different combinations of i, j we have

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad k = 1, \dots, n_{ij}.$$

Finally we may add an interaction term between the two:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}.$$

Also here we have constraints:

$$\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{s} \beta_j = \sum_{i=1}^{r} \gamma_{ij} = \sum_{j=1}^{s} \gamma_{ij} = 0.$$

5.2 Two level factorial design

We suppose we have k main factors $x_1, \dots, x_k$ making up a model of the form:

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_k \boldsymbol{x}_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

We only allow for our covariates x_i to be at *low level* and *high level*. This is moddeled as $-1, +1$, and is the first of the following assumptions. The second assumption will ensure simple computation and the most precise inference about coefficients:

1. Each column has entries ± 1.
2. The colomns are orthogonal, i.e. $\mathbf{1}^T \boldsymbol{x}_i = \sum_{i=1}^{n} \boldsymbol{x}_{ij} = 0$ and $\boldsymbol{x}_i^T \boldsymbol{x}_j = n\delta_{ij}$.

This in particular implies that we have $\boldsymbol{X}^T \boldsymbol{X} = nI_n$. Using results from regression analysis, this significantly simplifies our estimators. We get $\widehat{\boldsymbol{\beta}} = \frac{1}{n} \boldsymbol{X}^T \boldsymbol{Y}$:

$$\beta_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} Y_j.$$

The estimate of the j'th coefficient only depents on the j'th covariate! Hence removal of covariates leaves the rest unchanged. We call the covariates $x_1, \dots, x_k$ the main factors. A central question in experimentation is what happens when a factor goes from high to low level. This motivates the definition:

Definition 4. The *main effect* of main factor j is defined as:

$$\begin{aligned} \text{effect}_j &= \text{expected response at high level} - \text{expected response at low level} \\ &= \mathbb{E}\left[Y | x_j = 1\right] - \mathbb{E}\left[Y | x_j = -1\right] = 2\beta_j. \end{aligned}$$

The estimated effect is naturally by:

$$2\widehat{\beta}_j = \frac{1}{n/2}\sum_{x_{ij}=1} Y_i - \frac{1}{n/2}\sum_{x_{ij}=-1} Y_i$$
$$= \text{estimated response at high level} - \text{estimated response at low level} =: \widehat{\text{effect}_j}$$

Another consequence of the second assumption is that $\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \frac{\sigma^2}{n} I\right)$. Hence the components are independent with the same variance σ^2/n.

We also remark that we may express the regression sum of squares as:

$$\begin{aligned}\text{SSR} &= \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}\left(\sum_{j=0}^{n} x_{ij}\widehat{\beta}_j - \widehat{\beta}_0\right)^2 \\ &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij}\widehat{\beta}_j + \overbrace{x_{i0}}^{=1}\widehat{\beta}_0 - \widehat{\beta}_0\right)^2 = (\boldsymbol{X}^*\widehat{\boldsymbol{\beta}}^*)^T(\boldsymbol{X}^*\widehat{\boldsymbol{\beta}}^*) \\ &= \widehat{\boldsymbol{\beta}}^{*T}\boldsymbol{X}^{*T}\boldsymbol{X}^*\widehat{\boldsymbol{\beta}}^* = \widehat{\boldsymbol{\beta}}^{*T} n I_k \widehat{\boldsymbol{\beta}}^* = n\sum_{j=1}^{n}\widehat{\beta}_j^2.\end{aligned}$$

Here $\boldsymbol{X}^*$ is the feature matrix without the first column $\mathbf{1}$. The expression shows that $n\widehat{\beta}_j$ may be interpreted as the amount of variability explained by factor j.

5.2.1 Modelling interactions

To go from this to a *2^k-design*, we take into account all interactions of the factors. These are modelled as products of main factors:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{1,2}x_1x_2 + \cdots + \beta_{k-1,k}x_{k-1}k_k + \cdots + \beta_{1,2,\ldots,k}x_1\cdots x_k.$$

We extend the design matrix accordingly, and note that we still satisfy the assumptions. Hence we still have $\widehat{\boldsymbol{\beta}} = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{Y}$.

We define the *interaction between two factors* as half the main effect of one when the other is at high level minus half the main effect when the other is at low. This encapsulates that if the change in respons between levels of factor A is not independent of the level of B there is an interaction. Given the definition, it is natural to get the *estimated interaction effect* of 1 and 2 as $2\widehat{\beta}_{1,2}$. To see that this is natural, rewrite to obtain:

$$2\widehat{\beta}_{1,2} = \frac{1}{1/2}\underbrace{\left(\frac{1}{n/2}\sum_{\substack{x_{i1}=1\\x_{i2}=1}} Y_i - \frac{1}{n/2}\sum_{\substack{x_{i1}=-1\\x_{i2}=1}} Y_i\right)}_{\text{Estimated effect of 1 with 2 held at high}} - \frac{1}{1/2}\underbrace{\left(\frac{1}{n/2}\sum_{\substack{x_{i1}=1\\x_{i2}=-1}} Y_i - \frac{1}{n/2}\sum_{\substack{x_{i1}=-1\\x_{i2}=-1}} Y_i\right)}_{\text{Estimated effect of 1 with 2 held at low}}$$

5.2.2 Inference about effect

We have seen that $\widehat{\text{Effect}_j} \sim N(\text{Effect}_j, \frac{4\sigma^2}{n})$. To make inference about the effect/coefficients we need to make inference about σ^2. We cannot use estimator from multiple linear regression since for MLR we have $\widehat{\sigma}^2 = \frac{\text{SSE}}{n-p}$ and here $n = p$. If we don't have replicates in our experiment, to increase n, we have to resort to one of two methods.

1. We can neclect m higher order factors where it is reasonable that $\text{Effect}_j = 0$. Then we have $\widehat{\text{Effect}}_j \sim N(0, \frac{4\sigma^2}{n})$. Letting $\sigma^2_{\text{effect}} = \frac{4\sigma^2}{n}$ we then estimate:

$$\widehat{\sigma}^2_{\text{effect}} = \frac{1}{m} \sum_{\text{negligable}} \widehat{\text{Effect}}_j^2$$

2. The idea of *Lenth's method* / *pseudo standard error* (PSE) is to compute:

$$C_1, \ldots, C_m \text{ - the estimated effects } \widehat{A}, \widehat{B}, \widehat{AB}, \ldots$$

We then do the following:

(a) Order $|C_j|$ in increasing order.

(b) Find the median M of $|C_1|, \ldots, |C_m|$ and compute $\delta_0 = 1.5M$.

(c) Remove the effects C_j with $|C_j| \geq 2.5\delta_0$ and find the median of the rest of $|C_j|$.

Having done this we compute the PSE as

$$PSE = 1.5 \cdot \text{median}\left\{|C_j| : |C_j| < 2.5\delta_0\right\}.$$

This is then used as our estimator; $\widehat{\sigma}_{PSE} = PSE$.

In the case of replications ($n \cdot 2^k$ experiment) we use results from multiple linear regression:

$$\widehat{\sigma}^2 = \frac{SSE}{n - 2^k}, \quad \frac{\widehat{\beta}_j}{\widehat{\sigma}/\sqrt{n}} \sim t_{n-2^k}.$$

5.3 Fractional factorial design (2^{k-r}-design)

One common fractional design is the *half fraction*, where we fullfill only half of the 2^k experiments of the full design. To choose experiments perform:

1. Choose $k-1$ main factors and consider their full 2^{k-1} design.
2. For the k-th main factor, assign the same values as one of the interactions in the above design.

Suppose for instance we are performing a 2^{4-1} design with main factors A, B, C and D. Then to the full design for A, B, C and let:

$$D = ABC.$$

We call the above the *generator of the design*. Multiplication of this equation by D gives:

$$1 = ABCD.$$

This is called the *defining relation*. From the generator or from the defining relation we may obtain the remaining *aliases* / *confounded factors* by multiplying both sides of $1 = ABCD$ by A, B, C, AB, AC, BC to obtain:

$$\begin{array}{lll} A = BCD & B = ACD & C = ABD \\ AB = CD & AC = BD & BC = AD. \end{array}$$

generator The minimal order of $LHS + RHS$ is called the *resolution* of the design. Here the resolution is IV. The greater the resolution the better. This is because at high resolution we ensure that the main factors are only confounded by higher order interactions. We note that when estimating effects, the estimator of confounded factors is the same, in our example for instance:

$$\widehat{D} = \widehat{ABC}.$$

This will encapsulate the joint effect of both D and ABC, which again shows that high resolution is good as we often neglect the higher order interactions anyways.

5.3.1 Blocking

The motivation for *blocking* is that the conditions of the experiments may change. Hence we want a clever way of deciding which experiments to to "today" etc. Suppose that there is a fixed change h to the response, i.e. that "tomorrow" we have

$$Y_i \mapsto Y_i + h.$$

If we choose experiments such that for a given factor, the number of high and low level (+ and −) are equal, the effect of this factor is unchanged by the added h. We usually consider the main factors of greatest importance. Consider for example a 2^3 experiment done over 2 days. We say it is divided in 2 *blocks*. Let block I be done with ABC at − and block II at +. We obtain the following:

Block I						
A	B	C	AB	AC	BC	ABC
+	+	-	+	-	-	-
+	-	+	-	+	-	-
-	+	+	-	-	+	-
-	-	-	+	+	+	-

Block II						
A	B	C	AB	AC	BC	ABC
+	+	+	+	+	+	+
+	-	-	-	-	+	+
-	-	+	+	-	-	+
-	+	-	-	+	-	+

Figure 5: The two blocks when partitioning by the sign of ABC.

In this example our choice is good. Only the 3-factor interaction is affected by the supposed change in response since all other have equally many + and −. It would be much wose if a main factor were confounded with the block effect.

Index