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 ... 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.

Course progress

- First reading
  - [x] Lecture 1-22
  - [ ] Lecture 23
  - [ ] Lecture 24
  - [ ] Lecture 25
- Gjennomgang
  - [x] Lecture 1-2
  - [ ] Lecture 3-4
  - [ ] Lecture 5-6
  - [x] Lecture 7-8
  - [x] Lecture 9-10
  - [x] Lecture 11-12
  - [ ] Lecture 13-14
  - [ ] Lecture 15-16
  - [ ] Lecture 17-18
  - [ ] Lecture 19-20
  - [ ] Lecture 21-22
  - [ ] Lecture 23

Exams

- [ ] May 2023
- [ ] June 2019
- [ ] May 2018
- [ ] May 2017
- [ ] June 2016
- [ ] May 2015
- [ ] May 2014
- [ ] August 2014

Topics

1

Multivariate distributions and expectations (HS 4.1-4.2).

Multivariate moments (HS 4.2 using HS 2.1-2.4).

Transformations (HS 4.3, 4.4)

PCA (HS 11.1-11.3).

Charactestic functions (HS 4.2).

2

Multivariate normal distribution (HS 4.4, 5.1).

Estimation in the multivariate normal distribution (HS 3.3, 4.5).

Quadratic forms and idempotent matrices (FKLM Appendix B, Th. B2, B8).

3

Multiple linear regression: model, parameter estimation (FKLM 3.1, 3.2).

Properties of estimators, fitted values, residuals (FKLM 3.2).

Inference about coefficients (FKLM 3.3).

Multiple linear regression: t-test about coefficients, ANOVA decomposition, coefficient of determination, F-test (FKLM 3.2, 3.3).

General F-test for regression coefficients (FKLM 3.2, 3.3, 3.5).

transformation of data (FKLM 3.2, 3.3, 3.4, 3.5).

4

Model analysis and model selection (FKLM 3.4).

Multiple hypothesis testing (HBL).

Examples.

5

ANOVA (HS 8.1.1).

Design of experiment (DOE): two-level factorial design (T).

Keywords to know

1 Multivariate Distribution and its Generalisations

Lecture 1

random vector - vector with RV's as components

random matrix

cumulative distribution function (CDF)

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

absolutelly continuous if there exists density function 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

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

$\boldsymbol{X}$ 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:

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

Marginal distribution

Let $\boldsymbol{X}_A, \boldsymbol{X}_B$ be two random vectors st $\boldsymbol{X} = (\boldsymbol{X}_A, \boldsymbol{X}_B)^T$ has cdf F. Then:

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

In absolutelly continuous case we find

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

Conditional distribution

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

Independence

Say $\boldsymbol{X}_A, \boldsymbol{X}_B$ 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 $f(\boldsymbol{x}_B|\boldsymbol{x}_A) = f_B(\boldsymbol{x}_B)$.

Similar definition for independence when $\boldsymbol{X}$ has N components and not just 2.

Multivariate expectations and moments

expectation defined as

$$\mathbb{E}\left[\boldsymbol{X}\right] = \left(\mathbb{E}\left[X_1\right],\ldots,\mathbb{E}\left[X_p\right]\right)^T$$

Lecture 2

We can show $\mathbb{E}\left[a\boldsymbol{X} + b\boldsymbol{Y}\right] = a\mathbb{E}\left[\boldsymbol{X}\right] + b\mathbb{E}\left[\boldsymbol{Y}\right]$

For (shape compatible) matrices $\boldsymbol{A},\boldsymbol{B}$ we have $\mathbb{E}\left[\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}\right] = \boldsymbol{A}\mathbb{E}\left[\boldsymbol{X}\right]\boldsymbol{B}$

Let $\boldsymbol{X},\boldsymbol{Y}$ be random matrices whose product is defined. Then $\mathbb{E}\left[\boldsymbol{X}\boldsymbol{Y}\right] = \mathbb{E}\left[\boldsymbol{X}\right]\mathbb{E}\left[\boldsymbol{Y}\right]$.

Covariance matrix

Let $\boldsymbol{X} = (X_1,\ldots,X_p)^T$ and $\mathbb{E}\left[\boldsymbol{X}\right] =: \boldsymbol{\mu} = (\mu_1,\ldots,\mu_p)$. We then define:

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

This natrix is symmetric. We can also show:

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

The correlation matrix (with ones on the diagonal) 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 correlation matrix by

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

1.0.1 Matrix algebra

symmetric if $\boldsymbol{A}^T = \boldsymbol{A}$

orthogonal if $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$

eigenvalue and *eigenvector* ... solution of $\det(\boldsymbol{A} - \lambda\boldsymbol{I})$

also have $\det A = \prod_{i=1}^p \lambda_i$.

Jordan decomposition of symmetric matrtix

$$\boldsymbol{A} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T, \quad \boldsymbol{\Gamma} = \textit{TODO :}$$

Quadtraic form: let $\boldsymbol{A}$ symmetric, $\boldsymbol{x}$ a $(p \times 1)$ vector:

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

Theorem 1. *Transforming* $\boldsymbol{y} = \boldsymbol{\Gamma}^T \boldsymbol{x}$ *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 $\geq$. 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*.*

From this we obtain two more usefull results. * If $A > 0$ the inverse exists and the determinant is > 0

* If $A > 0$ there exists a unique positive definite square root with decomposition:

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

* $\boldsymbol{\Sigma} \geq 0$

* $\boldsymbol{\Sigma}_{\boldsymbol{XY}} = \boldsymbol{\Sigma}_{\boldsymbol{YX}}^T$

* If $\boldsymbol{X} \sim (\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{XX}}), \boldsymbol{Y} \sim (\boldsymbol{\mu}_{\boldsymbol{Y}}, \boldsymbol{\Sigma}_{\boldsymbol{YY}})$ then $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})^T$ has

$$\boldsymbol{\Sigma}_{\boldsymbol{ZZ}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{XX}} & \boldsymbol{\Sigma}_{\boldsymbol{XY}} \\ \boldsymbol{\Sigma Y X} & \boldsymbol{\Sigma}_{\boldsymbol{YY}} \end{pmatrix}$$

* Independence of $\boldsymbol{X}, \boldsymbol{Y}$ implies $\mathrm{Cov}\left[\boldsymbol{X}, \boldsymbol{Y}\right] = \boldsymbol{0}$. (NB: the converse not true)

* $\mathrm{Var}\left[\boldsymbol{AX} + \boldsymbol{b}\right] = \boldsymbol{A}\mathrm{Var}\left[\boldsymbol{X}\right]\boldsymbol{A}^T$

* $\mathrm{Cov}\left[\boldsymbol{X} + \boldsymbol{Y}, \boldsymbol{Z}\right] = \mathrm{Cov}\left[\boldsymbol{X}, \boldsymbol{Z}\right] + \mathrm{Cov}\left[\boldsymbol{Y}, \boldsymbol{Z}\right]$

* $\mathrm{Var}\left[\boldsymbol{X} + \boldsymbol{Y}\right] = \mathrm{Var}\left[\boldsymbol{X}\right] + \mathrm{Cov}\left[\boldsymbol{X}, \boldsymbol{Y}\right] + \mathrm{Cov}\left[\boldsymbol{Y}, \boldsymbol{Z}\right] + \mathrm{Var}\left[\boldsymbol{Y}\right]$

* $\mathrm{Cov}\left[\boldsymbol{AX}, \boldsymbol{BY}\right] = \boldsymbol{A}\mathrm{Cov}\left[\boldsymbol{X}, \boldsymbol{Y}\right]\boldsymbol{B}^T$

Lecture 3

Transformations

Lecture J

Mahalanolis transformation

[Recall univariate, $x \sim (\mu, \sigma^2)$

Put $y = \frac{x-\mu}{\sigma} \rightsquigarrow y \sim (0,1)]$

Now, for the multivariate cause:

$$x = (x_1, \ldots, x_p)^\top, \quad x \sim (\mu, \Sigma), \Sigma \text{ now -single}$$

Would tine $y = \varphi(x)$ sit. $y \sim (0, I)$

This is:

$$y = \Sigma^{-1/2}(x-\mu)$$

Where $\Sigma^{-1/2} = \left(\Sigma^{1/2}\right)^{-1}$ and $\Sigma^{1/2}$ the unique pos-des square toot of Σ.
Fys-sth arse:

$$\Sigma = \operatorname{diaq}\left(\sigma_1^2, \ldots \sigma^2\right)$$

sine iid...tres, envier sh to do

$\cdot \frac{1}{\sigma}$

[proof that it works:

$$\begin{aligned}
&E(y) = E\left(\Sigma^{-1/2}(x-\mu)\right) = \Sigma^{-1/2}(E(x)-\mu) = 0 \\
&\operatorname{Var}(y) = \operatorname{Var}\left(\Sigma^{-1/2}(x-\mu)\right) = \operatorname{Var}\left(\Sigma^{-1/2}x\right) \\
&= \Sigma^{-1/2} \operatorname{Var}(x)\left(\Sigma^{-1/2}\right)^\top = \text{sm} \\
&\Sigma^{-1/2}
\end{aligned}$$

$$= \underbrace{\Sigma^{-1/2}\Sigma^{1/2}}_{I}\underbrace{\Sigma^{1/2}\Sigma^{-1/2}}_{I} = I$$

Principal components analysis (HS 11.1-11.J)

Observe realisations of a random vector

$$x = (x_1, \ldots, x_p)^\top, \quad x \sim (\mu, \Sigma)$$

Aim: reduction of dimensionality

Remove some components and heep comports with $\underbrace{\text{Lave in formation}}_{\rightsquigarrow \text{ Large variance}}$

Forest transform $x \to y = Ax + b_2 y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$ so that:

(1) $E(y) = 0$

(2) $\operatorname{Cov}(y_i, y_j) = 0, \quad i \neq j$

i.e. $\operatorname{Var}(y)$ is diagonal

(3) $\operatorname{Var}(y_1) \geq \operatorname{Var}(y_2) \geq \cdots \geq \operatorname{Ver}(y_r)$

We have (Jordan decomposition)

Lecture 4

Lecture 5

2 The Multivariate Normal Distribution

Lecture 7

2.1 Estimation of the multivariate normal distribution

2.1.1 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.

2.1.2 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 $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$. These make up the columns of the $(n \times p)$ *data matrix* or *feature matrix*

$\boldsymbol{X}$ given as:

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

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

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

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

The matrix $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$ 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 1. $\bar{\boldsymbol{X}} \sim N(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$

Proof. By characteristic function. □

Proposition 2. $\mathbb{E}\left[\boldsymbol{S}\right] = \frac{n-1}{n}\boldsymbol{\Sigma}$.

Proof. □

Hence we obtain an unbiased estimator as

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

Lecture 8

2.1.3 Quadratic forms

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

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

Theorem 3. (Trace formula) $\mathbb{E}\left[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\right] = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$

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

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

□

2.1.4 Idempotent matrices

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

Proposition 3. *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 idenpotent matrix is* $\boldsymbol{I}$.
4. *All eigenvalues of idempotent matrices are* 0 *or* 1.

Proof. We have:

1. $(\boldsymbol{I} - \boldsymbol{A})^2 = \boldsymbol{I}^2 - \boldsymbol{AI} - \boldsymbol{IA} + \boldsymbol{A}^2 = 0$.
2. Obvious
3. Later
4. $\lambda\boldsymbol{x} = \boldsymbol{Ax} = \boldsymbol{AAx} = \lambda^2\boldsymbol{x}$.

□

Before our next important result, we recall from linear algebra that:

1. Let $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ have eigenvalues $\lambda_1, \dots, \lambda_p$. Then:

$$\text{tr}(\boldsymbol{A}) = \sum \lambda_i, \quad \det(\boldsymbol{A}) = \prod \lambda_i.$$

2. The matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ of rank r has *spectral decomposition*:

$$\boldsymbol{A} = \boldsymbol{P\Lambda P}^T$$

   where P is the matrix with eigenvalues as columns and $\boldsymbol{\Lambda}$ the diagonal matrix of eigenvalues. Note that $\boldsymbol{P} \in \mathbb{R}^{p\times r}, \boldsymbol{\Lambda} \in \mathbb{R}^{r\times r}$. TODO: something about orthonormal

Note that for $\boldsymbol{A}$ symmetric and idempotent, we have $\boldsymbol{\Lambda} = \boldsymbol{I}_r$ in point 2 and $\text{rank}(\boldsymbol{A}) = \text{tr}(\boldsymbol{A}) = r$ due to point 1.

The following theorem will be used extensivelly in the sequel.

Theorem 4. $\boldsymbol{Z} \sim N(0, \boldsymbol{I}_p)$ *and let* $\boldsymbol{R}, \boldsymbol{S}$ *be symmetric and idempotent of rank* r, s *respectivelly. Suppose also* $\boldsymbol{RS} = 0$. *Then*

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

Proof. We have: TODO:

1.
2.
3. By definition.

□

3 Multiple Linear Regression

3.1 Model and assumptions

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

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

Intercept

Explanatory variables

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

A bunch of different scalars

It can be written on matrix form as:

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

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: $\text{Var}[\varepsilon_i] = \sigma^2 \quad \forall i$.
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*:

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

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):

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

or we may use the *maximum likelihood estimator* (MLE). It turns out that with our assumptions the result is the same. Differentiating and equating to zero yields:

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

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

Proposition 4. *For the hat matrix we have:*

1. *Symmetric*
2. *Idempotent*
3. *rank 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. The MLE is given by $\widehat{\boldsymbol{\sigma}}^2 = \frac{1}{n} \widehat{\boldsymbol{\varepsilon}}^T \widehat{\boldsymbol{\varepsilon}}$, but this is skewed as $\mathbb{E}\left[\widehat{\boldsymbol{\varepsilon}}^T \widehat{\boldsymbol{\varepsilon}}\right] = \sigma^2 (n - p)$. Hence our unbiased estimator is:

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

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 5. *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. $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$
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*

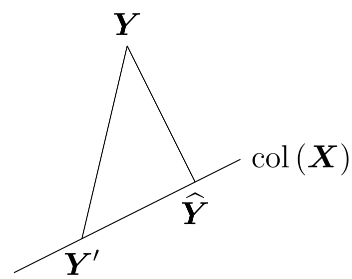

Figure 1: $\widehat{\boldsymbol{Y}}$ is the projection onto the column space of $\boldsymbol{X}$.

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. $\mathbb{P}[T \geq t_{\alpha,m}] = \alpha$ in tables.

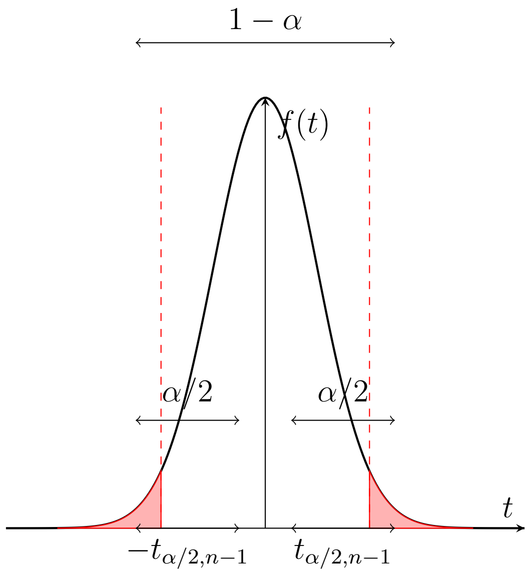

Figure 2: TODO: fix figure

We have seen that $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. Denote $(\boldsymbol{X}^T\boldsymbol{X})^{-1} = (e_{ij})_{i,j=1,\ldots,p}$. 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:

$$\boxed{\frac{\widehat{\beta}_j - \beta_j}{\sqrt{e_{jj}}\widehat{\sigma}} \sim t_{n-p}} \tag{1}$$

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:

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

3.3 Some notes on independence

TODO: questions about independence. Detour into sigma algebras etc ...

Theorem 5. *Suppose X, Y are independent random variables and that f, g are two measurable functions. Then $f(X), g(Y)$ are also independent.*

3.4 Analysis of variance (ANOVA)

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

Theorem 6. (ANOVA decomposition) *Assuming the necesarry assumptions,*

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

Proof. We first split the sum as:

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

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

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

□

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

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

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:

$$x_{i,k} = \alpha_0 + \alpha_1 x_{i,1} + \cdots + \alpha_{k-1}x_{i,k-1} + \delta_i.$$

As usual, we assume $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^T \sim N(0, \sigma^2 \boldsymbol{I})$. We may then estimate $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_{k-1})^T$ 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 $x_{ij}, j = 1, \dots, k-1$ we obtain $R_1^2, \dots, R_k^2$. It turns out that:

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

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:

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

For the residual sum of squares we also need the fact that $\boldsymbol{H x}_i = \boldsymbol{x}_i$ for all columns of $\boldsymbol{X}$. This follows readily as $\boldsymbol{HX} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}$. From this we have $\boldsymbol{H1} = \boldsymbol{1}$ as this is the first column of $\boldsymbol{X}$.

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

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

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

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

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

3.5 F-test

3.6 General F-test

TODO: important to have on yellow paper

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

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

Some special cases of this general setup are.

1. $r = 1, d = 0, A = (0, \dots, 1, \dots, 0)$ with 1 at index i, gives the test
$$\mathrm{H}_0 : \beta_i = 0, \qquad \mathrm{H}_1 : \beta_i \neq 0.$$
2. $r = 1, d = 0, A = (0, \dots, 1, \dots, -1, \dots, 0)$ with 1 at index i and -1 at index j, gives the test
$$\mathrm{H}_0 : \beta_i = \beta_j, \qquad \mathrm{H}_1 : \beta_i \neq \beta_j.$$
3. $r = k, d = \mathbf{0} \in \mathbb{R}^k, A = (\mathbf{0}, \mathrm{diag}(1)) \in \mathbb{R}^{k \times p}$, gives the test
$$\mathrm{H}_0 : \beta_i = 0 \quad \forall i \in \{1, \dots, k\}, \qquad \mathrm{H}_1 : \beta_i \neq 0 \text{ for some } i \in \{1, \dots, k\}.$$

Lecture 14

Let $\mathcal{B}$ be the space of $\boldsymbol{\beta}$ satisfying H_0. The restricted problem is:
$$\widehat{\boldsymbol{\beta}}^R = \underset{\boldsymbol{\beta} \in \mathcal{B}}{\arg\min} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Using lagrange multipliers and a bag of tricks, we obtain:
$$\widehat{\boldsymbol{\beta}}^R = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{A}^T (\boldsymbol{A}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{A}^T)^{-1} (\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{d}).$$

Denoting $\Delta = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^R$, we find:
$$\mathrm{SSE}^R = \mathrm{SSE} + \Delta^T \boldsymbol{X}^T \boldsymbol{X} \Delta$$

... IMPORTANT: the concrete expressions for the F statistic...

We claim that the under H_0, we have
$$\boxed{F = \frac{\mathrm{SSE}^R - \mathrm{SSE}/r}{\mathrm{SSE}/(n-p)} \sim F_{r,n-p}} \tag{2}$$

Proof. what the

□

Lecture 15

... example ...

3.7 Transformations of data

Motivation: ...

box cox transformation

variance stabilising transformation

Suppose $\mu = \boldsymbol{\varepsilon}(Y_i)$ and that $\mathrm{Var}\left[(\right] Y_i)$ depends on μ. ...

4 Model Analysis, Selection and Multiple Hypothesis Testing

Lecture 16

...

Lecture 17

Suppose k covariates. Then 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.

1. Adjusted coefficient of determination:

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

2.
3.
4.

example...

Multiple hypothesis testing

motivation ...

Lecture 18

. . .

FWER = probability of at least one false positive finding

... two representations

The *Bonferrony method*

The *Šidák method*

...

example 2019

. . .

example 2020

5 ANOVA and Design of Experiment

Lecture 19

Example with three groups and their means ... rewrite to regression problem ...

Analysis of varance (ANOVA)

p treatments, samples ...

Lecture 20

... cont ... + brief on two way ANOVA

Lecture 21

5.1 Two level factorial design

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

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

Further we suppose a feature matrix $\boldsymbol{X}$ satisfying:

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

TODO: expressions

Definition 2. The *main effect* of main factor j is defined as:

$$\text{effect}_j = \text{response at high level} - \text{response at low level} = 2\beta_j.$$

The estimated effect is naturally

$$\widehat{\text{effect}}_j = \text{estimated response at high level} - \text{estimated response at low level} = 2\widehat{\beta}_j.$$

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

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{1,2} x_1 x_2 + \cdots + \beta_{k-1,k} x_{k-1} k_k + \cdots + \beta_{1,2,\ldots,k} x_1 \cdots x_k.$$

We extend the design matrix accordingly, and note that we still satisfy the assumptions.

TODO: example ?

5.1.1 Inference about effect

Need inference about σ^2... cannot use estimator from multiple linear regression since for MLR we have $\widehat{\sigma}^2 = \frac{\text{SSE}}{n-p}$ and here $n = p$. We have to resort to one of two methods.

1. neglect some effects ... then these are normally dist ... use these as estimator ...
2. Lenth's method ...

Lecture 22

Resolution

resolution

Blocking

Vi tester en sitering [1].

References

[1] test. *test bok*. Ed. by Trond. UiO, 2030.

Index