Introduction to General and Generalized Linear Models

General Linear Models - part I

Henrik Madsen
Poul Thyregod

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Kgs. Lyngby

February 2012

Today

- The general linear model - intro
- The multivariate normal distribution
- Deviance
- Likelihood, score function and information matrix
- The general linear model - definition
- Estimation
- Fitted values
- Residuals
- Partitioning of variation
- Likelihood ratio tests
- The coefficient of determination

The general linear model - intro

- We will use the term *classical* GLM for the General linear model to distinguish it from GLM which is used for the Generalized linear model.
- The classical GLM leads to a unique way of describing the variations of experiments with a *continuous* variable.
- The classical GLM's include
  - Regression analysis
  - Analysis of variance - ANOVA
  - Analysis of covariance - ANCOVA
- The residuals are assumed to follow a multivariate normal distribution in the classical GLM.

The general linear model - intro

- Classical GLM's are naturally studied in the framework of the multivariate normal distribution.
- We will consider the set of n observations as a sample from a n-dimensional normal distribution.
- Under the normal distribution model, maximum-likelihood estimation of mean value parameters may be interpreted geometrically as *projection* on an appropriate subspace.
- The likelihood-ratio test statistics for model reduction may be expressed in terms of *norms* of these projections.

The multivariate normal distribution

Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ be a random vector with $Y_1, Y_2, \ldots, Y_n$ independent identically distributed (iid) $\mathrm{N}(0, 1)$ random variables.

Note that $\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{0}$ and the variance-covariance matrix $\mathrm{Var}[\boldsymbol{Y}] = \boldsymbol{I}$.

Definition (Multivariate normal distribution)

$\boldsymbol{Z}$ has an k-dimensional multivariate normal distribution if $\boldsymbol{Z}$ has the same distribution as $\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{b}$ for some n, some $k \times n$ matrix $\boldsymbol{A}$, and some k vector $\boldsymbol{b}$. We indicate the multivariate normal distribution by writing $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{A}^T)$.

Since $\boldsymbol{A}$ and $\boldsymbol{b}$ are fixed, we have $\mathrm{E}[\boldsymbol{Z}] = \boldsymbol{b}$ and $\mathrm{Var}[\boldsymbol{Z}] = \boldsymbol{A}\boldsymbol{A}^T$.

The multivariate normal distribution

Let us assume that the variance-covariance matrix is known apart from a constant factor, σ^2, i.e. $\mathrm{Var}[\boldsymbol{Z}] = \sigma^2 \boldsymbol{\Sigma}$.

The density for the k-dimensional random vector $\boldsymbol{Z}$ with mean $\boldsymbol{\mu}$ and covariance $\sigma^2 \boldsymbol{\Sigma}$ is:

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{(2\pi)^{k/2} \sigma^k \sqrt{\det \boldsymbol{\Sigma}}} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \right]$$

where $\boldsymbol{\Sigma}$ is seen to be (a) symmetric and (b) positive semi-definite.

We write $\boldsymbol{Z} \sim \mathrm{N}_k(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$.

The normal density as a statistical model

Consider now the n observations $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, and assume that a statistical model is

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}) \text{ for } \boldsymbol{y} \in \mathbb{R}^n$$

The variance-covariance matrix for the observations is called the *dispersion matrix*, denoted $\mathrm{D}[\boldsymbol{Y}]$, i.e. the dispersion matrix for $\boldsymbol{Y}$ is

$$\mathrm{D}[\boldsymbol{Y}] = \sigma^2 \boldsymbol{\Sigma}$$

Inner product and norm

Definition (Inner product and norm)

The bilinear form

$$\delta_\Sigma(\boldsymbol{y}_1, \boldsymbol{y}_2) = \boldsymbol{y}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_2$$

defines an *inner product* in $\mathbb{R}^n$. Corresponding to this inner product we can define *orthogonality*, which is obtained when the inner product is zero.

A *norm* is defined by

$$||\boldsymbol{y}||_\Sigma = \sqrt{\delta_\Sigma(\boldsymbol{y}, \boldsymbol{y})}.$$

Deviance for normal distributed variables

Definition (Deviance for normal distributed variables)

Let us introduce the notation

$$\mathrm{D}(\boldsymbol{y};\boldsymbol{\mu}) = \delta_{\Sigma}(\boldsymbol{y}-\boldsymbol{\mu},\boldsymbol{y}-\boldsymbol{\mu}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$$

to denote the quadratic norm of the vector $(\boldsymbol{y}-\boldsymbol{\mu})$ corresponding to the inner product defined by $\boldsymbol{\Sigma}^{-1}$.

For a normal distribution with $\boldsymbol{\Sigma} = \boldsymbol{I}$, the deviance is just the Residual Sum of Squares (RSS).

Deviance for normal distributed variables

Using this notation the normal density is expressed as a density defined on any finite dimensional vector space equipped with the inner product, δ_Σ:

$$f(\boldsymbol{y};\boldsymbol{\mu},\sigma^2) = \frac{1}{(\sqrt{2\pi})^n \sigma^n \sqrt{\det(\boldsymbol{\Sigma})}} \exp\left[-\frac{1}{2\sigma^2} \mathrm{D}(\boldsymbol{y};\boldsymbol{\mu})\right].$$

The likelihood and log-likelihood function

- The likelihood function is:

$$L(\boldsymbol{\mu}, \sigma^2; \boldsymbol{y}) = \frac{1}{(\sqrt{2\pi})^n \sigma^n \sqrt{\det(\boldsymbol{\Sigma})}} \exp\left[-\frac{1}{2\sigma^2}\,\mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu})\right]$$

- The log-likelihood function is (apart from an additive constant):

$$\begin{aligned}\ell_{\mu,\sigma^2}(\boldsymbol{\mu}, \sigma^2; \boldsymbol{y}) &= -(n/2)\log(\sigma^2) - \frac{1}{2\sigma^2}\,(\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \\ &= -(n/2)\log(\sigma^2) - \frac{1}{2\sigma^2}\,\mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu}).\end{aligned}$$

The score function, observed - and expected information for μ

- The score function wrt. $\boldsymbol{\mu}$ is

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ell_{\mu,\sigma^2}(\boldsymbol{\mu}, \sigma^2; \boldsymbol{y}) = \frac{1}{\sigma^2}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] = \frac{1}{\sigma^2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

- The observed information (wrt. $\boldsymbol{\mu}$) is

$$\boldsymbol{j}(\mu; \boldsymbol{y}) = \frac{1}{\sigma^2}\boldsymbol{\Sigma}^{-1}.$$

- It is seen that the observed information does not depend on the observations $\boldsymbol{y}$. Hence the expected information is

$$\boldsymbol{i}(\mu) = \frac{1}{\sigma^2}\boldsymbol{\Sigma}^{-1}.$$

The general linear model

In the case of a normal density the observation Y_i is most often written as

$$Y_i = \mu_i + \epsilon_i$$

which for all n observations $(Y_1, Y_2, \dots, Y_n)$ can be written on the matrix form

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}) \ \text{ for } \ \boldsymbol{y} \in \mathbb{R}^n$$

General Linear Models

- In the *linear model* it is assumed that $\boldsymbol{\mu}$ belongs to a linear (or affine) subspace Ω_0 of $\mathbb{R}^n$.
- The *full model* is a model with $\Omega_{full} = \mathbb{R}^n$ and hence each observation fits the model perfectly, i.e. $\widehat{\boldsymbol{\mu}} = \boldsymbol{y}$.
- The most restricted model is the *null model* with $\Omega_{null} = \mathbb{R}$. It only describes the variations of the observations by a common mean value for all observations.
- In practice, one often starts with formulating a rather comprehensive model with $\Omega = \mathbb{R}^k$, where $k < n$. We will call such a model a *sufficient model*.

The General Linear Model

Definition (The general linear model)

Assume that $Y_1, Y_2, \ldots, Y_n$ is normally distributed as described before. A *general linear model* for $Y_1, Y_2, \ldots, Y_n$ is a model where an affine hypothesis is formulated for $\boldsymbol{\mu}$. The hypothesis is of the form

$$\mathcal{H}_0 : \boldsymbol{\mu} - \boldsymbol{\mu}_0 \in \Omega_0,$$

where Ω_0 is a linear subspace of $\mathbb{R}^n$ of dimension k, and where $\boldsymbol{\mu}_0$ denotes a vector of *known offset values*.

Definition (Dimension of general linear model)

The dimension of the subspace Ω_0 for the linear model is the *dimension of the model*.

The design matrix

Definition (Design matrix for classical GLM)

Assume that the linear subspace $\Omega_0 = \text{span}\{x_1, \dots, x_k\}$, i.e. the subspace is spanned by k vectors ($k < n$).
Consider a general linear model where the hypothesis can be written as

$$\mathcal{H}_0 : \boldsymbol{\mu} - \boldsymbol{\mu}_0 = \boldsymbol{X}\boldsymbol{\beta} \text{ with } \boldsymbol{\beta} \in \mathbb{R}^k,$$

where $\boldsymbol{X}$ has full rank. The $n \times k$ matrix $\boldsymbol{X}$ of known deterministic coefficients is called the *design matrix*.
The i^{th} row of the design matrix is given by the *model vector*

$$\boldsymbol{x}_i^T = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}^T,$$

for the i^{th} observation.

Estimation of mean value parameters

Under the hypothesis

$$\mathcal{H}_0 : \boldsymbol{\mu} \in \Omega_0 \ ,$$

the maximum likelihood estimate for the set $\boldsymbol{\mu}$ is found as the orthogonal projection (with respect to δ_Σ), $p_0(\boldsymbol{y})$ of $\boldsymbol{y}$ onto the linear subspace Ω_0.

Theorem (ML estimates of mean value parameters)

For hypothesis of the form

$$\mathcal{H}_0 : \boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$$

the maximum likelihood estimated for $\boldsymbol{\beta}$ is found as a solution to the normal equation

$$\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

If $\boldsymbol{X}$ has full rank, the solution is uniquely given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$$

Properties of the ML estimator

Theorem (Properties of the ML estimator)

For the ML estimator we have

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_k(\boldsymbol{\beta}, \sigma^2 \, (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1})$$

Unknown $\boldsymbol{\Sigma}$

Notice that it has been assumed that $\boldsymbol{\Sigma}$ is known. If $\boldsymbol{\Sigma}$ is unknown, one possibility is to use the relaxation algorithm described in Madsen (2008) [a].

[a] Madsen, H. (2008) Time Series Analysis. Chapman, Hall

Fitted values

Fitted – or predicted – values

The *fitted* values $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is found as the projection of $\boldsymbol{y}$ (denoted $p_0(\boldsymbol{y})$) on to the subspace Ω_0 spanned by $\boldsymbol{X}$, and $\widehat{\boldsymbol{\beta}}$ denotes the local coordinates for the projection.

Definition (Projection matrix)

A matrix $\boldsymbol{H}$ is a *projection matrix* if and only if
(a) $\boldsymbol{H}^T = \boldsymbol{H}$ and
(b) $\boldsymbol{H}^2 = \boldsymbol{H}$, i.e. the matrix is *idempotent*.

The hat matrix

- The matrix
$$\boldsymbol{H} = \boldsymbol{X}[\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}$$
is a projection matrix.
- The projection matrix provides the predicted values $\widehat{\boldsymbol{\mu}}$, since
$$\widehat{\boldsymbol{\mu}} = p_0(\boldsymbol{y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y}$$
- It follows that the predicted values are normally distributed with
$$\mathrm{D}[\mathbf{X}\widehat{\boldsymbol{\beta}}] = \sigma^2\boldsymbol{X}[\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T = \sigma^2\boldsymbol{H}\boldsymbol{\Sigma}$$
- The matrix $\boldsymbol{H}$ is often termed the *hat matrix* since it transforms the observations $\boldsymbol{y}$ to their predicted values symbolized by a "hat" on the μ's.

Residuals

The observed *residuals* are

$$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

Orthogonality

The maximum likelihood estimate for $\boldsymbol{\beta}$ is found as the value of $\boldsymbol{\beta}$ which minimizes the *distance* $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||$.

The normal equations show that

$$\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0}$$

i.e. the *residuals* are orthogonal (with respect to $\boldsymbol{\Sigma}^{-1}$) to the subspace Ω_0.

The residuals are thus orthogonal to the fitted – or predicted – values.

Residuals

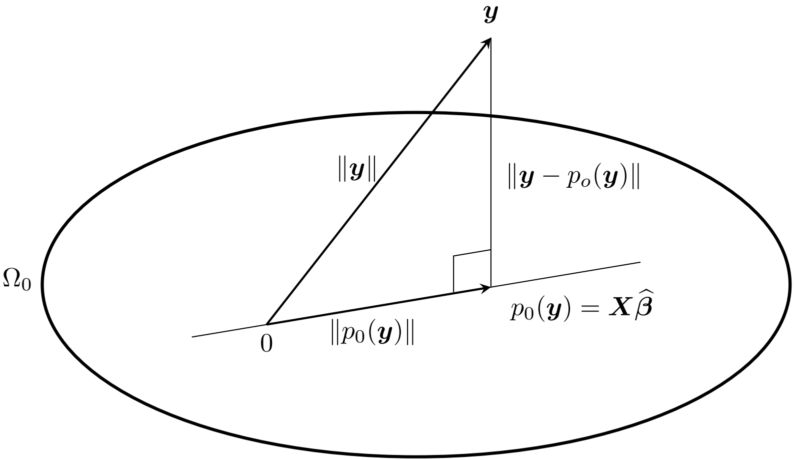

Figure: Orthogonality between the residual $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ and the vector $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

Residuals

- The residuals $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ are normally distributed with

$$\mathrm{D}[\boldsymbol{r}] = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

- The individual residuals do not have the same variance.
- The residuals are thus belonging to a subspace of dimension $n - k$, which is orthogonal to Ω_0.
- It may be shown that the distribution of the residuals $\boldsymbol{r}$ is independent of the fitted values $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

Cochran's theorem

Theorem (Cochran's theorem)

Suppose that $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$ *(i.e. standard multivariate Gaussian random variable)*

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_1\boldsymbol{Y} + \boldsymbol{Y}^T\boldsymbol{H}_2\boldsymbol{Y} + \cdots + \boldsymbol{Y}^T\boldsymbol{H}_k\boldsymbol{Y}$$

where $\boldsymbol{H}_i$ *is a symmetric* $n \times n$ *matrix with rank* n_i, $i = 1, 2, \ldots, k$. *Then any one of the following conditions implies the other two:*

- i *The ranks of the* $\boldsymbol{H}_i$ *adds to* n, *i.e.* $\sum_{i=1}^{k} n_i = n$
- ii *Each quadratic form* $\boldsymbol{Y}^T\boldsymbol{H}_i\boldsymbol{Y} \sim \chi^2_{n_i}$ *(thus the* H_i *are positive semidefinite)*
- iii *All the quadratic forms* $\boldsymbol{Y}^T\boldsymbol{H}_i\boldsymbol{Y}$ *are independent (necessary and sufficient condition).*

Partitioning of variation

Partitioning of the variation

$$
\begin{aligned}
\mathrm{D}(\boldsymbol{y};\boldsymbol{X}\boldsymbol{\beta}) &= \mathrm{D}(\boldsymbol{y};\boldsymbol{X}\widehat{\boldsymbol{\beta}}) + \mathrm{D}(\boldsymbol{X}\widehat{\boldsymbol{\beta}};\boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\
&+ (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
&\geq (\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})
\end{aligned}
$$

Partitioning of variation

χ^2-distribution of individual contributions

Under $\mathcal{H}_0$ it follows from the normal distribution of $\boldsymbol{Y}$ that

$$\mathrm{D}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \sim \sigma^2 \chi^2_n$$

Furthermore, it follows from the normal distribution of $\boldsymbol{r}$ and of $\widehat{\boldsymbol{\beta}}$ that

$$\mathrm{D}(\boldsymbol{y}; \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \sim \sigma^2 \chi^2_{n-k}$$
$$\mathrm{D}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}; \boldsymbol{X}\boldsymbol{\beta}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \sigma^2 \chi^2_k$$

moreover, the independence of $\boldsymbol{r}$ and $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ implies that $\mathrm{D}(\boldsymbol{y}; \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ and $\mathrm{D}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}; \boldsymbol{X}\boldsymbol{\beta})$ are independent.
Thus, the $\sigma^2\chi^2_n$-distribution on the left side is partitioned into two independent χ^2 distributed variables with $n-k$ and k degrees of freedom, respectively.

Estimation of the residual variance σ^2

Theorem (Estimation of the variance)

Under the hypothesis

$$\mathcal{H}_0 : \boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$$

the maximum marginal likelihood estimator for the variance σ^2 is

$$\widehat{\sigma}^2 = \frac{\mathrm{D}(\boldsymbol{y}; \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n-k} = \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n-k}$$

Under the hypothesis, $\widehat{\sigma}^2 \sim \sigma^2 \chi^2_f / f$ with $f = n - k$.

Summary: General Linear Model

- A general linear model is:

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

Consider the well known two way ANOVA:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{ i.i.d. } N(0, \sigma^2), \quad i = 1, 2, \quad j = 1, 2, 3.$$

An expanded view of this model is:

$$\begin{array}{llllllll}
y_{11} = \mu & + \alpha_1 & & + \beta_1 & & & + \varepsilon_{11} \\
y_{21} = \mu & & + \alpha_2 & + \beta_1 & & & + \varepsilon_{21} \\
y_{12} = \mu & + \alpha_1 & & & + \beta_2 & & + \varepsilon_{12} \\
y_{22} = \mu & & + \alpha_2 & & + \beta_2 & & + \varepsilon_{22} \\
y_{13} = \mu & + \alpha_1 & & & & + \beta_3 & + \varepsilon_{13} \\
y_{23} = \mu & & + \alpha_2 & & & + \beta_3 & + \varepsilon_{23}
\end{array} \tag{1}$$

The exact same in matrix notation:

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \\ y_{13} \\ y_{23} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix}}_{\boldsymbol{\varepsilon}} \tag{2}$$

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \\ y_{13} \\ y_{23} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

- $\mathbf{y}$ is the vector of all observations
- $\mathbf{X}$ is known as the *design matrix*
- $\boldsymbol{\beta}$ is the vector of parameters
- $\boldsymbol{\varepsilon}$ is a vector of independent $N(0, \sigma^2)$ "measurement noise"
  - The vector $\boldsymbol{\varepsilon}$ is said to follow a *multivariate normal distribution*
  - Mean vector $\mathbf{0}$
  - Covariance matrix $\sigma^2\mathbf{I}$
  - Written as: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ specifies the model, and everything can be calculated from $\mathbf{y}$ and $\mathbf{X}$.

In a general linear model (with both factors and covariates), it is surprisingly easy to construct the design matrix $\mathbf{X}$.

- For each factor: Add one column for each level, with ones in the rows where the corresponding observation is from that level, and zeros otherwise.
- For each covariate: Add one column with the measurements of the covariate.
- Remove linear dependencies (if necessary)

Example: linear regression:

$$y_i = \alpha + \beta \cdot x_i + \varepsilon$$

In matrix notation:

$$\mathbf{y} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \boldsymbol{\varepsilon}$$

Likelihood ratio tests

- In the classical GLM case the exact distribution of the likelihood ratio test statistic may be derived.
- Consider the following model for the data $\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$.
- Let us assume that we have the sufficient model

$$\mathcal{H}_1 : \boldsymbol{\mu} \in \Omega_1 \subset \mathbb{R}^n$$

  with $\dim(\Omega_1) = m_1$.
- Now we want to test whether the model may be reduced to a model where μ is restricted to some subspace of Ω_1, and hence we introduce $\Omega_0 \subset \Omega_1$ as a linear (affine) subspace with $\dim(\Omega_0) = m_0$.

Model reduction

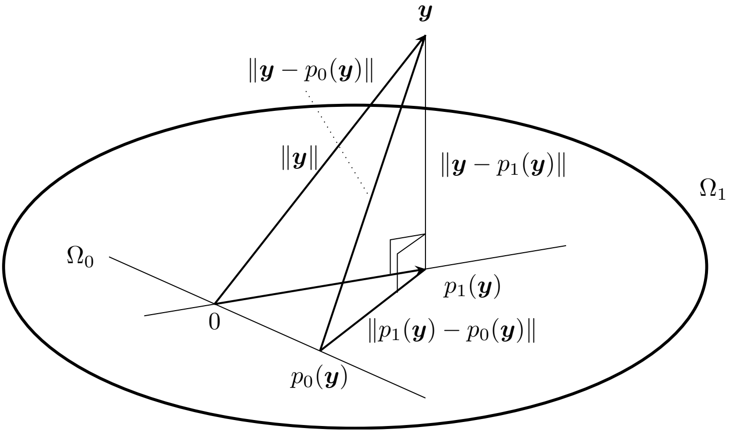

Figure: Model reduction. The partitioning of the deviance corresponding to a test of the hypothesis $\mathcal{H}_0 : \mu \in \Omega_0$ under the assumption of $\mathcal{H}_1 : \mu \in \Omega_1$.

Test for model reduction

Theorem (A test for model reduction)

The likelihood ratio test statistic for testing

$$\mathcal{H}_0 : \boldsymbol{\mu} \in \Omega_0 \ \text{ against the alternative } \ \mathcal{H}_1 : \boldsymbol{\mu} \in \Omega_1 \setminus \Omega_0$$

is a monotone function of

$$F(\boldsymbol{y}) = \frac{\mathrm{D}(p_1(\boldsymbol{y}); p_0(\boldsymbol{y}))/(m_1 - m_0)}{\mathrm{D}(\boldsymbol{y}; p_1(\boldsymbol{y}))/(n - m_1)}$$

where $p_1(\boldsymbol{y})$ and $p_0(\boldsymbol{y})$ denote the projection of $\boldsymbol{y}$ on Ω_1 and Ω_0, respectively. Under $\mathcal{H}_0$ we have

$$F \sim F(m_1 - m_0, n - m_1)$$

i.e. large values of F reflects a conflict between the data and $\mathcal{H}_0$, and hence lead to rejection of $\mathcal{H}_0$. The p-value of the test is found as $p = P[F(m_1 - m_0, n - m_1) \geq F_{obs}]$, where F_{obs} is the observed value of F given the data.

Test for model reduction

- The partitioning of the variation is presented in a Deviance table (or an *ANalysis Of VAriance table*, ANOVA).
- The table reflects the partitioning in the test for *model reduction*.
- The deviance between the variation of the model from the hypothesis is measured using the deviance of the observations from the model as a reference.
- Under $\mathcal{H}_0$ they are both χ^2 distributed, orthogonal and thus independent.
- This means that the ratio is F distributed.
- If the test quantity is large this shows evidence against the model reduction tested using $\mathcal{H}_0$.

Deviance table

Source	f	Deviance	Test statistic, F
Model versus hypothesis	$m_1 - m_0$	$\|\|p_1(\boldsymbol{y}) - p_0(\boldsymbol{y})\|\|^2$	$\dfrac{\|\|p_1(\boldsymbol{y}) - p_0(\boldsymbol{y})\|\|^2/(m_1 - m_0)}{\|\|\boldsymbol{y} - p_1(\boldsymbol{y})\|\|^2/(n - m_1)}$
Residual under model	$n - m_1$	$\|\|\boldsymbol{y} - p_1(\boldsymbol{y})\|\|^2$	
Residual under hypothesis	$n - m_0$	$\|\|\boldsymbol{y} - p_0(\boldsymbol{y})\|\|^2$	

Table: Deviance table corresponding to a test for model reduction as specified by $\mathcal{H}_0$. For $\Sigma = \boldsymbol{I}$ this corresponds to an analysis of variance table, and then 'Deviance' is equal to the 'Sum of Squared deviations (SS)'

Test for model reduction

The test is a conditional test

It should be noted that the test has been derived as a *conditional test*. It is a test for the hypothesis $\mathcal{H}_0 : \boldsymbol{\mu} \in \Omega_0$ under the assumption that $\mathcal{H}_1 : \boldsymbol{\mu} \in \Omega_1$ is true. The test does in no way assess whether $\mathcal{H}_1$ is in agreement with the data. On the contrary in the test the residual variation under $\mathcal{H}_1$ is used to estimate σ^2, i.e. to assess $\mathrm{D}(\boldsymbol{y}; p_1(\boldsymbol{y}))$.

The test does not depend on the particular parametrization of the hypotheses

Note that the test does only depend on the two sub-spaces Ω_1 and Ω_0, but not on how the subspaces have been parametrized (the particular choice of basis, i.e. the design matrix). Therefore it is sometimes said that the test is *coordinate free*.

Initial test for model 'sufficiency'

- In practice, one often starts with formulating a rather comprehensive model, a *sufficient model*, and then tests whether the model may be reduced to the *null model* with $\Omega_{null} = \mathbb{R}$, i.e. $\dim \Omega_{null} = 1$.
- The hypotheses are

$$\mathcal{H}_{null} : \boldsymbol{\mu} \in \mathbb{R}$$

$$\mathcal{H}_1 : \boldsymbol{\mu} \in \Omega_1 \setminus \mathbb{R}.$$

  where $\dim \Omega_1 = k$.
- The hypothesis is a hypothesis of "Total homogeneity", namely that all observations are satisfactorily represented by their common mean.

Deviance table

Source	f	Deviance	Test statistic, F
Model $\mathcal{H}_{null}$	$k-1$	$\lVert p_1(\boldsymbol{y}) - p_{null}(\boldsymbol{y})\rVert^2$	$\dfrac{\lVert p_1(\boldsymbol{y}) - p_{null}(\boldsymbol{y})\rVert^2/(k-1)}{\lVert \boldsymbol{y} - p_1(\boldsymbol{y})\rVert^2/(n-k)}$
Residual under $\mathcal{H}_1$	$n-k$	$\lVert \boldsymbol{y} - p_1(\boldsymbol{y})\rVert^2$	
Total	$n-1$	$\lVert \boldsymbol{y} - p_{null}(\boldsymbol{y})\rVert^2$	

Table: Deviance table corresponding to the test for model reduction to the null model.

Under $\mathcal{H}_{null}$, $F \sim F(k-1, n-k)$, and hence large values of F would indicate rejection of the hypothesis $\mathcal{H}_{null}$. The p-value of the test is $p = P[F(k-1, n-k) \geq F_{obs}]$.

Coefficient of determination, R^2

- The *coefficient of determination*, R^2, is defined as

$$R^2 = \frac{\mathrm{D}(p_1(\boldsymbol{y}); p_{null}(\boldsymbol{y}))}{\mathrm{D}(\boldsymbol{y}; p_{null}(\boldsymbol{y}))} = 1 - \frac{\mathrm{D}(\boldsymbol{y}; p_1(\boldsymbol{y}))}{\mathrm{D}(\boldsymbol{y}; p_{null}(\boldsymbol{y}))}, \quad 0 \le R^2 \le 1.$$

- Suppose you want to predict Y. If you do not know the x's, then the best prediction is $\overline{y}$. The variability corresponding to this prediction is expressed by the *total variation*.
- If the model is utilized for the prediction, then the prediction error is reduced to the *residual variation*.
- R^2 expresses the fraction of the total variation that is explained by the model.
- As more variables are added to the model, $\mathrm{D}(\boldsymbol{y}; p_1(\boldsymbol{y}))$ will decrease, and R^2 will increase.

Adjusted coefficient of determination, R^2_{adj}

- The *adjusted coefficient of determination* aims to correct that R^2 increases as more variables are added to the model.
- It is defined as:

$$R^2_{adj} = 1 - \frac{\mathrm{D}(\boldsymbol{y}; p_1(\boldsymbol{y}))/(n-k)}{\mathrm{D}(\boldsymbol{y}; p_{null}(\boldsymbol{y}))/(n-1)}.$$

- It charges a penalty for the number of variables in the model.
- As more variables are added to the model, $\mathrm{D}(\boldsymbol{y}; p_1(\boldsymbol{y}))$ decreases, but the corresponding degrees of freedom also decreases.
- The numerator in may increase if the reduction in the residual deviance caused by the additional variables does not compensate for the loss in the degrees of freedom.