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

October 2010
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
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.
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.
Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ be a random vector with $Y_1, Y_2, \ldots, Y_n$ independent identically distributed (iid) $N(0, 1)$ random variables.

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

**Definition (Multivariate normal distribution)**

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

Since $A$ and $\mathbf{b}$ are fixed, we have $E[\mathbf{Z}] = \mathbf{b}$ and $\text{Var}[\mathbf{Z}] = AA^T$. 
The multivariate normal distribution

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

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

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

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

We write $\mathbf{Z} \sim \mathcal{N}_k(\mu, \sigma^2 \mathbf{\Sigma})$. 
The normal density as a statistical model

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

\[
\mathbf{Y} \sim \mathcal{N}_n(\mathbf{\mu}, \sigma^2 \mathbf{\Sigma}) \quad \text{for} \quad \mathbf{y} \in \mathbb{R}^n
\]

The variance-covariance matrix for the observations is called the dispersion matrix, denoted \( D[\mathbf{Y}] \), i.e. the dispersion matrix for \( \mathbf{Y} \) is

\[
D[\mathbf{Y}] = \sigma^2 \mathbf{\Sigma}
\]
**Definition (Inner product and norm)**

The bilinear form

\[ \delta_\Sigma(y_1, y_2) = y_1^T \Sigma^{-1} y_2 \]

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

\[ \|y\|_\Sigma = \sqrt{\delta_\Sigma(y, y)}. \]
Deviance for normal distributed variables

Definition (Deviance for normal distributed variables)

Let us introduce the notation

\[ D(y; \mu) = \delta_{\Sigma}(y - \mu, y - \mu) = (y - \mu)^T \Sigma^{-1} (y - \mu) \]

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

For a normal distribution with \(\Sigma = I\), 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, \( \delta_\Sigma \):

\[
f(y; \mu, \sigma^2) = \frac{1}{(\sqrt{2\pi})^n \sigma^n \sqrt{\det(\Sigma)}} \exp \left[ -\frac{1}{2\sigma^2} D(y; \mu) \right].
\]
The likelihood and log-likelihood function

- The likelihood function is:

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

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

\[
l_{\mu,\sigma^2}(\mu, \sigma^2; y) = -(n/2) \log(\sigma^2) - \frac{1}{2\sigma^2} (y - \mu)^T \Sigma^{-1} (y - \mu)
= -(n/2) \log(\sigma^2) - \frac{1}{2\sigma^2} D(y; \mu).
\]
The score function, observed - and expected information for $\mu$

1. The score function wrt. $\mu$ is

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

2. The observed information (wrt. $\mu$) is

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

3. It is seen that the observed information does not depend on the observations $y$. Hence the expected information is

$$i(\mu) = \frac{1}{\sigma^2} \Sigma^{-1}.$$
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 $(Y_1, Y_2, \ldots, Y_n)$ can be written on the matrix form

$$Y = \mu + \epsilon$$

where

$$Y \sim N_n(\mu, \sigma^2 \Sigma) \text{ for } y \in \mathbb{R}^n$$
General Linear Models

- In the *linear model* it is assumed that \( \mu \) belongs to a linear (or affine) subspace \( \Omega_0 \) 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. \( \hat{\mu} = y \).

- 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*. 
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 $\mu$. The hypothesis is of the form

$$H_0 : \mu - \mu_0 \in \Omega_0,$$

where $\Omega_0$ is a linear subspace of $\mathbb{R}^n$ of dimension $k$, and where $\mu_0$ denotes a vector of known offset values.

Definition (Dimension of general linear model)

The dimension of the subspace $\Omega_0$ for the linear model is the dimension of the model.
The design matrix

Definition (Design matrix for classical GLM)

Assume that the linear subspace $\Omega_0 = \text{span}\{x_1, \ldots, x_k\}$, i.e. the subspace is spanned by $k$ vectors ($k < n$).

Consider a general linear model where the hypothesis can be written as

$$H_0 : \mu - \mu_0 = X\beta$$

with $\beta \in \mathbb{R}^k$,

where $X$ has full rank. The $n \times k$ matrix $X$ of known deterministic coefficients is called the design matrix.

The $i^{th}$ row of the design matrix is given by the model vector

$$x_i^T = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}^T$$

for the $i^{th}$ observation.
Estimation of mean value parameters

Under the hypothesis

\[ H_0 : \mu \in \Omega_0, \]

the maximum likelihood estimate for the set \( \mu \) is found as the orthogonal projection (with respect to \( \delta \Sigma \)), \( \rho_0(y) \) of \( y \) onto the linear subspace \( \Omega_0 \).

**Theorem (ML estimates of mean value parameters)**

*For hypothesis of the form*

\[ H_0 : \mu(\beta) = X\beta \]

*the maximum likelihood estimated for \( \beta \) is found as a solution to the normal equation*

\[ X^T \Sigma^{-1} y = X^T \Sigma^{-1} X \hat{\beta}. \]

*If \( X \) has full rank, the solution is uniquely given by*

\[ \hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y \]
Properties of the ML estimator

Theorem (Properties of the ML estimator)

For the ML estimator we have

\[ \hat{\beta} \sim N_k(\beta, \sigma^2 (X^T \Sigma^{-1} X)^{-1}) \]

Unknown \( \Sigma \)

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

Fitted values

Fitted – or predicted – values

The fitted values $\hat{\mu} = X\hat{\beta}$ is found as the projection of $y$ (denoted $p_0(y)$) on to the subspace $\Omega_0$ spanned by $X$, and $\hat{\beta}$ denotes the local coordinates for the projection.

Definition (Projection matrix)

A matrix $H$ is a projection matrix if and only if
(a) $H^T = H$ and
(b) $H^2 = H$, i.e. the matrix is idempotent.
The hat matrix

- The matrix

\[ H = X [X^T \Sigma^{-1} X]^{-1} X^T \Sigma^{-1} \]

is a projection matrix.

- The projection matrix provides the predicted values \( \hat{\mu} \), since

\[ \hat{\mu} = p_0(y) = X \hat{\beta} = Hy \]

- It follows that the predicted values are normally distributed with

\[ D[X\hat{\beta}] = \sigma^2 X [X^T \Sigma^{-1} X]^{-1} X^T = \sigma^2 H \Sigma \]

- The matrix \( H \) is often termed the *hat matrix* since it transforms the observations \( y \) to their predicted values symbolized by a ”hat” on the \( \mu \)’s.
Residuals

The observed residuals are

\[ r = y - X\hat{\beta} = (I - H)y \]

Orthogonality

The maximum likelihood estimate for \( \beta \) is found as the value of \( \beta \) which minimizes the distance \( \| y - X\beta \| \).

The normal equations show that

\[ X^T \Sigma^{-1} (y - X\hat{\beta}) = 0 \]

i.e. the residuals are orthogonal (with respect to \( \Sigma^{-1} \)) to the subspace \( \Omega_0 \).

The residuals are thus orthogonal to the fitted – or predicted – values.
Figure: Orthogonality between the residual \((y - X\hat{\beta})\) and the vector \(X\hat{\beta}\).
Residuals

The residuals \( r = (I - H)Y \) are normally distributed with

\[
D[r] = \sigma^2(I - H)
\]

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 \( \Omega_0 \).

It may be shown that the distribution of the residuals \( r \) is independent of the fitted values \( X\hat{\beta} \).
Cochran’s theorem

Theorem (Cochran’s theorem)

Suppose that $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ (i.e. standard multivariate Gaussian random variable)

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

where $\mathbf{H}_i$ 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 $\mathbf{H}_i$ adds to $n$, i.e. $\sum_{i=1}^{k} n_i = n$

ii. Each quadratic form $\mathbf{Y}^T \mathbf{H}_i \mathbf{Y} \sim \chi^2_{n_i}$ (thus the $\mathbf{H}_i$ are positive semidefinite)

iii. All the quadratic forms $\mathbf{Y}^T \mathbf{H}_i \mathbf{Y}$ are independent (necessary and sufficient condition).
Partitioning of variation

Partitioning of the variation

\[
D(y; X\beta) = D(y; \hat{X}\hat{\beta}) + D(\hat{X}\hat{\beta}; X\beta) \\
= (y - \hat{X}\hat{\beta})^T \Sigma^{-1} (y - \hat{X}\hat{\beta}) \\
+ (\hat{\beta} - \beta)^T X^T \Sigma^{-1} X (\hat{\beta} - \beta) \\
\geq (y - \hat{X}\hat{\beta})^T \Sigma^{-1} (y - \hat{X}\hat{\beta})
\]
Partitioning of variation

\(\chi^2\)-distribution of individual contributions

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

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

Furthermore, it follows from the normal distribution of \(r\) and of \(\hat{\beta}\) that

\[
D(y; X\hat{\beta}) = (y - X\hat{\beta})^T\Sigma^{-1}(y - X\hat{\beta}) \sim \sigma^2\chi^2_{n-k}
\]
\[
D(X\hat{\beta}; X\beta) = (\hat{\beta} - \beta)^T X^T\Sigma^{-1}X(\hat{\beta} - \beta) \sim \sigma^2\chi^2_k
\]

moreover, the independence of \(r\) and \(X\hat{\beta}\) implies that \(D(y; X\hat{\beta})\) and \(D(X\hat{\beta}; X\beta)\) are independent.

Thus, the \(\sigma^2\chi^2_n\)-distribution on the left side is partitioned into two independent \(\chi^2\) distributed variables with \(n - k\) and \(k\) degrees of freedom, respectively.
Theorem (Estimation of the variance)

Under the hypothesis

\[ \mathcal{H}_0 : \mu(\beta) = X\beta \]

the maximum marginal likelihood estimator for the variance \( \sigma^2 \) is

\[
\hat{\sigma}^2 = \frac{D(y; X\hat{\beta})}{n-k} = \frac{(y - X\hat{\beta})^T \Sigma^{-1} (y - X\hat{\beta})}{n-k}
\]

Under the hypothesis, \( \hat{\sigma}^2 \sim \sigma^2 \chi^2_f / f \) with \( f = n - k \).
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 \( Y \sim N_n(\mu, \sigma^2 \Sigma) \).

- Let us assume that we have the sufficient model

\[
\mathcal{H}_1 : \mu \in \Omega_1 \subset \mathbb{R}^n
\]

with \( \text{dim}(\Omega_1) = m_1 \).

- Now we want to test whether the model may be reduced to a model where \( \mu \) is restricted to some subspace of \( \Omega_1 \), and hence we introduce \( \Omega_0 \subset \Omega_1 \) as a linear (affine) subspace with \( \text{dim}(\Omega_0) = m_0 \).
Model reduction

\begin{equation}
\parallel y - p_0(y) \parallel
\end{equation}

\begin{equation}
\parallel y - p_1(y) \parallel
\end{equation}

\begin{equation}
\parallel p_1(y) - p_0(y) \parallel
\end{equation}

\begin{equation}
\Omega_0
\end{equation}

\begin{equation}
\Omega_1
\end{equation}

\begin{equation}
0
\end{equation}

\begin{equation}
p_0(y)
\end{equation}

\begin{equation}
p_1(y)
\end{equation}

\begin{equation}
y
\end{equation}

**Figure:** Model reduction. The partitioning of the deviance corresponding to a test of the hypothesis $H_0 : \mu \in \Omega_0$ under the assumption of $H_1 : \mu \in \Omega_1$. 
Theorem (A test for model reduction)

The likelihood ratio test statistic for testing

\[ H_0 : \mu \in \Omega_0 \quad \text{against the alternative} \quad H_1 : \mu \in \Omega_1 \setminus \Omega_0 \]

is a monotone function of

\[
F(y) = \frac{D(p_1(y); p_0(y))/(m_1 - m_0)}{D(y; p_1(y))/(n - m_1)}
\]

where \( p_1(y) \) and \( p_0(y) \) denote the projection of \( y \) on \( \Omega_1 \) and \( \Omega_0 \), respectively. Under \( H_0 \) 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 \( H_0 \), and hence lead to rejection of \( H_0 \). 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 $\chi^2$ 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

<table>
<thead>
<tr>
<th>Source</th>
<th>( f )</th>
<th>Deviance</th>
<th>Test statistic, ( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model versus hypothesis</td>
<td>( m_1 - m_0 )</td>
<td>( | p_1(y) - p_0(y) |^2 )</td>
<td>( \frac{| p_1(y) - p_0(y) |^2}{(m_1 - m_0)} )</td>
</tr>
<tr>
<td>Residual under model</td>
<td>( n - m_1 )</td>
<td>( | y - p_1(y) |^2 )</td>
<td>( \frac{| y - p_1(y) |^2}{(n - m_1)} )</td>
</tr>
<tr>
<td>Residual under hypothesis</td>
<td>( n - m_0 )</td>
<td>( | y - p_0(y) |^2 )</td>
<td></td>
</tr>
</tbody>
</table>

**Table**: Deviance table corresponding to a test for model reduction as specified by \( \mathcal{H}_0 \). For \( \Sigma = I \) 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 \( \mathcal{H}_0 : \mu \in \Omega_0 \) under the assumption that \( \mathcal{H}_1 : \mu \in \Omega_1 \) 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 \( \sigma^2 \), i.e. to assess \( D(y; p_1(y)) \).

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

Note that the test does only depend on the two sub-spaces \( \Omega_1 \) and \( \Omega_0 \), 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

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

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

  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

<table>
<thead>
<tr>
<th>Source</th>
<th>( f )</th>
<th>Deviance</th>
<th>Test statistic, ( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model ( H_{null} )</td>
<td>( k - 1 )</td>
<td>( | p_1(y) - p_{null}(y) |^2 )</td>
<td>( \frac{| p_1(y) - p_{null}(y) |^2}{| y - p_1(y) |^2 / (k - 1)} )</td>
</tr>
<tr>
<td>Residual under ( H_1 )</td>
<td>( n - k )</td>
<td>( | y - p_1(y) |^2 )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n - 1 )</td>
<td>( | y - p_{null}(y) |^2 )</td>
<td></td>
</tr>
</tbody>
</table>

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

Under \( H_{null} \), \( F \sim F(k - 1, n - k) \), and hence large values of \( F \) would indicate rejection of the hypothesis \( H_{null} \). The \( p \)-value of the test is \( p = P[F(k - 1, n - k) \geq F_{obs}] \).
The coefficient of determination, $R^2$, is defined as

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

Suppose you want to predict $Y$. If you do not know the $x$’s, then the best prediction is $\bar{y}$. 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, $D(y; p_1(y))$ will decrease, and $R^2$ will increase.
Adjusted coefficient of determination, $R_{adj}^2$

- The *adjusted coefficient of determination* aims to correct that $R^2$ increases as more variables are added to the model.
- It is defined as:
  
  $$R_{adj}^2 = 1 - \frac{D(y; p_1(y))/(n - k)}{D(y; p_{null}(y))/(n - 1)}.$$

- It charges a penalty for the number of variables in the model.
- As more variables are added to the model, $D(y; p_1(y))$ 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.