Introduction to General and Generalized Linear Models
Generalized Linear Models - part IIIa

Henrik Madsen
Poul Thyregod

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

March 18, 2012

Today

- The generalized linear model (summary)

  - Link function

  - Estimation

  - Fitted values

  - Residuals

- Likelihood ratio test

- Over-dispersion

- Examples – Eg. Over-dispersion and off-set.

The Generalized Linear Model

Definition (The generalized linear model)

Assume that $Y_1, Y_2, \ldots, Y_n$ are mutually independent, and the density can be described by an exponential dispersion model with the same variance function $V(\mu)$.

A *generalized linear model* for $Y_1, Y_2, \ldots, Y_n$ describes an affine hypothesis for $\eta_1, \eta_2, \ldots, \eta_n$, where

$$\eta_i = g(\mu_i)$$

is a transformation of the mean values $\mu_1, \mu_2, \ldots, \mu_n$.

The hypothesis is of the form

$$\mathcal{H}_0 : \boldsymbol{\eta} - \boldsymbol{\eta}_0 \in L,$$

where L is a linear subspace $\mathbb{R}^n$ of dimension k, and where $\boldsymbol{\eta}_0$ denotes a vector of *known off-set values*.

GLM vs GLM

General linear models	**Generalized linear models**
Normal distribution	Exponential dispersion family
Mean value linear	Function of mean value linear
Independent observations	Independent observations
Same variance	Variance function of mean
Easy to apply	Almost as easy to apply
Exact results	Approximate results

Dimension and design matrix

Definition (Dimension of the generalized linear model)

The dimension k of the subspace L for the generalized linear model is the *dimension of the model*

Definition (Design matrix for the generalized linear model)

Consider the linear subspace $L = \text{span}\{x_1, \ldots, x_k\}$, i.e. the subspace is spanned by k vectors ($k < n$), such that the hypothesis can be written

$$\boldsymbol{\eta} - \boldsymbol{\eta}_0 = \boldsymbol{X}\boldsymbol{\beta} \text{ with } \boldsymbol{\beta} \in \mathbb{R}^k,$$

where $\boldsymbol{X}$ has full rank. The $n \times k$ matrix $\boldsymbol{X}$ is called the *design matrix*. The ith row of the design matrix is given by the *model vector*

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

for the ith observation.

The link function

Definition (The link function)

The *link function*, $g(\cdot)$ describes the relation between the linear predictor η_i and the mean value parameter $\mu_i = \mathrm{E}[Y_i]$. The relation is

$$\eta_i = g(\mu_i)$$

The inverse mapping $g^{-1}(\cdot)$ thus expresses the mean value μ as a function of the linear predictor η:

$$\mu = g^{-1}(\eta)$$

that is

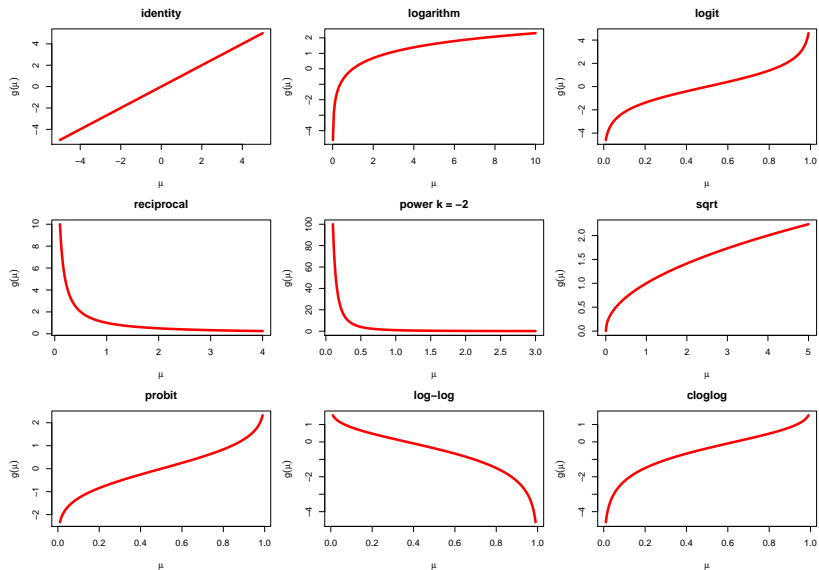$$\mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}) = g^{-1}\left(\sum_j x_{ij}\beta_j\right)$$

Link functions

The most commonly used link functions, $\eta = g(\mu)$, are :

Name	Link function $\eta = g(\mu)$	$\mu = g^{-1}(\eta)$
identity	μ	η
log	$\log(\mu)$	$\exp(\eta)$
logit	$\log(\mu/(1-\mu))$	$\exp(\eta)/[1 + \exp(\eta)]$
inverse	$1/\mu$	$1/\eta$
power	μ^k	$\eta^{1/k}$
sqrt	$\sqrt{\mu}$	η^2
probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$
log-log	$\log(-\log(\mu))$	$\exp(-\exp(\eta))$
cloglog	$\log(-\log(1-\mu))$	$1 - \exp(-\exp(\eta))$

Table: Commonly used link function.

Link functions

The canonical link

The canonical link is the function which transforms the mean to the canonical location parameter of the exponential dispersion family, i.e. it is the function for which $g(\mu) = \theta$. The canonical link function for the most widely considered densities are

Density	Link:$\eta = g(\mu)$	Name
Normal	$\eta = \mu$	identity
Poisson	$\eta = \log(\mu)$	log
Binomial	$\eta = \log[\mu/(1-\mu)]$	logit
Gamma	$\eta = 1/\mu$	inverse
Inverse Gauss	$\eta = 1/\mu^2$	`1/mu^2`

Table: Canonical link functions for some widely used densities.

Specification of a generalized linear model

a) Distribution / Variance function:
   Specification of the distribution – or the *variance function* $V(\mu)$.

b) Link function:
   Specification of the *link function* $g(\cdot)$, which describes a function of the mean value which can be described linearly by the explanatory variables.

c) Linear predictor:
   Specification of the linear dependency

$$g(\mu_i) = \eta_i = (\boldsymbol{x}_i)^T \boldsymbol{\beta}.$$

d) Precision (optional):
   If needed the precision is formulated as *known individual weights*, $\lambda_i = w_i$, or as a *common dispersion parameter*, $\lambda = 1/\sigma^2$, or a *combination* $\lambda_i = w_i/\sigma^2$.

Specification of a generalized linear model in R

```
> mice.glm <- glm(formula = resp ~ conc,
+                 family = binomial(link = logit),
+                 weights = NULL,
+                 data = mice
+                )
```

- formula; as in general linear models
- family
  - binomial(link = logit | probit | cauchit | log | cloglog)
  - gaussian(link = identity | log | inverse)
  - Gamma(link = inverse | identity | log)
  - inverse.gaussian(link = 1/mu^2 | inverse | identity | log)
  - poisson(link = log | identity | sqrt)
  - quasi(link = ... , variance = ...))
  - quasibinomial(link = logit | probit | cauchit | log | cloglog)
  - quasipoisson(link = log | identity | sqrt)

The score function w.r.t. the canonical parameter

The log likelihood function w.r.t. the canonical parameter

$$\ell_\theta(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^n w_i(\theta_i y_i - \kappa(\theta_i))$$

The *score function* w.r.t. the canonical parameter, θ:

$$\frac{\partial}{\partial \theta_i} \ell_\theta(\boldsymbol{\theta}; \boldsymbol{y}) = w_i(y_i - \tau(\theta_i))$$

or in matrix form

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_\theta(\boldsymbol{\theta}; \boldsymbol{y}) = \operatorname{diag}(\boldsymbol{w})(\boldsymbol{y} - \boldsymbol{\tau}(\boldsymbol{\theta}))$$

$\operatorname{diag}(\boldsymbol{w})$ denotes a diagonal matrix where i^{th} element is w_i, and

$$\boldsymbol{\tau}(\boldsymbol{\theta}) = \left\{ \begin{array}{c} \tau(\theta_1) \\ \tau(\theta_2) \\ \vdots \\ \tau(\theta_n) \end{array} \right\}$$

The score function w.r.t. the mean value parameter

The log likelihood function w.r.t. the mean value parameter

$$\ell_{\mu}(\boldsymbol{\mu}; \boldsymbol{y}) = -\frac{1}{2} \sum_{i=1}^{n} w_i \, d(y_i; \mu_i)$$

The score function w.r.t. the mean value parameter is:

$$l'_{\mu}(\boldsymbol{\mu}; \boldsymbol{y}) = \text{diag} \left\{ \frac{w_i}{V(\mu_i)} \right\} (\boldsymbol{y} - \boldsymbol{\mu})$$

which shows that the ML-estimate for $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = \boldsymbol{y}$.

For a single coordinate we have the log-likelihood:
$\ell(\mu_i; y_i) = -\frac{1}{2} w_i 2 \int_{\mu_i}^{y_i} \frac{y_i - u}{V(u)} \, du = w_i \int_{y_i}^{\mu_i} \frac{y_i - u}{V(u)} \, du$
So we get the score:
$\frac{\partial}{\partial \mu_i} \ell(\mu_i; y_i) = w_i \int_{y_i}^{\mu_i} f(u) \, du = w_i \frac{\partial}{\partial \mu_i} (F(\mu_i) - F(y_i)) = w_i f(\mu_i) = w_i \frac{y_i - \mu_i}{V(\mu_i)}$

Maximum likelihood estimation

Theorem (Estimation in generalized linear models)

Consider the generalized linear model as defined on slide 3 for the observations $Y_1, \ldots, Y_n$ and assume that $Y_1, \ldots, Y_n$ are mutually independent with densities, which can be described by an exponential dispersion model with the variance function $V(\cdot)$, dispersion parameter σ^2, and optionally the weights w_i.

Assume that the linear predictor is parameterized with $\boldsymbol{\beta}$ corresponding to the design matrix $\boldsymbol{X}$, then the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is found as the solution to

$$[\boldsymbol{X}(\boldsymbol{\beta})]^T \boldsymbol{i}_\mu(\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0} \ ,$$

where $\boldsymbol{X}(\boldsymbol{\beta})$ denotes the local design matrix and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ given by

$$\mu_i(\boldsymbol{\beta}) = g^{-1}(\boldsymbol{x}_i{}^T \boldsymbol{\beta}) \ ,$$

denotes the fitted mean values corresponding to the parameters $\boldsymbol{\beta}$, and $\boldsymbol{i}_\mu(\boldsymbol{\mu})$ is the expected information with respect to $\boldsymbol{\mu}$.

The estimates must be found by an iterative procedure.

Proof

$$\ell'_{\beta}(\boldsymbol{\beta}, \boldsymbol{y}) = \left[\frac{\partial \mu}{\partial \beta}\right]^T \ell'_{\mu}(\boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{y})$$

$$= \left[\frac{\partial \mu}{\partial \beta}\right]^T \mathsf{diag}\left\{\frac{w_i}{V(\mu_i)}\right\}(\boldsymbol{y} - \boldsymbol{\mu})$$

$$= \left[\frac{\partial \mu}{\partial \beta}\right]^T \boldsymbol{i}(\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})$$

And $\frac{\partial \mu}{\partial \beta} = \left[\frac{\partial \mu}{\partial \eta}\right]^T \frac{\partial \eta}{\partial \beta}$ is called the local design matrix.

Maximum likelihood estimation

The score function for the canonical link

For the canonical link the local design matrix $\boldsymbol{X}(\boldsymbol{\beta})$ is $\mathrm{diag}\{V(\mu_i)\}\boldsymbol{X}$, and hence the score function becomes

$$\boldsymbol{X}^T \mathrm{diag}\{V(\mu_i)\} \mathrm{diag}\left\{\frac{w_i}{V(\mu_i)}\right\}(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})),$$

or

$$\boldsymbol{X}^T \mathrm{diag}\{w_i\}\boldsymbol{y} = \boldsymbol{X}^T \mathrm{diag}\{w_i\}\boldsymbol{\mu}(\boldsymbol{\beta})$$

The equation is called the *mean value equation*, or *the normal equation*
For an unweighted model the mean value equation simply becomes

$$\boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{\mu}(\boldsymbol{\beta})$$

Properties of the ML estimator

Theorem (Asymptotic distribution of the ML estimator)

Under the hypothesis $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ we have asymptotically

$$\frac{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sqrt{\sigma^2}} \in N_k(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where the dispersion matrix $\boldsymbol{\Sigma}$ for $\widehat{\boldsymbol{\beta}}$ is

$$\mathrm{D}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\Sigma} = [\boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}) \boldsymbol{X}]^{-1}$$

with

$$\boldsymbol{W}(\boldsymbol{\beta}) = \mathrm{diag}\left\{\frac{w_i}{[g'(\mu_i)]^2 V(\mu_i)}\right\},$$

Linear prediction for the generalized linear model

Definition (Linear prediction for the generalized linear model)

The linear prediction $\widehat{\boldsymbol{\eta}}$ is defined as the values

$$\widehat{\boldsymbol{\eta}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

with the linear prediction corresponding to the i'th observation is

$$\widehat{\eta}_i = \sum_{j=1}^{k} x_{ij}\widehat{\beta}_j = (\boldsymbol{x}_i)^T\widehat{\boldsymbol{\beta}}.$$

The linear predictions $\widehat{\boldsymbol{\eta}}$ are approximately normally distributed with

$$\mathrm{D}[\widehat{\boldsymbol{\eta}}] \approx \widehat{\sigma}^2 \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T$$

where $\boldsymbol{\Sigma}$ is the dispersion matrix for $\widehat{\boldsymbol{\beta}}$.

Fitted values for the generalized linear model

Definition (Fitted values for the generalized linear model)

The fitted values are defined as the values

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}) \ ,$$

where the ith value is given as

$$\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i)$$

with the fitted value $\widehat{\eta}_i$ of the linear prediction.

The fitted values $\widehat{\boldsymbol{\mu}}$ are approximately normally distributed with

$$\mathrm{D}[\widehat{\boldsymbol{\mu}}] \approx \widehat{\sigma}^2 \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\right]^2 \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T$$

where $\boldsymbol{\Sigma}$ is the dispersion matrix for $\widehat{\boldsymbol{\beta}}$.

Residual deviance

Definition (Residual deviance)

Consider the generalized linear model defined on slide 3. The *residual deviance* corresponding to this model is

$$\mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) = \sum_{i=1}^{n} w_i\, d(y_i; \widehat{\mu}_i)$$

with $d(y_i; \widehat{\mu}_i)$ denoting the unit deviance corresponding the observation y_i and the fitted value $\widehat{\mu}_i$ and where w_i denotes the weights (if present). If the model includes a dispersion parameter σ^2, the *scaled* residual deviance is

$$\mathrm{D}^*(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) = \frac{\mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))}{\sigma^2}.$$

Residuals

Residuals represents the difference between the data and the model. In the classical GLM the residuals are $r_i = y_i - \widehat{\mu}_i$. These are called response residuals for GLM's. Since the variance of the response is not constant for most GLM's we need some modification. We will look at:

- Deviance residuals
- Pearson residuals

Residuals

Definition (Deviance residual)

Consider the generalized linear model from for the observations $Y_1, \ldots Y_n$.

The *deviance residual* for the i'th observation is defined as

$$r_i^D = r_D(y_i; \widehat{\mu}_i) = \mathsf{sign}(y_i - \widehat{\mu}_i)\sqrt{w_i\, d(y_i, \widehat{\mu}_i)}$$

where $\mathsf{sign}(x)$ denotes the *sign function* $\mathsf{sign}(x) = 1$ for $x > 0$ og $\mathsf{sign}(x) = -1$ for $x < 0$, and with w_i denoting the weight (if relevant), $d(y; \mu)$ denoting the unit deviance and $\widehat{\mu}_i$ denoting the fitted value corresponding to the i'th observation.

Assessments of the deviance residuals is in good agreement with the likelihood approach as the deviance residuals simply express differences in log-likelihood.

Residuals

Definition (Pearson residual)

Consider again the generalized linear model from for the observations $Y_1, \ldots Y_n$.

The *Pearson residuals* are defined as the values

$$r_i^P = r_P(y_i; \widehat{\mu}_i) = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)/w_i}}$$

The Pearson residual is thus obtained by scaling the response residual with $\sqrt{\mathrm{Var}[Y_i]}$. Hence, the Pearson residual is the response residual normalized with the estimated standard deviation for the observation.

Likelihood ratio tests

- The approximative normal distribution of the ML-estimator implies that many distributional results from the classical GLM-theory are carried over to generalized linear models as approximative (asymptotic) results.

- An example of this is the likelihood ratio test.

- In the classical GLM case it was possible to derive the exact distribution of the likelihood ratio test statistic (the F-distribution).

- For generalized linear models, this is not possible, and hence we shall use the asymptotic results for the logarithm of the likelihood ratio.

Likelihood ratio test

Theorem (Likelihood ratio test)

Consider the generalized linear model. Assume that the model

$$\mathcal{H}_1 : \boldsymbol{\eta} \in L \subset \mathbb{R}^k$$

holds with L parameterized as $\boldsymbol{\eta} = \boldsymbol{X}_1 \boldsymbol{\beta}$, and consider the hypotheses

$$\mathcal{H}_0 : \boldsymbol{\eta} \in L_0 \subset \mathbb{R}^m$$

where $\boldsymbol{\eta} = \boldsymbol{X}_0 \boldsymbol{\alpha}$ and $m < k$, and with the alternative $\mathcal{H}_1 : \boldsymbol{\eta} \in L \backslash L_0$.
Then the likelihood ratio test for $\mathcal{H}_0$ has the test statistic

$$-2 \log \lambda(\boldsymbol{y}) = \mathrm{D}\left(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}})\right) - \mathrm{D}\left(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)$$

When $\mathcal{H}_0$ is true, the test statistic will asymptotically follow a $\chi^2(k - m)$ distribution.

If the model includes a dispersion parameter, σ^2, then $\mathrm{D}\left(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))\right)$ will asymptotically follow a $\sigma^2 \chi^2(k - m)$ distribution.

Test for model 'sufficiency'

- In analogy with classical GLM's one often starts with formulating a rather comprehensive model, and then reduces the model by successive tests.

- In contrast to classical GLM's we may however test the goodness of fit of the initial model.

- The test is a special case of the likelihood ratio test.

Test for model 'sufficiency'

Test for model 'sufficiency'

Consider the generalized linear model, and assume that the dispersion $\sigma^2 = 1$.

Let $\mathcal{H}_{full}$ denote the *full*, or *saturated* model, i.e. $\mathcal{H}_{full} : \boldsymbol{\mu} \in \mathbb{R}^n$ and consider the hypotheses

$$\mathcal{H}_0 : \boldsymbol{\eta} \in L \subset \mathbb{R}^k$$

with L parameterized as $\boldsymbol{\eta} = \boldsymbol{X}_0 \boldsymbol{\beta}$.

Then, as the residual deviance under $\mathcal{H}_{full}$ is 0, the test statistic is the residual deviance $\mathrm{D}\left(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)$. When $\mathcal{H}_0$ is true, the test statistic is distributed as $\chi^2(n-k)$. The test rejects for large values of $\mathrm{D}\left(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)$.

Residual deviance measures goodness of fit

- The residual deviance $\mathrm{D}\left(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)$ is a reasonable measure of the goodness of fit of a model $\mathcal{H}_0$.

- When referring to a hypothesized model $\mathcal{H}_0$, we shall sometimes use the symbol $G^2(\mathcal{H}_0)$ to denote the residual deviance $\mathrm{D}\left(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)$.

- Using that convention, the partitioning of residual deviance may be formulated as

$$G^2(\mathcal{H}_0|\mathcal{H}_1) = G^2(\mathcal{H}_0) - G^2(\mathcal{H}_1)$$

with $G^2(\mathcal{H}_0|\mathcal{H}_1)$ interpreted as the goodness fit test statistic for $\mathcal{H}_0$ conditioned on $\mathcal{H}_1$ being true, and $G^2(\mathcal{H}_0)$ and $G^2(\mathcal{H}_1)$, denoting the unconditional goodness of fit statistics for $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively.

Analysis of deviance table

- The initial test for goodness of fit of the initial model is often represented in an *analysis of deviance table* in analogy with the ANOVA table for classical GLM's.

- In the table the goodness of fit test statistic corresponding to the initial model $G^2(\mathcal{H}_1) = \mathrm{D}\left(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)$ is shown in the line labelled "Error".

- The statistic should be compared to percentiles in the $\chi^2(n-k)$ distribution.

- The table also shows the test statistic for $\mathcal{H}_{null}$ under the assumption that $\mathcal{H}_1$ is true.

- The test investigates whether the model is necessary at all, i.e. whether at least some of the coefficients differ significantly from zero.

Analysis of deviance table

- Note, that in the case of a generalized linear model, we can start the analysis by using the residual (error) deviance to test whether the model may be maintained, at all.

- This is in contrast to the classical GLM's where the residual sum of squares around the initial model $\mathcal{H}_1$ served to estimate σ^2, and therefore we had no reference value to compare with the residual sum of squares.

- In the generalized linear models the variance is a known function of the mean, and therefore in general there is no need to estimate a separate variance.

Analysis of deviance table

Source	f	Deviance	Mean deviance	Goodness of fit interpretation
Model $\mathcal{H}_{null}$	$k-1$	$\mathrm{D}\left(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \widehat{\boldsymbol{\mu}}_{null}\right)$	$\dfrac{\mathrm{D}\left(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \widehat{\boldsymbol{\mu}}_{null}\right)}{k-1}$	$G^2(\mathcal{H}_{null}\vert\mathcal{H}_1)$
Residual (Error)	$n-k$	$\mathrm{D}\left(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)$	$\dfrac{\mathrm{D}\left(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)}{n-k}$	$G^2(\mathcal{H}_1)$
Corrected total	$n-1$	$\mathrm{D}\left(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}_{null}\right)$		$G^2(\mathcal{H}_{null})$

Table: Initial assessment of goodness of fit of a model $\mathcal{H}_0$. $\mathcal{H}_{null}$ and $\widehat{\boldsymbol{\mu}}_{null}$ refer to the *minimal model*, i.e. a model with all observations having the same mean value.

Overdispersion

- It may happen that even if one has tried to fit a rather comprehensive model (i.e. a model with many parameters), the fit is not satisfactory, and the residual deviance $D\left(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)$ is larger than what can be explained by the χ^2-distribution.

- An explanation for such a poor model fit could be an improper choice of linear predictor, or of link or response distribution.

- If the residuals exhibit a random pattern, and there are no other indications of misfit, then the explanation could be that the variance is larger than indicated by $V(\mu)$.

- We say that the data are *overdispersed*.

Overdispersion

- When data are *overdispersed*, a more appropriate model might be obtained by including a *dispersion parameter*, σ^2, in the model, i.e. a distribution model of the form with $\lambda_i = w_i/\sigma^2$, and σ^2 denoting the overdispersion, $\mathrm{Var}[Y_i] = \sigma^2 V(\mu_i)/w_i$.

- As the dispersion parameter only would enter in the score function as a constant factor, this does not affect the estimation of the mean value parameters $\boldsymbol{\beta}$.

- However, because of the larger error variance, the distribution of the test statistics will be influenced.

- If, for some reasons, the parameter σ^2 had been known beforehand, one would include this known value in the weights, w_i.

- Most often, when it is found necessary to choose a model with overdispersion, σ^2 shall be estimated from the data.

The dispersion parameter

- For the normal distribution family, the dispersion parameter is just the variance σ^2.

- In the case of a gamma distribution family, the shape parameter α acts as dispersion parameter.

- The maximum likelihood estimation of the shape parameter is not too complicated for the normal and the gamma distributions but for other exponential dispersion families, ML estimation of the dispersion parameter is more tricky.

- The problem is that the dispersion parameter enters in the likelihood function, not only as a factor to the deviance, but also in the normalizing factor $a(y_i, w_i/\sigma^2)$.

- It is necessary to have an explicit expression for this factor as function of σ^2 (as in the case of the normal and the gamma distribution families) in order to perform the maximum likelihood estimation.

The dispersion parameter

Approximate moment estimate for the dispersion parameter

It is common practice to use the residual deviance $D(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$ as basis for the estimation of σ^2 and use the result that $D(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$ is approximately distributed as $\sigma^2 \chi^2(n-k)$. It then follows that

$$\widehat{\sigma}_{dev}^2 = \frac{D(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))}{n-k}$$

is asymptotically unbiased for σ^2.

Alternatively, one would utilize the corresponding Pearson goodness of fit statistic

$$X^2 = \sum_{i=1}^{n} w_i \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}$$

which likewise follows a $\sigma^2 \chi^2(n-k)$-distribution, and use the estimator

$$\widehat{\sigma}_{Pears}^2 = \frac{X^2}{n-k}.$$

Deviance table in the case of overdispersion

Source	f	Deviance	Scaled deviance
Model $\mathcal{H}_{null}$	$k-1$	$\mathrm{D}\left(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \widehat{\boldsymbol{\mu}}_{null}\right)$	$\frac{\mathrm{D}(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \widehat{\boldsymbol{\mu}}_{null})/(k-1)}{\mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))/(n-k)}$
Residual (Error)	$n-k$	$\mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$	
Corrected total	$n-1$	$\mathrm{D}\left(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}_{null}\right)$	

Table: Example of Deviance table in the case of overdispersion. It is noted that the scaled deviance is equal to the model deviance scaled by the error deviance.

- The *scaled deviance*, D^*, i.e. deviance divided by $\widehat{\sigma}^2$ is used in the tests instead of the crude deviance in case of overdispersion.

- For calculation of p-values etc. the asymptotic χ^2-distribution of the scaled deviance is used.

Test for model reduction

- The principles for model reduction in generalized linear models are essentially the same as the principles for classical GLM's.

- In classical GLM's the deviance is calculated as a (weighted) sum of squares, and in generalized linear models the deviance is calculated using the expression for the unit deviance.

- Besides this, the major difference is that instead of the exact F-tests used for classical GLM's the tests in generalized linear models are only approximate tests using the χ^2-distribution.

- In particular does the principles of successive testing in hypotheses chains using a type I, or type III partition of the deviance carry over to generalized linear models.

Test of individual parameters β_j

Theorem (Test of individual parameters β_j - Wald test)

A hypothesis $\mathcal{H} : \beta_j = \beta_j^0$ related to specific values of the parameters is tested by means of the test statistic

$$u_j = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\widehat{\sigma^2}\widehat{\sigma}_{jj}}},$$

where $\widehat{\sigma^2}$ indicates the estimated dispersion parameter (if relevant), and $\widehat{\sigma}_{jj}$ denotes the j'th diagonal element in $\widehat{\boldsymbol{\Sigma}}$.

Under the hypothesis is u_j approximately distributed as a standardized normal distribution. The test statistic is compared with quantiles of a standardized normal distribution (some software packages use a $t(n - k)$ distribution). The hypothesis is rejected for large values of $|u_j|$. The p-value is found as $p = 2(1 - \Phi(|u_j|))$.

Test of individual parameters β_j

Theorem (Test of individual parameters β_j - Wald test)

In particular is the test statistic for the hypothesis $\mathcal{H} : \beta_j = 0$

$$u_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \hat{\sigma}_{jj}}}.$$

An equivalent test is obtained by considering the test statistic

$$z_j = u_j^2$$

and reject the hypothesis for for $z_j > \chi^2_{1-\alpha}(1)$.

Confidence intervals

Wald - interval for individual parameters

An approximate $100(1 - \alpha)$ % Wald-type confidence interval is obtained as

$$\widehat{\beta}_j \pm u_{1-\alpha/2}\sqrt{\widehat{\sigma^2 \widehat{\sigma}_{jj}}}$$

Confidence intervals for fitted values

An approximate $100(1 - \alpha)\%$ confidence interval for the linear prediction is obtained as

$$\widehat{\eta}_i \pm u_{1-\alpha/2}\sqrt{\widehat{\sigma^2 \widehat{\sigma}_{ii}}}$$

with $\widehat{\sigma}_{ii}$ denoting the i'th diagonal element in $\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T$.

The corresponding interval for the fitted value $\widehat{\mu}_i$ is obtained by applying the inverse link transformation $g^{-1}(\cdot)$ to the confidence limits.

Example: Link functions for binary response regression

An experiment testing the insulation effect of a gas (SF_6) was conducted. In the experiment a gaseous insulation was subjected to 100 high voltage pulses with a specified voltage, and it was recorded whether the insulation broke down (spark), or not. After each pulse the insulation was reestablished. The experiment was repeated at twelve voltage levels from 1065 kV to 1135 kV.

Voltage (kV)	1065	1071	1075	1083	1089	1094
Breakdowns	2	3	5	11	10	21
Trials	100	100	100	100	100	100

Voltage (kV)	1100	1107	1111	1120	1128	1135
Breakdowns	29	48	56	88	98	99
Trials	100	100	100	100	100	100

Table: The insulation effect of a gas (SF_6)

Example: Link functions for binary response regression

- As the insulation was restored after each voltage application it seems reasonable to assume that the trials were independent.

- At each trial the response is binary (Breakdown/Not), and therefore it seems appropriate to use a binomial distribution model for the experiment.

- We shall assume that the data are stored in an R object `dat` with the variables `Volt`, `Breakd`, `Trials`

- Let Z_i denote the number of breakdowns at the i'th trial at the voltage x_i. We shall then use the model $Z_i \sim B(n_i, p_i)$ with $n_i = 100$, and $p_i = p(x_i)$, where $p(x)$ is some suitable dose-response function.

Logit transformation - logistic regression

The logistic regression is of the form

$$g(p) = \eta = \ln\left(\frac{p}{1-p}\right) = \beta_1 + \beta_2 x$$

$$p(x) = \frac{\exp(\eta)}{1 + \exp(\eta)} = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)}$$

We use the following commands to fit the model:

```
dat$Resp<-cbind(Breakd,(Trials-Breakd))
fit1<-glm(Resp~Volt,family=binomial(link=logit),data=dat)
```

Logit link

```
> summary(fit1)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7572  -0.8518   0.9359   1.2977   2.3466

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.277e+02  7.061e+00  -18.08   <2e-16 ***
Volt         1.155e-01  6.396e-03   18.05   <2e-16 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 783.122  on 11  degrees of freedom
Residual deviance:  21.018  on 10  degrees of freedom
AIC: 70.613

Number of Fisher Scoring iterations: 4
```

Logit link

From the output we can make a deviance table:

Source	f	Deviance	Mean deviance
Model $\mathcal{H}_M$	1	762.10	762.10
Residual (Error)	10	21.018	2.102
Corrected total	11	783.12	71.193

The p-value corresponding to the goodness of fit statistic
$\mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) = 21.018$
is assessed by calculating

```
> pval <- 1- pchisq(21.01776,10)
```

leading to pval = 0.02097. Thus, $\mathcal{H}_{logist}$ is rejected at any significance
level greater than 2 %.

Logit link

Also, a look at the *deviance residuals*:

```
residuals(logist.glm)
       1        2        3        4         5         6
1.016293 0.8554846 1.22456 1.586976 -0.7922914 0.1608718

        7        8        9       10       11       12
-1.030256 -1.083212 -1.757153 1.204659 2.346577 1.517043
```

They indicate underestimation in the tails, and overestimation in the central part of the curve.

The probit link

The transformation

$$g(p) = \eta = \Phi^{-1}(p) = \beta_1 + \beta_2 x$$
$$p(x) = \Phi(\eta) = \Phi(\beta_1 + \beta_2 x)$$

with $\Phi(\cdot)$ denoting the cumulative distribution function for the standardized normal distribution is termed the *probit*-transformation.

The function tends towards 0 and 1 for $x \to \mp\infty$, respectively. The convergence is faster than for the logistic transformation.

There is a long tradition in biomedical literature for using the probit transformation.

We fit the model with:

```
fit2<-glm(Resp~Volt,family=binomial(link=probit),data=dat)
```

The probit link

```
> summary(fit2)
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.653  -1.252   1.250   1.421   2.395

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -71.105019   3.451897  -20.60   <2e-16 ***
Volt          0.064325   0.003129   20.56   <2e-16 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 783.122  on 11  degrees of freedom
Residual deviance:  26.215  on 10  degrees of freedom
AIC: 75.81

Number of Fisher Scoring iterations: 5
```

The probit link

From the output we can make a deviance table:

Source	f	Deviance	Mean deviance
Model $\mathcal{H}_M$	1	756.907	756.907
Residual (Error)	10	26.215	2.622
Corrected total	11	783.122	71.193

The p-value corresponding to the goodness of fit statistic
$\mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) = 26.215$
is assessed by calculating

```
> pval <- 1- pchisq(26.215,10)
```

leading to pval = 0.00346. Thus, $\mathcal{H}_{probit}$ is rejected at any significance level greater than 0.3 %. The fit is not satisfactory.

The probit link

Again, a look at the *deviance residuals*:

```
residuals(prob.glm)
        1         2        3         4          5          6
 1.666163  1.238981  1.3965  1.26034  -1.357938  -0.5152912
        7         8        9        10         11         12
-1.562847 -1.216113 -1.653013 1.492566 2.394602 1.280603
```

They indicate systematic underestimation in both tails, and overestimation in the central part.

Complementary log-log link

The transformation

$$g(p) = \eta = \ln(-\ln(1-p)) = \beta_0 + \beta_1 x$$
$$p(x) = 1 - \exp[-\exp(\beta_0 + \beta_1 x)]$$

is termed the *complementary log-log* transformation.

The response function is *asymmetrical*. It increases slowly away from 0, whereas it approaches 1 in a rather steep manner.

We fit the model with:

```
fit3<-glm(Resp~Volt,family=binomial(link=cloglog),data=dat)
```

Complementary log-log link

```
> summary(fit3)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.36540  -0.37472   0.04829   0.33605   1.00952

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -91.106296   4.601803  -19.80   <2e-16 ***
Volt          0.081900   0.004147   19.75   <2e-16 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 783.122  on 11  degrees of freedom
Residual deviance:   5.671  on 10  degrees of freedom
AIC: 55.266

Number of Fisher Scoring iterations: 4
```

Complementary log-log link

From the output we can make a deviance table:

Source	f	Deviance	Mean deviance
Model $\mathcal{H}_M$	1	777.451	777.451
Residual (Error)	10	5.671	0.567
Corrected total	11	783.122	71.193

The p-value corresponding to the goodness of fit statistic
$\mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) = 5.671$
is assessed by calculating

```
pval <- 1- pchisq(5.671,10),
```

leading to pval = 0.8421. Thus, data do not provide any evidence
against the cloglog model.

Complementary log-log link

This is further supported by the deviance residuals:

```
residuals(clog.glm)
            1           2          3          4          5          6
-0.02716415 -0.1760891 0.2060522 0.8231332 -1.11485 0.2832367
            7           8          9         10         11         12
 -0.2986162   0.123737 -0.6030393 1.009516 0.4944959 -1.365395
```

There is no systematic pattern in the residuals, and all residuals are in the interval ± 2.
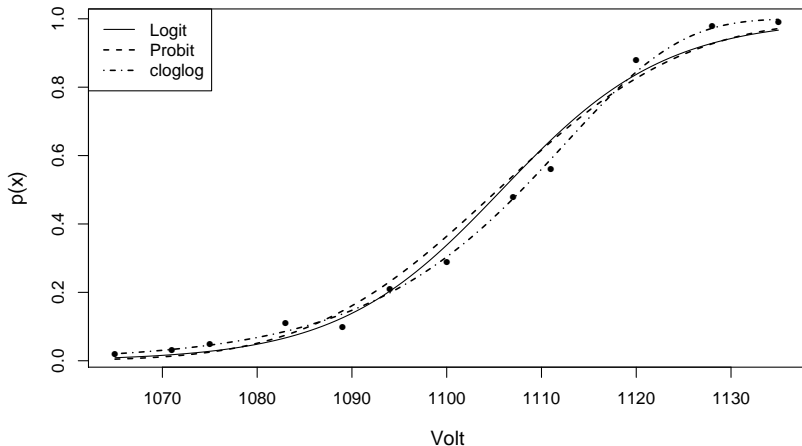
Logit/probit/cloglog

Figure: Probability of breakdown for an insulator as function of applied pulse voltage. The curves correspond to different assumptions on the functional form of the relation.

Example: Poisson regression – Use of off-set

The number of accidents with personal injury during daylight hours in January quarter during the years 1987 to 1990.

Year	1987	1988	1989	1990
Accidents, y_i	57	67	54	59
Traf.index, x_i	100	111	117	120

It is standard practice in road research to model accident counts Y_i as Poisson distributed random variables, $Y_i \sim Pois(\mu_i)$.

Poisson regression (continued)

We shall assume that the accident count is proportional to the amount of traffic,

$$\mu_i = \gamma x_i \tag{1}$$

Cf. Table 4.4 the canonical link for the Poisson distribution is

$$g(\mu) = \log(\mu) \ ,$$

therefore, in terms of the canonical link the model is

$$\eta_i - \log(x_i) = \beta_1$$

with $\beta_1 = \log(\gamma)$, or

$$\eta_i = \beta_1 + \log(x_i)$$

viz. a model with *offset* equal to $\log(x_i)$, and only one term, the intercept, β_1. The accident *rate* $\gamma = \exp(\beta_1)$.

Poisson regression (continued)

Since

$$\mu_i(\beta_1) = \exp\left(\beta_1 + \log(x_i)\right),$$

the mean value equation (4.38) becomes

$$\mathbf{1}^T \boldsymbol{y} = \mathbf{1}^T \begin{pmatrix} \exp\left(\beta_1 + \log(x_1)\right) \\ \vdots \\ \exp\left(\beta_1 + \log(x_4)\right) \end{pmatrix}$$

ie.

$$\sum_{i=1}^4 y_i = \sum_{i=1}^4 \exp(\beta_0) x_i$$

and, hence

$$\widehat{\gamma} = \exp(\widehat{\beta_1}) = \frac{\sum_{i=1}^4 y_i}{\sum_{i=1}^4 x_i} = \frac{237}{448} = 0.529 \tag{2}$$

Poisson regression – R-implementation

Assuming that the data from the table above is in a dataset `acc` with the further variable `lninx` calculated as the logarithm of the traffic index then the R command

```
glm(formula = Acc ~ offset(lninx), family = poisson(link = lo
```

results in the output

```
Coefficients:
(Intercept)
    -0.6367
Degrees of Freedom: 3 Total (i.e. Null);  3 Residual
Null Deviance:      2.803
Residual Deviance: 2.803        AIC: 28.48
```

The goodness of fit statistic $D(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) = 2.80$ with $P[\chi(3)^2 \geq 2.80] = 0.42$.

Thus, the data do not contradict the assumptions underlying the model.