Introduction to General and Generalized Linear Models

Generalized Linear Models - part I

Henrik Madsen
Poul Thyregod

DTU Informatics
Technical University of Denmark
DK-2800 Kgs. Lyngby

March 4, 2012

Today

- Classical GLM vs. GLM
- Motivating example
- Exponential families of distributions

General linear model - classical GLM

In the classical GLM it is assumed that:

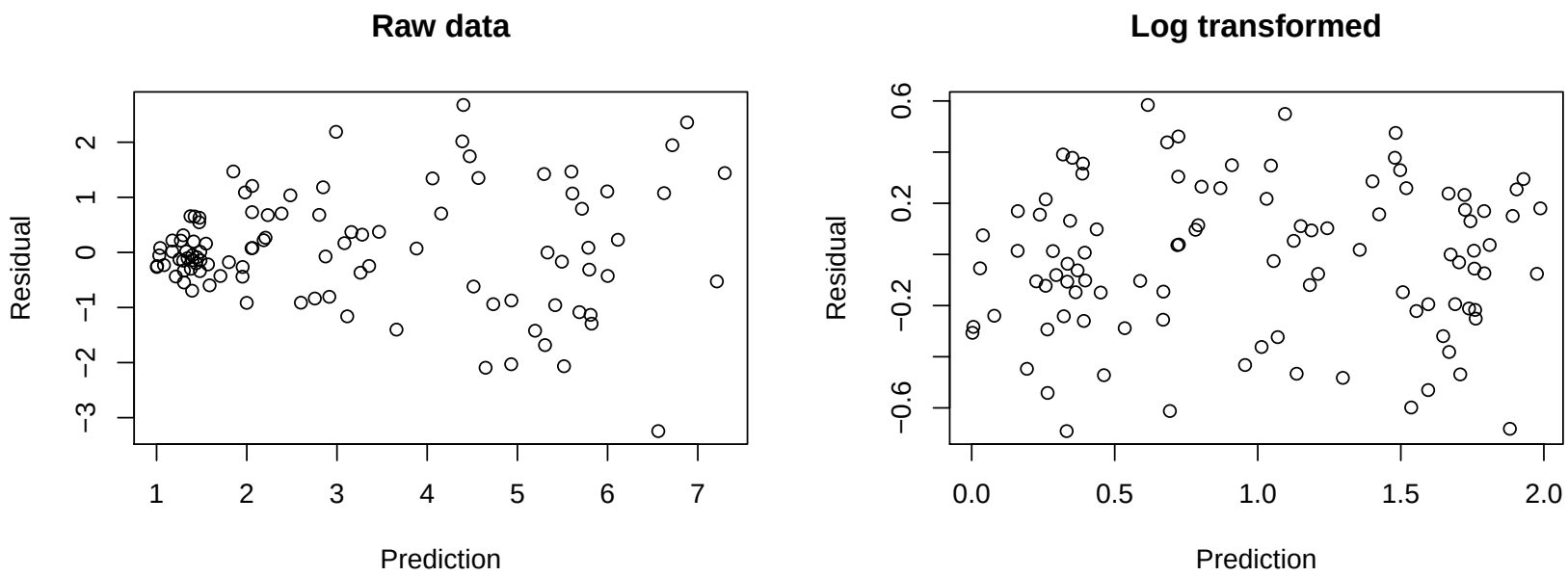
- The errors are normally distributed.
- The error variances are constant and independent of the mean.
- Systematic effects combine additively.
- The general linear model can be summarized as:

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

Often these assumptions may be justifiable but there are situations where these assumptions are far from being satisfied.

Transformation of data into normality

- Sometimes it is possible to transform data, such that it matches a general linear model.
- For instance if the variance is increasing with the mean



- Then we have seen that a log transformation often is appropriate

Generalized linear models - GLM

- Some types of observations can never be transformed into normality
- For a wide class of distributions, the so called exponential family, we can use generalized linear models
- Introduced by Nelder and Wedderburn in 1972.
- Formulate linear models for a transformation of the mean value.
- Do not transform the observations thereby preserving the distributional properties of the observations.
- Allows easy use for instance via the `glm()` function in R, similar to `lm()`

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

Types of response variables

i Count data ($y_1 = 57, \ldots, y_n = 59$ accidents) - Poisson distribution.

ii Binary response variables ($y_1 = 0, y_2 = 1, \ldots, y_n = 0$), or proportion of counts ($y_1 = 15/297, \ldots, y_n = 144/285$) - Binomial distribution.

iii Count data, waiting times - Negative Binomial distribution.

iv Multiple ordered categories "Unsatisfied", "Neutral", "Satisfied" - Multinomial distribution.

v Count data, multiple categories.

vi Continuous responses, constant variance ($y_1 = 2.567, \ldots, y_n = 2.422$) - Normal distribution.

vii Continuous positive responses with constant coefficient of variation - Gamma distribution.

viii Continuous positive highly skewed - Inverse Gaussian.

Motivating example

The generalized linear model will be introduced in the following example. The generalized linear model will then be explained in detail in this and the following lectures.

In toxicology it is usual practice to assess developmental effects of an agent by administering specified doses of the agent to pregnant mice, and assess the proportion of stillborn as a function of the concentration of the agent.

The quantity of interest is the *fraction*, y, of stillborn pups as a function of the concentration x of the agent.

A natural distributional assumption is the binomial distribution

$$Y \sim B(n_i, p_i)/n_i$$

.

Motivating example

The assumptions for the classical GLM are not satisfied in this case:

- For p close to 0 or 1 the distribution of Y is highly skewed violating the normality assumption.
- The variance, $Var[Y_i] = p_i(1-p_i)/n_i$ depends on the mean value p_i, the quantity we want to model violating the homoscedasticity assumption.
- A linear model on the form: $p_i = \beta_i + \beta_2 x_i$, will violate the natural restriction $0 < p_i < 1$.
- A model formulation of the form $y_i = p_i + \epsilon_i$ (mean plus noise) is not adequate - if such a model should satisfy $0 \leq y_i \leq 1$, then the distribution of ϵ_i would have to be dependent on p_i.

Motivating example

In a study of developmental toxicity of a chemical compound, a specified amount of an ether was dosed daily to pregnant mice, and after 10 days all fetuses were examined. The size of each litter and the number of stillborns were recorded:

Index	Number of stillborn, z_i	Number of fetuses, n_i	Fraction still-born, y_i	Concentration [mg/kg/day], x_i
1	15	297	0.0505	0.0
2	17	242	0.0702	62.5
3	22	312	0.0705	125.0
4	38	299	0.1271	250.0
5	144	285	0.5053	500.0

Table: Results of a dose-response experiment on pregnant mice. Number of stillborn fetuses found for various dose levels of a toxic agent.

Motivating example

Let Z_i denote the number of stillborns at dose concentration x_i.

We shall assume $Z_i \sim B(n_i, p_i)$, that is a binomial distribution corresponding to n_i independent trials (fetuses), and the probability, p_i, of stillbirth being the same for all n_i fetuses.

We want to model $Y_i = Z_i/n_i$, and in particular we want a model for $\mathrm{E}[Y_i] = p_i$.

Motivating example

We shall use a linear model for a function of p, the link function. The canonical link for the binomial distribution is the logit transformation

$$g(p) = \ln\left(\frac{p}{1-p}\right),$$

and we will formulate a linear model for the transformed mean values

$$\eta_i = \ln\left(\frac{p_i}{1-p_i}\right), \quad i = 1, 2, \ldots, 5.$$

The linear model is

$$\eta_i = \beta_1 + \beta_2 x_i, \quad i = 1, 2, \ldots, 5,$$

The inverse transformation, which gives the probabilities, p_i, for stillbirth is the logistic function

$$p_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}, \quad i = 1, 2, \ldots, 5$$

Motivating example - R

```
> mice<-data.frame(
+   stillb=c(15, 17, 22, 38, 144),
+   total=c(297, 242, 312, 299, 285),
+   conc=c(0, 62.5, 125, 250, 500)
+ )
> mice$resp <- cbind(mice$stillb,mice$total-mice$stillb)
```

Note the response variable is conposed by the vector of the number of stillborns, z_i, and the number of live fetuses, $n_i - z_i$.
We use the function `glm` to fit the model:

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

Motivating example - R

```
> anova(mice.glm)
Analysis of Deviance Table

Model: binomial, link: logit

Response: resp

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    4    259.107
conc  1   253.33         3      5.777
```

Motivating example - R

```
> summary(mice.glm)

Call:
glm(formula = resp ~ conc, family = binomial(link = logit), data = mice)

Deviance Residuals:
      1        2        3        4        5
 1.1317   1.0174  -0.5968  -1.6464   0.6284

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.2479337  0.1576602  -20.60   <2e-16 ***
conc         0.0063891  0.0004348   14.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 259.1073  on 4  degrees of freedom
Residual deviance:   5.7775  on 3  degrees of freedom
AIC: 35.204

Number of Fisher Scoring iterations: 4
```

Motivating example - R

The linear predictions and their estimated standard errors:

```
> predict(mice.glm,type='link',se.fit=TRUE)
$fit
          1           2           3           4           5
-3.24793371 -2.84861691 -2.44930011 -1.65066652 -0.05339932

$se.fit
         1          2          3          4          5
0.15766019 0.13490991 0.11411114 0.08421903 0.11382640

$residual.scale
[1] 1
```

The fitted values and their estimated standard errors:

```
> predict(mice.glm,type='response',se.fit=TRUE)
$fit
         1          2          3          4          5
0.03740121 0.05475285 0.07948975 0.16101889 0.48665334

$se.fit
          1           2           3           4           5
0.005676138 0.006982260 0.008349641 0.011377301 0.028436323

$residual.scale
```

Motivating example - R

The response residuals:

```
> residuals(mice.glm,type="response")
```

```
          1           2           3           4           5
0.013103843 0.015495079 -0.008976925 -0.033928587 0.018609817
```

The deviance residuals:

```
> residuals(mice.glm,type="deviance")
```

```
        1         2         3         4         5
1.1316578 1.0173676 -0.5967859 -1.6464253 0.6284281
```

The Pearson residuals:

```
> residuals(mice.glm,type="pearson")
```

```
        1         2         3         4         5
1.1901767 1.0595596 -0.5861854 -1.5961984 0.6285637
```

Motivating example

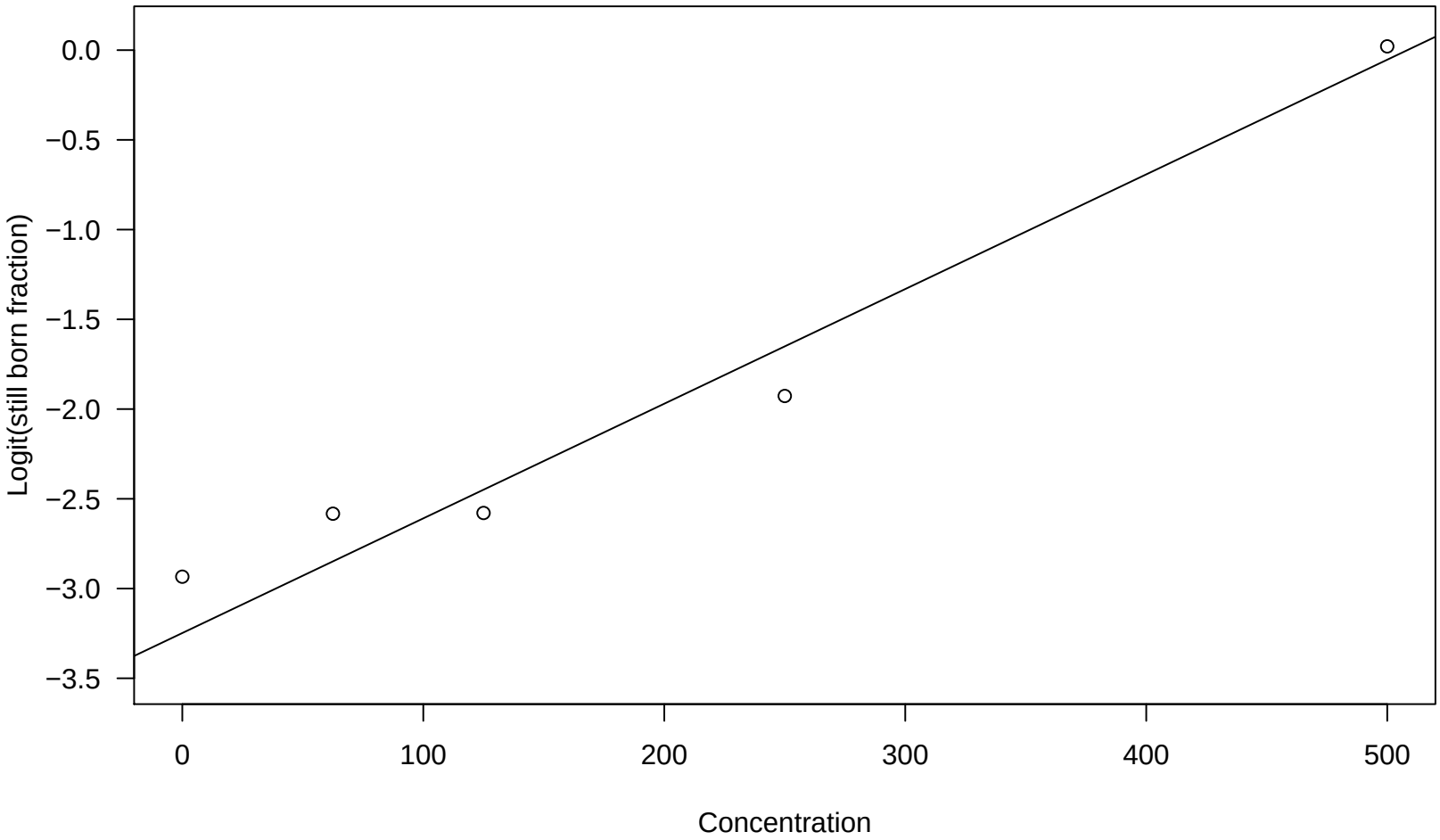


Figure: Logit transformed observations and corresponding linear predictions for dose response assay.

Motivating example

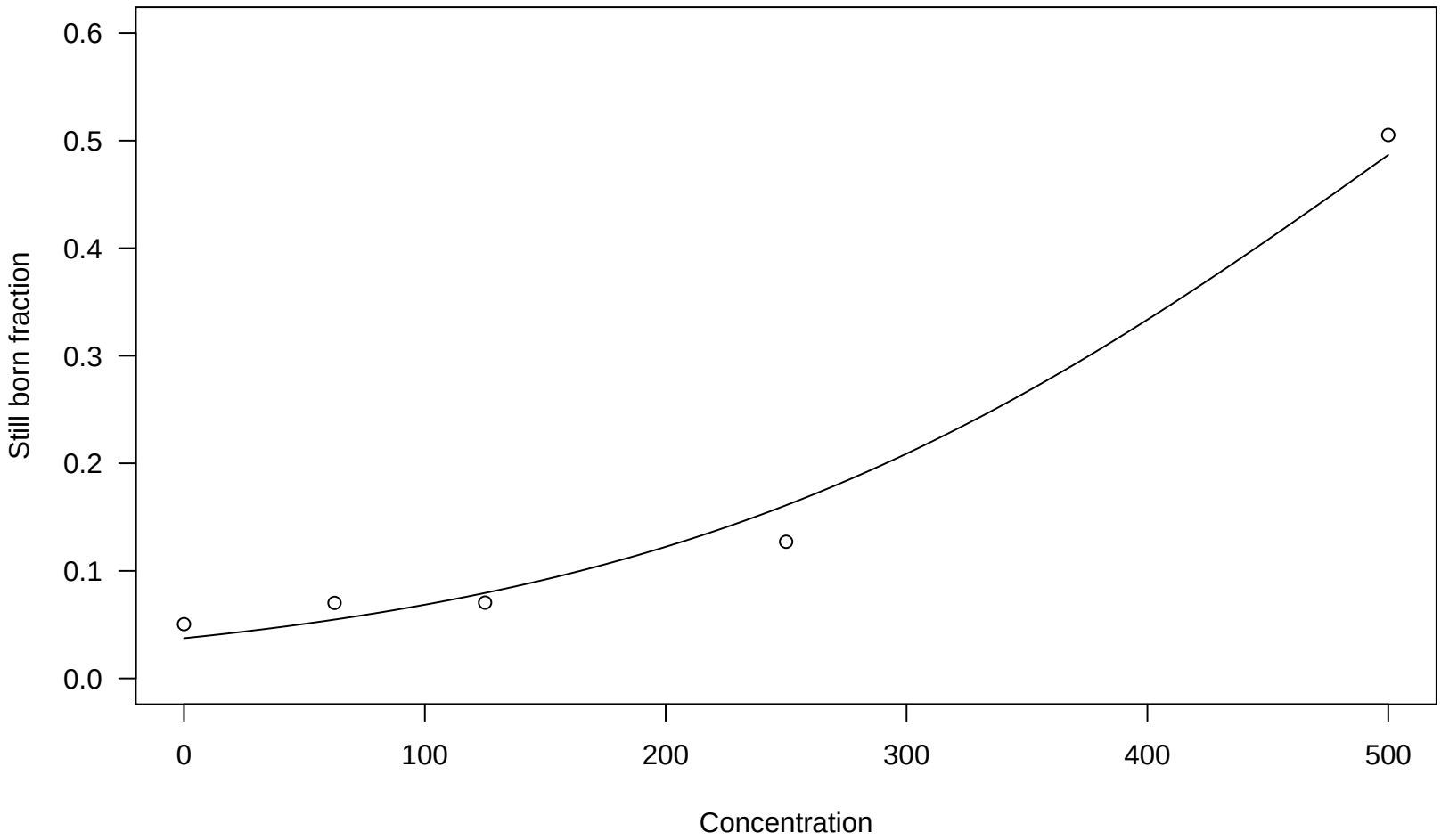


Figure: Observed fraction stillborn and corresponding fitted values under logistic regression for dose response assay.

Exponential families of distributions

Consider a univariate random variable Y with a distribution described by a family of densities $f_Y(y;\theta),\ \ \theta \in \Omega$.

Definition (A natural exponential family)

A family of probability densities which can be written on the form

$$f_Y(y;\theta) = c(y)\exp(\theta y - \kappa(\theta)),\ \ \theta \in \Omega$$

is called a *natural exponential family* of distributions. The function $\kappa(\theta)$ is called the *cumulant generator*. This representation is called the *canonical parametrization* of the family, and the parameter θ is called the *canonical parameter*.

Exponential families of distributions

Definition (An exponential dispersion family)

A family of probability densities which can be written on the form

$$f_Y(y;\theta) = c(y,\lambda)\exp(\lambda\{\theta y - \kappa(\theta)\})$$

is called an *exponential dispersion family* of distributions. The parameter $\lambda > 0$ is called the *precision parameter*.

- Basic idea: separate the mean value related distributional properties described by the *cumulant generator* $\kappa(\theta)$ from features as sample size, common variance, or common over-dispersion.
- In some cases the precision parameter represents a known number of observations as for the binomial distribution, or a known shape parameter as for the gamma (or χ^2-) distribution.
- In other cases the precision parameter represents an unknown dispersion like for the normal distribution, or an over-dispersion that is not related to the mean.

Example: Poisson distribution

Consider $Y \sim \text{Pois}(\mu)$. The probability function for Y is:

$$\begin{aligned} f_Y(y;\mu) &= \frac{\mu^y e^{-\mu}}{y!} \\ &= \frac{1}{y!}\exp\{y\log(\mu) - \mu\} \end{aligned}$$

Comparing with the equation for the natural exponential family it is seen that $\theta = \log(\mu)$ which means that $\mu = \exp(\theta)$.

Thus the Poisson distribution is a special case of a natural exponential family with canonical parameter $\theta = \log(\mu)$, cumulant generator $\kappa(\theta) = \exp(\theta)$ and $c(y) = 1/y!$.

The natural exponential family: $f_Y(y;\theta) = c(y)\exp(\theta y - \kappa(\theta))$

Example: Normal distribution

Consider $Y \sim N(\mu, \sigma^2)$. The probability function for Y is:

$$\begin{aligned} f_Y(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{1}{\sigma^2}\left(\mu y - \frac{\mu^2}{2}\right) - \frac{y^2}{2\sigma^2}\right\} \end{aligned}$$

Thus the normal distribution belongs to the exponential dispersion family with $\theta = \mu$, $\kappa(\theta) = \theta^2/2$ and $\lambda = 1/\sigma^2$. The canonical parameter space is $\Omega = \mathbb{R}$.

The exponential dispersion family: $f_Y(y; \theta) = c(y, \lambda) \exp(\lambda\{\theta y - \kappa(\theta)\})$

Example: Binomial distribution

Consider $Z \sim \text{Bin}(n, p)$. The probability function for Z is:

$$\begin{aligned} f_Z(n,p) &= \left(\begin{array}{c} n \\ z \end{array} \right) p^z (1-p)^{n-z} \\ &= \left(\begin{array}{c} n \\ z \end{array} \right) \exp\left(z \log\left(\frac{p}{1-p} \right) + n \log(1-p) \right) \end{aligned}$$

Thus the binomial distribution belongs to the natural exponential family with $\theta = \log\left(\frac{p}{1-p}\right)$ i.e. $p = \frac{\exp(\theta)}{1+\exp(\theta)}$, $\kappa(\theta) = n\log(1+\exp(\theta))$ and $\lambda = 1$.

The natural exponential family: $f_Y(y;\theta) = c(y)\exp(\theta y - \kappa(\theta))$

Example: Binomial distribution

Consider $Y = Z/n$ where $Z \sim \text{Bin}(n, p)$. The probability function for Y is:

$$f_Y(n, p) = \left(\begin{array}{c} n \\ yn \end{array} \right) p^{yn}(1-p)^{n-yn}$$
$$= \left(\begin{array}{c} n \\ yn \end{array} \right) \exp\left(n \left\{ y \log\left(\frac{p}{1-p} \right) + \log(1-p) \right\} \right)$$

Now we see that $\theta = \log\left(\frac{p}{1-p}\right)$ i.e. $p = \frac{\exp(\theta)}{1+\exp(\theta)}$, $\kappa(\theta) = \log(1+\exp(\theta))$ and $\lambda = n$.

In this case the precision parameter λ represents the (known) number of observations.

The exponential dispersion family: $f_Y(y;\theta) = c(y,\lambda)\exp(\lambda\{\theta y - \kappa(\theta)\})$

Mean and variance

- The properties of the exponential dispersion family are mainly determined by the cumulant generator $\kappa(\cdot)$.
- If Y a distribution belonging to the exponential dispersion family then:

$$\mathrm{E}[Y] = \kappa'(\theta)$$
$$\mathrm{Var}[Y] = \frac{\kappa''(\theta)}{\lambda}$$

- The function

$$\tau(\theta) = \kappa'(\theta)$$

defines an one to one mapping $\mu = \tau(\theta)$ of the parameter space, Ω, for the canonical parameter θ on to a subset, $\mathcal{M}$, of the real line, called the *mean value space*.

(Unit) variance function

(Unit) variance function

We have seen that the variance operator is: $\mathrm{Var}[Y] = \frac{\kappa''(\theta)}{\lambda}$. $\kappa''(\theta)$ is called the *variance function* and by using $\theta = \tau^{-1}(\mu)$ we get

$$V(\mu) = \kappa''(\tau^{-1}(\mu))$$

Variance operator and variance function

Note the distinction between the variance *operator*, $\mathrm{Var}[Y]$, which calculates the variance in the probability distribution of a random variable, Y, and the variance *function*, which is a function, $V(\mu)$, that describes the variance as a function of the mean value for a given family of distributions.

The deviance

Definition (The unit deviance)

As a mean for comparing observations, y, with μ, according to some model, we define the *unit deviance* as

$$d(y;\mu) = 2\int_{\mu}^{y} \frac{y-u}{V(u)}\, du \quad ,$$

where $V(\cdot)$ denotes the variance function.

The density for the exponential dispersion family in terms of μ

The density for the exponential dispersion family may be expressed in terms of the mean value parameter, μ as

$$g_Y(y;\mu,\lambda) = a(y,\lambda)\, \exp\left\{-\frac{\lambda}{2}\, d(y;\mu)\right\}.$$

Example: Unit deviance for the normal distribution

$$
\begin{aligned}
& 2\int_{\mu}^{y} \frac{y-u}{V(u)}\, du \\
&= 2\int_{\mu}^{y} y-u\ du \\
&= 2y\int_{\mu}^{y} 1\ du - 2\int_{\mu}^{y} u\ du \\
&= 2y(y-\mu) - 2\frac{1}{2}(y^2-\mu^2) \\
&= (y-\mu)^2
\end{aligned}
$$

Alternative definition of the deviance

Alternative definition of the deviance

Let $\ell(y;\mu)$ denote the log likelihood of the current model. Then apart from λ, the unit deviance may be defined as

$$d(y;\mu) = 2\max_{\mu} \ell(\mu;y) - 2\ell(\mu;y) \ .$$

The definition corresponds to considering a *normalized, or relative likelihood* for μ corresponding to the observation y:

$$R(\mu;y) = \frac{L(\mu;y)}{\max_{\mu} L(\mu;y)}$$

Then $d(y;\mu) = -2\log(R(\mu;y))$.

For the normal distribution with $\boldsymbol{\Sigma} = \boldsymbol{I}$, the deviance is just the residual sum of squares (RSS).

Variance function, unit deviance and λ

Family	$\mathcal{M}$	$\mathrm{Var}(\mu)$	Unit deviance $d(y;\mu)$	λ	θ
Normal	$(-\infty,\infty)$	1	$(y-\mu)^2$	$1/\sigma^2$	μ
Poisson	$(0,\infty)$	μ	$2\left[y\ln\left(\frac{y}{\mu}\right)-(y-\mu)\right]$	- [1]	$\ln(\mu)$
Gamma	$(0,\infty)$	μ^2	$2\left[\frac{y}{\mu}-\ln\left(\frac{y}{\mu}\right)-1\right]$	α [2]	$1/\mu$
Bin	(0,1)	$\mu(1-\mu)$	$2\left[y\ln\left(\frac{y}{\mu}\right)+(1-y)\ln\left(\frac{1-y}{1-\mu}\right)\right]$	n [3]	$\ln\left(\frac{\mu}{1-\mu}\right)$
Neg Bin	(0,1)	$\mu(1+\mu)$	$2\left[y\ln\left(\frac{y(1+\mu)}{\mu(1+y)}\right)+\ln\left(\frac{1+\mu}{1+y}\right)\right]$	r [4]	$\ln(\mu)$
I Gauss	$(0,\infty)$	μ^3	$\frac{(y-\mu)^2}{y\mu^2}$		$1/\mu^2$

Table: Mean value space, unit variance function and unit deviance for exponential dispersion families.

[1] The precision parameter λ can not be distinguished from the mean value.

[2] Gamma distribution with shape parameter α and scale parameter μ/α.

[3] $Y = Z/n$, where Z is the number of successes in n independent Bernoulli trials.

[4] $Y = Z/r$, where Z is the number of successes until the rth failure in independent Bernoulli. trials.

Exponential dispersion family

There are two equivalent representations for an exponential dispersion family:

i By the cumulant generator, $\kappa(\cdot)$ and parametrized by the canonical (or natural) parameter, $\theta \in \Omega$, and the precision parameter λ

ii By the variance function $V(\cdot)$ specifying the variance as a function of the mean value parameter, $\mu \in \mathcal{M}$, and further parametrized by the precision parameter λ.

The two parametrizations supplement each other:

- The parametrization in terms of the canonical parameter, θ has the advantage that the parameter space is the real line and therefore well suited for linear operations,
- The parametrization in terms of the mean value parameter, μ has the advantage that the fit of the model can be directly assessed as the mean value is measured in the same units as the observations, Y.

Exponential family densities as a statistical model

Consider n independent observations $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)^T$, and assume that they belong to the same exponential dispersion family with the cumulant generator, $\kappa(\cdot)$, and the precision parameter is a known weight, $\lambda_i = w_i$, and the density is on the form:

$$f_Y(y;\theta) = c(y,\lambda)\exp(\lambda\{\theta y - \kappa(\theta)\})$$

which can also be written as:

$$g_Y(y;\mu,\lambda) = a(y,\lambda)\ \exp\left\{-\frac{\lambda}{2}\ d(y;\mu)\right\}.$$

Exponential family densities as a statistical model

Then the joint density, using the canonical parameter, is

$$f(\boldsymbol{y};\boldsymbol{\theta}) = \exp\left[\sum_{i=1}^{n} w_i(\theta_i y_i - \kappa(\theta_i))\right] \prod_{i=1}^{n} c(y_i, w_i)$$

or, by introducing the mean value parameter, $\mu = \tau(\theta)$ we find, the equivalent joint density

$$g(\boldsymbol{y};\boldsymbol{\mu}) = \prod_{i=1}^{n} g_Y(y_i;\mu_i, w_i) = \exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i\, d(y_i;\mu_i)\right] \prod_{i=1}^{n} c(y_i, w_i)$$

Log-likelihood functions

The log likelihood function in the two cases are

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

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

where

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

Appendix: Remark 4.4 on page 93

$$
\begin{aligned}
d(y,\mu) &= 2\int_{\mu}^{y} \frac{y-u}{V(u)}\, du \\
&= 2\int_{\mu}^{y} \frac{y-u}{\kappa''(\tau^{-1}(u))}\, du \\
&= 2\int_{\mu}^{y} \frac{y-u}{\kappa''(\kappa'^{-1}(u))}\, du \\
&= 2\int_{\kappa'^{-1}(\mu)}^{\kappa'^{-1}(y)} y-\kappa'(x)\, dx \qquad \left[\begin{array}{ll} x & = \kappa'^{-1}(u) \\ dx & = \frac{1}{\kappa''(\kappa'^{-1}(u))}\, du \end{array}\right] \\
&= 2\left\{ y(\kappa'^{-1}(y) - \kappa'^{-1}(\mu)) - (\kappa(\kappa'^{-1}(y)) - \kappa(\kappa'^{-1}(\mu))) \right\} \\
&= 2\left\{ -y\theta + \kappa(\theta) + y\tau^{-1}(y) - \kappa(\tau^{-1}(y)) \right\}
\end{aligned}
$$