Generalized linear models

Juha Karvanen

April 30, 2009

Contents

Preface

This document contains short lecture notes for the course Generalized linear models, University of Helsinki, spring 2009. A more detailed treatment of the topic can be found from

- P. McCullagh and John A. Nelder, Generalized linear models. Second edition 1989. Chapman & Hall.
- A. J. Dobson, An introduction to generalized linear models. Second edition 2002. Third edition 2008. Chapman & Hall/CRC.
- lecture notes 2008. `http://www.rni.helsinki.fi/~jmh/glm08/`
- lecture notes 2005 (in Finnish). `http://www.rni.helsinki.fi/~jmh/glm05/glm05.pdf`.

1 What is a generalized linear model?

1.1 Model

Mathematical view: A statistical model is a set of probability distributions on the sample space $\mathcal{S}$. A parameterized statistical model is a parameter set Θ together with a function $P : \Theta \to P(\mathcal{S})$, which assigns to each parameter point $\theta \in \Theta$ a probability distribution P_θ on $\mathcal{S}$. A Bayesian model requires an additional component in the form of a prior distribution on Θ. [P. McCullagh (2002). What is a statistical model. The Annals of Statistics. Vol. 30, No. 5, 1225-1310.]

Applied view: Statistical model is a description of the probability distribution of random variables which can be assumed to represent a real world phenomenon.

Which of these are statistical models?

a) $X \sim N(\mu, \sigma^2)$

b) "The height of Finnish men follows a normal distribution."

c)

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) \propto \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(g_i) p_{\boldsymbol{\psi}}(x_i \mid g_i) p_{\boldsymbol{\theta}}(y_i \mid g_i, x_i),$$

d) "The risk of smokers to die to cardiovascular diseases is about twice the risk of non-smokers."

e) `glm(y ~ x, family=binomial(link = "logit"), data=doseresponse)`

1.2 Linear model

A simple linear model that describes the relationship of a single covariate x and a continuous response variable Y can be written as

$$Y_i = \alpha + \beta x_i + \epsilon_i, \tag{1}$$

where α is the intercept term, β is the regression coefficient for X and ϵ_i is an error term. Further assumptions are needed for the error term. For instance, we may assume that the error terms are mutually independent and $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \dots, n$. A less restrictive assumption is to specify only the first two moments $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$, i.e. the variance does not depend on x. Note that in model (1), the error term ϵ_i is written explicitly. It is also possible to write the same model without explicitly specifying ϵ_i

$$\mathrm{E}(Y_i \mid x_i) = \mu_i = \alpha + \beta x_i. \tag{2}$$

Model (2) tells on the expected value of Y_i on the condition of x. As a such, model (2) does not specify how the values of Y_i vary around the expected value $\mathrm{E}(Yi \mid x_i)$. Defining $\mathrm{Var}(Y_i) = \sigma^2$ we obtain a model equivalent to model (1). If the variation of Y_i is normally distributed, it can be also written $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$.

The linearity of linear model means linearity respect to the parameters. In other words, the model $\mu_i = \alpha + \beta x_i^3$ is also a linear model.

1.3 Generalized linear model

The linear model (2) can be transformed to a generalized linear model by replacing μ_i by $g(\mu_i)$

$$g(\mu_i) = \alpha + \beta x_i = \eta_i, \tag{3}$$

where g is a real-valued monotonic and differentiable function called link function and the term η_i is called linear predictor. In the other words, μ_i is the expected value of the response, η_i is a linear combination of the covariates and $g()$ defines the relationship between μ_i and η_i. Because $g()$ is monotonic, the relationship of μ_i and η_i is also monotonic. With the inverse of $g()$ we may write

$$\mu_i = g^{-1}(\eta_i), \tag{4}$$

which provides an alternative way to define GLM. Linear model is a special case of GLM where $g(\mu_i) = \mu_i$.

With multiple covariates the GLM is defined as

$$g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}. \tag{5}$$

The assumptions of the GLM are given in Section 3.

Note that GLM is different from applying a nonlinear transformation to response variable. In GLM, the nonlinear transformation is applied to the expected value of the response.

Variance is defined by the variance function V that specifies the variance of Y_i as a function μ_i

$$\mathrm{Var}(Y_i) \propto V(\mu_i). \tag{6}$$

1.4 Motivating examples

Generalized linear models are needed because linear models are not appropriate for all situations. In linear model it is implicitly assumed that the response can be have all real values, which is not the case in many practical situations. Examples:

- The number of hospital visits in a certain year for an individual is a count response that can have values $0, 1, 2, \ldots$.
- Monthly alcohol consumption (liters of absolute alcohol) for an individual is a nonnegative response that has zeroes for some individuals.
- Gamma-glutamyltransferase (GGT) measured from serum blood is a positive response.
- Daily rainfall is a nonnegative response.
- Presence or absence of a voltage peak in switching measurements of superconducting Josephson Junctions is a binary response.
- Fatality (fatal/non-fatal) of myocardial infarction (heart attack) is a binary response.
- Level of education (primary school, secondary school, B.Sc., M.Sc., PhD) is an ordinal response.
- The date of an event of coronary heart disease measured for a cohort of people is a time-to-event (or survival) response.

There are also situations where a linear model may be suitable although strictly speaking the response has an inappropriate distribution.

- Height of an adult is positive but can be modeled by linear model because all values are far from zero.

- The daily number of customers in a big supermarket is actually a count response but could be modeled by linear model because all values are far from zero and the number of possible values of the response is high.

1.5 Link functions

The choice of the link function $g()$ depends on the data, especially on the type of the response variable. If the response is a count, i.e. an integer, log-link $g(\mu_i) = \log(\mu_i)$ may be used. Log-link leads multiplicative model

$$\mu_i = \exp(\eta_i) = e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} \cdots e^{\beta_p x_{ip}} \tag{7}$$

If the response Y_i is a binary variable with possible values 0 and 1, it holds

$$\mu_i = \mathrm{E}(Y_i) = 1 \cdot P(Y_i = 1) + 0 \cdot P(Y_i = 0) = P(Y_i = 1). \tag{8}$$

The logit-link

$$g(\mu_i) = \mathrm{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \tag{9}$$

is maybe the most typical choice for binary response data. For positive continuous responses typical link functions are inverse link

$$\mu_i^{-1} = \eta_i \tag{10}$$

and inverse-squared link

$$\mu_i^{-2} = \eta_i. \tag{11}$$

1.6 Confusing terminology

1.6.1 Generalized linear model (GLM) and general linear model (GLM)

Unfortunately, the acronym GLM is sometimes used for general linear model. General linear model is a linear model. The word 'general' is used to indicate that the response **Y** may be multivariate and the covariates may include both continuous and categorical variables. In SAS, PROC GLM fits a general linear model, not a generalized linear model.

1.6.2 Names of X and Y

In different applications X and Y have various names that sometimes might be confusing. Examples are given below. Some of the names are synonyms and some have special emphasis in certain applications. Particularly, the terms ‘independent variable’ and ‘dependent variable’ may cause a confusion.

Names of X

- covariate
- explanatory variable
- factor
- risk factor
- exposure (variable)
- design variable
- controlled variable
- carrier variable
- regressor
- predictor
- input
- determinant
- *independent variable

Names of Y

- response
- explained variable
- outcome
- responding variable
- regressand
- experimental variable
- measured variable
- output
- *dependent variable

2 Generalized linear models in statistical software

2.1 Generalized linear models in R

In R (`www.r-project.org`) generalized linear models can be fitted using function `glm`. The syntax is

```
glm(formula, family = gaussian, data, weights, subset, na.action,
start = NULL, etastart, mustart, offset, control = glm.control(...),
model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, contrasts
= NULL, ...)
```

Arguments
Some important arguments are

`formula` an object of class ”formula” (or one that can be coerced to that class): a symbolic description of the model to be fitted.

`family` a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function.

`data` an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. `weights`] an optional vector of weights to be used in the fitting process.

`subset` an optional vector specifying a subset of observations to be used in the fitting process.

`offset` can be used to specify an a priori known component to be included in the linear predictor during fitting. This should be NULL or a numeric vector of length either one or equal to the number of cases. One or more offset terms can be included in the formula instead or as well, and if both are specified their sum is used. See model.offset.

`control` a list of parameters for controlling the fitting process.

Output
As an output an object of class “glm” is returned. A glm object is a list that contains the following components among the others:

`coefficients` a named vector of coefficients

`fitted.values` the fitted mean values, obtained by transforming the linear predictors by the inverse of the link function.

`deviance` up to a constant, minus twice the maximized log-likelihood. Where sensible, the constant is chosen so that a saturated model has deviance zero.

`aic` Akaike's An Information Criterion, minus twice the maximized log-likelihood plus twice the number of coefficients (so assuming that the dispersion is known).

`null.deviance` The deviance for the null model, comparable with deviance.

`iter` the number of iterations of IWLS used.

`df.residual` the residual degrees of freedom.

`df.null` the residual degrees of freedom for the null model.

`converged` logical. Was the IWLS algorithm judged to have converged?

Example: binomial family with logit-link (logistic regression)

```
set.seed(3000)
b<-3;
n<-500;
x<-rnorm(n);
y<-runif(n)<exp(b*x)/(1+exp(b*x))
m1<-glm(y~x,binomial(link = "logit"))
print(summary(m1))
```

Summary:

```
Call:
glm(formula = y ~ x, family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.66224  -0.53516   0.01267   0.45869   2.62460
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.01346    0.13572  -0.099    0.921
x            3.27787    0.29793  11.002   <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: 693.12  on 499  degrees of freedom
Residual deviance: 342.09  on 498  degrees of freedom
AIC: 346.09

Number of Fisher Scoring iterations: 6
```

2.2 Generalized linear models in SAS, Matlab and SPSS

There are several procedures in SAS for generalized linear models. PROC GLM (where G stands for ‘general’ not for ‘generalized’) can be used to fit and test linear models. Binary and categorical response data can be handled with PROC LOGISTIC, PROC PROBIT, PROC CATMOD and PROC GENMOD. PROC GENMOD is based on the philosophy of generalized linear models and allows user-defined link functions in addition to the commonly used link functions.

In Matlab, Statistics toolbox has function `glmfit` and `glmval`. SPSS Advanced Statistics contains the module GENLIN.

3 Theory of generalized linear models

3.1 Notation

The observed data set $(\mathbf{y}, \mathbf{X})$ contains n observations of $1 + p$ variables

$$\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \dots y_n \end{pmatrix}^T \tag{12}$$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots x_{1p} \\ x_{21} & x_{22} & \dots x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots x_{np} \end{pmatrix}. \tag{13}$$

Variable y is the response variable and variables $x_1, x_2, \dots x_p$ are explanatory variables or covariates. The observed value y_i is treated as a realization of a random variable Y_i. In experimental setup, the explanatory variables have fixed values set by the experimenter. In observational setup, the value x_{ij} can be understood to be a realization of a random variable X_{ij} but when distribution of Y_i is considered x_{ij} is taken as fixed.

The parameters include the regression coefficients

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \beta_2 & \dots & \beta_p \end{pmatrix}^T, \tag{14}$$

the linear predictors

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_1 & \eta_2 & \dots & \eta_n \end{pmatrix}^T, \tag{15}$$

the expected responses

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_n \end{pmatrix}^T, \tag{16}$$

and the canonical parameters

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 & \theta_2 & \dots & \theta_n \end{pmatrix}^T. \tag{17}$$

3.2 Model assumptions

1. The distribution of Y_i belongs to the exponential family. For the exponential family, the density function can be presented in the form

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left(\frac{a_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi/a_i) \right), \tag{18}$$

where

- θ_i, $i = 1, \ldots, n$ are unknown parameters (canonical parameters),
- ϕ is the dispersion parameter (scale parameter) that can be known or unknown,
- a_i, $i = 1, \ldots, n$ are known prior weights of each observation and
- $b()$ and $c()$ are known functions. The first derivative $b'()$ is monotonic and differentiable.

2. Random variables $Y_1, Y_2, \ldots, Y_n$ are mutually independent.

3. The expected value $\mu_i = \mathrm{E}(Y_i)$ depends on linear predictor $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$ through monotonic and differentiable link function g

$$g(\mu_i) = \eta_i. \tag{19}$$

For instance, normal, binomial, Poisson and gamma distributions belong to the exponential family. For exponential family (18) it holds

$$\mathrm{E}(Y_i) = b'(\theta_i) = \mu_i \tag{20}$$

and

$$\mathrm{Var}(Y_i) = \frac{b''(\theta_i)\phi}{a_i} = \frac{V(\mu_i)\phi}{a_i}. \tag{21}$$

As shown in section 3.8, the assumption on the exponential family can be relaxed.

3.3 Likelihood

The log-likelihood of $y_1, \ldots, y_n$ from an exponential family with known dispersion parameter ϕ can be written

$$l(\theta_1, \ldots, \theta_n; \phi, \mathbf{a}, \mathbf{y}) = \sum_{i=1}^{n} \left(\frac{a_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi/a_i) \right) \tag{22}$$

If there are no restrictions for parameters $\theta_1, \ldots, \theta_n$, the model is saturated, i.e. it has as many parameters as there are observations. In a GLM, the parameters $\theta_1, \ldots, \theta_n$ depend on $\mathbf{X}$ and the parameters $\beta_1, \ldots, \beta_p$ through functions $b()$ and $g()$

$$\sum_{j=1}^{p} \beta_j x_{ij} = \eta_i = g(\mu_i) = g(b'(\theta_i)). \tag{23}$$

Therefore, the log-likelihood can be written also a function of the parameters $\mu_1, \ldots, \mu_n$ or as a function of the parameters $\beta_1, \ldots, \beta_p$

$$l(\mu_1, \ldots, \mu_n; \phi, \mathbf{a}, \mathbf{y}) =$$
$$\sum_{i=1}^{n} \left(\frac{a_i(y_i(b')^{-1}(\mu_i) - b((b')^{-1}(\mu_i)))}{\phi} + c(y_i, \phi/a_i) \right), \tag{24}$$
$$l(\beta_1, \ldots, \beta_p; \phi, \mathbf{a}, \mathbf{y}) =$$
$$\sum_{i=1}^{n} \left(\frac{a_i(y_i(b')^{-1}(g^{-1}(\sum_{j=1}^{p} \beta_j x_{ij})) - b((b')^{-1}(g^{-1}(\sum_{j=1}^{p} \beta_j x_{ij}))))}{\phi} + c(y_i, \phi/a_i) \right). \tag{25}$$

3.4 Canonical link

The link function for which it holds $\eta_i = g(\mu_i) = \theta_i$ is called canonical link. Because $\mu_i = b'(\theta)$, it follows $g = (b')^{-1}$. The use of canonical link function simplifies calculations but this alone does not justify the use of canonical link. The link function should be selected on the basis of the data and prior knowledge on the problem.

3.5 Score function, observed information and expected information (Fisher information)

The partial derivative of log-likelihood with respect to some parameter is called score or score function. In the case of the exponential family (22) we obtain

$$
\begin{aligned}
\frac{\partial l}{\partial \theta_i} &= \frac{a_i(y_i - b'(\theta_i))}{\phi}, && (26)\\
\frac{\partial l}{\partial \mu_i} &= \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i} = \frac{a_i(y_i - b'(\theta_i))}{\phi}\frac{1}{V(\mu_i)}, && (27)\\
\frac{\partial l}{\partial \eta_i} &= \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} = \frac{a_i(y_i - b'(\theta_i))}{\phi}\frac{1}{V(\mu_i)}(g^{-1})'(\eta_i), && (28)\\
\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n}\frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n}\frac{a_i(y_i - b'(\theta_i))}{\phi}\frac{1}{V(\mu_i)}(g^{-1})'(\eta_i)x_{ij} = \\
&\quad \frac{1}{\phi}\sum_{i=1}^{n}\frac{a_i(y_i - \mu_i(\boldsymbol{\beta}))x_{ij}}{V(\mu_i(\boldsymbol{\beta}))g'(\mu_i(\boldsymbol{\beta}))} && (29)
\end{aligned}
$$

where the notation $\mu_i(\boldsymbol{\beta})$ emphasizes the fact that μ_i depends on $\boldsymbol{\beta}$.

The observed information is the negative of the matrix of second order partial derivatives of log-likelihood

$$
J(\boldsymbol{\beta}, \mathbf{y}) = -\frac{\partial^2 l(\boldsymbol{\beta}, \mathbf{y})}{\partial \boldsymbol{\beta}^2} = \begin{pmatrix} -\sum_{i=1}^{n}\frac{\partial^2 l(\boldsymbol{\beta}, y_i)}{\partial \beta_1^2} & \cdots & -\sum_{i=1}^{n}\frac{\partial^2 l(\boldsymbol{\beta}, y_i)}{\partial \beta_1 \partial \beta_p} \\ \vdots & \ddots & \vdots \\ -\sum_{i=1}^{n}\frac{\partial^2 l(\boldsymbol{\beta}, y_i)}{\partial \beta_p \partial \beta_1} & \cdots & -\sum_{i=1}^{n}\frac{\partial^2 l(\boldsymbol{\beta}, y_i)}{\partial \beta_p \partial \beta_p} \end{pmatrix} \quad (30)
$$

and the Fisher information or expected information is the expected value of observed information

$$
I(\boldsymbol{\beta}) = \mathrm{E}_{\mathbf{Y}}(J(\boldsymbol{\beta}, \mathbf{Y})) = \sum_{i=1}^{n}\mathrm{E}_{Y_i}(J(\boldsymbol{\beta}, Y_i)) = -\sum_{i=1}^{n}\mathrm{E}\left(\frac{\partial^2 l(\boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta}^2}\right). \quad (31)
$$

3.6 Estimation

The maximum likelihood estimate for $\boldsymbol{\beta}$ is obtained by solving score equations

$$
\frac{\partial l(\boldsymbol{\beta}, \mathbf{y})}{\partial \boldsymbol{\beta}} = 0. \quad (32)
$$

Usually the estimation requires numerical methods. Traditionally, the maximum likelihood estimation is carried out with Fisher scoring (also called iterative weighted least squares) which is a modification of the Newton-Raphson algorithm.

In Newton-Raphson update rule

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + J^{-1}\frac{\partial l(\boldsymbol{\beta}, \mathbf{y})}{\partial \boldsymbol{\beta}} \tag{33}$$

the observed information J is replaced by the expected information I. After some algebra, this leads to the update formula

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{z}^{(t)}, \tag{34}$$

where

$$\mathbf{W}^{(t)} = \begin{pmatrix} w_1^{(t)} & & \\ & \ddots & \\ & & w_1^{(t)} \end{pmatrix}, \tag{35}$$

$$w_i^{(t)} = \frac{a_i}{\left[g'\left(\mu_i(\hat{\boldsymbol{\beta}}^{(t)})\right)\right]^2 V\left(\mu_i(\hat{\boldsymbol{\beta}}^{(t)})\right)}, \tag{36}$$

$$\mathbf{z}^{(t)} = (z_1^{(t)} \dots z_n^{(t)})^T \tag{37}$$

$$z_i^{(t)} = \eta_i(\hat{\boldsymbol{\beta}}^{(t)}) + (y_i - \mu_i(\hat{\boldsymbol{\beta}}^{(t)}))g'\left(\mu_i(\hat{\boldsymbol{\beta}}^{(t)})\right). \tag{38}$$

It can be seen that the updating rule depends on the distribution of Y_i only through the variance function V.

When the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ exists, it is consistent and asymptotically normal with expected value $\boldsymbol{\beta}$ and covariance matrix $\phi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$.

The dispersion parameter ϕ can estimated by the deviance (see Section 3.7) estimator

$$\hat{\phi} = \frac{D}{n-p} \tag{39}$$

or the moment estimator

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{a_i(y_i - \mu_i(\hat{\boldsymbol{\beta}}))^2}{V(\mu_i(\hat{\boldsymbol{\beta}}))}. \tag{40}$$

3.7 Deviance

Deviance is defined as

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\phi(l(\mathbf{y}; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}; \mathbf{y})) \tag{41}$$

where $l(\mathbf{y};\mathbf{y})$ is the log-likelihood of the saturated model (full model). In the saturated model, the number of parameters equals the number of observations and likelihood obtains its maximum for the model class. Scaled deviance is defined as

$$D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = \frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{\phi} \tag{42}$$

As seen in Section 4.5, deviance is closely related to the likelihood ratio test.

3.8 Quasi-likelihood

GLMs allow defining the variance function independently from the link function. The assumption that the distribution of Y_i belongs to the exponential family can be replaced by an assumption that concerns only the variance of Y_i

$$\mathrm{Var}(Y_i) = \frac{\phi V(\mu_i)}{a_i}. \tag{43}$$

Parameters can be estimated maximizing quasilikelihood

$$Q(\boldsymbol{\beta};\mathbf{y}) = \frac{1}{\phi}\sum_{i=1}^{n}\int_{y_i}^{\mu_i}\frac{a(y_i-t)}{V(t)}dt. \tag{44}$$

The form of quasilikelihood function is chosen so that partial derivatives

$$\frac{\partial Q(\boldsymbol{\beta};\mathbf{y})}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{a_i(y_i-\mu_i(\boldsymbol{\beta}))x_{ij}}{V(\mu_i(\boldsymbol{\beta}))g'(\mu_i(\boldsymbol{\beta}))}. \tag{45}$$

are similar to the partial derivatives of likelihood function and consequently the parameters can be estimated by Fisher scoring.

4 Modeling

4.1 Process of modeling

1. Study design
2. Data collection
3. Selection of model class
4. Estimation
5. Model checking
6. Conclusions
7. Reporting

4.2 Residuals

Residuals can be used to check the model fit. For GLMs different kind of residuals can be defined:

Raw residuals (response residuals)

$$r_i = y_i - \hat{\mu}_i \tag{46}$$

Pearson residuals

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/a_i}} \tag{47}$$

whose squared sum

$$\sum_{i=1}^{n} r_{P,i}^2 = X^2 \tag{48}$$

is the Pearson chi-squared goodness-of-fit statistic.

Deviance residuals

$$r_{D,i} = \operatorname{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}, \tag{49}$$

where

$$d_i = 2a_i\left(y_i\left(\theta_i(y_i) - \theta_i(\hat{\mu}_i)\right) - b\left(\theta_i(y_i)\right) + b\left(\theta_i(\hat{\mu}_i)\right)\right). \tag{50}$$

The deviance is the squared sum of the deviance residuals

$$\sum_{i=1}^{n} r_{D,i}^2 = D(\mathbf{y}; \hat{\boldsymbol{\mu}}) \tag{51}$$

Anscombe residuals where y_i's and μ_i's are transformed so that the residuals become approximately normally distributed.

In R, method `residuals` for class `glm` can compute raw, Pearson and deviance residuals.

Influential observations can be identified, for instance, by calculating differences

$$\Delta_i \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)} \tag{52}$$

where $\hat{\boldsymbol{\beta}}^{(i)}$ is estimated from data without observation i.

4.3 Nonlinear terms

GLMs allow inclusion of known transformations of the covariates as far as the linear predictor η_i can be presented as a sum of transformed covariates. For instance, the design matrix may be defined as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{11}^2 & x_{11}^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n1}^2 & x_{n1}^3 \end{pmatrix}. \tag{53}$$

to fit a GLM with a third order polynomial for covariate x_1.

4.4 Interactions

Interaction terms are nonlinear transformations of two or more covariates. The type of interaction can synergistic (the joint effect is stronger than the additive effect) or antagonist (the joint effect is weaker than the additive effect). It is usually a bad idea to include interaction terms without the corresponding main effects.

4.5 Hypothesis testing

4.5.1 Single test

The test of hypothesis

$$\begin{aligned} H_0 : &\quad \beta_j = 0 \\ H_1 : &\quad \beta_j \neq 0 \end{aligned}$$

for a certain regression coefficient β_j of a GLM can be based on the likelihood ratio test

$$\Lambda = \frac{\sup\{L(\beta_1, \ldots, \beta_p; \mathbf{y}) : \beta_j = 0\}}{\sup\{L(\beta_1, \ldots, \beta_p; \mathbf{y})\}} \tag{54}$$

Using log-likelihoods the test statistic can be written

$$-2 \log \Lambda = 2(l(\hat{\boldsymbol{\beta}}; \mathbf{y}) - l(\hat{\boldsymbol{\beta}}_0; \mathbf{y})) \tag{55}$$

where $\hat{\boldsymbol{\beta}} = \hat{\beta}_1, \ldots, \hat{\beta}_p$ and $\hat{\boldsymbol{\beta}}_0 = \hat{\beta}_1, \ldots, \beta_j = 0, \ldots, \hat{\beta}_p$ are the maximum likelihood estimates under the two models. The statistic $-2 \log \Lambda$ follows asymptotically χ^2_1 distribution The test can written also in terms of deviance

$$-2 \log \Lambda = \frac{D(\mathbf{y}; \hat{\boldsymbol{\beta}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\beta}})}{\phi}. \tag{56}$$

The likelihood ratio test for more than one parameter is similar but the test statistic follows asymptotically χ^2 distribution with degrees of freedom equal to the difference in dimensionality of β and β_0. If the dispersion parameter is not known, the test statistics

$$\frac{D(\mathbf{y}; \hat{\boldsymbol{\beta}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\beta}})}{\hat{\phi}(p - q)} \tag{57}$$

where q is the dimensionality of β follows asymptotically F-distribution $F_{p-q,n-p}$.

4.5.2 Multiple tests

Let $p_1, p_2, \ldots, p_m$ be the nominal p-values from m tests. Family-wise error rate (FWER) is the probability that at least one true null hypothesis is falsely rejected. Several approaches for controlling FWER exist: a simple approach is the Bonferroni correction where the nominal p-values are compared to the α/m where α is the significance level. If the tests are dependent, the Bonferroni correction is too conservative and the actual significance level is smaller than α.

False discovery rate (FDR) is the expected proportion of incorrectly rejected null hypothesis in a set of hypotheses. The FDR analysis has been used e.g. in genome wide association (GWA) studies where the number of tests can be one million.

4.6 Model selection

Multiple models. Competing models are fitted and the estimated model parameters are reported for each model. The properties of the models are discussed. This is actually not a formal model selection method but a commonly used practical approach to the problem. The approach is feasible only if the number of the competing models is small.

Likelihood ratio test can be used to compare nested models.

Stepwise regression. In forward selection, the procedure starts with a null model and covariates are added one by one. The procedure continues until the newly added covariate does not improve the model. The improvement of the model defined e.g. by the p-value of the likelihood ratio test. In backward elimination, the procedure starts with the full model and covariates are removed one by one. The procedure continues until the removal of a covariate makes the model worse. The lasso (least absolute shrinkage and selection operator, `http://www-stat.stanford.edu/~tibs/lasso.html`) can be understood as a modernized version of stepwise regression (not based on likelihood). Stepwise methods cannot guarantee that the best model will be selected. Automated methods should not replace careful thinking.

Information criteria: Akaike information criterion (AIC), Bayesian information criterion (BIC), Bayes factor, crossvalidation, etc. AIC and BIC are straightforward to compute

$$\begin{aligned} AIC &= -2l(\boldsymbol{\beta}; \mathbf{y}) + 2p, \\ BIC &= -2l(\boldsymbol{\beta}; \mathbf{y}) + p\log(n), \end{aligned}$$

where p is the number of parameters in the model. The model with smallest value of AIC (or BIC if that is used) will be considered the best. Both AIC and BIC penalize models for a higher number of parameters. In BIC, the penalty depends also on the number of observations.

4.7 Experimental and observational studies

Experimental data origin from data generating mechanism where the experimenter selects the values of some variables. In observational data, all values are recorded as observed. The same GLMs can be used for both types of data. The analysis follows the same lines but the interpretation of the results may differ. In general, only experimental data allows causal inference. With observational data, the possibility of confounders and alternative causal explanations must be accounted.

4.8 Missing data

Usually there are missing observations in real world data. A statistician has the following options:

Ignore the missing observations and analyze only the complete cases. This is applicable if only few observations are missing.

Impute the missing values. Multiple imputation is preferred over single imputation. The challenges lie in the definition of the imputation model.

Model the data. The likelihood becomes an integral over the missing values. The results are sensitive to model misspecification and estimation may require a lot of computational resources.

4.9 Few words on independence

Term “independence” may have different meanings depending on the context. In statistics, the term refers to independence of events or to independence of random variables. Events A and B are independent if

$$P(A \text{ and } B) = P(A)P(B). \tag{58}$$

or equivalently, using conditional probabilities

$$P(A \,|\, B) = P(A) \tag{59}$$

or

$$P(B \mid A) = P(B). \tag{60}$$

Random variables X and Y are independent (marked $X \perp\!\!\!\perp Y$) if

$$F_{X,Y}(X, Y) = F_X(X)F_Y(Y). \tag{61}$$

The term "linear independence of random variables" is sometimes used to indicate that the random variables are uncorrelated but this usage is not recommended. In general, zero correlation does not imply independence.

In linear algebra, linear independence of a family of vectors means that none of the vectors can be presented as a linear combination of the other vectors. A matrix whose columns are linearly independent has full rank.

The concept of conditional independence is important when causality is considered. Random variables X and Y are independent on the condition of Z (notation $X \underset{Z}{\perp\!\!\!\perp} Y$ or $X \perp\!\!\!\perp Y \mid Z$ may be used) when

$$F_{X,Y \mid Z}(X, Y \mid Z) = F_{X \mid Z}(X \mid Z)F_{Y \mid Z}(Y \mid Z) \tag{62}$$

or equivalently

$$F_{X, \mid Y,Z}(X \mid Y, Z) = F_{X \mid Z}(X \mid Z). \tag{63}$$

5 Binary response

5.1 Representations of binary response data

In binary response data, the response Y_i has two possible values, for instance, 0 and 1. Binary response data can be presented in different formats:

Data matrix

x	y
250	0
250	1
350	1
300	0
250	0
300	1
$\vdots$	$\vdots$

Weighted data matrix

x	y	frequency
250	0	23
250	1	12
300	1	21
300	0	19
350	0	7
350	1	13

Frequency table (crosstabulation)

	$Y = 0$	$Y = 1$
$x = 250$	23	12
$x = 300$	21	19
$x = 350$	7	13

The response Y_i can be either a Bernoulli random variable (binary response) or a sum of Bernoulli random variable (binomial response). In the latter case, the observational units with the identical covariate values belong to the same covariate class. For the ith covariate class m_i binary responses are recorded and the number of responses 1 is denoted by K_i. The binomial response is defined

$$Y_i = \frac{K_i}{m_i}. \tag{64}$$

5.2 Link functions for binary data

If the possible values of Y_i are 0 and 1, it holds

$$P(Y_i = 1) = \mathrm{E}(Y_i) = g^{-1}(\eta_i), \tag{65}$$

where the possible values of the inverse link function $g^{-1}()$ belong to the interval $(0, 1)$. Any cumulative distribution function defines the inverse of a link function. The commonly used link functions are the logit link

$$g(\mu_i) = \mathrm{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right), \tag{66}$$

the probit link

$$g(\mu_i) = \mathrm{probit}(\mu_i) = \Phi^{-1}(\mu_i), \tag{67}$$

where Φ^{-1} is the inverse of cumulative distribution function (cdf) of the standard normal distribution and the complementary log-log link

$$g(\mu_i) = \mathrm{cloglog}(\mu_i) = \log(-\log(1-\mu_i)). \tag{68}$$

5.3 Odds and log-odds

It is often interesting to compare the estimated responses for different values of covariates. Denote

$$p_A = P(Y_i = 1 \mid \eta_A) \tag{69}$$
$$p_B = P(Y_i = 1 \mid \eta_B) \tag{70}$$

where η_A and η_B are the linear predictors for certain values of covariates. Now the odds ratio is defined as

$$\frac{p_A/(1-p_A)}{p_B/(1-p_B} \tag{71}$$

and the logarithm of the odds ratio becomes

$$\log\left(\frac{p_A/(1-p_A)}{p_B/(1-p_B}\right) = \log\left(\frac{p_A}{1-p_A}\right) - \log\left(\frac{p_B}{1-p_B}\right) = \mathrm{logit}(p_A) - \mathrm{logit}(p_B), \tag{72}$$

which in the case of logit link simplifies

$$\operatorname{logit}(p_A) - \operatorname{logit}(p_B) = \eta_A - \eta_B. \tag{73}$$

5.4 Latent variables

Consider an example where the effectiveness of an insecticide to mosquitos is studied. Mosquitos have different resistance to the insecticide. A mosquito dies ($Y = 1$) if the amount of insecticide x is higher than a threshold value T, which varies in the population. Because T cannot be directly measured, it is called latent variable. If T follows the normal distribution with mean $-\alpha/\beta$ and variance $1/\beta^2$ we obtain for a mosquito randomly chosen from the population

$$P(Y = 1) = P(T \leq x) = \Phi\left(\frac{x - (-\alpha/\beta)}{1/\beta}\right) = \Phi(\alpha + \beta x). \tag{74}$$

In other words, the use of normal distributed latent variable led to the probit model. If T follows logistic distribution, we will end up with the logistic model. If T follows Gumbel distribution, we will end up with the GLM with cloglog link.

5.5 Overdispersion

Overdispersion means that the variance in the data is greater than the variance assumed in the model. The sum of independent Bernoulli random variables

$$K = Y_1 + Y_2 + \ldots + Y_m \tag{75}$$

follows binomial distribution $K \sim \operatorname{Bin}(m, \mu)$ where $\operatorname{E}(Y_i) = \mu$. It follows that $\operatorname{Var}(K) = m\mu(1 - \mu)$. In real world datasets, however, the assumption of independence is often unrealistic and $\operatorname{Var}(K) > m\mu(1 - \mu)$. This is called overdispersion.

5.6 Non-existence of maximum likelihood estimates

Maximum likelihood estimates do not exist if the data can be perfectly separated on the basis of covariate values, for example, response 1 is always obtained if $x > 100$ and response 0 is always obtained if $x < 100$.

5.7 Example: Switching measurements

The Josephson junction (JJ) circuits are important non-linear components of superconducting electronics. The strong dependence of the physical parameters of JJ circuits as function of changes in environmental variables, for instance, temperature, electric noise, and magnetic field makes the JJ circuits to have several applications as ultra-sensitive sensors. Moreover, certain JJ circuits are promising candidates for realization of quantum computation. An experiment called switching measurement is a common way to probe the properties of a JJ circuit sample. In the experiment, sequences of current pulses are applied to the sample, while the voltage over the structure is monitored. Switching measurements are ideal applications for design of experiments in sense that the underlying parametric model for the switching dynamics of a single JJ can be derived directly from the laws of physics. With quantum mechanical arguments, it can be shown that the probability of the voltage response can be approximated by

$$\begin{aligned} P(Y=1) &= 1 - e^{-\exp(ax+b)} \\ P(Y=0) &= e^{-\exp(ax+b)}, \end{aligned} \tag{76}$$

where a and b are unknown parameters to be estimated and x is the height of the current pulse. It follows that the measurement data can be modeled by a GLM with cloglog link function.

In an experiment carried out in Low Temperature Laboratory, Helsinki University of Technology in August 2005, a sample consisting of aluminium–aluminium oxide–aluminium Josephson junction circuit in a dilution refrigerator at 20 millikelvin temperature was connected to computer controlled measurement electronics in order to apply the current pulses and record the resulting voltage pulses. The resistance of the sample at room temperature suggested that a pulse of 300 nA always causes a switching (response 1), which gave the upper limit for the initial estimation. The lower limit for the initial estimation, 200 nA was roughly estimated from the dimensions of the Josephson junction by an experienced physicist. The experiment was carried out sequentially so that the height of pulse for stage was determined using the measurement data recorded on the earlier stages.

6 Count response

6.1 Representations of count response data

In the count data, the possible values the response variable Y_i are integers $0, 1, 2, \ldots$. In practical situations there is sometimes an upper limit for Y_i but this can be ignored if $\mathrm{E}(Y_i)$ is far below the the upper limit. It is often assumed that the counted events arise from a Poisson process whose intensity depends on the covariates.

When the number of events for each individual are observed, the data can be presented in the form of data matrix

x_1	x_2	y
250	A	7
250	A	1
350	B	1
300	A	0
250	B	8
300	B	3
⋮	⋮	⋮

or weighted data matrix

x_1	x_2	y	frequency
250	A	0	4
250	A	1	2
250	A	3	1
250	B	0	4
250	B	1	5
250	B	2	3
250	B	3	2
300	A	0	6
⋮	⋮	⋮	⋮

When the number of events are observed only for each covariate class, the data can be presented in the form of frequency table (crosstabulation)

	$x_2 = A$	$x_2 = B$
$x_1 = 250$	23	12
$x_1 = 300$	21	19
$x_1 = 350$	7	13

6.2 Link functions for count data

Count data can be modeled by a GLM with the logarithmic link function

$$\log(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}. \tag{77}$$

These models are also called Poisson regression or log-linear models.

6.3 Likelihood

The log-likelihood Poisson distributed response has the form

$$\begin{aligned} l(\mathbf{y}; \boldsymbol{\beta}) = & \sum_{i=1}^{n} y_i \log(\mu_i) - \mu_i - \log(y_i!) = \\ & \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \beta_j x_{ij} y_i - \exp\left(\sum_{j=1}^{p} \beta_j x_{ij} \right) - \log(y_i!) \right). \end{aligned} \tag{78}$$

Example: Full likelihood for missing data Count response y_i, continuous covariate x_{i1} and binary covariate x_{i2} are recorded for the sample $i = 1, 2, \ldots, 1000$. y_i and x_{i1} are observed for all i but x_{i2} is missing for $i = 1, 2, \ldots, 250$. We can assume that the values x_{i2} are missing at random (MAR), i.e. the missingness does not depend on the unobserved value itself but it may depend on the observed response or on the other covariate. The full likelihood for the data can be written

$$\begin{aligned} & L(\mathbf{y}; \beta_1, \beta_2) = \\ & \prod_{i=1}^{250} \Bigg(p(X_{i2} = 1) p(x_{i1} | X_{i2} = 1) \frac{\exp(\beta_1 x_{i1} + \beta_2)^{y_i} \exp(-\exp(\beta_1 x_{i1} + \beta_2))}{y_i!} + \\ & p(X_{i2} = 0) p(x_{i1} | X_{i2} = 0) \frac{\exp(\beta_1 x_{i1})^{y_i} \exp(-\exp(\beta_1 x_{i1}))}{y_i!} \Bigg) \\ & \times \prod_{i=251}^{1000} p(x_{i2}) p(x_{i1} | x_{i2}) \frac{\exp(\beta_1 x_{i1} + \beta_2 x_{i2})^{y_i} \exp(-\exp(\beta_1 x_{i1} + \beta_2 x_{i2}))}{y_i!}. \end{aligned} \tag{79}$$

In order to calculate this likelihood we need to specify the marginal distribution $p(x_2)$ and the conditional distribution $p(x_1|x_2)$. This may require specification of some additional parameters that are considered as nuisance parameters and are estimated together with the parameters of interest β_1 and β_2.

6.4 Offset

The intensity of a process is defined in the form events/time and the observed number of events depends on the system has been followed up. In the GLM, the time is taken into account adding an offset term

$$\log(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij} + \log(t_i), \tag{80}$$

where t_i is the offset, i.e. the follow-up time or the time under exposure. The regression coefficient of the offset is fixed to 1. In R, this can done with the model term `offset`.

6.5 Overdispersion

For the Poisson distribution the variance is equal to expected value. In real-world data, it is common that the observed variance is higher, i.e. there is overdispersion.

6.6 Example: Follow-up for cardiovascular diseases

Cohort studies are important in epidemiological research. A population cohort represents the population of specified age range living in a certain geographical area. The FINRISK cohort 1982 consists of men and women who were 25-64 years old at baseline year 1982 (the beginning of the follow-up). The cohort has been followed up for deaths and fatal and non-fatal events of cardiovascular diseases until the end of year 2006. The example data set contains the number of recorded events grouped by year, age group, sex and area (Eastern Finland / Western Finland). The number of events should be considered in proportion to the person years of follow-up.

7 Nominal and ordinal response

7.1 Representations of nominal response data

Data matrix

x	y
250	A
250	C
350	C
300	A
250	B
300	B
⋮	⋮

Weighted data matrix

x	y	frequency
250	A	23
250	B	12
250	C	4
300	A	21
300	B	19
300	C	7
350	A	7
350	B	13
350	C	3

Frequency table (crosstabulation)

	$Y =' A'$	$Y =' B'$	$Y =' C'$
$x = 250$	23	12	4
$x = 300$	21	19	7
$x = 350$	7	13	3

In nominal response data, the response is one of the categories. In ordinal response data, the categories have a natural order. Binomial response is a special case of both nominal and ordinal response.

When there are more than two categories, nominal and ordinal response have a multivariate nature. For the ith covariate class m_i responses are recorded and the number of responses in the categories $1, 2, \ldots, Q$ is denoted by a vector $\big(K_{i1}K_{i2}\dots K_{iQ}\big)$. The response is then defined as vector

$$\mathbf{Y}_i = \left(\tfrac{K_{i1}}{m_i} \quad \tfrac{K_{i2}}{m_i} \dots \tfrac{K_{iQ}}{m_i}\right). \tag{81}$$

7.2 Multinomial distribution

Multinomial distribution is a generalization of binomial distribution in the case when there are more than two categories

$$f_i(k_{i1}, k_{i2}, \dots, k_{iQ}; m_i, \pi_{i1}, \pi_{i2}, \dots, \pi_{iQ}) = \frac{m_i!}{k_{i1}!k_{i2}!\cdots k_{iQ}!}\pi_{i1}^{k_{i1}}\pi_{i2}^{k_{i2}}\cdots\pi_{iQ}^{k_{iQ}}, \tag{82}$$

where π_{iq}'s are category probabilities for which $\sum_{q=1}^{Q}\pi_{iq} = 1$.

Multinomial distribution does not belong to the exponential family defined in equation (18) but it can derived from the Poisson distribution. Let $K_1, K_2, \dots, K_Q$ be independent Poisson distributed random variables with means $\lambda_1, \lambda_2, \dots, \lambda_Q$. The sum $m = K_1 + K_2 + \dots + K_Q$ is a random variable that follows Poisson distribution with the mean $\lambda = \lambda_1 + \lambda_2 + \dots + \lambda_Q$. Then the conditional distribution

$$\begin{aligned} f(k_1, k_2, \dots, k_Q; m) =& \frac{\prod_{q=1}^{Q}\lambda_q^{k_q}e^{-\lambda_q}}{k_q} \Bigg/ \frac{\lambda^m e^{-\lambda}}{m} = \\ & \left(\frac{\lambda_1}{\lambda}\right)^{k_1}\left(\frac{\lambda_2}{\lambda}\right)^{k_2}\cdots\left(\frac{\lambda_Q}{\lambda}\right)^{k_Q}\frac{m!}{k_1!k_2!\cdots k_Q!} \end{aligned} \tag{83}$$

has the form of the multinomial distribution.

7.3 Regression models for nominal and ordinal response

There are alternative ways to parameterize regression models for nominal and ordinal response. For nominal response data, one of the categories is typically chosen as the reference category and the model has form

$$\log\left(\frac{\pi_{iq}}{\pi_{i1}}\right) = \beta_{0q} + \sum_{j=1}^{p}\beta_{jq}x_{ij}, \tag{84}$$

where the first category serves as reference. For ordinal response data, the model can be written using cumulative probabilities $\gamma_{iq} = \pi_{i1} + \pi_{i2} + \dots + \pi_{iq}$

$$\log\left(\frac{\gamma_{iq}}{1-\gamma_{iq}}\right) = \beta_{0q} + \sum_{j=1}^{p}\beta_{jq}x_{ij}. \tag{85}$$

7.4 Proportional odds model

In proportional odds model it is assumed that the effect of the covariates is the same between all response categories on the logarithmic scale. Model (84)is then written as

$$\log\left(\frac{\pi_{iq}}{\pi_{i1}}\right) = \beta_{0q} + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{86}$$

and model (85) as

$$\log\left(\frac{\gamma_{iq}}{1-\gamma_{iq}}\right) = \beta_{0q} + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{87}$$

7.5 Latent variable interpretation for ordinal regression

In some situations, the actual variable of interest is a continuous response that is difficult or impossible to measure. Instead, an ordinal response variable is measured. It can be assumed that there are unobserved cutpoints that divide the continuous response into categories.

7.6 Nominal and ordinal response data in R

Because of the multivariate nature of the response, function `glm` cannot be directly applied to nominal or ordinal response data in the general case. When the data can be presented in the form a frequency table, log-linear models can be fitted using `glm(...,family=poisson(link=log))`. Function `multinom` from the package `nnet` can be used to fit multinomial log-linear models via neural networks. Function `lopr` from the package `MASS` can be used to fit proportional odds models with logit, probit or cloglog links. Function `vglm` from the package `VGAM` fits a large variety of vector GLMs including multinomial logit models and proportional odds models.

8 Positive response

8.1 Characteristics of positive response data

The response variable is continuous and obtains only positive values. Non-negative response may also obtain value 0. Typically, the distribution of the response is skewed. We may identify three situations:

1. All observed values are positive and “far” from zero.
2. All observed values are positive and some values are relatively close to zero.
3. All observed values are non-negative and a number of them are exactly zero.

The first situation is the easiest to handle whereas the third situation often requires two models, one for the probability of zero response and one for the positive response.

Models for positive response data often assume that the coefficient of variation (CV), the ratio of the the standard deviation to the expectation,

$$CV = \frac{\sqrt{\mathrm{Var}(Y)}}{\mathrm{E}(Y)}. \tag{88}$$

is constant. This is equivalent to assuming that the variance is proportional to square of the expectation

$$\mathrm{Var}(Y_i) \propto \mu_i^2. \tag{89}$$

8.2 Gamma distribution

The gamma distribution is a distribution with a constant coefficient of variation. The gamma distribution has the density function

$$f(y; \lambda, \nu) = \frac{1}{\lambda^\nu \Gamma(\nu)} y^{\nu-1} e^{-y/\lambda}, \quad y > 0, \tag{90}$$

where $\nu > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. Alternative parameterizations exist. Exponential distribution is a special case of the gamma distribution with the shape parameter $\nu = 1$. The gamma distribution belongs to the exponential family (18) with $\theta_i = 1/(\nu\lambda_i)$, $b(\theta) = \log(-\theta)$ and $\phi = 1/\nu$ and has the mean $\mathrm{E}(Y_i) = \nu\lambda_i$ and variance $\mathrm{Var}(Y_i) = \nu\lambda_i^2$. Usually the dispersion parameter ϕ is not known and need to be estimated.

8.3 Link functions for gamma distributed response

The canonical link for the gamma distribution is the inverse link (reciprocal)

$$g(\mu_i) = \frac{1}{\mu_i}. \tag{91}$$

Because $g(\mu_i)$ is always positive, the regression parameters $\boldsymbol{\beta}$ need to be restricted so that the linear predictor is positive. Identity link can be also for the gamma distribution. Restrictions for the regression parameters $\boldsymbol{\beta}$ are needed also the identity link.

Log-link is often used for gamma distributed response. The use of log-link implies multiplicative effect of the covariates. Restrictions for the regression parameters are not needed.

8.4 Lognormal distribution

An alternative to the GLM with gamma distributed response is to take the logarithm of the response and assume that $\log(Y)$ follows normal distribution. Then Y follows log-normal distribution

$$f(y;\mu,\sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right), \quad y > 0, \tag{92}$$

where μ is the mean of $\log(Y)$ and σ^2 is the variance of $\log(Y)$.

8.5 Inverse Gaussian distribution

The inverse Gaussian distribution is another standard distribution for positive responses in GLM. The pdf has form

$$f(y;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(\frac{-\lambda(y-\mu)^2}{2\mu^2 y}\right), \quad y > 0, \tag{93}$$

where $\mu > 0$ is location parameter and $\lambda > 0$ is the shape parameter. The inverse Gaussian distribution belongs to the exponential family (18) and has mean μ and variance μ^3/λ. When λ tends to infinity the pdf of the inverse Gaussian distribution approaches the pdf of the normal distribution.

Link functions used with the inverse Gaussian distribution include identity, log, inverse and inverse squared link,

$$g(\mu_i) = \frac{1}{\mu_i^2}, \tag{94}$$

which is the canonical link.

8.6 Compound Poisson model

In the compound Poisson model, the response is a sum of K independent identically distributed random variables and K follows Poisson distribution. The compound Poisson model can be used, for instance, to model the total amount of claims for an insurance company.

8.7 Weibull distribution

The Weibull distribution has the pdf

$$f(y; \nu, \lambda) = \frac{\nu}{\lambda}\left(\frac{y}{\lambda}\right)^{\nu-1} \exp\left(-\left(\frac{y}{\lambda}\right)^{\nu}\right), \quad y \geq 0, \tag{95}$$

where $\nu > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. The Weibull distribution belongs to the exponential family only if the shape parameter is known. The special case $\nu = 1$ is the exponential distribution.

8.8 Pareto distribution

The Pareto distribution has the pdf

$$f(y; \nu, c) = \frac{\nu c^{\nu}}{y^{\nu+1}}, \quad y \geq c, \tag{96}$$

where $\nu > 0$ is the shape parameter and $c > 0$ is the scale parameter.

9 Time-to-event response

9.1 Representations of time-to-event data

Data matrix – one time variable: the time under follow-up

x	t	d
5.4	10	0
6.0	7.7	1
4.3	10	0
4.4	10	0
5.1	4.5	1
5.9	10	0
⋮	⋮	

Data matrix – two time variables: the start and the end of the follow-up

x	t_1	t_2	d
5.4	41.3	51.3	0
6.0	60.3	68.0	1
4.3	54.4	64.4	0
4.4	62.9	72.9	0
5.1	45.0	49.5	1
5.9	58.7	68.7	0
⋮	⋮		

Analysis of time-to-event data is also known by names survival analysis, lifetime data analysis, failure time analysis, reliability analysis and duration analysis.

9.2 Censoring and truncation

In time-to-event data, the exact event times are not always available for all observations:

Right censoring : It is only known that $T_i > c_i$. c_i is observed.

Left censoring : It is only known that $T_i < c_i$. c_i is observed.

Interval censoring : It is only known that $c_{li} < T_i < c_{ui}$. c_{li} and c_{ui} are observed.

Some observations may be completely missing:

Left truncation : If $T < c$ the observation is not present in the data set. c is known.

Right truncation : If $T > c$ the observation is not present in the data set. c is known.

9.3 Prospective and retrospective studies

In a prospective study, a cohort of subjects is followed up for the future events. In a retrospective study, data on the past events are collected. The study illustrated in Figure 1 is a prospective study with some retrospective characteristics.

9.4 Survival function and hazard function

Consider event time T with the cdf $F(t)$ and the pdf $f(t)$. Survival (or survivor) function gives the probability the time of the event is later than a specified time

$$S(t) = 1 - F(t) = P(T > t). \tag{97}$$

Hazard function is defined as the event rate at time t conditional on survival until time t or later

$$\lambda(t) = \lim_{h \to 0^+} \frac{P(t < T < t + h \mid T \geq t)}{h} = \frac{f(t)}{S(t)}. \tag{98}$$

Integration over time leads to cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(u) du = -\log(S(t)). \tag{99}$$

For example, exponential function with the pdf $f(t) = \lambda e^{-\lambda t}$ and the cdf $F(t) = 1 - e^{-\lambda t}$ has the hazard function $\lambda(t) = \lambda$, i.e. the hazard does not depend on the time.

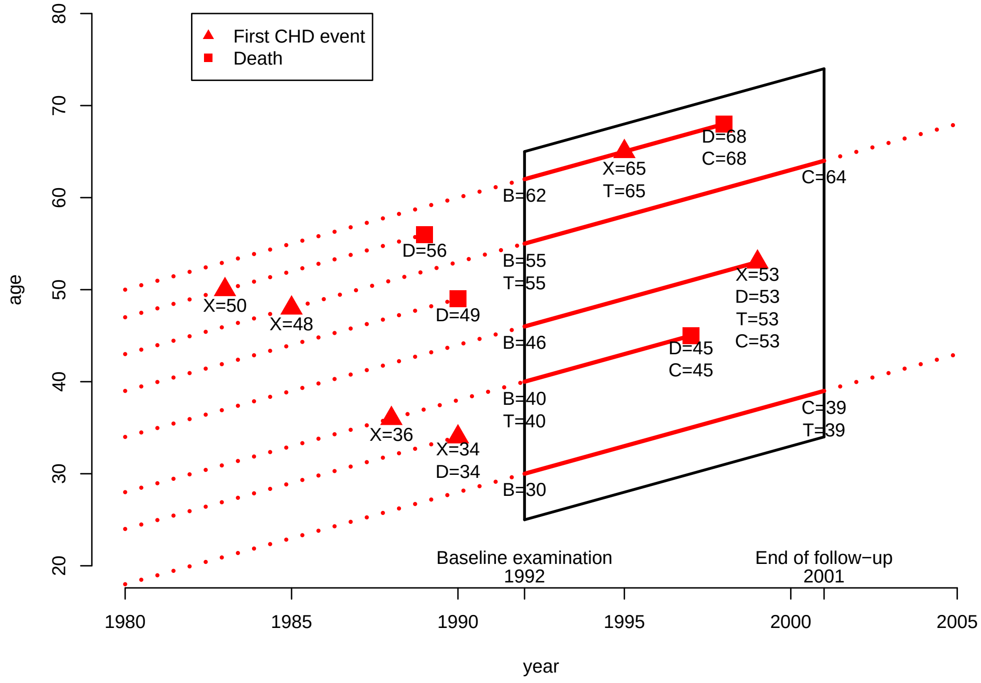

Figure 1: Illustration of a study design leading to left and right censored data with left truncation. The Lexis diagram of a cohort study is displayed. The follow-up period is from the year 1992 to the year 2001 and the age of the subjects is 25–65 years at the baseline examination. The following variables are presented: B = age at baseline examination, X = time of first coronary heart disease (CHD) event, C = censoring time, T = observed time and D = time of death. In the diagram, the data of eight subjects are presented. Two subjects have an event observed during the follow-up ($X = 65$ and $X = 53$). One of the events is fatal ($X = 53$ and $D = 53$) and the other is non-fatal ($X = 65$ and $D = 68$). One subject is right censored ($C = 39$). Two subjects have a left censored event ($X = 48$ and $X = 36$). At the baseline examination, the existence of a left censored event is recorded but the exact time of an event remains unknown. One of the subjects with left censored event dies during the follow-up period ($D = 45$); the other survives up to the end of follow-up ($C = 64$). Three subjects are completely unobserved ($D = 56$, $D = 49$ and $D = 34$). One of them had fatal CHD event ($X = 34$ and $D = 34$), one had a non-fatal event ($X = 50$) and died later ($D = 56$) and one died ($D = 49$) without a preceding CHD event.

9.5 Proportional hazards model

In proportional hazards model, the covariates have a multiplicative effect on the hazard function

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}), \tag{100}$$

where $\lambda_0(t)$ defines the baseline hazard. The baseline hazard function may be defined parametrically, for example, the hazard function of the Weibull distribution is often used. However, the most popular choice is the Cox model where $\lambda_0(t)$ is specified as a function that changes only at observed events times $\{t_i : d_i = 1\}$. The Cox model can be estimated via partial likelihood. The likelihood contribution of an event at time t_i equals

$$\frac{R_i(t_i)\lambda_i(t_i)}{\sum_{j=1}^{N} R_j(t_i)\lambda_j(t_i)}, \tag{101}$$

where $R_i(t)$ is the at-risk indicator. Censored events contribute to the partial likelihood only through their presence in the at-risk set.

In R, Cox models can be fitted with `cph` from the package `Design` or `coxph` from the package `survival` and Weibull models can be fitted with `weibreg` from the package `eha`. The response is defined as a survival object which can be created with the function `Surv`.

10 Extensions and related models

10.1 Beyond exponential family

The standard GLM assumptions on the exponential family, independence and the link function where presented in Section 3. Quasi-likelihood considered in Section 3.8 allows defining the variance function independently from the link function for binomial and count response. The disadvantage is that the quasi-likelihood is not a likelihood and the likelihood based theory does not apply directly. Multinomial response considered in Section 7 does not directly fit to the framework of exponential family. Cox model considered in Section 9.5 is a semi-parametric model where the time-to-event response does not belong to the exponential family.

10.2 Dependent responses

The assumption on the independence of the responses $Y_1, Y_2, \ldots, Y_n$ is unrealistic in many situations. In **longitudinal data**, the measurements are done for the same individuals at several time points. Usually, the measurements of the same individual are dependent. **Repeated measurements** are collected also in various other situations. For example, repeated measurements data are obtained when the diameters of trees are measured at different heights. In **clustered data**, the dependence follows from hierarchical structure of the data. For instance, the children from the same family are more similar than children from different families.

10.2.1 Generalized linear mixed models (GLMM)

Mixed models have both fixed covariate effects and random covariate effects. Random effects are considered as random variables. Often the main interest lies in the fixed effects and the parameters for the random effects are nuisance parameters. In a typical situation with repeated measurements, the random effect term represents all individual characteristics that are not measured.

10.2.2 Generalized estimation equations (GEE)

In the case of independent responses, the estimates $\hat{\boldsymbol{\beta}}$ are the solutions to the score equations

$$U_j = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0, \quad j = 1, \ldots, p. \tag{102}$$

These estimation equations can be generalized for dependent responses. Let $\mathbf{y}_i$ denote the vector of responses for subject i with and let $\mathbf{D}_i$ be the matrix of derivatives $\partial \boldsymbol{\mu}_i / \partial \beta_j$. The estimates $\hat{\boldsymbol{\beta}}$ are iteratively solved from the generalized estimation equations

$$\mathbf{U} = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{103}$$

where the matrix $\mathbf{V}_i$ is the covariance matrix of $\mathbf{Y}_i$.

10.3 Nonlinear covariate effects

10.3.1 Generalized additive models (GAM)

In generalized additive models, the covariates may have nonmonotonic nonlinear effect

$$g(\mu_i) = \sum_{j=1}^{p} f_j(x_{ij}), \tag{104}$$

where the functions f_j are estimated from the data. Typically these function are smooth splines (piecewise polynomials), where the smoothness is controlled by degree of freedom.

10.3.2 Neural networks

Artificial neural networks are highly nonlinear statistical models. The structure of the feedforward neural networks resembles GLM/GAM. In neural network jargon, the covariates are called input and the response is called output. The network consists of the input layer, one or more hidden layers and the output layer. The hidden layer may be defined as

$$v_{ik} = \varphi \left(\sum_{j=1}^{p} \beta_{kj}^{(1)} x_{ij} \right), \quad k = 1, 2, \ldots, q \tag{105}$$

where x_i is the input and φ is an activation function that has a similar role as the inverse link function has in GLMs. The output is then a nonlinear transformation of weighted output of the hidden layer

$$y_{ik} = \varphi\left(\sum_{j=1}^{q} \beta_{kj}^{(2)} v_{ij}\right), \quad k = 1, 2, \ldots, r. \tag{106}$$

In neural networks, the emphasis is usually on prediction, not on interpretation.

10.4 Bayesian estimation of GLM

The likelihood expressions can be applied to both Bayesian and frequentist analysis. The Bayesian inference requires also the priors of the model parameters to be specified. In R functions for Bayesian estimation of GLMs are available in the package `arm` more complicated models can estimated using BUGS (`http://mathstat.helsinki.fi/openbugs/` and `http://www.mrc-bsu.cam.ac.uk/bugs/`) or user-made software (usually a C code).