Generalized Linear Models*

Marlene Müller

Fraunhofer Institute for Industrial Mathematics (ITWM)
P.O. Box 3049, D–67663 Kaiserslautern (Germany)
marlene.mueller@gmx.de

January 6, 2004

1 Introduction

Generalized linear models (GLM) extend the concept of the well understood linear regression model. The linear model assumes that the conditional expectation of Y (the dependent or response variable) is equal to a linear combination $\boldsymbol{X}^\top\boldsymbol{\beta}$, i.e.

$$E(Y|\boldsymbol{X}) = \boldsymbol{X}^\top\boldsymbol{\beta}.$$

This could be equivalently written as $Y = \boldsymbol{X}^\top\boldsymbol{\beta} + \varepsilon$. Unfortunately, the restriction to linearity cannot take into account a variety of practical situations. For example, a continuous distribution of the error ε term implies that the response Y must have a continuous distribution as well. Hence, the linear regression model may fail when dealing with binary Y or with counts.

Example 1 *(Bernoulli responses)*
Let us illustrate a binary response model (Bernoulli Y) using a sample on credit worthiness. For each individual in the sample we know if the granted loan has defaulted or not. The responses are coded as

$$Y = \begin{cases} 1 & \textit{loan defaults}, \\ 0 & \textit{otherwise}. \end{cases}$$

The term of interest is how credit worthiness depends on observable individual characteristics $\boldsymbol{X}$ (age, amount and duration of loan, employment, purpose of loan, etc.). Recall that for a Bernoulli variable $P(Y = 1|\boldsymbol{X}) = E(Y|\boldsymbol{X})$ holds. Hence, the default probability $P(Y = 1|\boldsymbol{X})$ equals a regression of Y on $\boldsymbol{X}$. A useful approach is the following logit model:

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{x}^\top\boldsymbol{\beta})}.$$

Here the function of interest $E(Y|\boldsymbol{X})$ is linked to a linear function of the explanatory variables by the logistic cumulative distribution function (cdf) $F(u) = 1/(1 + e^{-u}) = e^u/(1 + e^u)$. □

The term *generalized linear models* (GLM) goes back to Nelder and Wedderburn (1972) and McCullagh and Nelder (1989) who show that if the distribution of the dependent variable Y is a member of the exponential family, then the class of models which connects the expectation of Y

*Prepared for J. Gentle, W. Härdle, Y. Mori (eds): *Handbook of Computational Statistics (Volume I). Concepts and Fundamentals*, Springer-Verlag, Heidelberg, 2004

to a linear combination of the variables $\boldsymbol{X}^\top\boldsymbol{\beta}$ can be treated in a unified way. In the following sections we denote the function which relates $\mu = E(Y|\boldsymbol{X})$ and $\eta = \boldsymbol{X}^\top\boldsymbol{\beta}$ by $\eta = G(\mu)$ or

$$E(Y|\boldsymbol{X}) = G^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta}).$$

This function G is called *link function*. For all considered distributions of Y there exists at least one canonical link function and typically a set of frequently used link functions.

2 Model Characteristics

The generalized linear model is determined by two components:

- the distribution of Y,
- the link function.

In order to define the GLM methodology as a specific class of nonlinear models (for a general approach to nonlinear regression see Chapter III.8), we assume that the distribution of Y is a member of the *exponential family*. The exponential family covers a large number of distributions, for example discrete distributions as the Bernoulli, binomial and Poisson which can handle binary and count data or continuous distributions as the normal, Gamma or Inverse Gaussian distribution.

2.1 Exponential Family

We say that a distribution is a member of the exponential family if its probability mass function (if Y discrete) or its density function (if Y continuous) has the following form:

$$f(y,\theta,\psi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\psi)} + c(y,\psi)\right\}. \tag{1}$$

The functions $a(\bullet)$, $b(\bullet)$ and $c(\bullet)$ will vary for different Y distributions. Our parameter of interest is θ, which is also called the *canonical parameter* (McCullagh and Nelder, 1989). The additional parameter ψ, that is only relevant for some of the distributions, is considered as a nuisance parameter.

Example 2 *(Normal distribution)*
Suppose Y is normally distributed with $Y \sim N(\mu,\sigma^2)$. The probability density function $f(y) = \exp\{-(y-\mu)^2/(2\sigma^2)\}/(\sqrt{2\pi}\sigma)$ can be written as in (1) by setting $\theta = \mu$ and $\psi = \sigma$ and $a(\psi) = \psi^2$, $b(\theta) = \theta^2/2$, and $c(y,\psi) = -y^2/(2\psi^2) - \log(\sqrt{2\pi}\psi)$. □

Example 3 *(Bernoulli distribution)*
If Y is Bernoulli distributed its probability mass function is

$$P(Y=y) = \mu^y(1-\mu)^{1-y} = \begin{cases} \mu & \text{if } y=1, \\ 1-\mu & \text{if } y=0. \end{cases}$$

This can be transformed into $P(Y=y) = \exp(y\theta)/(1+e^\theta)$ using the logit transformation $\theta = \log\{\mu/(1-\mu)\}$ equivalent to $\mu = e^\theta/(1+e^\theta)$. Thus we obtain an exponential family with $a(\psi) = 1$, $b(\theta) = -\log(1-\mu) = \log(1+e^\theta)$, and $c(y,\psi) = 0$. □

Table 1 lists some probability distributions that are typically used for a GLM. For the binomial and negative binomial distribution the additional parameter k is assumed to be known. Note also that the Bernoulli, geometric and exponential distributions are special cases of the binomial, negative binomial and Gamma distributions, respectively.

Table 1: GLM distributions.

	Range of y	$f(y)$	$\mu(\theta)$	Variance terms $V(\mu)$	$a(\psi)$
Bernoulli $B(\mu)$	$\{0,1\}$	$\mu^y(1-\mu)^{1-y}$	$\frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$	1
Binomial $B(k,\mu)$	$\{0,\ldots,k\}$	$\binom{k}{y}\mu^y(1-\mu)^{k-y}$	$\frac{ke^\theta}{1+e^\theta}$	$\mu\left(1-\frac{\mu}{k}\right)$	1
Poisson $P(\mu)$	$\{0,1,2,\ldots\}$	$\frac{\mu^y}{y!}e^{-\mu}$	$\exp(\theta)$	μ	1
Geometric $Geo(\mu)$	$\{0,1,2,\ldots\}$	$\left(\frac{\mu}{1+\mu}\right)^y\left(\frac{1}{1+\mu}\right)$	$\frac{e^\theta}{1-e^\theta}$	$\mu+\mu^2$	1
Negative Binomial $NB(\mu,k)$	$\{0,1,2,\ldots\}$	$\binom{k+y-1}{y}\left(\frac{\mu}{k+\mu}\right)^y\left(\frac{k}{k+\mu}\right)$	$\frac{ke^\theta}{1-e^\theta}$	$\mu+\frac{\mu^2}{k}$	1
Exponential $Exp(\mu)$	$(0,\infty)$	$\frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right)$	$-1/\theta$	μ^2	1
Gamma $G(\mu,\psi)$	$(0,\infty)$	$\frac{1}{\Gamma(\psi)}\left(\frac{\psi}{\mu}\right)^\psi\exp\left(-\frac{\psi y}{\mu}\right)y^{\psi-1}$	$-1/\theta$	μ^2	$\frac{1}{\psi}$
Normal $N(\mu,\psi^2)$	$(-\infty,\infty)$	$\frac{\exp\left\{-(y-\mu)^2/(2\psi^2)\right\}}{\sqrt{2\pi}\psi}$	θ	1	ψ^2
Inverse Gaussian $IG(\mu,\psi^2)$	$(0,\infty)$	$\frac{\exp\left\{-(y-\mu)^2/(2\mu^2y\psi^2)\right\}}{\sqrt{2\pi y^3}\psi}$	$\frac{1}{\sqrt{-2\theta}}$	μ^3	ψ^2

Table 2: Characteristics of GLMs.

	Canonical link $\theta(\mu)$	Deviance $D(\boldsymbol{y}, \boldsymbol{\mu})$
Bernoulli $B(\mu)$	$\log\left(\frac{\mu}{1-\mu}\right)$	$2\sum\left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\mu_i}\right)\right]$
Binomial $B(k, \mu)$	$\log\left(\frac{\mu}{k-\mu}\right)$	$2\sum\left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (k-y_i)\log\left(\frac{k-y_i}{k-\mu_i}\right)\right]$
Poisson $P(\mu)$	$\log(\mu)$	$2\sum\left[y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i)\right]$
Geometric $Geo(\mu)$	$\log\left(\frac{\mu}{1+\mu}\right)$	$2\sum\left[y_i \log\left(\frac{y_i + y_i\mu_i}{\mu_i + y_i\mu_i}\right) - \log\left(\frac{1+y_i}{1+\mu_i}\right)\right]$
Negative Binomial $NB(\mu, k)$	$\log\left(\frac{\mu}{k+\mu}\right)$	$2\sum\left[y_i \log\left(\frac{y_i k + y_i\mu_i}{\mu_i k + y_i\mu_i}\right) - k\log\left\{\frac{k(k+y_i)}{k(k+\mu_i)}\right\}\right]$
Exponential $Exp(\mu)$	$\frac{1}{\mu}$	$2\sum\left[\frac{y_i - \mu_i}{\mu_i} - \log\left(\frac{y_i}{\mu_i}\right)\right]$
Gamma $G(\mu, \psi)$	$\frac{1}{\mu}$	$2\sum\left[\frac{y_i - \mu_i}{\mu_i} - \log\left(\frac{y_i}{\mu_i}\right)\right]$
Normal $N(\mu, \psi^2)$	μ	$2\sum\left[(y_i - \mu_i)^2\right]$
Inverse Gaussian $IG(\mu, \psi^2)$	$\frac{1}{\mu^2}$	$2\sum\left[\frac{(y_i - \mu_i)^2}{y_i \mu_i^2}\right]$

2.2 Link Function

After having specified the distribution of Y, the link function G is the second component to choose for the GLM. Recall the model notation $\eta = \boldsymbol{X}^\top \boldsymbol{\beta} = G(\mu)$. In the case that the canonical parameter θ equals the linear predictor η, i.e. if

$$\eta = \theta,$$

the link function is called the *canonical link* function. For models with a canonical link the estimation algorithm simplifies as we will see in Subsection 3.3. Table 2 shows in its second column the canonical link functions of the exponential family distributions presented in Table 1.

Example 4 *(Canonical link for Bernoulli Y)*
For Bernoulli Y we have $\mu = e^\theta/(1+e^\theta)$, hence the canonical link is given by the logit transformation $\eta = \log\{\mu/(1-\mu)\}$. □

What link functions could we choose apart from the canonical? For most of the models exists a number of specific link functions. For Bernoulli Y, for example, any smooth cdf can be used. Typical links are the logistic and standard normal (Gaussian) cdfs which lead to logit and *probit* models, respectively. A further alternative for Bernoulli Y is the complementary log–log link $\eta = \log\{-\log(1-\mu)\}$.

A flexible class of link functions for positive Y observations is the class of power functions. These links are given by the Box-Cox transformation (Box and Cox, 1964), i.e. by $\eta = (\mu^\lambda - 1)/\lambda$ or $\eta = \mu^\lambda$ where we set in both cases $\eta = \log(\mu)$ for $\lambda = 0$.

3 Estimation

Recall that the least squares estimator for the ordinary linear regression model is also the maximum-likelihood estimator in the case of normally distributed error terms. By assuming that the distribution of Y belongs to the exponential family it is possible to derive maximum-likelihood estimates for the coefficients of a GLM. Moreover we will see that even though the estimation needs a numerical approximation, each step of the iteration can be given by a weighted least squares fit. Since the weights are varying during the iteration the likelihood is optimized by an *iteratively reweighted least squares* algorithm.

3.1 Properties of the Exponential Family

To derive the details of the maximum-likelihood algorithm we need to discuss some properties of the probability mass or density function $f(\bullet)$. For the sake of brevity we consider f to be a density function in the following derivation. However, the conclusions will hold for a probability mass function as well.

First, we start from the fact that $\int f(y,\theta,\psi)\,dy = 1$. Under suitable regularity conditions (it is possible to exchange differentiation and integration) this implies

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int f(y,\theta,\psi)\,dy = \int \frac{\partial}{\partial\theta} f(y,\theta,\psi)\,dy \\ &= \int \left\{\frac{\partial}{\partial\theta}\log f(y,\theta,\psi)\right\} f(y,\theta,\psi)\,dy = E\left\{\frac{\partial}{\partial\theta}\ell(y,\theta,\psi)\right\}, \end{aligned}$$

where $\ell(y,\theta,\psi) = \log f(y,\theta,\psi)$ denotes the *log-likelihood* function. The function derivative of ℓ with respect to θ is typically called the *score* function for which it is known that

$$E\left\{\frac{\partial^2}{\partial\theta^2}\ell(y,\theta,\psi)\right\} = -E\left\{\frac{\partial}{\partial\theta}\ell(y,\theta,\psi)\right\}^2.$$

This and taking first and second derivatives of (1) results in

$$0 = E\left\{\frac{Y - b'(\theta)}{a(\psi)}\right\}, \quad \text{and} \quad E\left\{\frac{-b''(\theta)}{a(\psi)}\right\} = -E\left\{\frac{Y - b'(\theta)}{a(\psi)}\right\}^2,$$

such that we can conclude

$$\begin{aligned} E(Y) &= \mu = b'(\theta), && (2) \\ \text{Var}(Y) &= V(\mu)a(\psi) = b''(\theta)a(\psi). && (3) \end{aligned}$$

Note that as a consequence from (1) the expectation of Y depends only on the parameter of interest θ. We also assume that the factor $a(\psi)$ is identical over all observations.

3.2 Maximum-Likelihood and Deviance Minimization

As pointed out before the estimation method of choice for $\boldsymbol{\beta}$ is maximum-likelihood. As an alternative the literature refers to the minimization of the *deviance*. We will see during the following derivation that both approaches are identical.

Suppose that we have observed a sample of independent pairs $(Y_i, \boldsymbol{X}_i)$ where $i = 1, \ldots, n$. For a more compact notation denote now the vector of all response observations by $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ and their conditional expectations (given $\boldsymbol{X}_i$) by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$. Recall that we study

$$E(Y_i|\boldsymbol{X}_i) = \mu_i = G(\boldsymbol{X}_i^\top \boldsymbol{\beta}) = G(\eta_i).$$

The sample log-likelihood of the vector $\boldsymbol{Y}$ is then given by

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} \ell(Y_i, \theta_i, \psi). \quad (4)$$

Here θ_i is a function of $\eta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$ and we use $\ell(Y_i, \theta_i, \psi) = \log f(Y_i, \theta_i, \psi)$ to denote the individual log-likelihood contributions for all observations i.

Example 5 *(Normal log-likelihood)*
For normal responses $Y_i \sim N(\mu_i, \sigma^2)$ we have $\ell(Y_i, \theta_i, \psi) = -(Y_i - \mu_i)^2/(2\sigma^2) - \log\left(\sqrt{2\pi}\sigma\right)$. This gives the sample log-likelihood

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \sigma) = n \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_i)^2. \quad (5)$$

Obviously, maximizing this log-likelihood is equivalent to minimizing the least squares criterion. □

Example 6 *(Bernoulli log-likelihood)*
The calculation in Example 3 shows that the individual log-likelihoods for the binary responses equal $\ell(Y_i, \theta_i, \psi) = Y_i \log(\mu_i) + (1 - Y_i)\log(1 - \mu_i)$. This leads to

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} \{Y_i \log(\mu_i) + (1 - Y_i)\log(1 - \mu_i)\} \quad (6)$$

for the sample version. □

The deviance defines an alternative objective function for optimization. Let us first introduce the *scaled deviance* which is defined as

$$D(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = 2\left\{\ell(\boldsymbol{Y}, \boldsymbol{\mu}^{max}, \psi) - \ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi)\right\}. \quad (7)$$

Here $\boldsymbol{\mu}^{max}$ (which typically equals $\boldsymbol{Y}$) is the vector that maximizes the saturated model, i.e. the function $\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi)$ without imposing any restriction on $\boldsymbol{\mu}$. Since the term $\ell(\boldsymbol{Y}, \boldsymbol{\mu}^{max}, \psi)$ does not depend on the parameter $\boldsymbol{\beta}$ we see that indeed the minimization of the scaled deviance is equivalent to the maximization of the sample log-likelihood (4).

If we now plug-in the exponential family form (1) into (4) we obtain

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\psi)} - c(Y_i, \psi) \right\}. \tag{8}$$

Obviously, neither $a(\psi)$ nor $c(Y_i, \psi)$ depend on the unknown parameter vector $\boldsymbol{\beta}$. Therefore, it is sufficient to consider

$$\sum_{i=1}^{n} \{Y_i \theta_i - b(\theta_i)\} \tag{9}$$

for the maximization. The deviance analog of (9) is the (non-scaled) deviance function

$$D(\boldsymbol{Y}, \boldsymbol{\mu}) = D(\boldsymbol{Y}, \boldsymbol{\mu}, \psi)\, a(\psi). \tag{10}$$

The (non-scaled) deviance $D(\boldsymbol{Y}, \boldsymbol{\mu})$ can be seen as the GLM equivalent of the *residual sum of squares* (RSS) in linear regression as it compares the log-likelihood ℓ for the "model" μ with the maximal achievable value of ℓ.

3.3 Iteratively Reweighted Least Squares Algorithm

We will now minimize the deviance with respect to $\boldsymbol{\beta}$. If we denote the gradient of (10) by

$$\nabla(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \left[-2 \sum_{i=1}^{n} \{Y_i \theta_i - b(\theta_i)\} \right] = -2 \sum_{i=1}^{n} \{Y_i - b'(\theta_i)\} \frac{\partial}{\partial \boldsymbol{\beta}} \theta_i \,, \tag{11}$$

our optimization problem consists in solving

$$\nabla(\boldsymbol{\beta}) = 0. \tag{12}$$

Note that this is (in general) a nonlinear system of equations in $\boldsymbol{\beta}$ and an iterative solution has to be computed. The smoothness of the link function allows us to compute the *Hessian* of $D(\boldsymbol{Y}, \boldsymbol{\mu})$, which we denote by $\mathcal{H}(\boldsymbol{\beta})$. Now a *Newton–Raphson* algorithm can be applied which determines the optimal $\widehat{\boldsymbol{\beta}}$ using the following iteration steps:

$$\widehat{\boldsymbol{\beta}}^{new} = \widehat{\boldsymbol{\beta}}^{old} - \left\{ \mathcal{H}\left(\widehat{\boldsymbol{\beta}}^{old}\right) \right\}^{-1} \nabla\left(\widehat{\boldsymbol{\beta}}^{old}\right).$$

A variant of the Newton–Raphson is the *Fisher scoring* algorithm that replaces the Hessian by its expectation with respect to the observations Y_i:

$$\widehat{\boldsymbol{\beta}}^{new} = \widehat{\boldsymbol{\beta}}^{old} - \left\{ E\mathcal{H}\left(\widehat{\boldsymbol{\beta}}^{old}\right) \right\}^{-1} \nabla\left(\widehat{\boldsymbol{\beta}}^{old}\right).$$

To find simpler representations for these iterations, recall that we have $\mu_i = G(\eta_i) = G(\boldsymbol{X}_i^\top \boldsymbol{\beta}) = b'(\theta_i)$. By taking the derivative of the right hand term with respect to $\boldsymbol{\beta}$ this implies

$$b'(\theta_i) \frac{\partial}{\partial \boldsymbol{\beta}} \theta_i = G(\boldsymbol{X}_i^\top \boldsymbol{\beta})\, \boldsymbol{X}_i.$$

Using that $b''(\theta_i) = V(\mu_i)$ as established in (3) and taking derivatives again, we finally obtain

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} \theta_i &= \frac{G'(\eta_i)}{V(\mu_i)} \boldsymbol{X}_i \\ \frac{\partial^2}{\partial \boldsymbol{\beta} \boldsymbol{\beta}^\top} \theta_i &= \frac{G''(\eta_i) V(\mu_i) - G'(\eta_i)^2 V'(\mu_i)}{V(\mu_i)^2} \boldsymbol{X}_i \boldsymbol{X}_i^\top. \end{aligned}$$

From this we can express the gradient and the Hessian of the deviance by

$$\begin{aligned}\nabla(\boldsymbol{\beta}) &= -2\sum_{i=1}^{n}\{Y_i-\mu_i\}\frac{G'(\eta_i)}{V(\mu_i)}\boldsymbol{X}_i\\ \mathcal{H}(\boldsymbol{\beta}) &= 2\sum_{i=1}^{n}\left\{\frac{G'(\eta_i)^2}{V(\mu_i)}-\{Y_i-\mu_i\}\frac{G''(\eta_i)V(\mu_i)-G'(\eta_i)^2V'(\mu_i)}{V(\mu_i)^2}\right\}\boldsymbol{X}_i\boldsymbol{X}_i^\top.\end{aligned}$$

The expectation of $\mathcal{H}(\boldsymbol{\beta})$ in the Fisher scoring algorithm equals

$$E\mathcal{H}(\boldsymbol{\beta}) = 2\sum_{i=1}^{n}\left\{\frac{G'(\eta_i)^2}{V(\mu_i)}\right\}\boldsymbol{X}_i\boldsymbol{X}_i^\top.$$

Let us consider only the Fisher scoring algorithm for the moment. We define the weight matrix

$$\mathbf{W} = \operatorname{diag}\left(\frac{G'(\eta_1)^2}{V(\mu_1)},\ldots,\frac{G'(\eta_n)^2}{V(\mu_n)}\right)$$

and the vectors $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_1,\ldots,\widetilde{Y}_n)^\top$, $\boldsymbol{Z} = (Z_1,\ldots,Z_n)^\top$ by

$$\widetilde{Y}_i = \frac{Y_i-\mu_i}{G'(\eta_i)}, \quad Z_i = \eta_i + \widetilde{Y}_i = \boldsymbol{X}_i^\top\boldsymbol{\beta}^{old} + \frac{Y_i-\mu_i}{G'(\eta_i)}.$$

Denote further by $\mathbf{X}$ the design matrix given by the rows $x_i^\top$. Then, the Fisher scoring iteration step for $\boldsymbol{\beta}$ can be rewritten as

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} + (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\widetilde{\boldsymbol{Y}} = (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\boldsymbol{Z}. \tag{13}$$

This immediately shows that each Fisher scoring iteration step is the result of a weighted least squares regression of the *adjusted dependent variables* Z_i on the explanatory variables $\boldsymbol{X}_i$. Since the weights are recalculated in each step we speak of the *iteratively reweighted least squares* (IRLS) algorithm. For the Newton–Raphson algorithm a representation equivalent to (13) can be found, only the weight matrix $\mathbf{W}$ differs.

The iteration will be stopped when the parameter estimate and/or the deviance do not change significantly anymore. We denote the final parameter estimate by $\widehat{\boldsymbol{\beta}}$.

3.4 Remarks on the Algorithm

Let us first note two special cases for the algorithm:

- In the linear regression model, where we have $G' \equiv 1$ and $\mu_i = \eta_i = \boldsymbol{X}_i^\top\boldsymbol{\beta}$, no iteration is necessary. Here the ordinary least squares estimator gives the explicit solution of (12).
- In the case of a canonical link function we have $b'(\theta_i) = G(\theta_i) = G(\eta_i)$ and hence $b''(\theta_i) = G'(\eta_i) = V(\mu_i)$. Therefore the Newton–Raphson and the Fisher scoring algorithms coincide.

There are several further remarks on the algorithm which concern in particular starting values and the computation of relevant terms for the statistical analysis:

- Equation (13) implies that in fact we do not need a starting value for $\boldsymbol{\beta}$. Indeed the adjusted dependent variables Z_i can be equivalently initialized by using appropriate values for $\eta_{i,0}$ and $\mu_{i,0}$. Typically, the following initialization is used (McCullagh and Nelder, 1989):
  - ⋆ For all but binomial models set $\mu_{i,0} = Y_i$ and $\eta_{i,0} = G(\mu_{i,0})$.

⋆ For binomial models set $\mu_{i,0} = (Y_i + \frac{1}{2})/(k+1)$ and $\eta_{i,0} = G(\mu_{i,0})$. (Recall that this holds with $k = 1$ in the Bernoulli case.)

The latter definition is based on the observation that G can not be applied to binary data. Therefore a kind of smoothing is used to obtain $\mu_{i,0}$ in the binomial case.

- During the iteration the convergence can be controlled by checking the relative change in the coefficients

$$\sqrt{\frac{(\boldsymbol{\beta}^{new} - \boldsymbol{\beta}^{old})^\top (\boldsymbol{\beta}^{new} - \boldsymbol{\beta}^{old})}{\boldsymbol{\beta}^{old\top} \boldsymbol{\beta}^{old}}} < \epsilon$$

and/or the relative change in the deviance

$$\left| \frac{D(\boldsymbol{Y}, \boldsymbol{\mu}^{new}) - D(\boldsymbol{Y}, \boldsymbol{\mu}^{old})}{D(\boldsymbol{Y}, \boldsymbol{\mu}^{old})} \right| < \epsilon \,.$$

- An estimate $\widehat{\psi}$ for the dispersion parameter ψ can be obtained from either the Pearson χ^2 statistic

$$\widehat{a}(\psi) = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{14}$$

or using deviance

$$\widehat{a}(\psi) = \frac{D(\boldsymbol{Y}, \boldsymbol{\mu})}{n-p} \,. \tag{15}$$

Here we use p for the number of estimated parameters and $\widehat{\mu}_i$ for the estimated regression function at the ith observation. Similarly, $\widehat{\boldsymbol{\mu}}$ is the estimated $\boldsymbol{\mu}$. Both estimators for $a(\psi)$ coincide for normal linear regression and follow an exact χ^2_{n-p} distribution then. The number $n-p$ (number of observations minus number of estimated parameters) is denoted as the *degrees of freedom* of the deviance.

- Typically, software for GLM allows for offsets and weights in the model. For details on the inclusion of weights we refer to Subsection 5.1. Offsets are deterministic components of η which can vary over the observations i. The model that is then fitted is

$$E(Y_i | \boldsymbol{X}_i) = G(\boldsymbol{X}_i^\top \boldsymbol{\beta} + o_i).$$

Offsets may be used to fit a model where a part of the coefficients is known. The iteration algorithm stays unchanged except for the fact that the optimization is only necessary with respect to the remaining unknown coefficients.

- Since the variance of Y_i will usually depend on $\boldsymbol{X}_i$ we cannot simply analyze residuals of the form $Y_i - \widehat{\mu}_i$. Instead, appropriate transformations have to be used. Classical proposals are Pearson residuals

$$r_i^P = \frac{Y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}} \,,$$

deviance residuals

$$r_i^D = \text{sign}(Y_i - \widehat{\mu}_i) \sqrt{d_i} \,,$$

where d_i is the contribution of the ith observation to the deviance, and Anscombe residuals

$$r_i^A = \frac{A(Y_i) - A(\widehat{\mu}_i)}{A'(\widehat{\mu}_i)\sqrt{V(\widehat{\mu}_i)}} \,,$$

where $A(\mu) = \int^\mu V^{-1/3}(u)\, du$.

3.5 Model Inference

The resulting estimator $\widehat{\boldsymbol{\beta}}$ has an asymptotic normal distribution (except of course for the normal linear regression case when this is an exact normal distribution).

Theorem 1
Under regularity conditions we have for the estimated coefficient vector

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(0, \boldsymbol{\Sigma}) \quad \text{as } n \to \infty\,.$$

As a consequence for the scaled deviance and the log-likelihood approximately hold $D(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}, \psi) \sim \chi^2_{n-p}$ *and* $2\{\ell(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}, \psi) - \ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi)\} \sim \chi^2_p$. □

For details on the necessary conditions see for example Fahrmeir and Kaufmann (1984). Note also that the asymptotic covariance $\boldsymbol{\Sigma}$ for the coefficient estimator $\widehat{\boldsymbol{\beta}}$ is the inverse of the Fisher information matrix, i.e.

$$\mathbf{I} = -E\left\{\frac{\partial^2}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T}\ell(Y, \mu, \psi)\right\}.$$

Since $\mathbf{I}$ can be estimated by the negative Hessian of the log-likelihood or its expectation, this suggests the estimator

$$\widehat{\boldsymbol{\Sigma}} = a(\widehat{\psi})\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{G'(\eta_{i,last})^2}{V(\mu_{i,last})}\right\}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]^{-1}.$$

Using the estimated covariance we are able to test hypotheses about the components of $\boldsymbol{\beta}$.

For model choice between two nested models a likelihood ratio test (LR test) is used. Assume that $\mathcal{M}_0$ (p_0 parameters) is a submodel of the model $\mathcal{M}$ (p parameters) and that we have estimated them as $\widehat{\boldsymbol{\mu}}_0$ and $\widehat{\boldsymbol{\mu}}$. For one-parameter exponential families (without a nuisance parameter ψ) we use that asymptotically

$$D(\boldsymbol{Y}, \boldsymbol{\mu}_0) - D(\boldsymbol{Y}, \boldsymbol{\mu}) \sim \chi^2_{p-p_0}. \tag{16}$$

The left hand side of (16) is a function of the ratio of the two likelihoods deviance difference equals minus twice the log-likelihood difference. In a two-parameter exponential family (ψ is to be estimated) one can approximate the likelihood ratio test statistic by

$$\frac{(n-p)\{D(\boldsymbol{Y}, \boldsymbol{\mu}_0) - D(\boldsymbol{Y}, \boldsymbol{\mu})\}}{(p-p_0)D(\boldsymbol{Y}, \boldsymbol{\mu})} \sim F_{p-p_0, n-p} \tag{17}$$

using the analog to the normal linear regression case (Venables and Ripley, 2002, Chapter 7).

Model selection procedures for possibly non-nested models can be based on Akaike's information criterion (Akaike, 1973)

$$AIC = D(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}, \widehat{\psi}) + 2p,$$

or Schwarz' Bayes information criterion (Schwarz, 1978)

$$BIC = D(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}, \widehat{\psi}) + \log(n)p,$$

where again p denotes the number of estimated parameters. For a general overview on model selection techniques see also Chapter III.1 of this handbook.

4 Practical Aspects

To illustrate the GLM in practice we recall Example 1 on credit worthiness. The credit data set that we use (Fahrmeir and Tutz, 1994) contains $n = 1000$ observations on consumer credits and a variety of explanatory variables. We have selected a subset of eight explanatory variables for the following examples.

The model for credit worthiness is based on the idea that default can be predicted from the individual and loan characteristics. We consider criteria as age, information on previous loans, savings, employment and house ownership to characterize the credit applicants. Amount and duration of the loan are prominent features of the granted loans. Some descriptive statistics can be found in Table 3. We remark that we have categorized the durations (months) into intervals since most of the realizations are multiples of 3 or 6 months.

Table 3: Credit data.

Variable	Yes	No	(in %)	
Y (observed default)	30.0	70.0		
PREVIOUS (no problem)	38.1	61.9		
EMPLOYED (≥ 1 year)	93.8	6.2		
DURATION $(9, 12]$	21.6	78.4		
DURATION $(12, 18]$	18.7	81.3		
DURATION $(18, 24]$	22.4	77.6		
DURATION ≥ 24	23.0	77.0		
SAVINGS	18.3	81.7		
PURPOSE (buy a car)	28.4	71.6		
HOUSE (owner)	15.4	84.6		
	Min.	Max.	Mean	Std.Dev.
AMOUNT (in DM)	250	18424	3271.248	2822.752
AGE (in years)	19	75	35.542	11.353

We are at the first place interested in estimating the probability of credit default in dependence of the explanatory variables $\boldsymbol{X}$. Recall that for binary Y it holds $P(Y = 1|\boldsymbol{X}) = E(Y|\boldsymbol{X})$. Our first approach is a GLM with logit link such that $P(Y = 1|\boldsymbol{X}) = \exp(\boldsymbol{X}^\top\boldsymbol{\beta})/\{1 + \exp(\boldsymbol{X}^\top\boldsymbol{\beta})\}$.

Example 7 *(Credit default on AGE)*
We initially estimate the default probability solely related to age, i.e. the model

$$P(Y = 1|AGE) = \frac{\exp(\beta_0 + \beta_1 AGE)}{1 + \exp(\beta_0 + \beta_1 AGE)}$$

or equivalently $\text{logit}\{P(Y = 1|AGE)\} = \beta_0 + \beta_1 AGE$. *The resulting estimates of the constant* β_0 *and the slope parameter* β_1 *are displayed in Table 4 together with summary statistics on the model fit.*

From the table we see that the estimated coefficient of AGE has a negative sign. Since the link function and its inverse are strictly monotone increasing, we can conclude that the probability of default must thus be decreasing with increasing AGE. Figure 1 shows on the left frequency barplots of AGE separately for $Y = 1$ *and* $Y = 0$. *From the observed frequencies we can recognize clearly the decreasing propensity to default. The right graph in Figure 1 displays the estimated probabilities* $P(Y = 1|AGE)$ *using the fitted logit model which are indeed decreasing.*

The t-values $(\sqrt{n}\,\widehat{\beta}_j/\sqrt{\widehat{\Sigma}_{jj}})$ *show that the coefficient of AGE is significantly different from* 0 *while the estimated constant is not. The test that is used here is an approximative t-test such*

Table 4: Credit default on AGE (logit model).

Variable	Coefficient	t-value
constant	-0.1985	-0.851
AGE	-0.0185	-2.873
Deviance	1213.1	
df	998	
AIC	1217.1	
Iterations	4	

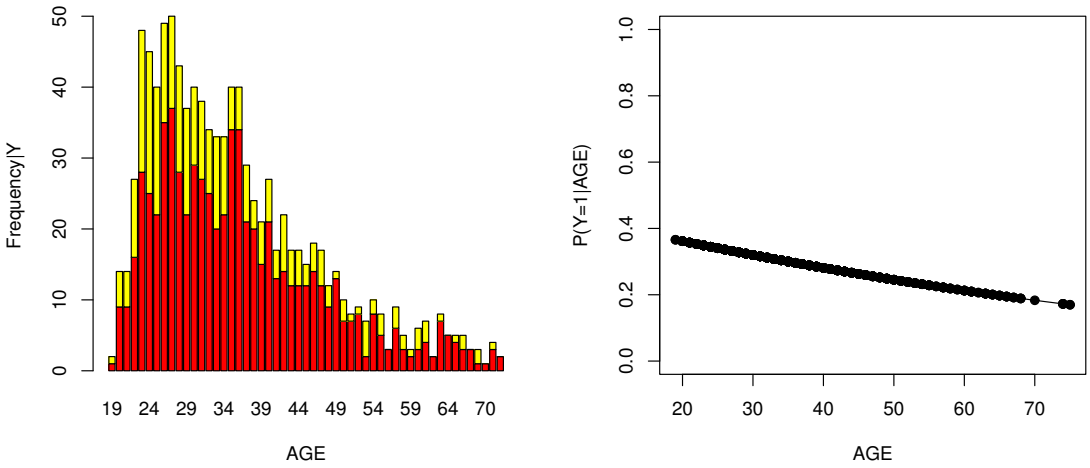

Figure 1: Credit default on AGE, left: frequency barplots of AGE for $Y = 1$ (yellow) and $Y = 0$ (red), right: estimated probabilities.

that $z_{1-\alpha/2}$-quantile of the standard normal can be used as critical value. This implies that at the usual 5% level we compare the absolute value of the t-value with $z_{0.975} \approx 1.96$.

A more general approach to test for the significance of AGE is to compare the fitted model with a model that involves only a constant default probability. Typically software packages report the deviance of this model as null deviance or similar. In our case we find a null deviance of 1221.7 at 999 degrees of freedom. If we apply the LR test statistic (16) to compare the null deviance to the model deviance of 1213.1 at 998 degrees of freedom, we find that constant model is clearly rejected at a significance level of 0.33%. □

Models using different link functions cannot be directly compared as the link functions might be differently scaled. In our binary response model for example a logit or a probit link function may be reasonable. However, the variance parameter of the standard logistic distribution is $\pi^2/3$ whereas that of the standard normal is 1. We therefore need to rescale one of the link functions in order to compare the resulting model fits. Figure 2 shows the standard logistic cdf (the inverse logit link) against the cdf of $N(0, \pi^2/3)$. The functions in the left graph of Figure 2 are hardly distinguishable. If we zoom in (right graph) we see that the logistic cdf vanishes to zero at the left boundary at a lower rate. This holds similarly for the right boundary and explains the ability of logit models to (slightly) better handle the case of extremal observations.

Example 8 *(Probit versus logit)*
If we want to compare the estimated coefficients from a probit to that of the logit model we need to rescale the probit coefficients by $\pi/\sqrt{3}$. Table 5 shows the results of a probit for credit default

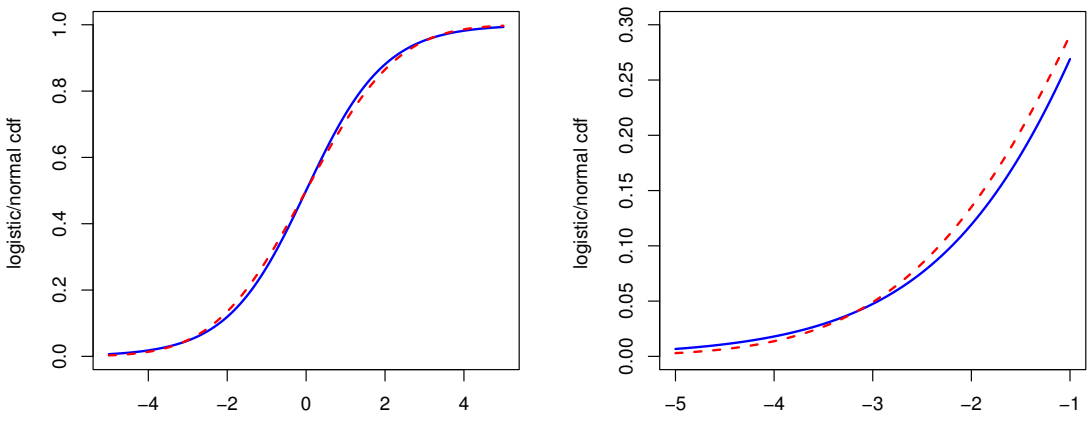

Figure 2: Logit (solid blue) versus appropriately rescaled probit link (dashed red), left: on the range $[-5, 5]$, right: on the range of $[-5, -1]$.

on AGE. The resulting rescaled coefficient for AGE in is of similar size as that for the logit model (cf. Table 4) while the constant is not significantly different from 0 in both fits. The deviance and the AIC of the probit fit are slightly larger.

A Newton–Raphson iteration (instead of the Fisher scoring reported in Table 5) does give somewhat different coefficients but returns nearly the same value of the deviance (1213.268 for Newton–Raphson versus 1213.265 for Fisher scoring). □

Table 5: Credit default on AGE (probit model), original and rescaled coefficients for comparison with logit.

Variable	Coefficient		t-value
	(original)	(rescaled)	
constant	-0.1424	-0.2583	-1.022
AGE	-0.0109	-0.0197	-2.855
Deviance	1213.3		
df	998		
AIC	1217.3		
Iterations	4	(Fisher Scoring)	

The next two examples intend to analyze if the fit could be improved by using a nonlinear function on AGE instead of $\eta = \beta_0 + \beta_1 \text{AGE}$. Two principally different approaches are possible:

- include higher order terms of AGE into η,
- categorize AGE in order to fit a stepwise constant η function.

Example 9 *(Credit default on polynomial AGE)*
We fit two logit models using second and third order terms in AGE. The estimated coefficients are presented in Table 6. A comparison of the quadratic fit and the linear fit from Example 7 using the LR test statistic (16) shows that the linear fit is rejected at a significance level of 3%*. A subsequent comparison of the quadratic against the cubic fit no significant improvement by the latter model. Thus, the quadratic term for AGE improves the fit whereas the cubic term*

does not show any further statistically significant improvement. This result is confirmed when we compare the AIC values of both models which are practically identical. Figure 3 shows the estimated default probabilities for the quadratic (left) and cubic AGE fits. We find that the curves are of similar shape. □

Table 6: Credit default on polynomial AGE (logit model).

Variable	Coefficient	t-value	Coefficient	t-value
constant	1.2430	1.799	0.4092	1.909
AGE	-0.0966	-2.699	-0.3240	-1.949
AGE**2	$9.56 \cdot 10^{-4}$	2.234	$6.58 \cdot 10^{-3}$	1.624
AGE**3	–	–	$-4.33 \cdot 10^{-5}$	-1.390
Deviance	1208.3		1206.3	
df	997		996	
AIC	1214.3		1214.3	
Iterations	4		4	

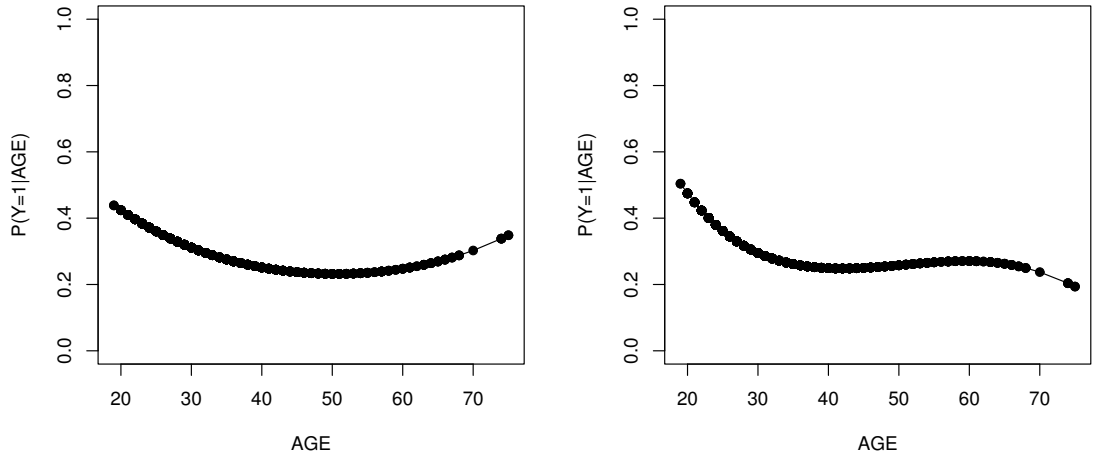

Figure 3: Credit default on polynomial AGE, left: estimated probabilities from quadratic function, right: estimated probabilities from cubic function.

To incorporate a possible nonlinear impact of a variable in the index function, we can alternatively categorize this variable. Another term for this is the construction of *dummy* variables. The most classical form of the categorization consists in using a design matrix that sets a value of 1 in the column corresponding to the category if the category is true and 0 otherwise. To obtain a full rank design matrix we omit one column for the reference category. In our example we leave out the first category which means that all estimated coefficients have to be compared to the zero coefficient of the reference category. Alternative categorization setups are given by omitting the constant, the sum coding (restrict the coefficients to sum up to 0), and the Helmert coding.

Example 10 *(Credit default on categorized AGE)*
We have chosen the intervals $(18, 23]$, $(23, 28]$, $\ldots$, $(68, 75]$ *as categories. Except for the last interval all of them are of the same length. The first interval* $(18, 23]$ *is chosen for the reference such that we will estimate coefficients only for the remaining 10 intervals.*

Frequency barplots for the intervals and estimated default probabilities are displayed in Figure 4. The resulting coefficients for this model are listed in Table 7. We see here that all coefficient estimates are negative. This means, keeping in mind that the group of youngest credit applicants is the reference, that all applicants from other age groups have an (estimated) lower default probability. However, we do not have a true decrease in the default probabilities with AGE since the coefficients do not form a decreasing sequence. In the range from age 33 to 63 we find two local minima and maxima for the estimated default probabilities.

Table 7: Credit default on categorized AGE (logit model).

Variable	Coefficients	t-values
constant	-0.4055	-2.036
AGE (23,28]	-0.2029	-0.836
AGE (28,33]	-0.3292	-1.294
AGE (33,38]	-0.9144	-3.320
AGE (38,43]	-0.5447	-1.842
AGE (43,48]	-0.6763	-2.072
AGE (48,53]	-0.8076	-2.035
AGE (53,58]	-0.5108	-1.206
AGE (58,63]	-0.4055	-0.864
AGE (63,68]	-0.7577	-1.379
AGE (68,75]	-1.3863	-1.263
Deviance	1203.2	
df	989	
AIC	1225.2	
Iterations	4	

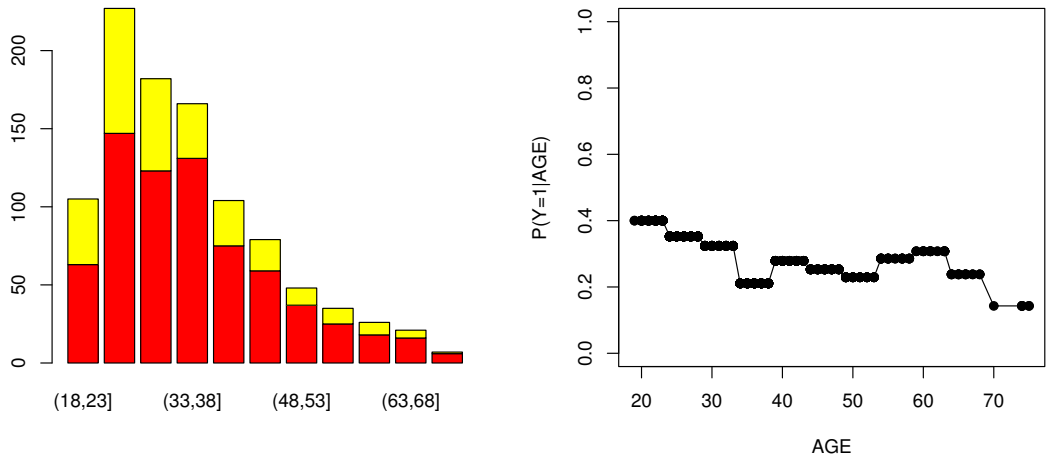

Figure 4: Credit default on categorized AGE, left: frequency barplots of categorized AGE for $Y = 1$ (yellow) and $Y = 0$ (red), right: estimated probabilities.

It is interesting to note that the deviance of the categorized AGE fit is the smallest that we obtained up to now. This is explained by the fact that we have fitted the most flexible model here. Unfortunately, this flexibility pays with the number of parameters. The AIC criterion as a compromise between goodness-of-fit and number of parameters states that all previous fitted

models are preferable. Nevertheless, categorization is a valuable tool to explore if there are nonlinear effects. A related technique is local regression smoothing which is shortly reviewed in Subsection 5.8. □

The estimation of default probabilities and the prediction of credit default should incorporate more than only one explanatory variable. Before fitting the full model with all available information, we discuss the modeling of interaction effects.

Example 11 *(Credit default on AGE and AMOUNT)*
The variable AMOUNT is the second continuous explanatory variable in the credit data set. (Recall that duration is quantitative as well but quasi-discrete.) We will therefore use AGE and AMOUNT to illustrate the effects of the simultaneous use of two explanatory variables. A very simple model is of course $logit\{P(Y=1|AGE,AMOUNT)\} = \beta_0 + \beta_1 AGE + \beta_2 AMOUNT$. *This model, however, separates the impact of AGE and AMOUNT into additive components. The effect of having both characteristics simultaneously is modeled by adding the multiplicative interaction term AGE∗AMOUNT. On the other hand we have seen that at least AGE should be complemented by a quadratic term. For that reason we compare the linear interaction model* $logit\{P(Y=1|AGE,AMOUNT)\} = \beta_0 + \beta_1 AGE + \beta_2 AMOUNT + \beta_3 AGE * AMOUNT$ *with a specification using quadratic terms and a third model specification using both, quadratic and interaction terms.*

Table 8: Credit default on AGE and AMOUNT (logit model).

Variable	Coefficient	t-value	Coefficient	t-value	Coefficient	t-value
constant	0.0159	-0.044	1.1815	1.668	1.4864	2.011
AGE	-0.0350	-3.465	-0.1012	-2.768	-0.1083	-2.916
AGE**2	–	–	$9.86\cdot 10^{-4}$	2.251	$9.32\cdot 10^{-4}$	2.100
AMOUNT	$-2.80\cdot 10^{-5}$	-0.365	$-7.29\cdot 10^{-6}$	-0.098	$-1.18\cdot 10^{-4}$	-1.118
AMOUNT**2	–	–	$1.05\cdot 10^{-8}$	1.753	$9.51\cdot 10^{-9}$	1.594
AGE*AMOUNT	$3.99\cdot 10^{-6}$	1.951	–	–	$3.37\cdot 10^{-6}$	1.553
Deviance	1185.1		1180.2		1177.7	
df	996		995		994	
AIC	1193.1		1190.2		1189.7	
Iterations	4		4		4	

Table 8 shows the results for all three fitted models. The model with quadratic and interaction terms has the smallest AIC of the three fits. Pairwise LR tests show, however, that the largest of the three models is not significantly better than the model without the interaction term. The obtained surface on AGE and AMOUNT from the quadratic+interaction fit is displayed in Figure 5. □

Let us remark that interaction terms can also be defined for categorical variables. In this case interaction is modeled by including dummy variables for all possible combinations of categories. This may largely increase the number of parameters to estimate.

Example 12 *(Credit default on the full set of explanatory variables)*
In a final analysis we present now the results for the full set of variables from Table 3. We first estimated a logit model using all variables (AGE and AMOUNT also with quadratic and interaction terms). Most of the estimated coefficients in the second column of Table 9 have the expected sign. For example, the default probability decreases if previous loan were paid back without problems, the credit applicant is employed and has some savings, and the loan is used

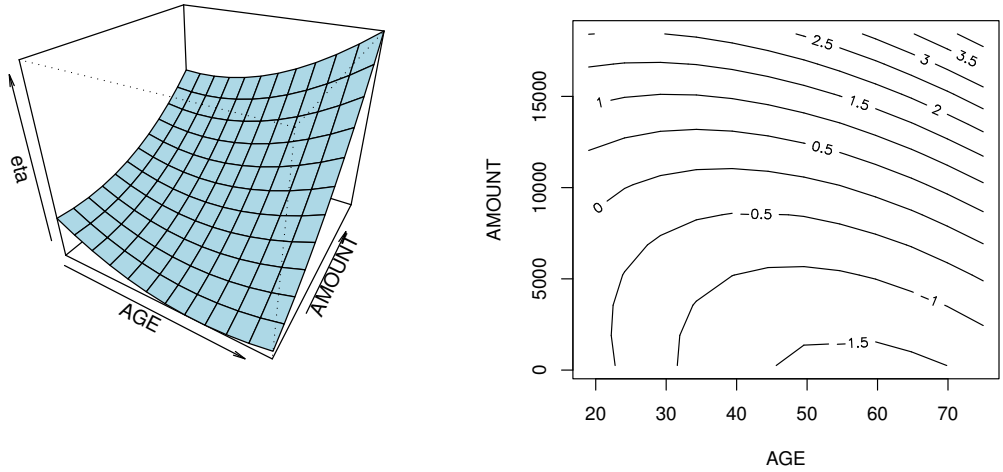

Figure 5: Credit default on AGE and AMOUNT using quadratic and interaction terms, left: surface and right: contours of the fitted η function.

to buy a car (rather than to invest the loan into goods which cannot serve as a security). A bit surprising is the fact that house owners seem to have higher default probabilities. This might be explained by the fact that house owners usually have additional obligations. The DURATION variable is categorized as described above. Again we have used the first category (loans up to 9 months) as reference. Since the series of DURATION coefficients is monotone increasing, we can conclude that longer duration increases the default probability. This is also plausible.

*After fitting the full model we have run an automatic stepwise model selection based on AIC. This reveals that the insignificant terms AGE*AMOUNT and EMPLOYED should be omitted. The fitted coefficients for this final model are displayed in the fourth column of Table 9.* □

5 Complements and Extensions

For further reading on GLM we refer to the textbooks of Dobson (2001), McCullagh and Nelder (1989) and Hardin and Hilbe (2001) (the latter with a special focus on STATA). Venables and Ripley (2002, Chapter 7) and Gill (2000) present the topic of generalized linear models in a very compact form. Collett (1991), Agresti (1996), Cox and Snell (1989), and Bishop et al. (1975) are standard references for analyzing categorical responses. We recommend the monographs of Fahrmeir and Tutz (1994) and Lindsey (1997) for a detailed introduction to GLM with a focus on multivariate, longitudinal and spatial data. In the following sections we will shortly review some specific variants and enhancements of the GLM.

5.1 Weighted Regression

Prior weights can be incorporated to the generalized linear model by considering the exponential density in the form

$$f(y_i, \theta_i, \psi) = \exp\left[\frac{w_i\{y\theta - b(\theta)\}}{a(\psi)} + c(y, \psi, w_i)\right].$$

This requires to optimize the sample log-likelihood

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} w_i \left\{\frac{Y_i\theta_i - b(\theta_i)}{a(\psi)} - c(Y_i, \psi, w_i)\right\}$$

Table 9: Credit default on full set of variables (logit model).

Variable	Coefficient	t-value	Coefficient	t-value
constant	1.3345	1.592	0.8992	1.161
AGE	-0.0942	-2.359	-0.0942	-2.397
AGE**2	$8.33 \cdot 10^{-4}$	1.741	$9.35 \cdot 10^{-4}$	1.991
AMOUNT	$-2.51 \cdot 10^{-4}$	-1.966	$-1.67 \cdot 10^{-4}$	-1.705
AMOUNT**2	$1.73 \cdot 10^{-8}$	2.370	$1.77 \cdot 10^{-8}$	2.429
AGE*AMOUNT	$2.36 \cdot 10^{-6}$	1.010	–	–
PREVIOUS	-0.7633	-4.652	-0.7775	-4.652
EMPLOYED	-0.3104	-1.015	–	–
DURATION $(9, 12]$	0.5658	1.978	0.5633	1.976
DURATION $(12, 18]$	0.8979	3.067	0.9127	3.126
DURATION $(18, 24]$	0.9812	3.346	0.9673	3.308
DURATION ≥ 24	1.5501	4.768	1.5258	4.710
SAVINGS	-0.9836	-4.402	-0.9778	-4.388
PURPOSE	-0.3629	-2.092	-0.3557	-2.051
HOUSE	0.6603	3.155	0.7014	3.396
Deviance	1091.5		1093.5	
df	985		987	
AIC	1121.5		1119.5	
Iterations	4		4	

or its equivalent, the deviance.

The weights w_i can be 0 or 1 in the simplest case that one wants to exclude specific observations from the estimation. The typical case of applying weights is the case of repeated independent realizations.

5.2 Overdispersion

Overdispersion may occur in one-parameter exponential families where the variance is supposed to be a function of the mean. This concerns in particular the binomial or Poisson families where we have $EY = \mu$ and $\mathrm{Var}(Y) = \mu(1 - \mu/k)$ or $\mathrm{Var}(Y) = \mu$, respectively. Overdispersion means that the actually observed variance from the data is larger than the variance imposed by the model. The source for this may be a lack of independence in the data or a misspecification of the model. One possible approach is to use alternative models that allows for a nuisance parameter in the variance, as an example think of the negative binomial instead of the Poisson distribution. For detailed discussions on overdispersion see Collett (1991) and Agresti (1990).

5.3 Quasi- or Pseudo-Likelihood

Let us remark that in the case that the distribution of Y itself is unknown but its two first moments can be specified, the quasi-likelihood function may replace the log-likelihood function. This means we still assume that

$$\begin{aligned} E(Y) &= \mu, \\ \mathrm{Var}(Y) &= a(\psi)\, V(\mu). \end{aligned}$$

The quasi-likelihood function is defined through

$$\ell(y,\theta,\psi) = \frac{1}{a(\psi)} \int_{\mu(\theta)}^{y} \frac{(s-y)}{V(s)}\, ds\,, \tag{18}$$

cf. Nelder and Wedderburn (1972). If Y comes from an exponential family then the derivatives of the log-likelihood and quasi-likelihood function coincide. Thus, (18) establishes in fact a generalization of the likelihood approach.

5.4 Multinomial Responses

A multinomial model (or nominal logistic regression) is applied if the response for each observation i is one out of more than two alternatives (categories). For identification one of the categories has to be chosen as reference category; without loss of generality we use here the first category. Denote by π_j the probability $P(Y = j|\boldsymbol{X})$, then we can consider the logits with respect to the first category, i.e.

$$\text{logit}(\pi_j) = \log\left(\frac{\pi_j}{\pi_1}\right) = \boldsymbol{X}_j^\top \boldsymbol{\beta}_j.$$

The terms $\boldsymbol{X}_j$ and $\boldsymbol{\beta}_j$ indicate that the explanatory variables and their corresponding coefficients may depend on category j. Equivalently we can define the model by

$$\begin{aligned} P(Y=1|\boldsymbol{X}) &= \frac{1}{1+\sum_{k=2}^{J}\exp(\boldsymbol{X}_k^\top\boldsymbol{\beta}_k)} \\ P(Y=j|\boldsymbol{X}) &= \frac{\boldsymbol{X}_j^\top\boldsymbol{\beta}}{1+\sum_{k=2}^{J}\exp(\boldsymbol{X}_k^\top\boldsymbol{\beta}_k)}\,. \end{aligned}$$

It is easy to recognize that the logit model is a special case of the multinomial model for exactly two alternatives.

If the categories are ordered in some natural way then this additional information can be taken into account. A latent variable approach leads to the cumulative logit model or the ordered probit model. We refer here to Dobson (2001, Section 8.4) and Greene (2000, Chapter 21) for ordinal logistic regression and ordered probit analysis, respectively.

5.5 Contingency Tables

The simplest form of a contingency table

Category	1	2	...	J	$\sum$
Frequency	Y_1	Y_2	...	Y_J	n

with one factor and a predetermined sample size n of observations is appropriately described by a multinomial distribution and can hence be fitted by the multinomial logit model introduced in Subsection 5.4. We could be for instance be interested in comparing the trivial model $EY_1 = \ldots = EY_J = \mu$ to the model $EY_2 = \mu_2, \ldots, EY_J = \mu_J$ (again we use the first category as reference). As before further explanatory variables can be included into the model.

Two-way or higher dimensional contingency tables involve a large variety of possible models. Let explain this with the help of the following two-way setup:

Category	1	2	...	J	$\sum$
1	Y_{11}	Y_{12}	...	Y_{1J}	$n_{1\bullet}$
2	Y_{21}	Y_{22}	...	Y_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
K	Y_{K1}	Y_{K2}	...	Y_{KJ}	$n_{K\bullet}$
$\sum$	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet J}$	n

Here we assume to have two factors, one with realizations $1, \ldots, J$, the other with realizations $1, \ldots, K$. If the Y_{jk} are independent Poisson variables with parameters μ_{jk}, then their sum is a Poisson variable with parameter $E(n) = \mu = \sum \mu_{jk}$. The Poisson assumption implies that the number of observations n is a random variable. Conditional on n, the joint distribution of the Y_{jk} is the multinomial distribution. Without additional explanatory variables, one is typically interested in estimating models of the type

$$\log(EY_{jk}) = \beta_0 + \beta_j + \beta_k$$

in order to compare this with the saturated model $\log(EY_{jk}) = \beta_0 + \beta_j + \beta_k + \beta_{jk}$. If the former model holds then the two factors are independent. Another hypothetical model could be of the form $\log(EY_{jk}) = \beta_0 + \beta_j$ to check whether the second factor matters at all. As in the multinomial case, further explanatory variables can be included. This type of models is consequently termed log-linear model. For more details see for example Dobson (2001, Chapter 9) and McCullagh and Nelder (1989, Chapter 6).

5.6 Survival Analysis

Survival data are characterized by non-negative observations which typically have a skewed distribution. An additional complication arises due to the fact that the observation period may end before the individual fails such that censored data may occur. The exponential distribution with density $f(y, \theta) = \theta e^{-\theta y}$ is a very simple example for a survival distribution. In this special case the survivor function (the probability to survive beyond y) is given by $S(y) = e^{-\theta y}$ and the hazard function (the probability of death within y and $y + dy$ after survival up to y) equals $h(y, \theta) = \theta$. Given additional explanatory variables this function is typically modeled by

$$h(y, \theta) = \exp(\boldsymbol{X}^\top \boldsymbol{\beta}).$$

Extensions of this model are given by using the Weibull distribution leading to non-constant hazards and Cox' proportional hazards model (Cox, 1972) which uses a semiparametric approach. More material on survival analysis can be found in Chapter III.12.

5.7 Clustered Data

Clustered data in relation to regression models mean that data from known groups ("clusters") are observed. Often these are the result of repeated measurements on the same individuals at different time points. For example, imagine the analysis of the effect of a medical treatment on patients or the repeated surveying of households in socio-economic panel studies. Here, all observations on the same individual form a cluster. We speak of longitudinal or panel data in that case. The latter term is typically used in the econometric literature.

When using clustered data we have to take into account that observations from the same cluster are correlated. Using a model designed for independent data may lead to biased results or at least significantly reduce the efficiency of the estimates.

A simple individual model equation could be written as follows:

$$E(Y_{ij}|\boldsymbol{X}_{ij}) = G^{-1}(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta}_j).$$

Here i is used to denote the ith individual observation in the jth cluster. Of course more complex specifications, for example with hierarchical clusters, can be formulated as well.

There is a waste amount of literature which deals with many different possible model specifications. A comprehensive resource for linear and nonlinear mixed effect models (LME, NLME) for continuous responses is Pinheiro and Bates (2000). The term "mixed" here refers to the fact that these models include additional random and/or fixed effect components to allow for correlation within and heterogeneity between the clusters.

For generalized linear mixed models (GLMM), i.e. clustered observations with responses from GLM-type distribution, several approaches are possible. For repeated observations, Liang and Zeger (1986) and Zeger and Liang (1986) propose to use generalized estimating equations (GEE) which result in a quasi-likelihood estimator. They show that the correlation matrix of $\boldsymbol{Y}_j$, the response observations from one cluster, can be replaced by a "working correlation" as long as the moments of $\boldsymbol{Y}_j$ are correctly specified. Useful working correlations depend on a small number of parameters. For longitudinal data an autoregressive working correlation can be used for example. For more details on GEE see also the monograph by Diggle et al. (2002). In the econometric literature longitudinal or panel data are analyzed with a focus on continuous and binary responses. Standard references for econometric panel data analyses are Hsiao (1990) and Arellano (2003). Models for clustered data with complex hierarchical structure are often denoted as multilevel models. We refer to the monograph of Goldstein (2003) for an overview.

5.8 Semiparametric Generalized Linear Models

Nonparametric components can be incorporated into the GLM at different places. For example, it is possible to estimate a single index model

$$E(Y|\boldsymbol{X}) = g(\boldsymbol{X}^\top \boldsymbol{\beta})$$

which differs from the GLM by its unknown smooth link function $g(\bullet)$. The parameter vector $\boldsymbol{\beta}$ in this model can then be only identified up to scale. The estimation of such models has been studied e.g. by Ichimura (1993), Weisberg and Welsh (1994) and Gallant and Nychka (1987).

Local regression in combination with likelihood-based estimation is introduced in Loader (1999). This concerns models of the form

$$E(Y|\boldsymbol{X}) = G^{-1}\left\{m(\boldsymbol{X})\right\},$$

where m is an unknown smooth (possibly multidimensional) function. Further examples of semiparametric GLM are generalized additive and generalized partial linear models (GAM, GPLM). These models are able to handle (additional) nonparametric components in the function η. For example, the GAM is specified in this simplest form by

$$E(Y|\boldsymbol{X}) = G^{-1}\left\{\beta_0 + \sum_{j=1}^{p} m_j(X_j)\right\}.$$

Here the m_j denote univariate (or low dimensional) unknown smooth functions which have to be estimated. For their identification is should be assumed, that $Em(X_j) = 0$. The generalized partial linear model combines a linear and a nonparametric function in the function η and is specified as

$$E(Y|\boldsymbol{X}) = G^{-1}\left\{\boldsymbol{X}_1^\top \boldsymbol{\beta} + m(\boldsymbol{X}_2)\right\}.$$

Example 13 *(Semiparametric credit model)*
We have fitted a generalized partial linear model as a variant of the final model from Example 12. The continuous variables AGE and AMOUNT were used as arguments for the nonparametric component. All other variables of the final model have been included to the linear part of the index function η. Figure 6 shows the estimated nonparametric function of AGE and AMOUNT. Although the stepwise model selection in Example 12 indicated that there is no interaction between AGE and AMOUNT, we see now that this interaction could be in fact of some more sophisticated form. The estimation was performed using a generalization of the Speckman (1988) estimator to generalized models. The local kernel weights are calculated from a Quartic (Biweight) kernel function using bandwidths approximately equal to 33.3% of the ranges of AGE and AMOUNT, respectively. Details on the used kernel based estimation can be found in Severini and Staniswalis (1994) and Müller (2001). □

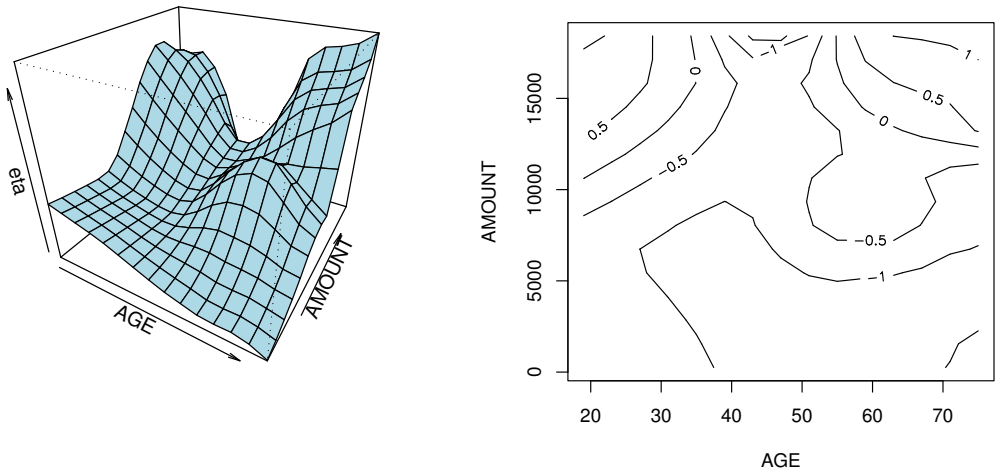

Figure 6: Credit default on AGE and AMOUNT using a nonparametric function, left: surface and right: contours of the fitted function on AGE and AMOUNT.

Some more material on semiparametric regression can be found in Chapters III.5 and III.10 of this handbook. For a detailed introduction to semiparametric extensions of GLM we refer to the textbooks by Hastie and Tibshirani (1990), Härdle et al. (2004), Ruppert et al. (1990), and Green and Silverman (1994).

References

Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csàdki, F., editors, *Second Internationl Symposium on Information Theory*, pages 267–281. Akademiai Kiadó, Budapest.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

Box, G. and Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26:211–243.

Collett, D. (1991). *Modelling Binary Data*. Chapman and Hall, London.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 74:187–220.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*, volume 32 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 2 edition.

Diggle, P., Heagerty, P., Liang, K.-L., and Zeger, S. (2002). *Analysis of Longitudinal Date*. Oxford University Press, second edition.

Dobson, A. J. (2001). *An Introduction to Generalized Linear Models*. Chapman and Hall, London, second edition.

Fahrmeir, L. and Kaufmann, H. (1984). Consistency and asymptotic normality of the maximum-likelihood estimator in generalized linear models. *Annals of Statistics*, 13:342–368.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer.

Gallant, A. and Nychka, D. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55(2):363–390.

Gill, J. (2000). *Generalized Linear Models: A Unified Approach*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-134, Thousand Oaks, CA.

Goldstein, H. (2003). *Multilevel Statistical Models*. Hodder Arnold, London.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, volume 58 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

Greene, W. H. (2000). *Econometric Analysis*. Prentice Hall, Upper Saddle River, New Jersey, 4 edition.

Hardin, J. and Hilbe, J. (2001). *Generalized Linear Models and Extensions*. Stata Press.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004). *Nonparametric and Semiparametric Modeling: An Introduction*. Springer, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

Hsiao, C. (1990). *Analysis of Panel Data*. Econometric Society Monographs No. 11. Cambridge University Press.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single–index models. *Journal of Econometrics*, 58:71–120.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Lindsey, J. K. (1997). *Applying Generalized Linear Models*. Springer, New York.

Loader, C. (1999). *Local Regression and Likelihood*. Springer, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 2 edition.

Müller, M. (2001). Estimation and testing in generalized partial linear models — a comparative study. *Statistics and Computing*, 11:299–309.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

Ruppert, D., Wand, M. P., and Carroll, R. J. (1990). *Semiparametric Regression*. Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89:501–511.

Speckman, P. E. (1988). Regression analysis for partially linear models. *Journal of the Royal Statistical Society, Series B*, 50:413–436.

Turlach, B. A. (1994). Computer-aided additive modeling. Doctoral Thesis, Université Catholique de Louvain, Belgium.

Venables, W. N. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Weisberg, S. and Welsh, A. H. (1994). Adapting for the missing link. *Annals of Statistics*, 22:1674–1700.

Zeger, S. L. and Liang, K. Y. a. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130.