Subsections

7.3 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.

7.3.1 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 $\int f(y,\theta,\psi)\,\mathrm{d}y = 1$. Under suitable regularity conditions (it is possible to exchange differentiation and integration) this implies

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

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 (7.1) 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

$$E(Y) = \mu = b'(\theta), \tag{7.2}$$

$$\mathbf{Var}(Y) = V(\mu)a(\psi) = b''(\theta)a(\psi). \tag{7.3}$$

Note that as a consequence from (7.1) 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.

7.3.2 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

$$\ell(\boldsymbol{Y},\boldsymbol{\mu},\psi) = \sum_{i=1}^{n}\ell(Y_i,\theta_i,\psi). \tag{7.4}$$

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*

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \sigma) = n \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_i)^2. \tag{7.5}$$

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*

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} \{Y_i \log(\mu_i) + (1 - Y_i)\log(1 - \mu_i)\} \tag{7.6}$$

for the sample version.

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

$$D(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = 2\left\{\ell(\boldsymbol{Y}, \boldsymbol{\mu}^{\max}, \psi) - \ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi)\right\}. \tag{7.7}$$

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 (7.4).

If we now plug-in the exponential family form (7.1) into (7.4) we obtain

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\psi)} - c(Y_i, \psi) \right\}. \tag{7.8}$$

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

$$\sum_{i=1}^{n} \{Y_i \theta_i - b(\theta_i)\} \tag{7.9}$$

for the maximization. The deviance analog of (7.9) is the (non-scaled) deviance function

$$D(\boldsymbol{Y}, \boldsymbol{\mu}) = D(\boldsymbol{Y}, \boldsymbol{\mu}, \psi)\, a(\psi). \tag{7.10}$$

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.

7.3.3 Iteratively Reweighted Least Squares Algorithm

We will now minimize the deviance with respect to $\boldsymbol{\beta}$. If we denote the gradient of (7.10) by

$$\nabla(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \left[-2 \sum_{i=1}^{n} \{Y_i \theta_i - b(\theta_i)\} \right] = -2 \sum_{i=1}^{n} \{Y_i - b'(\theta_i)\} \frac{\partial}{\partial \boldsymbol{\beta}} \theta_i \,, \tag{7.11}$$

our optimization problem consists in solving

$$\nabla(\boldsymbol{\beta}) = 0. \tag{7.12}$$

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:

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

A variant of the Newton-Raphson is the *Fisher scoring* algorithm that replaces the Hessian by its expectation with respect to the observations $\boldsymbol{Y}_i$:

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

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 (7.3) and taking derivatives again, we finally obtain

$$\frac{\partial}{\partial \boldsymbol{\beta}} \theta_i = \frac{G'(\eta_i)}{V(\mu_i)} \boldsymbol{X}_i$$

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

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

$$\nabla(\boldsymbol{\beta}) = -2 \sum_{i=1}^{n} \{Y_i - \mu_i\} \frac{G'(\eta_i)}{V(\mu_i)} \boldsymbol{X}_i$$

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

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

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

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

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

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

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} + (\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\widetilde{\boldsymbol{Y}} = (\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{Z}. \tag{7.13}$$

This immediately shows that each Fisher scoring iteration step is the result of a weighted least squares regression of the *adjusted dependent variables* $\boldsymbol{Z}_i$ 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 (7.13) can be found, only the weight matrix $\boldsymbol{W}$ 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 $\widehat{\boldsymbol{\beta}}$.

7.3.4 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 (7.12).
- 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 (7.13) 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 ([27]):

  * 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

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

  and/or the relative change in the deviance

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

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

$$\widehat{a}(\psi) = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{7.14}$$

  or using deviance

$$\widehat{a}(\psi) = \frac{D(\boldsymbol{Y}, \boldsymbol{\mu})}{n - p}. \tag{7.15}$$

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 Sect. 7.5.1. 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 $A(\mu) = \int^\mu V^{-1/3}(u)\,\mathrm{d}u$.

7.3.5 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 [12]. 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.

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

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

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

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

$$D(\boldsymbol{Y}, \boldsymbol{\mu}_0) - D(\boldsymbol{Y}, \boldsymbol{\mu}) \sim \chi^2_{p-p_0}. \tag{7.16}$$

The left hand side of (7.16) 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

$$\frac{(n-p)\{D(\boldsymbol{Y},\boldsymbol{\mu}_0)-D(\boldsymbol{Y},\boldsymbol{\mu})\}}{(p-p_0)D(\boldsymbol{Y},\boldsymbol{\mu})} \sim F_{p-p_0,n-p} \tag{7.17}$$

using the analog to the normal linear regression case ([36]), Chap. 7.

Model selection procedures for possibly non-nested models can be based on Akaike's information criterion [3]

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

or Schwarz' Bayes information criterion [32]

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

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