Appendix A
Background Material

All who debate on matters of uncertainity, should be free from prejudice, partiality, anger, or compassion.
—Caius Sallustius Crispus

A.1 Some Classical Likelihood Theory

Most of the VGLM/VGAM framework is infrastructure directed towards maximizing a full-likelihood model, therefore it is useful to summarize some supporting results from classical likelihood theory. The following is a short summary of a few selected topics serving the purposes of this book. The focus is on aspects of direct relevance to the practitioners and users of the software. The presentation is informal and nonrigorous; rigorous treatments, including justification and proofs, can be found in the texts listed in the bibliographic notes. The foundation of this subject was developed by Fisher a century ago (around the decade of WW1), and he is regarded today as the father of modern statistics.

A.1.1 Likelihood Functions

The usual starting point is to let Y be a random variable with density function $f(y; \boldsymbol{\theta})$ depending on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, a multidimensional unknown parameter. Values that Y can take are denoted in lower-case, i.e., y. By 'density function', here is meant a probability (mass) function for a discrete-valued Y, and probability density function for continuous Y. We shall refer to f as simply the density function, and use integration rather than summation to denote quantities such as expected values, e.g., $E(Y) = \int f(y)\, dy$ where the range of integration is over the *support* of the distribution, i.e., those values of y where $f(y) > 0$ (called $\mathcal{Y}$).

A lot of statistical practice centres upon making inference about $\boldsymbol{\theta}$, having observed $Y = y$. As well as obtaining an estimate $\widehat{\boldsymbol{\theta}}$, it is customary to cite some measure of accuracy or plausibility of the estimate, usually in the form of its

T.W. Yee, *Vector Generalized Linear and Additive Models*,
Springer Series in Statistics, DOI 10.1007/978-1-4939-2818-7

standard error, $\mathrm{SE}(\widehat{\boldsymbol{\theta}})$. It is also common to conduct hypothesis tests, e.g., for a one-parameter model, test the null hypothesis $H_0 : \theta = \theta_0$ for some known and fixed value θ_0.

Let $\boldsymbol{\Omega}$ be the parameter space, which is the set of possible values that $\boldsymbol{\theta}$ can take. For example, if $Y \sim N(\mu,\ \sigma^2)$ where $\boldsymbol{\theta} = (\mu, \sigma)^T$, then $\boldsymbol{\Omega} = \mathbb{R} \times (0, \infty) = \mathbb{R} \times \mathbb{R}_+$. Another simple example is the beta distribution having positive shape parameters $\boldsymbol{\theta} = (s_1, s_2)^T$, therefore $\boldsymbol{\Omega} = \mathbb{R}_+^2$. Clearly, $\boldsymbol{\Omega} \subseteq \mathbb{R}^p$.

In the wider VGLM/VGAM framework, some of our responses $\boldsymbol{y}_i$ may be multivariate, therefore let $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T)^T$ be a random vector of n observations, each $\boldsymbol{Y}_i$ being a random vector. We observe $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ in totality.

Each $\boldsymbol{y}_i$ can be thought of as being a realization from some statistical distribution with joint density function $f(\boldsymbol{y}_i; \boldsymbol{\theta})$. With n observations, the joint density function can be written $f(\boldsymbol{y}; \boldsymbol{\theta})$. We say that a (parametric) statistical model is a set of possible density functions indexed by $\boldsymbol{\theta} \in \boldsymbol{\Omega}$, i.e.,

$$\mathcal{M}_{\boldsymbol{\theta}} \;=\; \{f(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Omega}\},$$

which may be simplified to just $\mathcal{M}$.

The approach considered in this book is to assume that the user knows such a family of distributions. Often this strong assumption is groundless, and therefore parametric models may give misleading results. A method that lies in between the fully-parametric method adopted in this book and nonparametric methods is based on using an empirical likelihood (Owen, 2001), which gives the best of both worlds. The empirical likelihood supplies information at a sufficient rate that reliable confidence intervals/regions and hypothesis tests can be constructed.

Of course, parameterizations are not unique, e.g., for many distributions in Chap. 12, the scale parameter b is used so that the form y/b appears in the density, whereas some practitioners prefer to use its reciprocal, called the rate, and then the densities have the term λy. Two other examples, from Table 12.11, are the beta and beta-binomial distributions which are commonly parameterized in terms of the shape parameters, otherwise the mean and a dispersion parameter.

Regardless of the parameterization chosen, the parameter must be *identifiable*. This means that each element of $\mathcal{M}_{\boldsymbol{\theta}}$ corresponds to exactly one value of $\boldsymbol{\theta}$. Stated another way, if $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2 \in \boldsymbol{\Omega}$ with $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ then the densities $\mathcal{M}_{\boldsymbol{\theta}_1}(\boldsymbol{y}) \neq \mathcal{M}_{\boldsymbol{\theta}_2}(\boldsymbol{y})$. As an example, consider the multinomial logit model (1.25) described in Sect. 14.2. We can have

$$P(Y = j|\boldsymbol{x}) = \frac{e^c}{e^c}\ \frac{\exp\{\eta_j\}}{\sum_{k=1}^{M+1} \exp\{\eta_k\}} \;=\; \frac{\exp\{\eta_j + c\}}{\sum_{k=1}^{M+1} \exp\{\eta_k + c\}},$$

for any constant c, hence the $M+1$ η_js are non-identifiable. In practice, we choose $c = -\eta_t$ for some t, and then redefine the η_j. The family function `multinomial()` chooses $t = M+1$, by default, as the reference group so that $\eta_{M+1} \equiv 0$, but $t = 1$ is another popular software default.

A.1.2 Maximum Likelihood Estimation

Maximum likelihood estimation is the most widely used general-purpose estimation procedure in statistics. It centres on the *likelihood function* for $\boldsymbol{\theta}$, based on the observation of $\boldsymbol{Y} = \boldsymbol{y}$:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \;=\; f(\boldsymbol{y}; \boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \boldsymbol{\Omega}. \tag{A.1}$$

With a philosophical twist, two quantities can be seen contrasted here: the likelihood function that is a function of the parameter $\boldsymbol{\theta}$, given the data $\boldsymbol{y}$, cf. the density that is a function of the data $\boldsymbol{y}$, given the parameter $\boldsymbol{\theta}$. The likelihood function is thus the probability of observing what we got ($\boldsymbol{y}$) as a function of $\boldsymbol{\theta}$ based on our model. Clearly, this holds for discrete responses, but it can be easily justified for continuous responses too (see below). Thus maximum likelihood estimation treats the data as being fixed and given, and it determines $\boldsymbol{\theta}$ which makes our observed data most probable.

It is much more convenient to work on a log-scale. One major reason for this monotone transformation is that data is very commonly assumed to be independent, so we can obtain additivity of log-likelihood contributions. Also, rather than having a single observation $Y = y$, it is more general to have $\boldsymbol{Y}_i = \boldsymbol{y}_i$ for $i = 1, \ldots, n$, where n is the sample size. Putting these two properties together,

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \;=\; f(\boldsymbol{y}; \boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} f(\boldsymbol{y}_i; \boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} L_i, \tag{A.2}$$

where the data is $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$.

Now, taking the logarithm of this joint distribution gives the *log-likelihood* function

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) \;=\; \sum_{i=1}^{n} \log f(\boldsymbol{y}_i; \boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} \ell_i. \tag{A.3}$$

The fact that this is a sum will enable us later to state large sample properties of ML estimators by application of the law of large numbers.

Maximum likelihood estimation involves maximizing L, or equivalently, ℓ. We can write

$$\widehat{\boldsymbol{\theta}} \;=\; \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Omega}} \ell(\boldsymbol{\theta}; \boldsymbol{y}),$$

and the solution need not be unique or even exist. Unless $\widehat{\boldsymbol{\theta}}$ is on the boundary, we obtain $\widehat{\boldsymbol{\theta}}$ by solving $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \boldsymbol{0}$. Iterative methods (Sect. A.1.2.4) are commonly employed to obtain the *maximum likelihood estimate* $\widehat{\boldsymbol{\theta}}$, because no closed-form expression can be obtained.

In maximizing ℓ, it is the relative values of $\ell(\boldsymbol{\theta})$ that matter, not their values in absolute terms. Hence, some authors omit any additive constants not involving $\boldsymbol{\theta}$ from ℓ but still use "=" in (A.3). This actually holds implicitly for continuous responses $\boldsymbol{Y}$ because the probability that $\boldsymbol{Y} = \boldsymbol{y}$ is actually 0, hence fundamentally, (A.1) is actually of the form $f(\boldsymbol{y}; \boldsymbol{\theta}) \cdot \varepsilon$ which is a 'real' probability—it represents the chances of observing a value in a small set of volume centred at $\boldsymbol{y}$. Then (A.2) involves a $\propto$ because the width of the volume, as measured by ε, does not depend on $\boldsymbol{\theta}$, and therefore (A.3) is equality up to a constant. For families such as `posbinomial()`, it is necessary to set the argument `omit.constant` to `TRUE` when comparing nested models that have different normalizations (Sect. 17.2.1).

ML estimators are functions of quantities known as *sufficient statistics*. A statistic is simply a function of the sample space $\mathcal{S}$, and it will be denoted here by T. Sufficient statistics are statistics that reduce the data into two parts: a useful part and an irrelevant part. The sufficient statistic contains all the information about $\boldsymbol{\theta}$ that is contained in $\boldsymbol{Y}$, and it is not unique. By considering only the useful part, sufficient statistics allow for a form of data reduction. The usual definition of a statistic T that is sufficient for $\mathcal{M}_{\boldsymbol{\theta}}$ of $\boldsymbol{Y}$ is that the conditional distribution $f(\boldsymbol{Y}|T = t)$ does not depend on $\boldsymbol{\theta}$, for all values of t. However, this definition is not as useful as one would like. Fortunately, there is a famous result called the factorization theorem that is more useful than the original definition, because it provides a method for testing whether a statistic T is sufficient, as well as obtaining T in the first place. It can be stated as follows.

Factorization Theorem A statistic $T(\boldsymbol{Y})$ is sufficient for $\mathcal{M}_{\boldsymbol{\theta}}$ iff there exist non-negative functions $g(\cdot;\boldsymbol{\theta})$ and h such that

$$f(\boldsymbol{y};\boldsymbol{\theta}) \;=\; g(T(\boldsymbol{y});\boldsymbol{\theta}) \cdot h(\boldsymbol{y}). \tag{A.4}$$

Then clearly maximizing a likelihood via f is equivalent to maximizing g only, because h is independent of $\boldsymbol{\theta}$.

Some well-known examples of sufficient statistics are as follows.

(i) If $Y_i \sim$ Poisson(μ) independently, then $\sum_i Y_i$ is sufficient for $\theta = \mu$. Similarly, if $Y_i \sim$ Binomial($n = 1, \mu$) is a sequence of independent Bernoulli random variables, then $\sum_i Y_i$ is also sufficient for $\theta = \mu$. In both cases, there is a reduction of n values down to one value.
(ii) If the Y_i are a random sample from an $N(\mu,\ \sigma^2)$ distribution, then $(\overline{y}, s^2)$ are sufficient for $\boldsymbol{\theta} = (\mu, \sigma)^T$. This is reduction of an n-vector down to 2 values.

A.1.2.1 Notation

The standard notation

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \;=\; \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p}\right)^T \;=\; \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)^T,$$

$$\frac{\partial \boldsymbol{b}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \;=\; \left[\left(\frac{\partial b_j(\boldsymbol{\theta})}{\partial \theta_k}\right)\right] \;=\; \left(\frac{\partial \boldsymbol{b}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T, \;\text{and}\; \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T} \;=\; \left[\left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_j\,\partial \theta_k}\right)\right]$$

is adopted.

Before describing the Fisher scoring algorithm which is central to this book, it is necessary to define some standard quantities first. Let the *score* (or *gradient*) vector be defined as

$$\boldsymbol{U}(\boldsymbol{\theta}) \;=\; \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{A.5}$$

the *Hessian* as

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) \;=\; \frac{\partial \boldsymbol{U}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \;=\; \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T}, \tag{A.6}$$

and the *(observed) information matrix* as

$$\boldsymbol{\mathcal{I}}_O(\boldsymbol{\theta}) \;=\; -\,\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) \;=\; -\,\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T}. \tag{A.7}$$

Sometimes it is necessary to distinguish between the true value of $\boldsymbol{\theta}$ (called $\boldsymbol{\theta}_*$) and $\boldsymbol{\theta}$ itself. If not, then $\boldsymbol{\theta}$ is used for both meanings.

The acronym "MLE" is used loosely to stand for: maximum likelihood estimation, maximum likelihood estimator, and maximum likelihood estimate.

A.1.2.2 Regularity Conditions

To formalize the method of MLE more adequately, some mathematical properties required of $\mathcal{M}_{\boldsymbol{\theta}}$ must be established. These are called regularity conditions. A distribution satisfying them is called *regular*, otherwise it is nonregular.

Regularity Condition I The dimension of $\boldsymbol{\theta}$ is fixed. A counterexample is a problem used commonly to motivate James-Stein estimation, which is that $Y_i \sim N(\mu_i,\ \sigma^2 = 1)$ independently. Then $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_n)^T$ grows with increasing n. Neyman and Scott (1948) showed that MLEs could be inconsistent when the number of parameters increased with n. In such cases, a method to eliminate unnecessary parameters is often sought, e.g., by integrating or conditioning them out.

Regularity Condition II The parameter $\boldsymbol{\theta}$ is identifiable.

Regularity Condition III The distributions $\mathcal{M}_{\boldsymbol{\theta}}$ have a common support, i.e., are independent of $\boldsymbol{\theta}$. Here are some counterexamples.

(i) The simplest is $Y_i \sim \text{Unif}(0, \theta)$, so that its support is a function of θ.

(ii) Another common type of example is a 3-parameter density parameterized by a location (a), scale (b) and shape (s) parameter, and whose support is defined on (a, ∞). A specific example of this that has received considerable attention is the 3-parameter Weibull distribution, whose CDF can be written as $1 - \exp\{-[(y-a)/b]^s\}$ for $y > a$, and 0 otherwise. Another example of this sort is the 3-parameter lognormal distribution where $\log(Y - a) \sim N(\mu,\ \sigma^2)$ so that $a < Y < \infty$.

(iii) The generalized extreme value distribution (GEV; Sect. 16.2) depends on the unknown parameter values. This problem is studied in depth in Smith (1985), who also considered the 3-parameter Weibull distribution.

Regularity Condition IV $\boldsymbol{\Omega}$ is an open set (of $\mathbb{R}^p$).

Regularity Condition V The true value $\boldsymbol{\theta}_*$ lies in the interior of $\boldsymbol{\Omega}$.

Regularity Condition VI The first three derivatives of ℓ exist on an open set containing $\boldsymbol{\theta}_*$ (call it $\mathcal{A}$, say), and $\partial^3 \log f(y; \boldsymbol{\theta})/(\partial\theta_s\, \partial\theta_t\, \partial\theta_u) \leq M(y)$ uniformly for $\boldsymbol{\theta} \in \mathcal{A}$, where $0 < E(M(y)) < \infty$.

The next condition addresses the interchange of the order of double differentiation with respect to $\boldsymbol{\theta}$ and integration over $\mathcal{S}$.

Regularity Condition VII For all $\boldsymbol{y} \in \mathcal{Y}$ and $\boldsymbol{\theta} \in \boldsymbol{\Omega}$, ℓ is twice-differentiable with

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{\theta})\, d\boldsymbol{y} \;=\; \int_{\mathcal{Y}} \frac{\partial}{\partial \boldsymbol{\theta}}\, f(\boldsymbol{y}; \boldsymbol{\theta})\, d\boldsymbol{y},$$

and

$$\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T} \int_{\mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{\theta})\, d\boldsymbol{y} \;=\; \int_{\mathcal{Y}} \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T}\, f(\boldsymbol{y}; \boldsymbol{\theta})\, d\boldsymbol{y}.$$

A commonly used counterexample of regularity conditions VI–VII is the double exponential (Laplace) distribution (Sect. 15.3.2), whose derivative does not exist at the location parameter.

A.1.2.3 Fisher Information

A very important quantity in MLE theory is the *Fisher information*, which can manifest itself in the form of the *Fisher information matrix*, or *expected information matrix* (EIM). This measures the average amount of information about the parameter $\boldsymbol{\theta}$ over all possible observations, not just those actually observed. Intuitively, it measures the average amount of curvature of ℓ at the MLE $\widehat{\boldsymbol{\theta}}$. If the data provides a lot of information about $\boldsymbol{\theta}$, then the peak at the MLE will be sharp, not flat, because the parameter has a large effect on the likelihood function. Flatness, or a lack of steepness, denotes a lot of uncertainty in the estimated parameter. The EIM can be defined as

$$\boldsymbol{\mathcal{I}}_E(\boldsymbol{\theta}) \;=\; \mathrm{Var}\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right). \tag{A.8}$$

The Fisher information has some basic properties:

1. For independent observations, it is *additive*; and for i.i.d. random variables, this can be written as
$$\boldsymbol{\mathcal{I}}_E(\boldsymbol{\theta}) \;=\; n\,\boldsymbol{\mathcal{I}}_{E1}(\boldsymbol{\theta}),$$
where $\boldsymbol{\mathcal{I}}_{E1}(\boldsymbol{\theta})$ is the EIM for the first observation. This makes intuitive sense, because increasing n ought to increase the amount of information there is about $\boldsymbol{\theta}$. That the total Fisher information is the sum of each observation's Fisher information will be shown later to imply that the amount of uncertainty in $\widehat{\boldsymbol{\theta}}$ should decrease with increasing n, i.e., $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$ should decrease in a matrix sense.
2. It is *positive-semidefinite*. Practically, for us it is positive-definite over a large part of $\boldsymbol{\Omega}$, though singular EIMs can occur as extreme cases in likelihood theory.
3. It changes under transformations, and the EIM under monotonic transformations is readily available, as follows. Let $g_j(\boldsymbol{\theta})$ be a set of p invertible functions that are differentiable. Then
$$\boldsymbol{\mathcal{I}}_E(\boldsymbol{g}) \;=\; \frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{g}}\, \boldsymbol{\mathcal{I}}_E(\boldsymbol{\theta})\, \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{g}^T} \tag{A.9}$$
where $\boldsymbol{g} = (g_1(\boldsymbol{\theta}), \ldots, g_p(\boldsymbol{\theta}))^T$. This result is used much in this book, both directly and indirectly, e.g., the variance-covariance matrix (A.27) for the delta method, and it lurks in the background of (18.6), (18.9) and (18.11).
As a simple example, if $\tau = g(\theta)$ where g is smooth and $g'(\theta) \neq 0$, then $\boldsymbol{\mathcal{I}}_{E1}(\tau) = \boldsymbol{\mathcal{I}}_{E1}(\theta)/[g'(\theta)]^2$. Applied specifically to $Y \sim \mathrm{Poisson}(\lambda)$, then $\boldsymbol{\mathcal{I}}_{E1}(\lambda) = 1/\lambda$, and for $\tau = \sqrt{\lambda}$, $\boldsymbol{\mathcal{I}}_{E1}(\tau) = (4\lambda)/4 = 4$, which is independent of λ (this is known as the Poisson variance-stabilizing transformation).
4. For some models with $p > 1$, it is possible for the (j, k) EIM element to be equal to 0 ($j \neq k$). If so, then θ_j and θ_k are said to be *orthogonal*, and this implies *asymptotic independence* between them. An important consequence of two parameters being orthogonal is that the MLE of one parameter varies only slowly with the other parameter. Indeed, for some models where several

parameterizations have been proposed, it is not uncommon to prefer ones with orthogonal parameters because of the stability they produce. Computationally, it can lead to faster convergence and be numerically well-conditioned. And in the case of VGAM, less storage may arise because of the matrix-band format used to represent EIMs (Sect. 18.3.5), e.g., for the bivariate odds ratio model it has the form (McCullagh and Nelder, 1989, p.228)

$$\begin{pmatrix} \times & \times & 0 \\ \times & \times & 0 \\ 0 & 0 & \times \end{pmatrix}$$

so that the working weights can be stored in an $n \times 4$ matrix, which is a saving of $2n$ doubles compared to n general 3×3 working weight matrices. If necessary, one might reorder the θ_j so that the non-zero values cluster about the diagonal band; this idea holds for family function `posbernoulli.tb()` (Ex. 17.5). For more details, see Cox and Reid (1987) and Young and Smith (2005).

Some Examples of EIMs

The VGAM package implements Fisher scoring on most parts, therefore each model must have EIMs that are tractable or can be approximated. In the latter case, Sect. 9.2 describes some methods. We now illustrate the former case by considering simple distributions that have closed-form expressions for the EIM elements. These examples come from the VGAM package.

1. `betaR()` The standard beta density, as implemented by `[dpqr]beta()`, parameterizes the density in terms of the two positive shape parameters, and it is

$$f(y; s_1, s_2) \;=\; \frac{y^{s_1-1}\,(1-y)^{s_2-1}}{Be(s_1, s_2)} \;=\; \frac{y^{s_1-1}\,(1-y)^{s_2-1}\;\Gamma(s_1+s_2)}{\Gamma(s_1)\,\Gamma(s_2)}$$

for $y \in (0,1)$. For one observation, $\ell = (s_1-1)\log y + (s_2-1)\log(1-y) + \log\Gamma(s_1+s_2) - \log\Gamma(s_1) - \log\Gamma(s_2)$, from which the derivatives are

$$\begin{aligned}
\frac{\partial \ell}{\partial s_1} &= \log y + \psi(s_1+s_2) - \psi(s_1), \\
\frac{\partial \ell}{\partial s_2} &= \log(1-y) + \psi(s_1+s_2) - \psi(s_2), \\
-\frac{\partial^2 \ell}{\partial s_j^2} &= \psi'(s_j) - \psi'(s_1+s_2), \qquad j = 1, 2, \\
-\frac{\partial^2 \ell}{\partial s_1\,\partial s_2} &= -\psi'(s_1+s_2).
\end{aligned}$$

The second derivatives are not functions of y, and therefore the OIM and EIM coincide, both being

$$\begin{pmatrix} \psi'(s_1) - \psi'(s_1+s_2) & -\psi'(s_1+s_2) \\ -\psi'(s_1+s_2) & \psi'(s_2) - \psi'(s_1+s_2) \end{pmatrix}.$$

2. `rayleigh()` Sometimes the property $E[\partial\ell/\partial\theta_j] = 0$ can be used to good effect when working out elements of the EIM, as the following simple

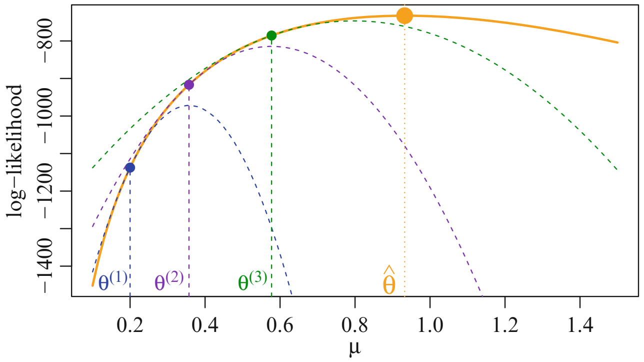

Fig. A.1 The first few Newton-like iterations for a Poisson regression fitted to the `V1` data set. The *solid orange curve* is $\ell(\theta)$ with $\theta = \mu$. The initial value is $\theta^{(1)} = 0.2$. Each iteration $\theta^{(a)}$ corresponds to the maximum of the quadratic (*dashed curves*) from the previous iteration.

example shows. From Table 12.8, the density of the Rayleigh distribution is $y \cdot \exp\{-2^{-1}(y/b)^2\}/b^2$ for positive y and positive scale parameter b. Then, for one observation, $\ell = \log y - 2^{-1}(y/b)^2 - 2\log b$ so that $\ell' = ([y/b]^2 - 2)/b$. Equating this to 0 implies that $E(Y^2) = 2b^2$. Then $-\ell'' = (3y^2 - 2b^2)/b^4$ so that the EIM is $(3 \times 2b^2 - 2b^2)/b^4 = 4/b^2$.

A.1.2.4 Newton-Like Algorithms

Given that an iterative method will be used to solve for the MLE, let's expand ℓ in a first-order Taylor series about the current estimate at iteration $a - 1$:

$$\ell(\boldsymbol{\theta}^{(a)}) \;\approx\; \ell(\boldsymbol{\theta}^{(a-1)}) + (\boldsymbol{\theta}^{(a)} - \boldsymbol{\theta}^{(a-1)})^T \, \frac{\partial \ell(\boldsymbol{\theta}^{(a-1)})}{\partial \boldsymbol{\theta}}.$$

Now take the first derivatives: $\partial\ell/\partial\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}^{(a)}$ is equal to

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\theta}^{(a)})}{\partial \boldsymbol{\theta}} &= \frac{\partial \ell(\boldsymbol{\theta}^{(a-1)})}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \ell(\boldsymbol{\theta}^{(a-1)})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T} \left(\boldsymbol{\theta}^{(a)} - \boldsymbol{\theta}^{(a-1)}\right) \qquad \text{(A.10)} \\ &= \boldsymbol{U}(\boldsymbol{\theta}^{(a-1)}) + \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}^{(a-1)}) \left(\boldsymbol{\theta}^{(a)} - \boldsymbol{\theta}^{(a-1)}\right). \end{aligned}$$

Ideally, the next iteration will be very good, or even better, it will be optimal. If so, then $\boldsymbol{\theta}^{(a)}$ will have the value $\widehat{\boldsymbol{\theta}}$, which is the MLE—and then its score vector will be $\mathbf{0}$. Thus we will be totally optimistic and set the LHS of (A.10) to $\mathbf{0}$. Upon rearrangement, this leads to the Newton-Raphson step

$$\boldsymbol{\theta}^{(a)} \;=\; \boldsymbol{\theta}^{(a-1)} - \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}^{(a-1)})^{-1}\, \boldsymbol{U}(\boldsymbol{\theta}^{(a-1)}). \qquad \text{(A.11)}$$

The algorithm converges quickly at a quadratic convergence rate, provided that ℓ is well-behaved (close to a quadratic) in a neighbourhood of the maximum, and if the starting value is close enough to the solution. By a 'quadratic convergence rate', it is meant that

$$\lim_{a\to\infty} \frac{\|\boldsymbol{\theta}^{(a)} - \widehat{\boldsymbol{\theta}}\|}{\|\boldsymbol{\theta}^{(a-1)} - \widehat{\boldsymbol{\theta}}\|^2} \;=\; c$$

for some positive c. What this means in practice is that the number of correct decimal places doubles at each iteration near the solution.

Figure A.1 illustrates the idea behind Newton-like algorithms for a simple one-parameter problem involving a Poisson regression fitted to the `V1` data set. Starting at $\theta^{(0)} = 0.2$, successive quadratics are fitted to approximate ℓ and obtain the next iteration $\theta^{(a)}$. These quadratics match the derivatives $\ell^{(\nu)}(\theta^{(a-1)})$ for $\nu = 0, 1, 2$.

The Newton-Raphson algorithm requires the inversion of an order-p matrix, which is $O(p^3)$ and therefore expensive for very large p, and it does require the programming of the $p(p+1)/2$ unique elements of $\boldsymbol{\mathcal{H}}$. And a Newton-Raphson step is not guaranteed to be an improvement: $\ell(\boldsymbol{\theta}^{(a)}) < \ell(\boldsymbol{\theta}^{(a-1)})$ is a possibility. There have been many modifications proposed to the plain Newton-Raphson algorithm, but that is beyond the scope of this book; for more details see, e.g., Dennis and Schnabel (1996), Nocedal and Wright (2006), Weihs et al. (2014).

An alternative procedure proposed by Fisher is to replace the OIM by the EIM. The result is

$$\boldsymbol{\theta}^{(a)} \;=\; \boldsymbol{\theta}^{(a-1)} + \boldsymbol{\mathcal{I}}_E^{-1}(\boldsymbol{\theta}^{(a-1)})\, \boldsymbol{U}(\boldsymbol{\theta}^{(a-1)}), \tag{A.12}$$

which is known as *Fisher's method of scoring*, or just *Fisher scoring*. This method usually possesses only a linear convergence rate, meaning

$$\lim_{a\to\infty} \frac{\|\boldsymbol{\theta}^{(a)} - \widehat{\boldsymbol{\theta}}\|}{\|\boldsymbol{\theta}^{(a-1)} - \widehat{\boldsymbol{\theta}}\|} \;=\; c,$$

for some $0 < c < 1$, however typically $c \approx 0$ so that the convergence rate is quite acceptable. As the n EIMs are usually positive-definite, this means that each step is in an ascent direction, and half-stepping can be used to guarantee an improvement at each step (Sect. 3.5.4).

Fisher scoring is implemented by VGAM mainly for two reasons. The first is that, for most models, the EIMs are positive-definite over a large portion of the parameter space $\boldsymbol{\Omega}$, in contrast to OIMs which tend to be positive-definite in a smaller subset. As an example, consider the Rayleigh distribution above. Clearly, $-\ell''$ is positive for $y > \sqrt{2/3}\, b$, whereas the EIM is positive for all b. As mentioned elsewhere, IRLS requires *each* of the n EIMs to be positive-definite, not just their sum. The second reason is that EIMs are often simpler than the OIM. Fisher scoring may be performed by using the iteratively reweighted (generalized) least squares algorithm—see Sect. 3.2 for details. For GLMs with a canonical link, the OIM equals the EIM, therefore Newton-Raphson and Fisher scoring coincide.

How can one know whether one has reached the true solution? We say that $\widehat{\boldsymbol{\theta}}$ is a *stationary point* if $\boldsymbol{U}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{0}$. Iterative numerical methods may converge to a stationary point called a *local* maximum, e.g., when ℓ is multimodal such as Fig. 12.1 and the initial values are not very good. Also, if $\boldsymbol{\mathcal{I}}_O(\widehat{\boldsymbol{\theta}})$ is positive-definite, then $\widehat{\boldsymbol{\theta}}$ is a relative maximum. Equivalently, all its eigenvalues are positive, but if $\widehat{\boldsymbol{\mathcal{H}}}$ has positive and negative eigenvalues, then $\widehat{\boldsymbol{\theta}}$ is known as a *saddle point*. For some models, it can be proven that ℓ is concave in $\boldsymbol{\theta}$. If so, then the MLE is unique, and any local solution is the global solution. For example, for several categorical regression models, see Pratt (1981).

Incidentally, another common Newton-like method known as the Gauss-Newton method is used, particularly in nonlinear regression. This approximates the Hessian by $\sum_i \boldsymbol{u}(\boldsymbol{\theta}^{(a-1)})\, \boldsymbol{u}(\boldsymbol{\theta}^{(a-1)})^T$. It has the advantage that only first derivatives are needed, however it can suffer from the so-called *large residual problem* that causes its convergence to be very slow.

A.1.3 Properties of Maximum Likelihood Estimators

Under regularity conditions, MLEs have many good properties. They are described as *asymptotic* because $n \to \infty$. We write $\widehat{\boldsymbol{\theta}}_n$ to emphasize the MLE is based on a sample of size n, because this is enlightening in the case of i.i.d. observations. Recall here that $\boldsymbol{\theta}_*$ is the true value of $\boldsymbol{\theta}$. The properties of MLEs include the following.

1. *Asymptotic consistency*: for all $\varepsilon > 0$ and $\boldsymbol{\theta}_* \in \boldsymbol{\Omega}$,

$$\lim_{n\to\infty} P[\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*\|_\infty > \varepsilon] \;=\; 0. \tag{A.13}$$

   That is, the distribution of $\widehat{\boldsymbol{\theta}}_n$ collapses around $\boldsymbol{\theta}_*$. Here, the maximum (infinity) norm is used to show that the usual plim definition (A.32) is applied element-by-element to $\boldsymbol{\theta}_n$. It is common to write $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_*$ (convergence in probability). This is called weak consistency; a stronger form based on almost sure convergence in probability can be defined.
2. *Asymptotic normality*: $\widehat{\boldsymbol{\theta}}_n$ is asymptotically $N_p(\boldsymbol{\theta}_*,\, \boldsymbol{\mathcal{I}}_E^{-1}(\boldsymbol{\theta}_*))$ as $n \to \infty$, i.e.,

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathcal{D}} N_p(\boldsymbol{\theta}_*,\, \boldsymbol{\mathcal{I}}_E^{-1}(\boldsymbol{\theta}_*)) \tag{A.14}$$

   (convergence in distribution). For i.i.d. data, this can be stated as

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*\right) \xrightarrow{\mathcal{D}} N_p(\mathbf{0},\, \boldsymbol{\mathcal{I}}_{E1}^{-1}(\boldsymbol{\theta}_*)). \tag{A.15}$$

   Thus under i.i.d. conditions, $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_*$ in distribution at a $\sqrt{n}$-rate. In consequence of the above,

$$(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*)^T\, \boldsymbol{\mathcal{I}}_E(\boldsymbol{\theta}_*)\, (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) \;\sim\; \chi^2_p \tag{A.16}$$

   as $n \to \infty$.
3. *Asymptotically unbiasedness*: $E(\widehat{\boldsymbol{\theta}}_n) \to \boldsymbol{\theta}_*$ as $n \to \infty$, for all $\boldsymbol{\theta}_* \in \boldsymbol{\Omega}$.
4. *Asymptotically efficiency*: If a most-efficient (unbiased) estimator exists, then it will be the MLE. See the Cramér-Rao inequality of Sect. A.1.3.1.
5. *Invariance*: Another fundamental property is that if $\widehat{\boldsymbol{\theta}}$ is the MLE, then under a different parameterization $g(\boldsymbol{\theta})$ (where g is some monotone function of $\boldsymbol{\theta}$), the MLE of $g(\boldsymbol{\theta})$ is $g(\widehat{\boldsymbol{\theta}})$. This means we can choose the most convenient parameterization, or one having superior properties. Maximum likelihood estimation is also invariant under transformation of the observations. This can be seen from (A.30): the LHS density is $f_Y(y;\boldsymbol{\theta})$ and the RHS is $f_X(x(y);\boldsymbol{\theta}) \cdot |dx/dy|$ where dx/dy is independent of $\boldsymbol{\theta}$.

6. Under mild regularity conditions,

$$E\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right) \;=\; \mathbf{0}, \tag{A.17}$$

$$\boldsymbol{\mathcal{I}}_E(\boldsymbol{\theta}) \;=\; E\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\,\frac{\partial \ell}{\partial \boldsymbol{\theta}^T}\right) \;=\; -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T}\right) \;=\; -\,E\left(\frac{\partial}{\partial \boldsymbol{\theta}^T}\,\boldsymbol{U}\right). \tag{A.18}$$

7. Under regularity conditions, the score itself is asymptotically normal. In particular,

$$\boldsymbol{U}(\boldsymbol{\theta}_*) \;\sim\; N_p(\mathbf{0},\, \boldsymbol{\mathcal{I}}_E(\boldsymbol{\theta}_*)) \tag{A.19}$$

as $n \to \infty$.

A.1.3.1 The Cramér-Rao Inequality

A simplified version of the famous Cramér-Rao inequality is stated as follows. Under regularity conditions and i.i.d. conditions, for all n and unbiased estimators $\widehat{\boldsymbol{\theta}}_n$,

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}_n) - \boldsymbol{\mathcal{I}}_E^{-1}(\boldsymbol{\theta}) \tag{A.20}$$

is positive-semidefinite. It is usually stated for the one-parameter case only, in which case

$$\frac{1}{n\,\mathcal{I}_{E1}(\theta)} \;=\; \mathcal{I}_E^{-1}(\theta) \;\leq\; \mathrm{Var}(\widehat{\theta}_n). \tag{A.21}$$

That is, the inverse of the EIM (known as the Cramér-Rao lower bound; CRLB) is a lower bound for the variance of an unbiased estimator; it is used as a benchmark to compare the performance of any unbiased estimator. An approximation to the multiparameter case (A.20) is to apply (A.21) to each diagonal element of $\boldsymbol{\mathcal{I}}_E^{-1}(\boldsymbol{\theta})$.

For some models, equality in (A.21) can be attained, therefore that estimator is (fully) efficient, or *best*, or a *minimum variance unbiased estimator* (MVUE). For other models, there exists no unbiased estimator that achieves the lower bound. Typically, the MLE achieves the CRLB.

While unbiasedness of an estimator is considered a good thing for many people, a viable option is to consider biased estimators which have a lower mean-squared error

$$\mathsf{MSE} \;=\; E\left[\sum_{j=1}^{p}(\widehat{\theta}_j - \theta_{*j})^2\right] \;=\; E[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*\|^2] \;=\; \mathrm{trace}\{E[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)^T]\}.$$

The decomposition

$$\mathsf{MSE}(\widehat{\boldsymbol{\theta}}) \;=\; \mathrm{trace}\{\mathrm{Var}(\widehat{\boldsymbol{\theta}})\} + \|\mathrm{Bias}(\widehat{\boldsymbol{\theta}})\|^2 \tag{A.22}$$

is in contrast to the variance of the estimator with its bias $E(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_*$.

A.1.4 Inference

Based on the above properties, MLE provides confidence intervals/regions for estimated quantities, tests of goodness-of-fit, and tests for the comparison of models. Loosely, one can view confidence intervals/regions and hypothesis testing as two sides of the same coin. Our summary here will separate out the two. Sometimes we partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ where $p_j = \dim(\boldsymbol{\theta}_j)$, and treat $\boldsymbol{\theta}_2$ as a nuisance parameter. Let the true value of $\boldsymbol{\theta}_1$ be $\boldsymbol{\theta}_{*1}$.

A.1.4.1 Confidence Intervals and Regions

There are two common methods, although three are listed here to parallel the hypothesis testing case.

1. **Wald Test** Based on (A.16),

$$\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*\right)^T \mathbf{V}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*\right) \ \dot{\sim}\ \chi_p^2$$

in large samples. Here, $\mathbf{V}^{-1}$ is commonly chosen to be one of the following: (a) $\boldsymbol{\mathcal{I}}_E(\widehat{\boldsymbol{\theta}})$, (b) $\boldsymbol{\mathcal{I}}_O(\widehat{\boldsymbol{\theta}})$. The idea behind these is to use any consistent estimator, and both choices are equivalent to 1st-order approximation. Based on 2nd-order approximations and conditional arguments, Efron and Hinkley (1978) argued that the OIM is superior as an estimator of variance. As VGAM implements Fisher scoring, type (a) serves as the basis for the estimated variance-covariance matrix.

Based on the above, an approximate normal-theory $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}_1$ is the ellipsoid defined as the set of all $\boldsymbol{\theta}_{1*}$ satisfying

$$(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1*})^T \, \boldsymbol{\mathcal{I}}_E(\widehat{\boldsymbol{\theta}}_1) \, (\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1*}) \ \leq\ \chi^2_{p_1}(\alpha).$$

For VGAM, an approximate $100(1-\alpha)\%$ confidence interval for θ_j is given by

$$\widehat{\theta}_j \ \pm\ z(\alpha/2)\,\mathrm{SE}(\widehat{\theta}_j), \qquad (A.23)$$

where the SE derives from the EIM, which is of the form $(\mathbf{X}_{\text{VLM}}^T \mathbf{W}\,\mathbf{X}_{\text{VLM}})^{-1}$ (Eq. (3.21); see Sect. 3.2 for details).

2. **Score Test** Like the Wald test, confidence regions may be proposed which are based on a quadratic approximation to ℓ. Consequently, parameterizations which improve this approximation will give more accurate results, e.g., with the aid of parameter link functions. However, since the score test method is the least common of the three, no details are given here apart from a small mention in the hypothesis testing situation below.

3. **Likelihood Ratio Test (LRT)** Let the *profile likelihood* for $\boldsymbol{\theta}_1$ be

$$R(\boldsymbol{\theta}_1) \ = \ \max_{\boldsymbol{\theta}_2}\ L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)/L(\widehat{\boldsymbol{\theta}}).$$

Then the LR subset statistic $-2\log R(\boldsymbol{\theta}_{*1}) \sim \chi^2_{p_1}$ asymptotically, therefore an approximate $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}_1$ is the set of all $\boldsymbol{\theta}_{1*}$ such that

$$-2 \log R(\boldsymbol{\theta}_{1*}) \;<\; \chi^2_{p_1}(\alpha).$$

For a simple 1-parameter model, this reduces to the set of all θ values satisfying

$$2\left[\ell(\widehat{\theta};\boldsymbol{y}) - \ell(\theta;\boldsymbol{y})\right] \;\leq\; \chi^2_1(\alpha). \tag{A.24}$$

The methods function `confint.glm()` in MASS computes confidence intervals for each coefficient of a fitted GLM, based on the method of profile likelihoods. More generally, we can write the profile log-likelihood of $\boldsymbol{\theta}_1$ as $\ell_P(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1))$, where $\widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)$ is the MLE of $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$. Being of lower dimension, ℓ_P is often used for inference, e.g., if $\widehat{\boldsymbol{\theta}}_2(\theta_1)$ is easy.

A.1.4.2 Hypothesis Testing

For hypothesis testing, there are three well-known ways for tests of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is known and fixed. None of the tests are uniformly better, although the LRT is considered superior in many problems. Another advantage of the LRT is that it is invariant under nonlinear reparameterizations—this is not so for the Wald test, and for the score test, invariance depends on the choice of $\mathbf{V}$.

1. **Wald Test** Based on (A.16) and under the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T \mathbf{V}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \;\;\dot{\sim}\;\; \chi^2_p \tag{A.25}$$

in large samples. Here, $\mathbf{V}^{-1}$ is commonly chosen to be one of the following: (a) $\boldsymbol{\mathcal{I}}_E(\widehat{\boldsymbol{\theta}})$, (b) $\boldsymbol{\mathcal{I}}_O(\widehat{\boldsymbol{\theta}})$, (c) $\boldsymbol{\mathcal{I}}_E(\boldsymbol{\theta}_0)$, (d) $\boldsymbol{\mathcal{I}}_O(\boldsymbol{\theta}_0)$. The idea behind (a)–(b) is to use any consistent estimator.

This result can be extended to arbitrary linear combinations of $\boldsymbol{\theta}$. In particular, for the linear combination $\boldsymbol{e}_j^T\boldsymbol{\theta} = \theta_j$, and $\boldsymbol{\theta}_0 = 0$, we usually take the square root and obtain the Wald statistic for $H_0 : \theta_j = 0$

$$z_0 \;=\; \frac{\widehat{\theta}_j - 0}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_j)}} \;=\; \frac{\widehat{\theta}_j}{\mathrm{SE}(\widehat{\theta}_j)},$$

which is treated as a Z-statistic (or a t-ratio for LMs). One-sided tests are then accommodated, e.g., $H_1 : \theta_j < 0$ or $H_1 : \theta_j > 0$, in which case the p-values are $\Phi(z_0)$ and $\Phi(-z_0)$ provided $\widehat{\theta}_j < 0$ and $\widehat{\theta}_j > 0$, respectively [and $2\,\Phi(-|z_0|)$ for the 2-sided alternative $H_1 : \theta_j \neq 0$]. Alternatively, Z^2 may be treated as having an approximate χ^2_1 distribution. For VGLMs, VGAM prints out Wald statistics (usually type (a)) with the methods function `summary()`. The 4-column table of estimates, SEs, Wald statistics and p-values can be obtained by, e.g.,

```
> coef(summary(vglmObject))   # Entire table
> coef(summary(vglmObject))[, "Pr(>|z|)"] # p-values
```

Given a fitted model (including an LM or GLM) that has $\widehat{\boldsymbol{\theta}}$ and some estimate $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}})$ obtainable by `coef()` and `vcov()`, the function `linearHypothesis()` in car can test a system of linear hypotheses based on the Wald test.

2. **Score Test** Using the result (A.19) and under $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$\boldsymbol{U}(\boldsymbol{\theta}_0)^T \, \boldsymbol{\mathcal{I}}_E^{-1}(\boldsymbol{\theta}_0) \, \boldsymbol{U}(\boldsymbol{\theta}_0) \quad \sim \quad \chi^2_p \tag{A.26}$$

asymptotically (Rao, 1948). The EIM is evaluated at the hypothesized value $\boldsymbol{\theta}_0$, but at the MLE $\widehat{\boldsymbol{\theta}}$ is an alternative. Both versions of the test are valid; in fact, they are asymptotically equivalent. One advantage of using $\boldsymbol{\theta}_0$ is that calculation of the MLE may be bypassed. One disadvantage is that the test can be inconsistent (Freedman, 2007). In spite of their simplicity, score tests are not as commonly used as Wald and LR tests. Further information about score tests is at, e.g., Rao (1973). The package mdscore implements a modified score test for GLMs that offers improvements in accuracy when n is small.

3. **Likelihood Ratio Test (LRT)** This test is based on a comparison of maximized likelihoods for nested models. Suppose we are considering two models, $\mathcal{M}_1$ and $\mathcal{M}_2$ say, such that $\mathcal{M}_1 \subseteq \mathcal{M}_2$. That is, $\mathcal{M}_1$ is a subset or special case of $\mathcal{M}_2$. For example, one may obtain a simpler model $\mathcal{M}_1$ by setting some of the θ_j in $\mathcal{M}_2$ to zero, and we want to test the hypothesis that those elements are indeed zero.

   The basic idea is to compare the maximized likelihoods of the two models. The maximized likelihood under the smaller model $\mathcal{M}_1$ is

$$\sup_{\boldsymbol{\theta} \in \mathcal{M}_1} L(\boldsymbol{\theta}; \boldsymbol{y}) \;=\; L(\widehat{\boldsymbol{\theta}}_{\mathcal{M}_1}; \boldsymbol{y}),$$

   where $\widehat{\boldsymbol{\theta}}_{\mathcal{M}_1}$ is the MLE of $\boldsymbol{\theta}$ under model $\mathcal{M}_1$. Likewise, the maximized likelihood under the larger model $\mathcal{M}_2$ has the same form

$$\sup_{\boldsymbol{\theta} \in \mathcal{M}_2} L(\boldsymbol{\theta}; \boldsymbol{y}) \;=\; L(\widehat{\boldsymbol{\theta}}_{\mathcal{M}_2}; \boldsymbol{y}),$$

   where $\widehat{\boldsymbol{\theta}}_{\mathcal{M}_2}$ is the MLE of $\boldsymbol{\theta}$ under model $\mathcal{M}_2$. The ratio of these quantities,

$$\lambda \;=\; \frac{L(\widehat{\boldsymbol{\theta}}_{\mathcal{M}_1}; \boldsymbol{y})}{L(\widehat{\boldsymbol{\theta}}_{\mathcal{M}_2}; \boldsymbol{y})},$$

   lies in $[0, 1]$. Values close to 0 indicate that the smaller model is not acceptable compared to the larger model, while values close to unity indicate that the smaller model is almost as good as the large model.

   Under regularity conditions, the *likelihood ratio test statistic*

$$-2\log\lambda \;=\; 2\log L(\widehat{\boldsymbol{\theta}}_{\mathcal{M}_2}; \boldsymbol{y}) - 2\log L(\widehat{\boldsymbol{\theta}}_{\mathcal{M}_1}; \boldsymbol{y}) \;\to\; \chi^2_{\nu}$$

   where $\nu = \dim(\mathcal{M}_2) - \dim(\mathcal{M}_1)$, the difference in the number of parameters in the two models. When applied to GLMs, the LRT is also known as the deviance test.

   LRTs may be performed using `lrtest()`, e.g., for the following two `vglm()` objects where the simpler model is a special case of the more complex model,

```
> # Models must be nested:
> lrtest(Complex.model, Simpler.model)
```

   returns the LRT statistic and p-value.

In the above, the Wald and score tests were for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, however, hypothesis tests involving only a subset of $\dim(\boldsymbol{\theta}_0)$ parameters are easily handled: replace p in (A.25) and (A.26) by p_0 and choose the relevant submatrix of $\mathbf{V}^{-1}$.

All three tests are asymptotically equivalent, and therefore can be expected to give similar results in large samples. In small samples, simulation studies have suggested that LRTs are generally the best. Note that the calculation of a LRT requires fitting two models ($\mathcal{M}_1$ and $\mathcal{M}_2$), compared to only one model for the Wald test ($\mathcal{M}_2$), and sometimes no model at all for the score test. However, note that the Hauck-Donner phenomenon (Sect. 2.3.6.2) may affect the Wald test but not the LRT.

The three test statistics have an elegant geometric interpretation that is illustrated in Fig. A.2a,b for the single-parameter case. In a nutshell, the pertinent features are the horizontal and vertical distances between $\ell(\theta_0)$ and $\ell(\widehat{\theta})$, and the slope $\ell'(\theta_0)$. This example comes from a negative binomial $\text{NB}(\mu, k)$ distribution fitted to the `machinists` data set. The two plots are for $\theta = k$ and $\theta = \log k$. Here, $H_0 : k = \frac{1}{3}$, chosen for illustrative purposes only.

- The Wald test statistic is a function of $|\widehat{\theta} - \theta_0|$. Heuristically, the justification is to expand $\ell(\theta_0)$ about $\widehat{\theta}$ in a Taylor series under the assumption that the null hypothesis is true:

$$\ell(\theta_0) \;\approx\; \ell(\widehat{\theta}) + \frac{1}{2}\,\ell''(\widehat{\theta})(\theta_0 - \widehat{\theta})^2$$

  because $\ell'(\widehat{\theta}) = 0$ and $H_0 : \theta_* = \theta_0$. Then the Wald test statistic

$$(\theta_0 - \widehat{\theta})\,[-\ell''(\widehat{\theta})]\,(\theta_0 - \widehat{\theta}) \;\approx\; 2\{\ell(\widehat{\theta}) - \ell(\theta_0)\}$$

  i.e., approximates the LRT statistic. Here, choice (b) in (A.25) provides the metric. Expanded the way it appears here, the Wald test statistic is the squared *horizontal* distance after some standardization.

- The score test is a function of $\ell'(\theta_0)$, i.e., its *slope*. If $\widehat{\theta}$ approaches θ_0, then this derivative gets closer to 0, hence we would tend to reject the null hypothesis if the slope is very different from zero. Heuristically, it can be justified by expanding $\ell'(\widehat{\theta})$ about θ_0 in a Taylor series under the assumption that the null hypothesis is true:

$$\ell'(\widehat{\theta}) \;=\; 0 \;\approx\; \ell'(\theta_0) + \ell''(\theta_0)(\widehat{\theta} - \theta_0) + \frac{1}{2}\,\ell'''(\theta_0)(\widehat{\theta} - \theta_0)^2$$

  so that $(\widehat{\theta} - \theta_0) \;\approx\; \ell'(\theta_0)/\{-\ell''(\theta_0)\}$. Choosing choice (d) in (A.25), we can write $\sqrt{\boldsymbol{\mathcal{I}}_O(\theta_0)}\,(\widehat{\theta} - \theta_0) \approx \ell'(\theta_0)/\sqrt{\boldsymbol{\mathcal{I}}_O(\theta_0)}$. Both sides are approximately standard normally distributed. Upon squaring both sides,

$$(\widehat{\theta} - \theta_0)\,\boldsymbol{\mathcal{I}}_O(\theta_0)\,(\widehat{\theta} - \theta_0) \;=\; U(\theta_0)\,\boldsymbol{\mathcal{I}}_O(\theta_0)\,U(\theta_0)$$

  which is a Wald test statistic expressed in terms of the gradient at the hypothesized value.

- The LRT statistic is a function of $\ell(\widehat{\theta}) - \ell(\theta_0)$, in fact, it is simply twice that. This corresponds to the labelled *vertical* distance.

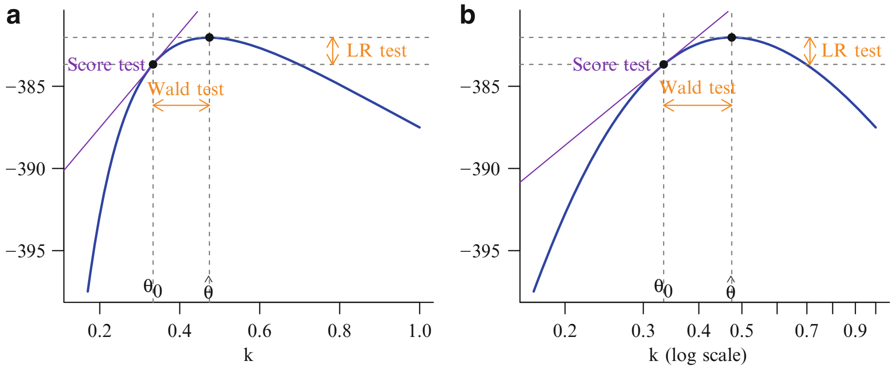

Fig. A.2 Negative binomial NB(μ, k) distribution fitted to the `machinists` data set. The y-axis is ℓ. Let $\theta = k$ and $\theta^* = \log k$. (**a**) $\ell(\theta)$ is the *solid blue curve*. (**b**) $\ell(\theta^*)$ is the *solid blue curve*. Note: for $H_0 : \theta = \theta_0$ (where $\theta_0 = \frac{1}{3}$), the likelihood-ratio test, score test and Wald test statistics are based on quantities highlighted with respect to ℓ. In particular, the score statistic is based on the tangent $\ell'(\theta_0)$.

Figure A.2b shows the same problem but under the reparameterization $\theta = \log k$. The log-likelihood is now more symmetric about $\widehat{\theta}$, i.e., its quadratic approximation is improved, therefore we would expect inferences to be more accurate compared to the first parameterization.

A.1.4.3 Delta Method

The *delta method* is a general method for obtaining approximate standard errors of functions of the parameter. Its basic idea is local linearization via derivatives. Let $\phi = g(\boldsymbol{\theta})$ be some function of the parameter. Apply a Taylor-series expansion about the true value:

$$\widehat{\phi} \;=\; g(\widehat{\boldsymbol{\theta}}) \;=\; g(\boldsymbol{\theta}_*) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)^T \, \frac{\partial g(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}} + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)^T \, \frac{\partial^2 g(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T} \, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) + \cdots ,$$

hence

$$\sqrt{n}\, \Big(g(\widehat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}_*)\Big) \;\approx\; \sqrt{n}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)^T (\partial g(\boldsymbol{\theta}_*)/\partial \boldsymbol{\theta}).$$

Consequently, from (A.14),

$$g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_*) \xrightarrow{\mathcal{D}} N_p\Big(\mathbf{0},\; (\partial g(\boldsymbol{\theta}_*)/\partial \boldsymbol{\theta}^T)\, \boldsymbol{\mathcal{I}}_E^{-1}(\boldsymbol{\theta}_*)\, (\partial g(\boldsymbol{\theta}_*)/\partial \boldsymbol{\theta})\Big). \quad \text{(A.27)}$$

To make use of this result, all quantities are computed at the MLE: for large n,

$$\mathrm{SE}(\widehat{\phi}) \;\approx\; \left\{ \sum_{j=1}^{p} \sum_{k=1}^{p} \frac{\partial g}{\partial \theta_j} \, \frac{\partial g}{\partial \theta_k} \, \widehat{v}_{jk} \right\}^{\frac{1}{2}} \;=\; \left\{ \frac{\partial g(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^T} \, \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}) \, \frac{\partial g(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right\}^{\frac{1}{2}}, \quad \text{(A.28)}$$

i.e., all the partial derivatives are evaluated at $\widehat{\boldsymbol{\theta}}$. In the case of $p = 1$ parameter, (A.28) reduces to

$$\mathrm{SE}(\widehat{\phi}) \;\approx\; \left|\frac{\mathrm{d}g}{\mathrm{d}\theta}\right| \sqrt{\widehat{v}_{11}}, \qquad \text{(A.29)}$$

where $\mathrm{d}g/\mathrm{d}\theta$ is evaluated at $\widehat{\theta}$.

For simple intercept-only models, VGAM uses the delta method in calls of the form `vcov(vglmObject, untransform = TRUE)`. This is possible because (A.29) is readily computed for models having the form $\eta_j = g_j(\theta_j) = \beta_{(j)1}$ for simple links. The accuracy of the method depends on the functional form of g_j and the precision of $\widehat{\theta}_j$.

A.2 Some Useful Formulas

A.2.1 Change of Variable Technique

Suppose that a random variable X has a known PDF $f_X(x)$, and $Y = g(X)$ is some transformation of X, where $g : \mathbb{R} \to \mathbb{R}$ is any differentiable monotonic function. That is, g is increasing or decreasing, therefore is invertible (one-to-one). Then the PDF of Y, by the change-of-variable formula, is

$$f_Y(y) \;=\; f_X\big(g^{-1}(y)\big) \cdot \left|\frac{d}{dy}\, g^{-1}(y)\right| \;=\; f_X(x(y)) \cdot \left|\frac{dx}{dy}\right|. \qquad \text{(A.30)}$$

A.2.2 Series Expansions

The following series expansions are useful, e.g., to work out the first and expected second derivatives of the GEV and GPD, as $\xi \to 0$:

$$\begin{aligned}
\log(1+z) &= z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \cdots \quad \text{for } |z| \leq 1 \;\text{ and } z \neq -1,\\
e^z &= \lim_{n\to\infty} \left(1 + \frac{z}{n}\right)^n,\\
(1+z)^\alpha &= 1 + \alpha z + \frac{\alpha(\alpha-1)}{2!} z^2 + \frac{\alpha(\alpha-1)(\alpha-2)}{3!} z^3 + \cdots, \quad \text{for } |z| \leq 1,\\
(1+x)^{-1} &= 1 - x + x^2 - x^3 + \cdots \quad \text{for } -1 < x < 1.
\end{aligned}$$

A.2.3 Order Notation

There are two types of Landau's O-notation which are convenient abbreviations for us.

A.2.3.1 For Algorithms

Here, the $O(\cdot)$ notation is mainly used to measure the approximate computational expense of algorithms, especially in terms of time and memory. For functions $f(n)$ and $g(n)$, we say $f(n) = O(g(n))$ if and only if there exists two (positive and finite) constants c and n_0 such that

$$|f(n)| \ \le \ c\,|g(n)| \tag{A.31}$$

for all $n \ge n_0$. For us, f and g are positive-valued, therefore (A.31) states that f does not increase faster than g. Saying that the computing time of an algorithm is $O(g(n))$ implies that its execution time takes no more than some constant multiplied by $g(n)$.

It can be shown that, e.g., $O(1) < O(\log\, n) < O(n) < O(n\, \log\, n) < O(n^2) < O(n^3) < O(2^n) < O(n!) < O(n^n)$. In any pairwise comparison, these inequalities usually do not hold in practice unless n is sufficiently large. As an example, the fastest known sorting algorithms for elements of a general n-vector cost $O(n \log n)$ whereas simpler algorithms such as bubble sort cost $O(n^2)$. Some people have suggested that usually an algorithm should be no more than $O(n \log n)$ to be practically manageable for very large data sets.

The so-called big-O notation, described above implicitly for integer n, is also useful and similarly defined for a real argument. For example, an estimator with an asymptotic bias of $O(h^2)$ has less asymptotic bias than another estimator whose asymptotic bias is $O(h)$, because $h \to 0^+$ as $n \to \infty$. Such considerations are made in, e.g., Sect. 2.4.6.2.

A.2.3.2 For Probabilities

In direct parallel with the above, the *order in probability* notation deals with convergence in probability of sets of random variables. A sequence of random variables $X_1, X_2, \ldots$ is said to *converge in probability* to the random variable X if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P[\,|X_n - X| > \varepsilon] \ = \ 0. \tag{A.32}$$

The random variable X is called the *probability limit* of X_n, and it is written plim $X_n = X$, or alternatively, as $X_n \xrightarrow{\mathcal{P}} X$.

Now if $\{X_n\}$ is a set of random variables and $\{a_n\}$ is a set of constants, then $X_n = O_p(a_n)$ if for all $\varepsilon > 0$, there exists a finite $N > 0$ such that

$$P\left[\left|\frac{X_n}{a_n}\right| > N\right] \ < \ \varepsilon, \tag{A.33}$$

for all n. If $X_n = O_p(a_n)$, then we say that X_n/a_n is *stochastically bounded*. As an example, we say that $\{X_n\}$ is at most of order in probability n^k if, for every $\varepsilon > 0$, there exists a real N so that $P[n^{-k}\,|X_n| > N] < \varepsilon$ for all n.

A.2.4 Conditional Expectations

Provided that all the expectations are finite, for random variables X and Y,

$$\begin{aligned} E(Y) &= E_X\{E(Y|X)\}\,, && \text{(A.34)} \\ E[g(Y)] &= E_X\{E[g(Y)|X]\} \;\; \text{(iterated expectation)}, && \text{(A.35)} \\ \mathrm{Var}(Y) &= E_X\{\mathrm{Var}(Y|X)\} + \mathrm{Var}_X\{E(Y|X)\} \;\; \text{(conditional variance)}. && \text{(A.36)} \end{aligned}$$

One application of these formulas is the beta-binomial distribution (Sect. 11.4).

A.2.5 Random Vectors

Here are some basic results regarding random vectors $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, i.e., vectors of random variables.

1. $E(\boldsymbol{X}) = \boldsymbol{\mu}_{\boldsymbol{X}}$, where the ith element of $\boldsymbol{\mu}_{\boldsymbol{X}}$ is $E(X_i)$. Similarly, $E(\boldsymbol{Y}) = \boldsymbol{\mu}_{\boldsymbol{Y}}$.
2. $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}})^T]$, with $\mathrm{Var}(\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X})$ ($= \boldsymbol{\Sigma}_{\boldsymbol{X}}$, say). We write $\boldsymbol{X} \sim (\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$.
3. $\mathrm{Cov}(\mathbf{A}\boldsymbol{X}, \mathbf{B}\boldsymbol{Y}) = \mathbf{A}\,\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})\,\mathbf{B}^T$ for conformable matrices $\mathbf{A}$ and $\mathbf{B}$ of constants.
4. $E[\boldsymbol{X}^T \mathbf{A} \boldsymbol{X}] = \boldsymbol{\mu}_{\boldsymbol{X}}^T \mathbf{A} \boldsymbol{\mu}_{\boldsymbol{X}} + \mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{X}})$.
5. $\mathrm{trace}(\mathbf{AB}) = \mathrm{trace}(\mathbf{BA})$ for conformable matrices $\mathbf{A}$ and $\mathbf{B}$.
6. $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}^T) = \mathrm{rank}(\mathbf{A}^T\mathbf{A}) = \mathrm{rank}(\mathbf{A}\mathbf{A}^T)$.
7. If $\mathbf{A}$ is $n \times n$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, then
$$\mathrm{trace}(\mathbf{A}) \;=\; \sum_{i=1}^{n} \lambda_i, \qquad \det(\mathbf{A}) \;=\; \prod_{i=1}^{n} \lambda_i.$$
8. A symmetric matrix $\mathbf{A}$ is positive-definite if $\boldsymbol{x}^T \mathbf{A} \boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \mathbf{0}$. Such matrices have positive eigenvalues, are invertible, and have a Cholesky decomposition that exists and is unique.

Some proofs for these can be found in, e.g., Seber and Lee (2003).

A.3 Some Linear Algebra

Least squares computations are usually based on orthogonal methods such as the QR factorization and singular value decomposition, because they are numerically more stable than naïve methods. They almost always give more accurate answers. A few details are given below.

A.3.1 Cholesky Decomposition

Given an $n \times n$ symmetric positive-definite matrix $\mathbf{A}$, its Cholesky decomposition $\mathbf{A} = \mathbf{U}^T\mathbf{U}$ where $\mathbf{U}$ is an upper-triangular matrix (i.e., $(\mathbf{U})_{ij} \equiv U_{ij} = 0$ for $i > j$) with positive diagonal elements. When $\mathbf{A}$ is 1×1, then $\mathbf{U}$ is just the square root of the element A_{11}. The computation of $\mathbf{U}$ might be written:

```
Iterate:    For $i = 1, \ldots, n$
  (i) $U_{ii} = \sqrt{A_{ii} - \sum_{k=1}^{i-1} U_{ki}^2}$
  (ii) Iterate:    For $j = i+1, \ldots, n$
          $U_{ij} = (A_{ij} - \sum_{k=1}^{i-1} U_{ki}\, U_{kj})/U_{ii}$
```

The first operation is to compute $U_{11} = \sqrt{A_{11}}$. The algorithm requires $\frac{1}{3}n^3 + O(n^2)$ flops, which is about half the cost of the more general **LU** decomposition (Gaussian elimination).

Solving the linear system of equations $\mathbf{A}\boldsymbol{x} = \boldsymbol{y}$ can be achieved by first solving $\mathbf{U}^T\boldsymbol{z} = \boldsymbol{y}$ by forward substitution, and then solving $\mathbf{U}\boldsymbol{x} = \boldsymbol{z}$ by backward substitution. Each of these steps requires $n^2 + O(n)$ flops. Forward substitution here might be written as

```
Iterate:    For $i = 1, \ldots, n$
    $z_i = (y_i - \sum_{k=1}^{i-1} U_{ki}\, z_k)/U_{ii}$
```

The first operation is to compute $z_1 = y_1/U_{11}$. Likewise, backward substitution here might be written as

```
Iterate:    For $i = n, \ldots, 1$
    $x_i = (z_i - \sum_{k=i+1}^{n} U_{ik}\, x_k)/U_{ii}$
```

The first operation is to compute $x_n = z_n/U_{nn}$.

A variant of the above is the *rational* Cholesky decomposition, which can be written $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T$, where $\mathbf{L}$ is a *unit* lower-triangular matrix, and $\mathbf{D}$ is a diagonal matrix with positive diagonal elements. By ‘unit’, we mean that the diagonal elements of $\mathbf{L}$ are all unity. This variant avoids computing n square roots in the usual algorithm, and should be used if $\mathbf{A}$ is banded with only a few bands, e.g., tridiagonal. (A matrix $\mathbf{T}$ is tridiagonal if $(\mathbf{T})_{ij} = 0$ for $|i - j| > 1$).

If $\mathbf{A}$ is a band matrix, with $(2m + 1)$ elements in its central band, then the Hutchinson and de Hoog (1985) algorithm is a method for computing the $2m + 1$ central bands of its inverse. The rational Cholesky decomposition of $\mathbf{A}$ has an $\mathbf{L}$ which is $(m + 1)$-banded, and the approximate cost is $\frac{1}{3}m^3 + nm^2 + O(m^2)$ flops. For cubic smoothing splines, $m = 2$ and the algorithm can be applied to compute the GCV.

Incidentally, a common method of measuring the width of a symmetric band matrix is by its half-bandwidth, e.g., $c = (2m+1)$ elements in its central band corresponds to a half-bandwidth of $(c+1)/2 = m+1$. Hence diagonal and tridiagonal matrices have half-bandwidths 1 and 2, etc.

A.3.2 Sherman-Morrison Formulas

If $\mathbf{A}$ is invertible, and $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors with $1 + \boldsymbol{v}^T \mathbf{A}^{-1} \boldsymbol{u} \neq 0$, then the Sherman-Morrison formula is

$$\left(\mathbf{A} + \boldsymbol{u}\boldsymbol{v}^T\right)^{-1} \;=\; \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\mathbf{A}^{-1}}{1 + \boldsymbol{v}^T\mathbf{A}^{-1}\boldsymbol{u}}. \tag{A.37}$$

If $\mathbf{A}$ is invertible, then the Sherman-Morrison-Woodbury formula is

$$(\mathbf{A} + \mathbf{U}\mathbf{V})^{-1} \;=\; \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{I} + \mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}\mathbf{A}^{-1}. \tag{A.38}$$

Incidentally, provided all inverses exist,

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} \;=\; \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{pmatrix} \tag{A.39}$$

where $\mathbf{A}^{11} = \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}$ or equivalently, $\mathbf{A}^{11} = \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1}$.

A.3.3 QR Method

The QR decomposition of an $n \times p$ matrix $\mathbf{X}$ with $n > p$ is

$$\mathbf{X} \;=\; \mathbf{Q}\mathbf{R} \;=\; (\mathbf{Q}_1\ \mathbf{Q}_2)\begin{pmatrix} \mathbf{R}_1 \\ \mathbf{O} \end{pmatrix} \;=\; \mathbf{Q}_1\mathbf{R}_1, \tag{A.40}$$

where $\mathbf{Q}$ $(n \times n)$ is orthogonal (i.e., $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_n$, or equivalently, $\mathbf{Q}^{-1} = \mathbf{Q}^T$) and $\mathbf{R}_1$ $(p \times p)$ is upper triangular.

In R, the function `qr()` computes the QR factorization, and there are associated functions such as `qr.coef()`, `qr.qty()`, `qr.Q()` and `qr.R()`. These functions are based on LINPACK by default, but there is a logical argument for `qr()` in the form of `LAPACK = FALSE` that can be set to `TRUE` to call LAPACK instead. One can think of LAPACK (Anderson et al., 1999) as a more modern version of LINPACK (Dongarra et al., 1979).

Given a rank-p model matrix $\mathbf{X}$, solving the normal equations (2.6) by the QR method means that the OLS estimate $\widehat{\boldsymbol{\beta}} = \mathbf{R}_1^{-1}\mathbf{Q}_1^T\boldsymbol{y}$ is easily computed because `qr.qty()` returns $\mathbf{Q}_1^T\boldsymbol{y}$, and back substitution can be then used. As $\mathbf{X}$ is of full column-rank, all the diagonal elements of $\mathbf{R}_1$ are nonzero (positive by convention, actually).

It is easily verified that if the diagonal elements of $\mathbf{R}_1$ are positive (trivially achieved by negating certain columns of $\mathbf{Q}_1$ if necessary) then $\mathbf{R}_1$ corresponds to the Cholesky decomposition of $\mathbf{X}^T\mathbf{X}$, i.e., $\mathbf{X}^T\mathbf{X} = \mathbf{R}_1^T\mathbf{R}_1$. But the QR decomposition is the preferred method for computing $\widehat{\boldsymbol{\beta}}$ because there is no need to compute the sum-of-squares and cross-products matrix $\mathbf{X}^T\mathbf{X}$—doing so squares the condition number, so that if the columns of $\mathbf{X}$ are almost linearly dependent, then there will be a loss of accuracy. In general, orthogonal methods do not exacerbate ill-conditioned matrices.

For large n and p, the cost of performing a QR decomposition on $\mathbf{X}$ using Householder reflections[1] is approximately $2np^2$ floating point operations. This is about twice the cost of solving the normal equations by Cholesky when $n \gg p$.

A.3.4 Singular Value Decomposition

The singular value decomposition (SVD) of $\mathbf{X}$ as above is

$$\mathbf{X} \;=\; \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{A.41}$$

where $\mathbf{U}$ $(n \times p)$ is such that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$, and $\mathbf{V}$ $(p \times p)$ is orthogonal, and $\mathbf{D}$ is a $p \times p$ diagonal matrix with non-negative elements d_{ii} (called the *singular values*). The matrix $\mathbf{U}$ here comprises the first p columns of an orthogonal matrix, much like $\mathbf{Q}_1$ does to $\mathbf{Q}$ in (A.40).

It is easy to show that the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are d_{ii}^2, and it is usual to sort the singular values so that $d_{11} \geq d_{22} \geq \cdots \geq d_{pp} \geq 0$. With this enumeration, the eigenvectors of $\mathbf{X}^T\mathbf{X}$ make up the columns of $\mathbf{V}$, and the first p eigenvectors of $\mathbf{X}\mathbf{X}^T$ make up the columns of $\mathbf{U}$. A common method for determining the rank of $\mathbf{X}$ is to count the number of nonzero singular values, however, comparisons with 0 are made in light of the machine precision, i.e, `.Machine$double.eps`. In R, `svd()` computes the SVD by LAPACK, and the cost is approximately $6np^2 + 11p^3$ flops—which can be substantially more expensive than the QR decomposition.

A special case of the SVD is when $\mathbf{X}$ is square, symmetric and positive-definite. Then its SVD can be written as

$$\mathbf{X} \;=\; \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T, \tag{A.42}$$

where $\boldsymbol{\Lambda}$ has the sorted eigenvalues of $\mathbf{X}$ along its diagonal, and $\mathbf{P}$ is orthogonal with the respective eigenvectors of $\mathbf{X}$ defining its columns. Equation (A.42) is known as the *spectral decomposition* or eigendecomposition of $\mathbf{X}$, and a useful consequence is that powers of $\mathbf{X}$ have the simple form

$$\mathbf{X}^s \;=\; \mathbf{P}\boldsymbol{\Lambda}^s\mathbf{P}^T, \tag{A.43}$$

e.g., $s = \pm\frac{1}{2}$ especially.

A.4 Some Special Functions

Many densities or their log-likelihoods are expressed in terms of special functions. A few of the more common ones are mentioned here.

[1] Ex. A.6; another common algorithm by Givens rotations entails an extra cost of about 50%

A.4.1 Gamma, Digamma and Trigamma Functions

The gamma function is defined for $x > 0$ as

$$\Gamma(x) \;=\; \int_0^\infty t^{x-1}\, e^{-t}\, dt \tag{A.44}$$

and can be computed by `gamma(x)`, and its logarithm by `lgamma(x)`. For positive integer a,

$$\Gamma(a+1) \;=\; a\,\Gamma(a) \;=\; a! \tag{A.45}$$

and Stirling's approximation for large x is

$$\Gamma(x+1) \;\sim\; \sqrt{2\pi x}\, x^x\, e^{-x}. \tag{A.46}$$

A useful limit is

$$\lim_{n\to\infty} \frac{\Gamma(n+\alpha)}{\Gamma(n)\, n^\alpha} \;=\; 1 \qquad \forall\alpha \in \mathbb{R}. \tag{A.47}$$

The incomplete gamma function

$$P(a,x) \;=\; \frac{1}{\Gamma(a)} \int_0^x t^{a-1}\, e^{-t}\, dt \tag{A.48}$$

may be evaluated by `pgamma(x, a)`.

Derivatives of the log-gamma function are often encountered in discrete and continuous distributions. For such, define $\psi(x) = \Gamma'(x)/\Gamma(x)$ as the digamma function, and $\psi'(x)$ as the trigamma function.

For the digamma function, since $\psi(x+1) = \psi(x) + x^{-1}$, it follows that for integer $a \geq 2$,

$$\psi(a) \;=\; -\gamma + \sum_{i=1}^{a-1} i^{-1} \;\text{ where }\; -\psi(1) \;=\; \gamma \;\approx\; 0.5772 \tag{A.49}$$

is the Euler–Mascheroni constant. For large x, a series expansion for the digamma function is

$$\psi(x) \;=\; \log x - \frac{1}{2\,x} + \sum_{k=1}^{\infty} \frac{B_{2k}}{2\,k\,x^{2k}} \;=\; \log x - \frac{1}{2\,x} - \frac{1}{12\,x^2} + \cdots\,, \tag{A.50}$$

where B_k is the kth Bernoulli number.

For the trigamma function, since $\psi'(x+1) = \psi'(x) - x^{-2}$, it follows that for integer $a \geq 2$, $\psi'(a) = \pi^2/6 - \sum_{i=1}^{a-1} i^{-2}$ because $\psi'(1) = \pi^2/6$. For large x, a series expansion for the trigamma function is

$$\psi'(x) \;=\; \frac{1}{x} + \frac{1}{2\,x^2} + \sum_{k=1}^{\infty} \frac{B_{2k}}{x^{2k+1}} \;=\; \frac{1}{x} + \frac{1}{2\,x^2} + \frac{1}{6\,x^3} - \frac{1}{30\,x^5} + \cdots\,. \tag{A.51}$$

Higher-order derivatives of $\psi(x)$ may be computed by `psigamma()`.

A.4.2 Beta Function

The beta function is defined as

$$Be(a,b) \;=\; \int_0^1 t^{a-1}\,(1-t)^{b-1}\,dt, \qquad 0<a,\;0<b. \tag{A.52}$$

Then

$$Be(a,b) \;=\; \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}. \tag{A.53}$$

The incomplete beta function is

$$I_x(a,b) \;=\; \frac{Be_x(a,b)}{Be(a,b)}, \tag{A.54}$$

where

$$Be_x(a,b) \;=\; \int_0^x t^{a-1}\,(1-t)^{b-1}\,dt. \tag{A.55}$$

The function $I_x(a,b)$ can be evaluated by `pbeta(x, a, b)`.

A.4.3 The Riemann Zeta Function

The Riemann zeta function is defined by

$$\zeta(s) \;=\; \sum_{n=1}^{\infty} n^{-s}, \qquad \Re(s)>1. \tag{A.56}$$

Analytic continuation via

$$\zeta(s) \;=\; 2^s\,\pi^{s-1}\,\sin(\pi s/2)\,\Gamma(1-s)\,\zeta(1-s)$$

implies that it can be defined for all $\Re(s)$, with $\zeta(1)=\infty$. Some special values are $\zeta(2)=\pi^2/6$, and $\zeta(4)=\pi^4/90$. Euler found that for integer $n\geq 2$, $\zeta(2n)=A_{2n}$ where A_{2n} is rational. Indeed, $A_{2n}=\frac{1}{2}(-1)^{n+1}\,B_{2n}\,(2\pi)^{2n}/(2n)!$ in terms of the Bernoulli numbers.

A.4.4 Erf and Erfc

The *error function*, `erf(x)`, is defined for all x as

$$\frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)\,dt, \tag{A.57}$$

therefore is closely related to the CDF $\Phi(\cdot)$ of the standard normal distribution. The inverse function is defined for $x\in[-1,1]$, i.e., `erf(x, inverse = TRUE)`.

The *complementary error function*, `erfc(x)`, is defined as `1-erf(x)`. Its inverse function is defined for $x \in [0, 2]$.

A.4.5 The Marcum Q-Function

The (generalized) Marcum Q-function is defined as

$$\begin{aligned} Q_m(a,b) &= \int_b^\infty x \left(\frac{x}{a}\right)^{m-1} \exp\left\{-\frac{x^2+a^2}{2}\right\} I_{m-1}(ax)\,dx \\ &= \exp\left\{-\frac{a^2+b^2}{2}\right\} \sum_{k=1-m}^{\infty} \left(\frac{a}{b}\right)^k I_k(ab) \end{aligned} \tag{A.58}$$

where $a \geq 0$, $b \geq 0$ and m is a positive integer. Here, I_{m-1} is a modified Bessel function of the first kind of order $m-1$ (as in Table A.1). The Marcum Q-function is used, e.g., as a CDF for noncentral chi-squared and Rice distributions, i.e., `price()`.

The case $m = 1$ is known as the ordinary Marcum Q-function.

A.4.6 Exponential Integral, Debye Function

The exponential integral, which is defined for real x, can be computed by `expint()` and is

$$Ei(x) = \int_{-\infty}^{x} t^{-1}\, e^{t}\, dt, \quad x \neq 0. \tag{A.59}$$

The function `expexpint()` computes $e^{-x}\, Ei(x)$, and `expint.E1()` computes

$$E_1(x) = \int_x^\infty t^{-1}\, e^{-t}\, dt, \quad x \geq 0. \tag{A.60}$$

The Debye function $D_n(x)$ is defined as

$$D_n(x) = \frac{n}{x^n} \int_0^x \frac{t^n}{e^t - 1}\, dt \tag{A.61}$$

for $x \geq 0$ and $n = 0, 1, 2, 3, \ldots$.

A.4.7 Bessel Functions

Bessel functions appear widely in probability and statistics, e.g., distributions for directional data such as those defined on circles and spheres, Poisson processes and distributions (the most notable being the difference of two Poisson distributions, called the Skellam distribution). Of the various kinds, Table A.1 lists the most relevant ones relating to Chaps. 11–12.

Table A.1 Bessel functions (modified and unmodified) of order ν. The order `nu` may be fractional.

Function	Formula	R function	Name
$I_\nu(x)$	$\sum_{m=0}^{\infty} \frac{1}{m!\,\Gamma(m+\nu+1)} \left(\frac{x}{2}\right)^{2m+\nu}$	`besselI(x, nu)`	Modified Bessel function of the first kind
$K_\nu(x)$	$\lim_{\lambda\to\nu} \frac{\pi}{2} \frac{I_{-\lambda}(x) - I_\lambda(x)}{\sin(\lambda\pi)}$	`besselK(x, nu)`	Modified Bessel function of the third kind
$J_\nu(x)$	$\sum_{m=0}^{\infty} \frac{(-1)^m}{m!\,\Gamma(m+\nu+1)} \left(\frac{x}{2}\right)^{2m+\nu}$	`besselJ(x, nu)`	Bessel function of the first kind
$Y_\nu(x)$	$\lim_{\lambda\to\nu} \frac{J_\lambda(x)\cos(\lambda\pi) - J_{-\lambda}(x)}{\sin(\lambda\pi)}$	`besselY(x, nu)`	Bessel function of the second kind (Weber's function)

Bibliographic Notes

There are multitudes of books covering statistical inference and likelihood theory in detail, e.g., Edwards (1972), Rao (1973), Cox and Hinkley (1974), Silvey (1975), Barndorff-Nielsen and Cox (1994), Lindsey (1996), Welsh (1996), Severini (2000), Owen (2001), Casella and Berger (2002), Young and Smith (2005), Boos and Stefanski (2013). Most texts on mathematical statistics include at least a chapter on MLE, e.g., Knight (2000), Bickel and Doksum (2001), Shao (2003). Another book on statistical inference, which is compact and is concentrated on concepts, is Cox (2006). Hypothesis testing is treated in detail in Lehmann and Romano (2005).

A readable and applied account of models based on ML estimation is Azzalini (1996). Another applied book based on likelihood is Clayton and Hills (1993). GLMs are covered in detail in McCullagh and Nelder (1989); see also Lindsey (1997), Dobson and Barnett (2008). There have been a number of extensions of GLMs proposed. One of them, called "multivariate GLMs" by Fahrmeir and Tutz (2001, Sect.3.1.4). Another is the idea of composite link functions (Thompson and Baker, 1981). Standard texts for GAMs are Hastie and Tibshirani (1990) and Wood (2006).

A comprehensive account on many aspects of linear algebra, both theoretically and numerically, is Hogben (2014). Another, Golub and Van Loan (2013), remains an authoritative reference on matrix computations.

Detailed treatments of many special functions can be found in, e.g., Abramowitz and Stegun (1964), Gil et al. (2007), Olver et al. (2010).

Exercises

Ex. A.1. Let $\mathbf{A}$ and $\mathbf{B}$ be general $n \times n$ matrices, and $\boldsymbol{x}$ and $\boldsymbol{y}$ be general n-vectors. Work out the cost (expressed in $O(\cdot)$ complexity) of computing the following quantities in terms of the number of multiplications and additions, e.g., $n(n-1) = n^2 + O(n)$ multiplications, $n-1 = n + O(1)$ additions.

(a) $\mathbf{A} + \mathbf{B}$,
(b) $5\,\mathbf{A}$,
(c) $\boldsymbol{x}^T \boldsymbol{y}$,
(d) $\mathbf{A}\,\boldsymbol{x}$,
(e) $\boldsymbol{x}^T \mathbf{A}\,\boldsymbol{x}$,
(f) $\mathbf{AB}$,
(g) $\mathrm{trace}(\mathbf{A})$,
(h) $\mathrm{trace}(\mathbf{A}^T\mathbf{A})$.
(i) Which is cheaper for computing $\mathbf{AB}\boldsymbol{x}$: $\mathbf{A}(\mathbf{B}\boldsymbol{x})$ or $(\mathbf{AB})\boldsymbol{x}$? By how much?

Ex. A.2. Prove that if $f_1 = O(g_1)$ and $f_2 = O(g_2)$ then $f_1 \cdot f_2 = O(g_1 \cdot g_2)$.

Ex. A.3. The R function `sort()`, by default, uses an algorithm called Shellsort. There are variants of this algorithm, but suppose the running time is $O(n^{4/3})$. Suppose it takes 2.4 seconds to sort 2 million (random) observations on a certain machine. Very crudely, how long might it be expected to sort 11 million (random) observations on that machine?

Ex. A.4. Use the results of Sect. A.2.4 to derive the mean and variance of Y_i^* for the beta-binomial distribution, i.e., (11.13).

Ex. A.5. From Sect. A.3.2, if $\mathbf{K}$ is a positive-definite matrix, show that

$$\left(\mathbf{I} + \mathbf{TKT}^T\right)^{-1} \;=\; \mathbf{I} - \mathbf{T}\left(\mathbf{K}^{-1} + \mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T. \tag{A.62}$$

Ex. A.6. **QR Factorization by the Householder Reflections**
Suppose $\mathbf{X}$ is $n \times p$ with $n > p$ and of rank p. A Householder matrix is of the form

$$\boldsymbol{\mathcal{P}} \;=\; \mathbf{I}_n - \frac{2\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}} \tag{A.63}$$

for some n-vector $\boldsymbol{v} \neq \mathbf{0}$.

(a) Show that $\boldsymbol{\mathcal{P}}$ is symmetric and orthogonal.
(b) If $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{y}$ with $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2$, show that $\boldsymbol{\mathcal{P}}\boldsymbol{x} = \boldsymbol{y}$.
(c) Let $\boldsymbol{x}_{(1)}$ be the first column of $\mathbf{X}$. Suppose we want to choose $\boldsymbol{v}$ so that $\boldsymbol{\mathcal{P}}\boldsymbol{x}_{(1)} = c\,\boldsymbol{e}_1$ for some $c \neq 0$. Show that selecting $\boldsymbol{v} = \boldsymbol{x}_{(1)} + \alpha\boldsymbol{e}_1$ with $\alpha = \pm\|\boldsymbol{x}_{(1)}\|_2$ will achieve this. Given the choice of the sign of α, why is $\alpha = \mathrm{sign}(x_{11}) \cdot \|\boldsymbol{x}_{(1)}\|_2$ the better choice?
(d) Now for the kth column of $\mathbf{X}$, suppose we want to annihilate elements below the kth diagonal, leaving elements above the kth diagonal unchanged. Let $\boldsymbol{x}_{(k)} = (\boldsymbol{x}_{(k)}^{*T}, \boldsymbol{x}_{(k)}^{**T})^T$ be the kth column of $\mathbf{X}$, for $k = 2, \ldots, p$, where the first element of $\boldsymbol{x}_{(k)}^{**}$ is the diagonal element x_{kk}. We want to choose $\boldsymbol{v}_k$

so that $\boldsymbol{\mathcal{P}}_k \, \boldsymbol{x}_{(k)} = (\boldsymbol{x}^{*T}_{(k)}, c_k, \mathbf{0}^T_{n-k})^T$ for some $c_k \neq 0$. Show that selecting $\boldsymbol{v}_k = (\mathbf{0}^T_{k-1}, x_{kk} + \alpha_k, \boldsymbol{x}^{**T}_{(k)[-1]})^T$ with $\alpha_k = \pm \|\boldsymbol{x}^{**}_{(k)}\|_2$ achieves this.

(e) Show that the product of two orthogonal matrices is orthogonal.

(f) Deduce that $\mathbf{Q}_1$ comprises the first p columns of the product $\boldsymbol{\mathcal{P}}_1 \boldsymbol{\mathcal{P}}_2 \cdots \boldsymbol{\mathcal{P}}_p$, and that $\mathbf{R} = \boldsymbol{\mathcal{P}}_p \cdots \boldsymbol{\mathcal{P}}_2 \boldsymbol{\mathcal{P}}_1 \mathbf{X}$, in the QR factorization (A.40) of $\mathbf{X}$.

Ex. A.7. QR Factorization and Hilbert Matrices
Hilbert matrices, which are defined by $(\mathbf{X})_{ij} = (i + j - 1)^{-1}$ for $i, j = 1, \ldots, n$, are notorious for being ill-conditioned for n as little as 8 or 9. Compute the QR decomposition of the 8×4 left submatrix of an order-8 Hilbert matrix by explicitly computing the Householder matrices $\boldsymbol{\mathcal{P}}_1, \boldsymbol{\mathcal{P}}_2, \boldsymbol{\mathcal{P}}_3, \boldsymbol{\mathcal{P}}_4$ described in the previous exercise. Then check your answer with `qr()`.

Ex. A.8. QR Method and Weighted Least Squares

(a) Extend the algorithm for estimating the OLS $\widehat{\boldsymbol{\beta}}$ by the QR method to handle WLS.

(b) For (a), how can $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})$ be computed?

Ex. A.9. Show that

(a) the inverse of a nonsingular upper triangular matrix is also upper triangular,

(b) the product of two upper triangular matrices is upper triangular.

Ex. A.10. Express the error function (A.57), and its inverse, in terms of $\Phi(\cdot)$ or $\Phi^{-1}(\cdot)$.

Ex. A.11. Consider the log-gamma function. Show that $\log \Gamma(y+a) - \log \Gamma(y) \sim a \log y$ as $y \to \infty$, where $0 < a \ll y$.

Ex. A.12. Digamma Function

(a) Verify the recurrence formula $\psi(z+1) = \psi(z) + z^{-1}$.

(b) The digamma function has a single root on the positive real line. Apply the Newton-Raphson algorithm (A.11) to compute this root to at least 10 decimal places.

Ex. A.13. Derive the score vector and EIM for the following distributions, to show that they involve digamma and trigamma functions.

(a) The log-F distribution (`logF()`).

(b) The Dirichlet distribution (`dirichlet()`).

Everything comes to an end which has a beginning.
—Marcus Fabius Quintilianus

Glossary

See Tables A.2, A.3, A.4, A.5.

T.W. Yee, *Vector Generalized Linear and Additive Models*,
Springer Series in Statistics, DOI 10.1007/978-1-4939-2818-7

Table A.2 Summary of some notation used throughout the book. Some R commands are given.

Notation	Comments
μ	Mean
$\widetilde{\mu}$	Median
$u_+ = \max(u, 0)$	Positive part of u, with $u_+^p = (u_+)^p$ and not $(u^p)_+$, `pmax(u, 0)`
$u_- = -\min(u, 0)$	Negative part of u, so that $u = u_+ - u_-$ & $\|u\| = u_+ + u_-$, `-pmin(u, 0)`
$\lfloor u \rfloor$	Floor of u, the largest integer not greater than u, e.g., $\lfloor 28.1 \rfloor = 28$, `floor(u)`
$\lceil u \rceil$	Ceiling of u, the smallest integer not less than u, e.g., $\lceil 28.1 \rceil = 29$, `ceiling(u)`
$\text{sign}(u)$	Sign of u, -1 if $u < 0$, $+1$ if $u > 0$, 0 if $u = 0$, `sign(u)`
$I(\text{statement})$	Indicator function, 1/0 if statement is true/false, `as.numeric(statement)`
$\mathbb{C}$	Complex plane (excluding infinity), with $\Re(z)$ = the real part of z
$\mathbb{N}^0$	Set of all nonnegative integers, $0(1)\infty$
$\mathbb{N}^+$	Set of all positive integers, $1(1)\infty$
$\mathbb{R}$	Real line (excluding infinity), i.e., $(-\infty, \infty)$
$\mathbb{Z}$	Set of all integers
$a(b)c$	$\{a, a+b, a+2b, \ldots, c\}$; `seq(a, c, by = b)`
$\|\boldsymbol{x}\|_p$	$(\sum_i \|x_i\|^p)^{1/p}$, the p-norm of $\boldsymbol{x}$, so that $\|\boldsymbol{x}\|_\infty = \max(\|x_1\|, \|x_2\|, \ldots)$. By default, $p = 2$ so that $\|\boldsymbol{x}\|$ is the length of $\boldsymbol{x}$
$\|\boldsymbol{x} - \boldsymbol{y}\|$	Euclidean distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e., $\sqrt{(\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y})}$, `norm(x - y, "2")`
$\mathbf{1}_M$	M-vector of 1s, `rep(1, M)`
$\mathbf{0}_n$	n-vector of 0s, `rep(0, n)`
$\boldsymbol{e}_i$	$(0, \ldots, 0, 1, 0, \ldots, 0)^T$, a vector of zeros, but with a one in the ith position, `diag(n)[, i, drop = FALSE]`
`ncol`$(\mathbf{A})$	Number of columns of matrix $\mathbf{A}$, `ncol(A)`. And $\mathcal{R}_k$ = `ncol`$(\mathbf{H}_k)$
$\text{vec}(\mathbf{A})$	Vectorization of matrix $\mathbf{A}$ by columns, $(\boldsymbol{a}_1^T, \ldots, \boldsymbol{a}_n^T)^T$, `c(A)`
$\boldsymbol{x}_{[-1]i}$	The vector $\boldsymbol{x}_i$ with the first element deleted, `x[-1]`
$\mathbf{B}_{[-1,]}$	The matrix $\mathbf{B}$ with the first row deleted, `B[-1, ]`
$\mathbf{B}_{[,-1]}$	The matrix $\mathbf{B}$ with the first column deleted, `B[, -1]`
$\otimes$	Kronecker product, $\mathbf{A} \otimes \mathbf{B} = [(a_{ij}\mathbf{B})]$, `kronecker(A, B)`
$\circ$	Hadamard (element-by-element) product, $(\mathbf{A} \circ \mathbf{B})_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}$, `A * B`

Table A.3 Summary of further notation used throughout the book.

Notation	Comments
$\sim$	Is distributed as
$\sim$	Is asymptotically equivalent to, or converges to (e.g., (2.75), (2.79))
$\dot{\sim}$	Is approximately distributed as
$\xrightarrow{\mathcal{D}}$	Convergence in distribution, i.e., $\{Y_i\} \xrightarrow{\mathcal{D}} Y$ if $\lim_{n\to\infty} F_n(y) = F_Y(y)$ for all y where F_Y is continuous (Y_i has CDF F_i)
$\xrightarrow{\mathcal{P}}$	Convergence in probability, (A.32)
$\phi(z)$	PDF of a standard normal, $N(\mu = 0, \sigma^2 = 1)$, $(2\pi)^{-\frac{1}{2}}\, e^{-\frac{1}{2}z^2}$ for $z \in \mathbb{R}$, `dnorm(z)`
$\Phi(z)$	CDF of a standard normal, `pnorm(z)`
$z(\alpha)$	$(1-\alpha)$-quantile of $N(0,1)$, i.e., `qnorm(1-alpha)`, `qnorm(alpha, lower.tail = FALSE)`
$\chi^2_\nu(\alpha)$	$(1-\alpha)$-quantile of a chi-square distribution with ν degrees of freedom, i.e., `qchisq(1-alpha, df = nu)`, `qchisq(alpha, df = nu, lower.tail = FALSE)`
$t_\nu(\alpha)$	$(1-\alpha)$-quantile of a Student t distribution with ν degrees of freedom, i.e., `qt(1-alpha, df = nu)`, `qt(alpha, df = nu, lower.tail = FALSE)`
iff	If and only if, i.e., a necessary and sufficient condition, $\Longleftrightarrow$
$\log x$	Natural logarithm, $\log_e$, ln, `log(x)`
$\Gamma(x)$	Gamma function $\int_0^\infty t^{x-1}\, e^{-t}\, dt$ for $x > 0$, Sect. A.4.1, `gamma(x)`
$\psi(x) = \Gamma'(x)/\Gamma(x)$	Digamma function, $d \log \Gamma(x)/dx$, `digamma(x)`
$\psi'(x)$	Trigamma function, `trigamma(x)`
$\gamma = -\psi(1)$	Euler–Mascheroni constant, ≈ 0.57722, `-digamma(1)`
Cauchy sequence	A sequence $\{\boldsymbol{x}_n\}$ in a vector space $\mathcal{V}$ satisfying: given any $\varepsilon > 0$, $\exists N \in \mathbb{N}^+$ such that $\lVert \boldsymbol{x}_m - \boldsymbol{x}_n \rVert \le \varepsilon$ whenever $m, n \ge N$
$\mathcal{L}_2(a,b)$	$\{f \;:\; f$ is a Lebesgue square integrable function on $(a,b)\}$, i.e., $\int_a^b \lvert f(t)\rvert^2\, dt < \infty$. For $(a,b) = \mathbb{R}$, we write $\mathcal{L}_2$
$\mathcal{C}^k[a,b]$	$\{f : f', f'', \dots, f^{(k)}$ all exist and are continuous on $[a,b]\}$. Note that $f \in \mathcal{C}^k[a,b]$ implies that $f \in \mathcal{C}^{k-1}[a,b]$. Also, $\mathcal{C}[a,b] \equiv \mathcal{C}^0[a,b] = \{f(t) : f(t)$ continuous and real valued for $a \le t \le b\}$
$\mathcal{W}_2^m[a,b]$	A *Sobolev space* of order m is $\{f : f^{(j)}, j = 0, \dots, m-1$, are absolutely continuous on $[a,b]$, and $f^{(m)} \in \mathcal{L}_2[a,b]\}$
f absolutely continuous on $[a,b]$	$\forall\, \varepsilon > 0$, $\exists\, \delta > 0$ such that $\sum\limits_{i=1}^n \lvert f(x_i') - f(x_i)\rvert < \varepsilon$ whenever $\{[x_i, x_i'] : i = 1, \dots, n\}$ is a finite collection of mutually disjoint subintervals of $[a,b]$ with $\sum\limits_{i=1}^n \lvert x_i - x_i'\rvert < \delta$. That is, f is differentiable almost everywhere and equals the integral of its derivative
$l_p(\mathbb{R}^n)$	$\{\boldsymbol{x} = (x_1, \dots, x_n)^T : (\sum_{i=1}^n \lvert x_i\rvert^p)^{1/p} < \infty$ for $1 \le p < \infty\}$
Convex function $f : \mathcal{X} \to \mathbb{R}$	$f(tx_1 + (1-t)x_2) \le t\, f(x_1) + (1-t)\, f(x_2)$ $\forall t \in [0,1]$ and $x_1, x_2 \in \mathcal{X}$, e.g., x^2 and e^x on $\mathbb{R}$. A sufficient condition is that $f''(x) > 0$ $\forall x \in \mathcal{X}$
Concave function $f : \mathcal{X} \to \mathbb{R}$	$f(tx_1 + (1-t)x_2) \ge t\, f(x_1) + (1-t)\, f(x_2)$ $\forall t \in [0,1]$ and $x_1, x_2 \in \mathcal{X}$, e.g., $\sqrt{x}$ and $\log x$ on $(0,\infty)$. A sufficient condition is that $f''(x) < 0$ $\forall x \in \mathcal{X}$

Table A.4 Summary of some quantities. Data is $(y_i, \boldsymbol{x}_i)$ for $i = 1, \ldots, n$. See also Table 8.5. The indices $i = 1, \ldots, n$, $j = 1, \ldots, M$, $k = 1, \ldots, p$, $s = 1, \ldots, S$, $q = 1, \ldots, Q$. Starred quantities are estimated, as well as $\mathbf{C}$ and $\mathbf{A}$.

Notation	Comments
S	Number of responses. If $S > 1$ then these are "multiple responses"
M_1	Number of η_j for a single response
M	Number of η_j (summed over all S responses), e.g., $M = M_1 S$
Q_1	$\dim(\boldsymbol{y}_i)$ for a single response, hence $Q = Q_1 S$
$\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T = (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(Q)})$	Response matrix, is $n \times Q$
$\mathbf{X} = \mathbf{X}_{\text{LM}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)})$	LM (model) matrix $[(x_{ik})]$, is $n \times p$ $(= n_{\text{LM}} \times p_{\text{LM}})$; $\boldsymbol{x}^{(1)} = \mathbf{1}_n$ if there is an intercept term
$\mathbf{X}_{\text{VLM}}$	VLM (model) matrix, $(nM) \times p_{\text{VLM}}$ $(= n_{\text{VLM}} \times p_{\text{VLM}})$, (3.18), (3.20)
$\boldsymbol{x} = (x_1, \ldots, x_p)^T = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)^T$	Vector of explanatory variables, with $x_1 = 1$ if there is an intercept term, $\boldsymbol{x}_1$ is $p_1 \times 1$, and $\boldsymbol{x}_2$ is $p_2 \times 1$. Sometimes $\boldsymbol{x} = (x_1, \ldots, x_d)^T$, especially when referring to additive models
$\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{ip}) = (\boldsymbol{x}_{1i}^T, \boldsymbol{x}_{2i}^T)$	ith row of $\mathbf{X}$
$\boldsymbol{x}_{ij} = (x_{i1j}, \ldots, x_{ipj})^T$	Vector of explanatory variables for $\eta_j(\boldsymbol{x}_{ij})$. Explanatory variables specific to η_j (see `xij` argument). Partitioned into $\boldsymbol{x}_i^*$ and $\boldsymbol{x}_{ij}^*$ as in (3.35)
$\mathbf{X}_{\texttt{form2}}$	LM (model) matrix for argument `form2`. Has n rows
$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^T$	Vector of linear/additive predictors, with $\boldsymbol{\eta}_i = (\eta_{1i}, \ldots, \eta_{Mi})^T$
$\mathbf{H}_k = \left(\boldsymbol{h}_k^{(1)}, \ldots, \boldsymbol{h}_k^{(\mathcal{R}_k)}\right) = (\boldsymbol{h}_{1k}, \ldots, \boldsymbol{h}_{Mk})^T$	Constraint matrix $(M \times \mathcal{R}_k)$ for x_k. Known, fixed and of full column-rank, (3.25)
$\boldsymbol{\eta}_i = \sum\limits_{k=1}^{p} \mathbf{H}_k \, \boldsymbol{\beta}_{(k)}^* \, x_{ik}$	Vector of linear predictors, (3.27)
$\boldsymbol{\eta}_i = \sum\limits_{k=1}^{d} \mathbf{H}_k \, \boldsymbol{f}_k^*(x_{ik})$	Vector of additive predictors, (3.25)
$\eta_j(\boldsymbol{x}_i) = \sum\limits_{k=1}^{p} \beta_{(j)k} \, x_{ik}$	jth linear predictor (without constraints), (1.1)
$\eta_j(\boldsymbol{x}_i) = \sum\limits_{k=1}^{d} f_{(j)k}(x_{ik})$	jth additive predictor (without constraints), (1.2)
$\boldsymbol{f}_k^*(x_k) = \left(f_{(1)k}^*(x_k), \ldots, f_{(\mathcal{R}_k)k}^*(x_k)\right)^T$	A $\mathcal{R}_k$-vector of smooth functions of x_k
$\mathbf{C} = (\boldsymbol{c}_{(1)}, \ldots, \boldsymbol{c}_{(R)}) = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{p_2})^T$	Matrix of constrained coefficients, (5.3)
$\mathbf{A} = (\boldsymbol{a}_{(1)}, \ldots, \boldsymbol{a}_{(R)}) = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M)^T$	Matrix of regression coefficients, (5.4)
$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_R)^T = \mathbf{C}^T \boldsymbol{x}_2$	Vector of R latent variables or gradients, (5.1)
$\boldsymbol{\nu}_i = (\nu_{i1}, \ldots, \nu_{iR})^T = \mathbf{C}^T \boldsymbol{x}_{2i}$	ith site score
$\boldsymbol{\beta}_{(k)}^* = (\beta_{(1)k}^*, \ldots, \beta_{(\mathcal{R}_k)k}^*)^T$	Coefficients for x_k to be estimated, (3.28)
$\mathbf{B} = (\boldsymbol{\beta}_1 \ \ \boldsymbol{\beta}_2 \ \ \cdots \ \ \boldsymbol{\beta}_M) = \left(\mathbf{H}_1 \, \boldsymbol{\beta}_{(1)}^* \mid \cdots \mid \mathbf{H}_p \, \boldsymbol{\beta}_{(p)}^*\right)^T$	Matrix of VLM/VGLM regression coefficients, $p \times M$, (1.32), (3.29)

Table A.5 Summary of some further quantities. See also Table 8.5.

Notation	Comments
$\mathbf{A}^T$	Transpose of $\mathbf{A}$, $(\mathbf{A}^T)_{ij} = (\mathbf{A})_{ji}$
$\boldsymbol{\beta}^\dagger$	$\text{vec}(\mathbf{B}) = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_M^T)^T$, (3.8)
$\boldsymbol{\theta}$	A generic vector of parameters to be estimated, often $(\theta_1, \ldots, \theta_p)^T$, can denote its true value
$\boldsymbol{\theta}_*$	The true value of $\boldsymbol{\theta}$, used occasionally when needed, p.536
$\boldsymbol{\mathcal{H}}$	Hat matrix, (2.10)
$\boldsymbol{\mathcal{H}}$	Hessian matrix, $[(\partial^2 \ell / (\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}^T))]$, (A.6)
$\boldsymbol{\mathcal{I}}_E$	Expected (Fisher) information matrix (EIM), (A.8)
$\boldsymbol{\mathcal{I}}_{E1}$	EIM for one observation
$\boldsymbol{\mathcal{I}}_O$	Observed information matrix, $-\boldsymbol{\mathcal{H}}$ (OIM), (A.7)
$\boldsymbol{\mathcal{P}}$	Householder matrix, (Ex. A.6)
$Y_{(i)}$	ith order statistic of $Y_1, Y_2, \ldots, Y_n$, so that $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$
$\overline{y}_{i\bullet}$	Mean of y_{ij} over all j, $\sum_{j=1}^{n_i} y_{ij}/n_i$, (Sect. 1.5.2.4)
D	Deviance, e.g., (3.53)

References

Abramowitz, M. and I. A. Stegun (Eds.) 1964. *Handbook of Mathematical Functions*. New York: Dover.

Adams, N., M. Crowder, D. J. Hand, and D. Stephens (Eds.) 2004. *Methods and Models in Statistics*. London: Imperial College Press.

Adler, J. 2010. *R in a Nutshell*. Sebastopol: O'Reilly.

Agresti, A. 2010. *Analysis of Ordinal Categorical Data* (2nd ed.). Hoboken: Wiley.

Agresti, A. 2013. *Categorical Data Analysis* (Third ed.). Hoboken: Wiley.

Agresti, A. 2015. *Foundations of Linear and Generalized Linear Models*. Hoboken: Wiley.

Ahn, S. K. and G. C. Reinsel 1988. Nested reduced-rank autoregressive models for multiple time series. *Journal of the American Statistical Association* 83(403):849–856.

Ahsanullah, M. H. and G. G. Hamedani 2010. *Exponential Distribution: Theory and Methods*. New York: Nova Science.

Aigner, D. J., T. Amemiya, and D. Poirer 1976. On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review* 17(2):377–396.

Aitkin, M., B. Francis, J. Hinde, and R. Darnell 2009. *Statistical Modelling in R*. Oxford: Oxford University Press.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki (Eds.), *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akadémiai Kaidó.

Albert, A. and J. A. Anderson 1984. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71(1):1–10.

Allison, P. 2004. Convergence problems in logistic regression. See Altman et al. (2004), pp. 238–252.

Altman, M., J. Gill, and M. P. McDonald 2004. *Numerical Issues in Statistical Computing for the Social Scientist*. Hoboken: Wiley-Interscience.

Altman, M. and S. Jackman 2011. Nineteen ways of looking at statistical software. *Journal of Statistical Software 42*(2), 1–12.

Amemiya, T. 1984. Tobit models: a survey. *Journal of Econometrics* 24(1–2):3–61.

Amemiya, T. 1985. *Advanced Econometrics*. Oxford: Blackwell.

Amodei, L. and M. N. Benbourhim 1991. A vector spline approximation with application to meteorology. In P. J. Laurent, A. Le Méhauté, and L. L. Schumaker (Eds.), *Curves and Surfaces*, pp. 5–10. Boston: Academic Press.

T.W. Yee, *Vector Generalized Linear and Additive Models*,
Springer Series in Statistics, DOI 10.1007/978-1-4939-2818-7

Amstrup, S. C., T. L. McDonald, and B. F. J. Manly 2005. *Handbook of Capture–Recapture Analysis*. Princeton: Princeton University Press.

Anderson, E., Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen 1999. *LAPACK Users' Guide* (Third ed.). Philadelphia: SIAM Publications.

Anderson, J. A. 1984. Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B* 46(1):1–30. With discussion.

Anderson, T. W. 1951. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics* 22(3):327–351.

Andrews, H. P., R. D. Snee, and M. H. Sarner 1980. Graphical display of means. *American Statistician* 34(4):195–199.

Arnold, B. C. 2015. *Pareto Distributions* (Second ed.). Boca Raton: Chapman & Hall/CRC.

Aronszajn, N. 1950. Theory of reproducing kernels. *Transactions of the American Mathematical Society* 68(3):337–404.

Ashford, J. R. and R. R. Sowden 1970. Multi-variate probit analysis. *Biometrics* 26(3):535–546.

Azzalini, A. 1996. *Statistical Inference: Based on the Likelihood*. London: Chapman & Hall.

Azzalini, A. 2014. *The Skew-normal and Related Families*. Cambridge: Cambridge University Press.

Baillargeon, S. and L.-P. Rivest 2007. Rcapture: Loglinear models for capture–recapture in R. *Journal of Statistical Software* 19(5):1–31.

Baker, F. B. and S.-H. Kim 2004. *Item Response Theory: Parameter Estimation Techniques* (Second ed.). New York: Marcel Dekker.

Balakrishnan, N. and A. P. Basu (Eds.) 1995. *The Exponential Distribution: Theory, Methods, and Applications*. Amsterdam: Gordon and Breach.

Balakrishnan, N. and C.-D. Lai 2009. *Continuous Bivariate Distributions* (Second ed.). New York: Springer.

Balakrishnan, N. and V. B. Nevzorov 2003. *A Primer on Statistical Distributions*. New York: Wiley-Interscience.

Banerjee, S. and A. Roy 2014. *Linear Algebra and Matrix Analysis for Statistics*. Boca Raton: CRC Press.

Barndorff-Nielsen, O. E. and D. R. Cox 1994. *Inference and Asymptotics*. London: Chapman & Hall.

Barrodale, I. and F. D. K. Roberts 1974. Solution of an overdetermined system of equations in the ℓ_1 norm. *Communications of the ACM* 17(6):319–320.

Beaton, A. E. and J. W. Tukey 1974. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* 16(2):147–185.

Beirlant, J., Y. Goegebeur, J. Segers, J. Teugels, D. De Waal, and C. Ferro 2004. *Statistics of Extremes: Theory and Applications*. Hoboken: Wiley.

Bellman, R. E. 1961. *Adaptive Control Processes*. Princeton: Princeton University Press.

Belsley, D. A., E. Kuh, and R. E. Welsch 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Berlinet, A. and C. Thomas-Agnan 2004. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Boston: Kluwer Academic Publishers.

Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hausman 1974. Estimation and inference in nonlinear structural models. *Ann. Econ. and Soc. Measur.* 3–4: 653–665.

Bickel, P. J. and K. A. Doksum 2001. *Mathematical Statistics: Basic Ideas and Selected Topics* (Second ed.). Upper Saddle River: Prentice Hall.

Bilder, C. M. and T. M. Loughin 2015. *Analysis of Categorical Data with R.* Boca Raton: CRC Press.

Birch, J. B. 1980. Some convergence properties of iterated least squares in the location model. *Communications in Statistics B* 9(4):359–369.

Bock, R. D. and M. Leiberman 1970. Fitting a response model for n dichotomously scored items. *Psychometrika* 35(2):179–197.

Boos, D. D. and L. A. Stefanski 2013. *Essential Statistical Inference.* New York: Springer.

Bowman, K. O. and L. R. Shenton 1988. *Properties of Estimators for the Gamma Distribution.* New York: Marcel Dekker.

Braun, W. J. and D. J. Murdoch 2008. *A First Course in Statistical Programming with R.* Cambridge: Cambridge University Press.

Buja, A., T. Hastie, and R. Tibshirani 1989. Linear smoothers and additive models. *The Annals of Statistics* 17(2):453–510. With discussion.

Burnham, K. P. and D. R. Anderson 2002. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach* (Second ed.). New York: Springer.

Byrd, R. H. and D. A. Pyne 1979. Some results on the convergence of the iteratively reweighted least squares. In *ASA Proc. Statist. Computat. Section*, pp. 87–90.

Cameron, A. C. and P. K. Trivedi 2013. *Regression Analysis of Count Data* (Second ed.). Cambridge: Cambridge University Press.

Cantoni, E. and T. Hastie 2002. Degrees-of-freedom tests for smoothing splines. *Biometrika* 89(2):251–263.

Carroll, R. J. and D. Ruppert 1988. *Transformation and Weighting in Regression.* New York: Chapman and Hall.

Casella, G. and R. L. Berger 2002. *Statistical Inference* (Second ed.). Pacific Grove: Thomson Learning.

Castillo, E., A. S. Hadi, N. Balakrishnan, and J. M. Sarabia 2005. *Extreme Value and Related Models with Applications in Engineering and Science.* Hoboken: Wiley.

Chambers, J. M. 1998. *Programming with Data: A Guide to the S Language.* New York: Springer.

Chambers, J. M. 2008. *Software for Data Analysis: Programming with R.* Statistics and Computing. New York: Springer.

Chambers, J. M. and T. J. Hastie (Eds.) 1991. *Statistical Models in S.* Pacific Grove: Wadsworth/Brooks Cole.

Cheney, W. and D. Kincaid 2012. *Numerical Mathematics and Computing* (Seventh ed.). Boston: Brooks/Cole.

Chotikapanich, D. (Ed.) 2008. *Modeling Income Distributions and Lorenz Curves.* New York: Springer.

Christensen, R. 1997. *Log-linear Models and Logistic Regression* (Second ed.). New York: Springer-Verlag.

Christensen, R. 2011. *Plane Answers to Complex Questions: The Theory of Linear Models* (4th ed.). New York: Springer-Verlag.

Christensen, R. H. B. 2013. *Analysis of ordinal data with cumulative link models—estimation with the R-package ordinal.* R package version 2013.9–30.

Claeskens, G. and N. L. Hjort 2008. *Model Selection and Model Averaging.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Clayton, D. and M. Hills 1993. *Statistical Models in Epidemiology.* Oxford: Oxford University Press.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74(368):829–836.

Cleveland, W. S. and S. J. Devlin 1988. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83(403):596–610.

Cleveland, W. S., E. Grosse, and W. M. Shyu 1991. Local regression models. See Chambers and Hastie (1991), pp. 309–376.

Cohen, Y. and J. Cohen 2008. *Statistics and Data with R: An Applied Approach Through Examples.* Chichester: John Wiley & Sons.

Coles, S. 2001. *An Introduction to Statistical Modeling of Extreme Values.* London: Springer-Verlag.

Consul, P. C. and F. Famoye 2006. *Lagrangian Probability Distributions.* Boston: Birkhäuser.

Cook, R. D. and S. Weisberg 1982. *Residuals and Influence in Regression.* Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Cox, D. R. 2006. *Principles of Statistical Inference.* Cambridge: Cambridge University Press.

Cox, D. R. and D. V. Hinkley 1974. *Theoretical Statistics.* London: Chapman & Hall.

Cox, D. R. and N. Reid 1987. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B* 49(1):1–39. With discussion.

Crawley, M. J. 2005. *Statistics: An Introduction using R.* Chichester: John Wiley & Sons.

Crowder, M. and T. Sweeting 1989. Bayesian inference for a bivariate binomial distribution. *Biometrika* 76(3):599–603.

Dalgaard, P. 2008. *Introductory Statistics with R* (Second ed.). New York: Springer.

Davino, C., C. Furno, and D. Vistocco 2014. *Quantile Regression: Theory and Applications.* Chichester: Wiley.

Davison, A. C. 2003. *Statistical Models.* Cambridge: Cambridge University Press.

Davison, A. C. and E. J. Snell 1991. Residuals and diagnostics. See Hinkley et al. (1991), pp. 83–106.

de Boor, C. 2001. *A Practical Guide to Splines (Revised Edition).* New York: Springer.

de Gruijter, D. N. M. and L. J. T. Van der Kamp 2008. *Statistical Test Theory for the Behavioral Sciences.* Boca Raton, FL, USA: Chapman & Hall/CRC.

de Haan, L. and A. Ferreira 2006. *Extreme Value Theory.* New York: Springer.

de Vries, A. and J. Meys 2012. *R for Dummies.* Chichester: Wiley.

De'ath, G. 1999. Principal curves: a new technique for indirect and direct gradient analysis. *Ecology* 80(7):2237–2253.

del Pino, G. 1989. The unifying role of iterative generalized least squares in statistical algorithms. *Statistical Science* 4(4):394–403.

Dempster, A. P., N. M. Laird, and D. B. Rubin 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B* 39(1):1–38. With discussion.

Dempster, A. P., N. M. Laird, and D. B. Rubin 1980. Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. In P. R. Krishnaiah (Ed.), *Multivariate Analysis–V: Proceedings of the Fifth International Symposium on Multivariate Analysis*, pp. 35–57. Amsterdam: North-Holland Publishing Company.

Dennis, J. E. and R. B. Schnabel 1996. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia: Society for Industrial and Applied Mathematics.

Devroye, L. 1986. *Non-Uniform Random Variate Generation*. New York: Springer-Verlag.

Dobson, A. J. and A. Barnett 2008. *An Introduction to Generalized Linear Models* (Third ed.). Boca Raton: Chapman & Hall/CRC Press.

Dongarra, J. J., J. R. Bunch, C. B. Moler, and G. W. Stewart 1979. *LINPACK User's Guide*. Philadelphia: SIAM Publications.

Edwards, A. W. F. 1972. *Likelihood*. Cambridge: Cambridge University Press.

Efron, B. 1986. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* 81(395):709–721.

Efron, B. 1991. Regression percentiles using asymmetric squared error loss. *Statistica Sinica* 1(1):93–125.

Efron, B. 1992. Poisson overdispersion estimates based on the method of asymmetric maximum likelihood. *Journal of the American Statistical Association* 87(417):98–107.

Efron, B. and D. V. Hinkley 1978. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* 65(3):457–487. With discussion.

Eilers, P. H. C. and B. D. Marx 1996. Flexible smoothing with *B*-splines and penalties. *Statistical Science* 11(2):89–121.

Elandt-Johnson, R. C. 1971. *Probability Models and Statistical Methods in Genetics*. New York: Wiley.

Embrechts, P., C. Klüppelberg, and T. Mikosch 1997. *Modelling Extremal Events for Insurance and Finance*. New York: Springer-Verlag.

Eubank, R. L. 1999. *Spline Smoothing and Nonparametric Regression* (Second ed.). New York: Marcel-Dekker.

Everitt, B. S. and D. J. Hand 1981. *Finite Mixture Distributions*. London: Chapman & Hall.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx 2011. *Regression: Models, Methods and Applications*. Berlin: Springer.

Fahrmeir, L. and G. Tutz 2001. *Multivariate Statistical Modelling Based on Generalized Linear Models* (Second ed.). New York: Springer-Verlag.

Fan, J. and I. Gijbels 1996. *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

Fan, J. and J. Jiang 2005. Nonparametric inferences for additive models. *Journal of the American Statistical Association* 100(471):890–907.

Fan, J. and Q. Yao 2003. *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

Faraway, J. J. 2006. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Boca Raton: Chapman and Hall/CRC.

Faraway, J. J. 2015. *Linear Models with R* (Second ed.). Boca Raton: Chapman & Hall/CRC.

Fessler, J. A. 1991. Nonparametric fixed-interval smoothing with vector splines. *IEEE Transactions on Signal Processing* 39(4):852–859.

Finkenstadt, B. and H. Rootzén (Eds.) 2003. *Extreme Values in Finance, Telecommunications and the Environment*. Boca Raton: Chapman & Hall/CRC.

Firth, D. 1991. Generalized linear models. See Hinkley et al. (1991), pp. 55–82.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80(1):27–38.

Firth, D. 2003. Overcoming the reference category problem in the presentation of statistical models. *Sociological Methodology* 33(1):1–18.

Firth, D. and R. X. de Menezes 2004. Quasi-variances. *Biometrika* 91(1):65–80.

Fishman, G. S. 1996. *Monte Carlo: Concepts, Algorithms, and Applications*. New York: Springer-Verlag.

Fitzenberger, B., R. Koenker, and J. A. F. Machado (Eds.) 2002. *Economic Applications of Quantile Regression*. Berlin: Springer-Verlag.

Forbes, C., M. Evans, N. Hastings, and B. Peacock 2011. *Statistical Distributions* (fouth ed.). Hoboken: John Wiley & Sons.

Fox, J. and S. Weisberg 2011. *An R Companion to Applied Regression* (Second ed.). Thousand Oaks: Sage Publications.

Freedman, D. A. 2007. How can the score test be inconsistent? *American Statistician* 61(4):291–295.

Freedman, D. A. and J. S. Sekhon 2010. Endogeneity in probit response models. *Political Analysis* 18(2):138–150.

Freund, J. E. 1961. A bivariate extension of the exponential distribution. *Journal of the American Statistical Association* 56(296):971–977.

Friedman, J. H. and W. Stuetzle 1981. Projection pursuit regression. *Journal of the American Statistical Association* 76(376):817–823.

Frühwirth-Schnatter, S. 2006. *Finite Mixture and Markov Switching Models*. New York: Springer.

Gabriel, K. R. and S. Zamir 1979. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* 21(4):489–498.

Gauch, Hugh G., J., G. B. Chase, and R. H. Whittaker 1974. Ordinations of vegetation samples by Gaussian species distributions. *Ecology* 55(6):1382–1390.

Gentle, J. E., W. K. Härdle, and Y. Mori 2012. *Handbook of Computational Statistics: Concepts and Methods* (Second ed.). Berlin: Springer.

Gentleman, R. 2009. *R Programming for Bioinformatics*. Boca Raton: Chapman & Hall/CRC.

Geraci, M. and M. Bottai 2007. Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* 8(1):140–154.

Gil, A., J. Segura, and N. M. Temme 2007. *Numerical Methods for Special Functions*. Philadelphia: Society for Industrial and Applied Mathematics.

Gill, J. and G. King 2004. What to do when your Hessian is not invertible: Alternatives to model respecification in nonlinear estimation. *Sociological Methods & Research* 33(1):54–87.

Gilleland, E., M. Ribatet, and A. G. Stephenson 2013. A software review for extreme value analysis. *Extremes* 16(1):103–119.

Goldberger, A. S. 1964. *Econometric Theory*. New York: Wiley.

Golub, G. H. and C. F. Van Loan 2013. *Matrix Computations* (Fourth ed.). Baltimore: Johns Hopkins University Press.

Gomes, M.I., and A. Guillou. 2015. Extreme value theory and statistics of univariate extremes: a review. *International Statistical Review* 83(2):263–292.

Goodman, L. A. 1981. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association* 76(374):320–334.

Gower, J. C. 1987. Introduction to ordination techniques. In P. Legendre and L. Legendre (Eds.), *Developments in Numerical Ecology*, pp. 3–64. Berlin: Springer-Verlag.

Green, P. J. 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society, Series B* 46(2):149–192. With discussion.

Green, P. J. and B. W. Silverman 1994. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.

Greene, W. H. 2012. *Econometric Analysis* (Seventh ed.). Upper Saddle River: Prentice Hall.

Greene, W. H. and D. A. Hensher 2010. *Modeling Ordered Choices: A Primer*. Cambridge: Cambridge University Press.

Gu, C. 2013. *Smoothing Spline ANOVA Models* (Second ed.). New York, USA: Springer.

Gumbel, E. J. 1958. *Statistics of Extremes*. New York, USA: Columbia University Press.

Gupta, A. K. and S. Nadarajah (Eds.) 2004. *Handbook of Beta Distribution and Its Applications*. New York, USA: Marcel Dekker.

Hao, L. and D. Q. Naiman 2007. *Quantile Regression*. Thousand Oaks, CA, USA: Sage Publications.

Härdle, W. 1987. *Smoothing Techniques With Implementation in S*. New York, USA: Springer-Verlag.

Härdle, W. 1990. *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Härdle, W., H. Liang, and J. Gao 2000. *Partially Linear Models*. New York, USA: Springer.

Härdle, W., M. Müller, S. Sperlich, and A. Werwatz 2004. *Nonparametric and Semiparametric Models*. Berlin: Springer.

Harezlak, J., D. Ruppert, and M.P. Wand. 2016. *Semiparametric regression in R*. New York: Springer.

Harper, W. V., T. G. Eschenbach, and T. R. James 2011. Concerns about maximum likelihood estimation for the three-parameter Weibull distribution: Case study of statistical software. *American Statistician* 65(1):44–54.

Harrell, F. E. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, USA: Springer.

Harville, D. A. 1997. *Matrix Algebra From a Statistician's Perspective*. New York, USA: Springer-Verlag.

Hastie, T. 1996. Pseudosplines. *Journal of the Royal Statistical Society, Series B* 58(2):379–396.

Hastie, T. and W. Stuetzle 1989. Principal curves. *Journal of the American Statistical Association* 84(406):502–516.

Hastie, T. and R. Tibshirani 1993. Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* 55(4):757–796.

Hastie, T. J. and D. Pregibon 1991. Generalized linear models. See Chambers and Hastie (1991), pp. 195–247.

Hastie, T. J. and R. J. Tibshirani 1990. *Generalized Additive Models*. London: Chapman & Hall.

Hastie, T. J., R. J. Tibshirani, and J. H. Friedman 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Second ed.). New York, USA: Springer-Verlag.

Hauck, J. W. W. and A. Donner 1977. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association* 72(360):851–853.

He, X. 1997. Quantile curves without crossing. *American Statistician* 51(2):186–192.

Heinze, G. and M. Schemper 2002. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 21(16):2409–2419.

Hensher, D. A., J. M. Rose, and W. H. Greene 2014. *Applied Choice Analysis* (Second ed.). Cambridge: Cambridge University Press.

Hilbe, J. M. 2009. *Logistic Regression Models*. Boca Raton, FL, USA: Chapman & Hall/CRC.

Hilbe, J. M. 2011. *Negative Binomial Regression* (Second ed.). Cambridge, UK; New York, USA: Cambridge University Press.

Hinkley, D. V., N. Reid, and E. J. Snell (Eds.) 1991. *Statistical Theory and Modelling. In Honour of Sir David Cox, FRS*, London. Chapman & Hall.

Hogben, L. (Ed.) 2014. *Handbook of Linear Algebra* (Second ed.). Boca Raton, FL, USA: Chapman & Hall/CRC.

Hörmann, W., J. Leydold, and G. Derflinger 2004. *Automatic Nonuniform Random Variate Generation*. Berlin: Springer.

Horvitz, D. G. and D. J. Thompson 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260):663–685.

Huber, P. J. 2011. *Data Analysis: What Can Be Learned From the Past 50 Years*. Hoboken, NJ, USA: Wiley.

Huber, P. J. and E. M. Ronchetti 2009. *Robust Statistics* (second ed.). New York, USA: Wiley.

Huggins, R. and W.-H. Hwang 2011. A review of the use of conditional likelihood in capture–recapture experiments. *International Statistical Review* 79(3):385–400.

Huggins, R. M. 1989. On the statistical analysis of capture experiments. *Biometrika* 76(1):133–140.

Huggins, R. M. 1991. Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics* 47(2):725–732.

Hui, F. K. C., S. Taskinen, S. Pledger, S. D. Foster, and D. I. Warton 2015. Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution* 6(4):399–411.

Hurvich, C. M. and C.-L. Tsai 1989. Regression and time series model selection in small samples. *Biometrika* 76(2):297–307.

Hutchinson, M. F. and F. R. de Hoog 1985. Smoothing noisy data with spline functions. *Numerische Mathematik* 47(1):99–106.

Hwang, W.-H. and R. Huggins 2011. A semiparametric model for a functional behavioural response to capture in capture–recapture experiments. *Australian & New Zealand Journal of Statistics* 53(4):403–421.

Ichimura, H. 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* 58(1–2):71–120.

Ihaka, R. and R. Gentleman 1996. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5(3):299–314.

Imai, K., G. King, and O. Lau 2008. Toward a common framework for statistical analysis and development. *Journal of Computational and Graphical Statistics* 17(4):892–913.

Izenman, A. J. 1975. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis* 5(2):248–264.

Izenman, A. J. 2008. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* New York, USA: Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani 2013. *An Introduction to Statistical Learning with Applications in R.* New York, USA: Springer.

Joe, H. 2014. *Dependence Modeling with Copulas.* Boca Raton, FL, USA: Chapman & Hall/CRC.

Johnson, N. L., A. W. Kemp, and S. Kotz 2005. *Univariate Discrete Distributions* (Third ed.). Hoboken, NJ, USA: John Wiley & Sons.

Johnson, N. L., S. Kotz, and N. Balakrishnan 1994. *Continuous Univariate Distributions* (Second ed.), Volume 1. New York, USA: Wiley.

Johnson, N. L., S. Kotz, and N. Balakrishnan 1995. *Continuous Univariate Distributions* (Second ed.), Volume 2. New York, USA: Wiley.

Johnson, N. L., S. Kotz, and N. Balakrishnan 1997. *Discrete Multivariate Distributions.* New York, USA: John Wiley & Sons.

Jones, M. C. 1994. Expectiles and M-quantiles are quantiles. *Statistics & Probability Letters* 20(2):149–153.

Jones, M. C. 2002. Student's simplest distribution. *The Statistician* 51(1):41–49.

Jones, M. C. 2009. Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology* 6(1):70–81.

Jones, O., R. Maillardet, and A. Robinson 2014. *Introduction to Scientific Programming and Simulation Using R* (Second ed.). Boca Raton, FL, USA: Chapman and Hall/CRC.

Jongman, R. H. G., C. J. F. ter Braak, and O. F. R. van Tongeren (Eds.) 1995. *Data Analysis in Community and Landscape Ecology.* Cambridge: Cambridge University Press.

Jørgensen, B. 1984. The delta algorithm and GLIM. *International Statistical Review* 52(3):283–300.

Jørgensen, B. 1997. *The Theory of Dispersion Models.* London: Chapman & Hall.

Jorgensen, M. 2001. Iteratively reweighted least squares. In A. H. El-Shaarawi and W. W. Piegorsch (Eds.), *Encyclopedia of Environmetrics*, Volume 2, pp. 1084–1088. Chichester, New York, USA: Wiley.

Kaas, R., M. Goovaerts, J. Dhaene, and M. Denuit 2008. *Modern Actuarial Risk Theory Using R* (Second ed.). Berlin: Springer.

Kateri, M. 2014. *Contingency Table Analysis. Methods and Implementation Using R.* New York, USA: Birkhäuser/Springer.

Kennedy, William J., J. and J. E. Gentle 1980. *Statistical Computing.* New York, USA: Marcel Dekker.

Keogh, R. H. and D. R. Cox 2014. *Case-Control Studies.* New York, USA: Cambridge University Press.

Kleiber, C. and S. Kotz 2003. *Statistical Size Distributions in Economics and Actuarial Sciences.* Hoboken, NJ, USA: Wiley-Interscience.

Kleiber, C. and A. Zeileis 2008. *Applied Econometrics with R.* New York, USA: Springer.

Klugman, S. A., H. H. Panjer, and G. E. Willmot 2012. *Loss Models: From Data to Decisions* (4th ed.). Hoboken, NJ, USA: Wiley.

Klugman, S. A., H. H. Panjer, and G. E. Willmot 2013. *Loss Models: Further Topics.* Hoboken, NJ, USA: Wiley.

Knight, K. 2000. *Mathematical Statistics.* Boca Raton, FL, USA: Chapman & Hall/CRC.

Kocherlakota, S. and K. Kocherlakota 1992. *Bivariate Discrete Distributions.* New York, USA: Marcel Dekker.

Koenker, R. 1992. When are expectiles percentiles? (problem). *Econometric Theory* 8(3):423–424.

Koenker, R. 2005. *Quantile Regression.* Cambridge: Cambridge University Press.

Koenker, R. 2013. Discussion: Living beyond our means. *Statistical Modelling* 13(4):323–333.

Koenker, R. and G. Bassett 1978. Regression quantiles. *Econometrica* 46(1):33–50.

Kohn, R. and C. F. Ansley 1987. A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM Journal on Scientific and Statistical Computing* 8(1):33–48.

Konishi, S. and G. Kitagawa 2008. *Information Criteria and Statistical Modeling.* Springer Series in Statistics. New York, USA: Springer.

Kooijman, S. A. L. M. 1977. Species abundance with optimum relations to environmental factors. *Annals of Systems Research* 6:123–138.

Kosmidis, I. 2014a. Bias in parametric estimation: reduction and useful side-effects. *WIREs Computational Statistics* 6:185–196.

Kosmidis, I. 2014b. Improved estimation in cumulative link models. *Journal of the Royal Statistical Society, Series B* 76(1):169–196.

Kosmidis, I. and D. Firth 2009. Bias reduction in exponential family nonlinear models. *Biometrika* 96(4):793–804.

Kosmidis, I. and D. Firth 2010. A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics* 4:1097–1112.

Kotz, S., T. J. Kozubowski, and K. Podgórski 2001. *The Laplace Distribution and Generalizations: a Revisit with Applications to Communications, Economics, Engineering, and Finance.* Boston, MA, USA: Birkhäuser.

Kotz, S. and S. Nadarajah 2000. *Extreme Value Distributions: Theory and Applications.* London: Imperial College Press.

Kotz, S. and J. R. van Dorp 2004. *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications.* Singapore: World Scientific.

Kozubowski, T. J. and S. Nadarajah 2010. Multitude of Laplace distributions. *Statistical Papers* 51(1):127–148.

Lange, K. 2002. *Mathematical and Statistical Methods for Genetic Analysis* (Second ed.). New York, USA: Springer-Verlag.

Lange, K. 2010. *Numerical Analysis for Statisticians* (Second ed.). New York, USA: Springer.

Lange, K. 2013. *Optimization* (Second ed.). New York, USA: Springer.

Lawless, J. F. 1987. Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* 15(3):209–225.

Lawless, J. F. 2003. *Statistical Models and Methods for Lifetime Data* (Second ed.). Hoboken, NJ, USA: John Wiley & Sons.

Leadbetter, M. R., G. Lindgren, and H. Rootzén 1983. *Extremes and Related Properties of Random Sequences and Processes.* New York, USA: Springer-Verlag.

Leemis, L. M. and J. T. McQueston 2008. Univariate distribution relationships. *American Statistician* 62(1):45–53.

Lehmann, E. L. and G. Casella 1998. *Theory of Point Estimation* (Second ed.). New York, USA: Springer.

Lehmann, E. L. and J. P. Romano 2005. *Testing Statistical Hypotheses* (3rd ed.). New York, USA: Springer.

Lesaffre, E. and A. Albert 1989. Partial separation in logistic discrimination. *Journal of the Royal Statistical Society, Series B* 51(1):109–116.

Libby, D. L. and M. R. Novick 1982. Multivariate generalized beta distributions with applications to utility assessment. *Journal of Educational and Statistics* 7(4):271–294.

Lindsay, B. G. 1995. *Mixture Models: Theory, Geometry and Applications*, Volume 5. Hayward CA, USA: NSF-CBMS Regional Conference Series in Probability and Statistics, IMS.

Lindsey, J. K. 1996. *Parametric Statistical Inference.* Oxford: Clarendon Press.

Lindsey, J. K. 1997. *Applying Generalized Linear Models.* New York, USA: Springer-Verlag.

Liu, H. and K. S. Chan 2010. Introducing COZIGAM: An R package for unconstrained and constrained zero-inflated generalized additive model analysis. *Journal of Statistical Software* 35(11):1–26.

Liu, I. and A. Agresti 2005. The analysis of ordered categorical data: An overview and a survey of recent developments. *Test* 14(1):1–73.

Lloyd, C. J. 1999. *Statistical Analysis of Categorical Data.* New York, USA: Wiley.

Loader, C. 1999. *Local Regression and Likelihood.* New York, USA: Springer.

Lopatatzidis, A. and P. J. Green 1998. Semiparametric quantile regression using the gamma distribution. *Unpublished manuscript.*

Maddala, G. S. 1983. *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Mai, J.-F. and M. Scherer 2012. *Simulating Copulas: Stochastic Models, Sampling Algorithms, and Applications.* London: Imperial College Press.

Maindonald, J. H. and W. J. Braun 2010. *Data Analysis and Graphics Using R: An Example-Based Approach* (Third ed.). Cambridge: Cambridge University Press.

Marra, G. and R. Radice 2010. Penalised regression splines: theory and application to medical research. *Statistical Methods in Medical Research* 19(2):107–125.

Marshall, A. W. and I. Olkin 2007. *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families.* New York, USA: Springer.

McCrea, R. S. and B. J. T. Morgan 2015. *Analysis of Capture–Recapture Data.* Boca Raton, FL, USA: Chapman & Hall/CRC.

McCullagh, P. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* 42(2):109–142. With discussion.

McCullagh, P. 1989. Some statistical properties of a family of continuous univariate distributions. *Journal of the American Statistical Association* 84(405):125–129.

McCullagh, P. and J. A. Nelder 1989. *Generalized Linear Models* (Second ed.). London: Chapman & Hall.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Conditional Logit Analysis of Qualitative Choice Behavior*, pp. 105–142. New York, USA: Academic Press.

McLachlan, G. J. and D. Peel 2000. *Finite Mixture Models.* New York, USA: Wiley.

Mikosch, T. 2006. Copulas: tales and facts (with rejoinder). *Extremes* 9(1): 3–20,55–62.

Miller, A. 2002. *Subset Selection in Regression* (Second ed.). Boca Raton, FL, USA: Chapman & Hall/CRC.

Miller, J. J. and E. J. Wegman 1987. Vector function estimation using splines. *Journal of Statistical Planning and Inference* 17:173–180.

Morris, C. N. 1982. Natural exponential families with quadratic variance functions. *The Annals of Statistics* 10(1):65–80.

Mosteller, F. and J. W. Tukey 1977. *Data Analysis and Regression.* Reading, MA, USA: Addison-Wesley.

Murthy, D. N. P., M. Xie, and R. Jiang 2004. *Weibull Models.* Hoboken, NJ, USA: Wiley.

Myers, R. H., D. C. Montgomery, G. G. Vining, and T. J. Robinson 2010. *Generalized Linear Models With Applications in Engineering and the Sciences* (Second ed.). Hoboken, NJ, USA: Wiley.

Nadarajah, S. and S. A. A. Bakar 2013. A new R package for actuarial survival models. *Computational Statistics* 28(5):2139–2160.

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

Nelsen, R. B. 2006. *An Introduction to Copulas* (Second ed.). New York, USA: Springer.

Newey, W. K. and J. L. Powell 1987. Asymmetric least squares estimation and testing. *Econometrica* 55(4):819–847.

Neyman, J. and E. L. Scott 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16(1):1–32.

Nocedal, J. and S. J. Wright 2006. *Numerical Optimization* (Second ed.). New York, USA: Springer.

Nosedal-Sanchez, A., C. B. Storlie, T. C. M. Lee, and R. Christensen 2012. Reproducing kernel Hilbert spaces for penalized regression: A tutorial. *American Statistician* 66(1):50–60.

Novak, S. Y. 2012. *Extreme Value Methods with Applications to Finance.* Boca Raton, FL, USA: CRC Press.

Olver, F. W. J., D. W. Lozier, R. F. Boisvert, and C. W. Clark (Eds.) 2010. *NIST Handbook of Mathematical Functions.* New York, USA: National Institute of Standards and Technology, and Cambridge University Press.

Osborne, M. R. 1992. Fisher's method of scoring. *International Statistical Review* 60(1):99–117.

Osborne, M. R. 2006. Least squares methods in maximum likelihood problems. *Optimization Methods and Software* 21(6):943–959.

Otis, D. L., K. P. Burnham, G. C. White, and D. R. Anderson 1978. Statistical inference from capture data on closed animal populations. *Wildlife Monographs* 62:3–135.

Owen, A. B. 2001. *Empirical Likelihood.* Boca Raton, FL, USA: Chapman & Hall/CRC.

Page, L. A., S. Hajat, and R. S. Kovats 2007. Relationship between daily suicide counts and temperature in England and Wales. *British Journal of Psychiatry* 191(2):106–112.

Pal, N., C. Jin, and W. K. Lim 2006. *Handbook of Exponential and Related Distributions for Engineers and Scientists.* Boca Raton, FL, USA: Chapman & Hall/CRC.

Palmer, M. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology* 74(8):2215–2230.

Palmgren, J. 1989. Regression models for bivariate binary responses. Technical Report 101, Biostatistics Dept, University of Washington, Seattle, USA.

Park, B. U., E. Mammen, Y. K. Lee, and E. R. Lee 2015. Varying coefficient regression models: a review and new developments. *International Statistical Review* 83(1):36–64.

Pickands, J. 1975. Statistical inference using extreme order statistics. *The Annals of Statistics* 3(1):119–131.

Plackett, R. L. 1965. A class of bivariate distributions. *Journal of the American Statistical Association* 60(310):516–522.

Poiraud-Casanova, S. and C. Thomas-Agnan 2000. About monotone regression quantiles. *Statistics & Probability Letters* 48(1):101–104.

Powers, D. A. and Y. Xie 2008. *Statistical Methods for Categorical Data Analysis* (Second ed.). Bingley, UK: Emerald.

Pratt, J. W. 1981. Concavity of the log likelihood. *Journal of the American Statistical Association* 76(373):103–106. Correction p.954, Vol 77.

Prentice, R. L. 1974. A log gamma model and its maximum likelihood estimation. *Biometrika* 61(3):539–544.

Prentice, R. L. 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* 81(394):321–327.

Prescott, P. and A. T. Walden 1980. Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika* 67(3):723–724.

Randall, J. H. 1989. The analysis of sensory data by generalized linear model. *Biometrics Journal* 31(7):781–793.

Rao, C. R. 1948. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* 44(1):50–57.

Rao, C. R. 1973. *Linear Statistical Inference and its Applications* (Second ed.). New York, USA: Wiley.

Rasch, G. 1961. On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 4:321–333.

Reinsch, C. H. 1967. Smoothing by spline functions. *Numerische Mathematik* 10(3):177–183.

Reinsel, G. C. and R. P. Velu 1998. *Multivariate Reduced-Rank Regression: Theory and Applications*. New York, USA: Springer-Verlag.

Reinsel, G. C. and R. P. Velu 2006. Partically reduced-rank multivariate regression models. *Statistica Sinica* 16(3):899–917.

Reiss, R.-D. and M. Thomas 2007. *Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields* (Third ed.). Basel, Switzerland: Birkhäuser.

Rencher, A. C. and G. B. Schaalje 2008. *Linear Models in Statistics* (second ed.). New York, USA: John Wiley & Sons.

Richards, F. S. G. 1961. A method of maximum-likelihood estimation. *Journal of the Royal Statistical Society, Series B* 23(2):469–475.

Richards, S. J. 2012. A handbook of parametric survival models for actuarial use. *Scandinavian Actuarial Journal* 2012(4):233–257.

Ridout, M. S. 1990. Non-convergence of Fisher's method of scoring—a simple example. *GLIM Newsletter* 20(6).

Rinne, H. 2009. *The Weibull Distribution*. Boca Raton, FL, USA: CRC Press.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Ripley, B. D. 2004. Selecting amongst large classes of models. See Adams et al. (2004), pp. 155–170.

Rose, C. and M. D. Smith 2002. *Mathematical Statistics with* Mathematica. New York, USA: Springer.

Rose, C. and M. D. Smith 2013. *Mathematical Statistics with* Mathematica. eBook.

Rubin, D. B. 2006. Iteratively reweighted least squares. In *Encyclopedia of Statistical Sciences*, Volume 6. Wiley.

Ruppert, D., M. P. Wand, and R. J. Carroll 2003. *Semiparametric Regression*. Cambridge: Cambridge University Press.

Ruppert, D., M. P. Wand, and R. J. Carroll 2009. Semiparametric regression during 2003–2007. *Electronic Journal of Statistics* 3(1):1193–1256.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa 1986. *Akaike Information Criterion Statistics*. Dordrecht, Netherlands: D. Reidel Publishing Company.

Schenker, N. and J. F. Gentleman 2001. On judging the significance of differences by examining the overlap between confidence intervals. *American Statistician* 55(3):182–186.

Schepsmeier, U. and J. Stöber 2014. Derivatives and Fisher information of bivariate copulas. *Statistical Papers* 55(2):525–542.

Schimek, M. G. (Ed.) 2000. *Smoothing and Regression: Approaches, Computation, and Application*. New York, USA: Wiley.

Schnabel, S. K. and P. H. C. Eilers 2009. Optimal expectile smoothing. *Computational Statistics & Data Analysis* 53(12):4168–4177.

Schumaker, L. L. 2007. *Spline Functions: Basic Theory* (Third ed.). Cambridge: Cambridge University Press.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2):461–464.

Seber, G. A. F. 2008. *A Matrix Handbook for Statisticians*. Hoboken, NJ, USA: Wiley.

Seber, G. A. F. and A. J. Lee 2003. *Linear Regression Analysis* (Second ed.). New York, USA: Wiley.

Seber, G. A. F. and C. J. Wild 1989. *Nonlinear Regression*. New York, USA: Wiley.

Self, S. G. and K.-Y. Liang 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82(398):605–610.

Senn, S. 2004. John Nelder: From general balance to generalised models (both linear and hierarchical). See Adams et al. (2004), pp. 1–12.

Severini, T. A. 2000. *Likelihood Methods in Statistics*. New York, USA: Oxford University Press.

Shao, J. 2003. *Mathematical Statistics* (Second ed.). New York, USA: Springer.

Shao, J. 2005. *Mathematical Statistics: Exercises and Solutions*. New York, USA: Springer.

Silverman, B. W. 1984. Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics* 12(3):898–916.

Silverman, B. W. 1985. Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society, Series B* 47(1):1–21. With discussion.

Silvey, S. D. 1975. *Statistical Inference*. London: Chapman & Hall.

Simonoff, J. S. 2003. *Analyzing Categorical Data*. New York, USA: Springer-Verlag.

Sklar, A. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris* 8:229–231.

Small, C. G. and D. L. McLeish 1994. *Hilbert Space Methods in Probability and Statistical Inference*. New York, USA: Wiley.

Smith, M. and R. Kohn 2000. Nonparametric seemingly unrelated regression. *Journal of Econometrics* 98(2):257–281.

Smith, R. L. 1985. Maximum likelihood estimation in a class of nonregular cases. *Biometrika* 72(1):67–90.

Smith, R. L. 1986. Extreme value theory based on the r largest annual events. *Journal of Hydrology* 86(1–2):27–43.

Smith, R. L. 2003. Statistics of extremes, with applications in environment, insurance and finance. See Finkenstadt and Rootzén (2003), pp. 1–78.

Smithson, M. and E. C. Merkle 2013. *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*. London: Chapman & Hall/CRC.

Smyth, G. K. 1989. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society, Series B* 51(1):47–60.

Smyth, G. K. 1996. Partitioned algorithms for maximum likelihood and other nonlinear estimation. *Statistics and Computing* 6(3):201–216.

Smyth, G. K., A. F. Huele, and A. P. Verbyla 2001. Exact and approximate REML for heteroscedastic regression. *Statistical Modelling* 1(3):161–175.

Spector, P. 2008. *Data Manipulation with R*. New York, USA: Springer Verlag.

Srivastava, V. K. and T. D. Dwivedi 1979. Estimation of seemingly unrelated regression equations: A brief survey. *Journal of Econometrics* 10(1):15–32.

Srivastava, V. K. and D. E. A. Giles 1987. *Seemingly Unrelated Regression Equations Models: Estimation and Inference*. New York, USA: Marcel Dekker.

Stacy, E. W. 1962. A generalization of the gamma distribution. *Annals of Mathematical Statistics* 33(3):1187–1192.

Takane, Y., H. Yanai, and S. Mayekawa 1991. Relationships among several methods of linearly constrained correspondence analysis. *Psychometrika* 56(4):667–684.

Tawn, J. A. 1988. An extreme-value theory model for dependent observations. *Journal of Hydrology* 101(1–4):227–250.

Taylor, J. W. 2008. Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics* 6(2):231–252.

Taylor, L. R. 1961. Aggregation, variance and the mean. *Nature* 189(4766): 732–735.

ter Braak, C. J. F. 1986. Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67(5):1167–1179.

ter Braak, C. J. F. 1995. Calibration. See Jongman et al. (1995), pp. 78–90.

ter Braak, C. J. F. and I. C. Prentice 1988. A theory of gradient analysis. In *Advances in Ecological Research*, Volume 18, pp. 271–317. London: Academic Press.

ter Braak, C. J. F. and P. F. M. Verdonschot 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences* 57(3):255–289.

ter Braak, C. J. F., and P. Šmilauer 2015. Topics in constrained and unconstrained ordination. *Plant Ecology* 216(5):683–696.

Thompson, R. and R. J. Baker 1981. Composite link functions in generalized linear models. *Journal of the Royal Statistical Society, Series C* 30(2):125–131.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58(1):267–288.

Titterington, D. M., A. F. M. Smith, and U. E. Makov 1985. *Statistical Analysis of Finite Mixture Distributions*. New York, USA: Wiley.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26(1):24–36.

Trivedi, P. K. and D. M. Zimmer 2005. Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics* 1(1):1–111.

Tutz, G. 2012. *Regression for Categorical Data*. Cambridge: Cambridge University Press.

van den Boogaart, K. G. and R. Tolosana-Delgado 2013. *Analyzing Compositional Data with R*. Berlin: Springer.

Venables, W. N. and B. D. Ripley 2002. *Modern Applied Statistics With S* (4th ed.). New York, USA: Springer-Verlag.

von Eye, A. and E.-E. Mun 2013. *Log-linear Modeling: Concepts, Interpretation, and Application*. Hoboken, NJ, USA: Wiley.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica* 57(2):307–333.

Wahba, G. 1982. Vector splines on the sphere, with application to the estimation of vorticity and divergence from discrete, noisy data. In W. S. K. Zeller (Ed.), *Multivarate Approximation Theory*, Volume 2, pp. 407–429. Birkhäuser: Verlag.

Wahba, G. 1990. *Spline models for observational data*, Volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM).

Wand, M. P. and M. C. Jones 1995. *Kernel Smoothing*. London: Chapman & Hall.

Wand, M. P. and J. T. Ormerod 2008. On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics* 50(2):179–198.

Wang, Y. 2011. *Smoothing Splines: Methods and Applications*. Boca Raton, FL, USA: Chapman & Hall/CRC.

Webb, M. H., S. Wotherspoon, D. Stojanovic, R. Heinsohn, R. Cunningham, P. Bell, and A. Terauds 2014. Location matters: Using spatially explicit occupancy models to predict the distribution of the highly mobile, endangered swift parrot. *Biological Conservation* 176:99–108.

Wecker, W. E. and C. F. Ansley 1983. The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association* 78(381):81–89.

Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61(3):439–447.

Wegman, E. J. 1981. Vector splines and the estimation of filter functions. *Technometrics* 23(1):83–89.

Weihs, C., O. Mersmann, and U. Ligges 2014. *Foundations of Statistical Algorithms: With References to R Packages*. Boca Raton, FL, USA: CRC Press.

Weir, B. S. 1996. *Genetic Data Analysis II*. Sunderland, MA, USA: Sinauer.

Welsh, A. H. 1996. Robust estimation of smooth regression and spread functions and their derivatives. *Statistica Sinica* 6:347–366.

Welsh, A. H., R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmayer 1996. Modelling the abundances of rare species: statistical models for counts with extra zeros. *Ecological Modelling* 88(1–3):297–308.

Welsh, A. H., D. B. Lindenmayer, and C. F. Donnelly 2013. Fitting and interpreting occupancy models. *PLOS One* 8(1):1–21.

Welsh, A. H. and T. W. Yee 2006. Local regression for vector responses. *Journal of Statistical Planning and Inference* 136(9):3007–3031.

Wickham, H. 2015. *Advanced R.* Boca Raton, FL, USA: Chapman & Hall/CRC.

Wild, C. J. and T. W. Yee 1996. Additive extensions to generalized estimating equation methods. *Journal of the Royal Statistical Society, Series B* 58(4):711–725.

Wilkinson, G. N. and C. E. Rogers 1973. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society, Series C* 22(3):392–399.

Williams, B. K., J. D. Nichols, and M. J. Conroy 2002. *Analysis and Management of Animal Populations.* London: Academic Press.

Williams, D. A. 1975. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* 31(4):949–952.

Winkelmann, R. 2008. *Econometric Analysis of Count Data* (5th ed.). Berlin: Springer.

Winkelmann, R. and S. Boes 2006. *Analysis of Microdata.* Berlin: Springer.

Withers, C. S. and S. Nadarajah 2009. The asymptotic behaviour of the maximum of a random sample subject to trends in location and scale. *Random Operators and Stochastic Equations* 17(1):55–60.

Wold, S. 1974. Spline functions in data analysis. *Technometrics* 16(1):1–11.

Wood, S. N. 2006. *Generalized Additive Models: An Introduction with R.* London: Chapman and Hall.

Wooldridge, J. M. 2006. *Introductory Econometrics: A Modern Approach* (5th ed.). Mason, OH, USA: South-Western.

Yanai, H., K. Takeuchi, and Y. Takane 2011. *Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition.* New York, USA: Springer.

Yang, H.-C. and A. Chao 2005. Modeling animals' behavioral response by Markov chain models for capture–recapture experiments. *Biometrics* 61(4):1010–1017.

Yasuda, N. 1968. Estimation of the interbreeding coefficient from phenotype frequencies by a method of maximum likelihood scoring. *Biometrics* 24(4):915–934.

Yatchew, A. 2003. *Semiparametric Regression for the Applied Econometrician.* Cambridge: Cambridge University Press.

Yee, T. W. 1998. On an alternative solution to the vector spline problem. *Journal of the Royal Statistical Society, Series B* 60(1):183–188.

Yee, T. W. 2000. Vector splines and other vector smoothers. In J. G. Bethlehem and P. G. M. van der Heijden (Eds.), *Proceedings in Computational Statistics COMPSTAT 2000*, pp. 529–534. Heidelberg: Physica-Verlag.

Yee, T. W. 2004a. A new technique for maximum-likelihood canonical Gaussian ordination. *Ecological Monographs* 74(4):685–701.

Yee, T. W. 2004b. Quantile regression via vector generalized additive models. *Statistics in Medicine* 23(14):2295–2315.

Yee, T. W. 2006. Constrained additive ordination. *Ecology* 87(1):203–213.

Yee, T. W. 2010a. The VGAM package for categorical data analysis. *Journal of Statistical Software* 32(10):1–34.

Yee, T. W. 2010b. VGLMs and VGAMs: an overview for applications in fisheries research. *Fisheries Research* 101(1–2):116–126.

Yee, T. W. 2014. Reduced-rank vector generalized linear models with two linear predictors. *Computational Statistics & Data Analysis* 71:889–902.

Yee, T. W. and A. F. Hadi 2014. Row-column interaction models, with an R implementation. *Computational Statistics* 29(6):1427–1445.

Yee, T. W. and T. J. Hastie 2003. Reduced-rank vector generalized linear models. *Statistical Modelling* 3(1):15–41.

Yee, T. W. and N. D. Mitchell 1991. Generalized additive models in plant ecology. *Journal of Vegetation Science* 2(5):587–602.

Yee, T. W. and A. G. Stephenson 2007. Vector generalized linear and additive extreme value models. *Extremes* 10(1–2):1–19.

Yee, T. W., J. Stoklosa, and R. M. Huggins 2015. The VGAM package for capture–recapture data using the conditional likelihood. *Journal of Statistical Software* 65(5):1–33.

Yee, T. W. and C. J. Wild 1996. Vector generalized additive models. *Journal of the Royal Statistical Society, Series B* 58(3):481–493.

Yeo, I.-K. and R. A. Johnson 2000. A new family of power transformations to improve normality or symmetry. *Biometrika* 87(4):954–959.

Young, G. A. and R. L. Smith 2005. *Essentials of Statistical Inference.* Cambridge: Cambridge University Press.

Yu, K. and J. Zhang 2005. A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics - Theory and Methods* 34(9–10):1867–1879.

Yu, P. and C. A. Shaw 2014. An efficient algorithm for accurate computation of the Dirichlet-multinomial log-likelihood function. *Bioinformatics* 30(11):1547–54.

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* 57(298):348–368.

Zhang, C. 2003. Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. *Journal of the American Statistical Association* 98(463):609–628.

Zhang, Y. and O. Thas 2012. Constrained ordination analysis in the presence of zero inflation. *Statistical Modelling* 12(6):463–485.

Zhu, M., T. J. Hastie, and G. Walther 2005. Constrained ordination analysis with flexible response functions. *Ecological Modelling* 187(4):524–536.

Zuur, A. F. 2012. *A Beginner's Guide to Generalized Additive Models with R.* Newburgh, UK: Highland Statistics Ltd.

Zuur, A. F., E. N. Ieno, and E. H. Meesters 2009. *A Beginner's Guide to R.* New York, USA: Springer.

Zuur, A. F., A. A. Saveliev, and E. N. Ieno 2012. *Zero Inflated Models and Generalized Linear Mixed Models with R.* Newburgh, UK: Highland Statistics Ltd.

Index

T.W. Yee, *Vector Generalized Linear and Additive Models*,
Springer Series in Statistics, DOI 10.1007/978-1-4939-2818-7