HHS Public Access

Author manuscript

Scand Stat Theory Appl. Author manuscript; available in PMC 2018 September 01.

Published in final edited form as:

Scand Stat Theory Appl. 2017 September ; 44(3): 636–665. doi:10.1111/sjos.12269.

Exact and Approximate Statistical Inference for Nonlinear Regression and the Estimating Equation Approach

Eugene Demidenko

Departament of Mathematics, Dartmouth College, USA

Abstract

The exact density distribution of the nonlinear least squares estimator in the one-parameter regression model is derived in closed form and expressed through the cumulative distribution function of the standard normal variable. Several proposals to generalize this result are discussed. The exact density is extended to the estimating equation (EE) approach and the nonlinear regression with an arbitrary number of linear parameters and one intrinsically nonlinear parameter. For a very special nonlinear regression model, the derived density coincides with the distribution of the ratio of two normally distributed random variables previously obtained by Fieller (1932), unlike other approximations previously suggested by other authors. Approximations to the density of the EE estimators are discussed in the multivariate case. Numerical complications associated with the nonlinear least squares are illustrated, such as nonexistence and/or multiple solutions, as major factors contributing to poor density approximation. The nonlinear Markov-Gauss theorem is formulated based on the near exact EE density approximation.

Keywords

Edgeworth approximation; exact statistical inference; Markov theorem; Michaelis-Menten model; saddlepoint approximation; small-sample property; partial least squares

1. Introduction

According to the JSTOR database, the search for “nonlinear regression” returns 47,988 published items (on December 6, 2016) with the first article published by G.U. Yule in 1909. Since then, nonlinear regression has been studied along four lines of research: (1) Development of nonlinear least squares optimization algorithms. It is worthwhile to mention that the algorithm developed by Levenberg (1944) & Marquardt (1963) till now is the dominant method for sum of squares minimization and is widely used in modern numerical packages. (2) Developing curvature measures as a characterization of the model nonlinearity, starting from the pioneering 1961 paper by Beale and then continued by Bates & Watts (1980). (3) Connection of the curvature measures to criteria for existence and uniqueness of the nonlinear least squares estimate studied in the work by Demidenko (1989, 2000, 2006). (4) Derivation of the density of the nonlinear least squares estimator in small samples, mostly developed in the mid-eighties. The present paper continues the quest for the exact

Eugene Demidenko, 7927 Rubin, DHMC, Lebanon, New Hampshire 03756, USA, eugened@dartmouth.edu.

probability density distribution of the NLS estimate as continuation of the work by Pazman (1984), Hougaard (1985) and Skovgaard (1985). Needless to say how important the exact distribution of any estimator might be: it can be used as the benchmark for other approximations, to construct accurate confidence intervals and to test hypotheses with the exact type I error.

Studying statistical properties of nonlinear models in small samples is the most formidable problem of statistics. Although several general techniques are available, such as the saddlepoint and Edgeworth approximations (Goutis & Casella, 1999; Barndorf-Nielsen & Cox, 1979) they are still approximations in scope. To illustrate, consider the popular Lugannani & Rice (1980) and Fraser *et al.* (1999) saddle point approximation of order $O(n^{-3/2})$ for the cumulative distribution function (cdf) of the standardized maximum likelihood estimator with the sample size n used in the recent survey paper by Brazzale & Davison (2008)

$$F_n(b) \simeq \Phi(b)+\phi(b)\left(\frac{1}{b}-\frac{1}{q_n(b)}\right),$$

where Φ and ϕ are the cdf and density of the standard normal distribution, and q_n is a positive function of the Fisher information (we do not present the exact formula for brevity). However the cdf $F_n(b)$ is not an increasing function of b which means that the density may become negative. In fact, many density approximations of the NLS estimator suggested by the previous authors may become negative as we learn from the next section. To the contrary, our exact density is always positive although derived under somewhat stringent conditions.

The study of the small-sample properties of nonlinear estimation tangles with numerical issues as nonexistence of the least squares solution and the presence of multiple solutions. Understandably, the previous authors who studied the distribution approximation for finite n even do not mention the possibility of the estimate nonexistence or multiplicity because their occurrence vanishes with $n \to \infty$.

Several authors recognized the possibility of existence multiple local minima of the sum of squares in nonlinear regression and more generally multiple solutions of the estimating equation (M-estimator) leading to the concept of the intensity as a substitute of the density function (Skovgaard, 1990). This approach has been further developed in the following up work by Jensen & Wood (1998) and Almudevar *et al.* (2000). In contrast to that line of research, we aim at derivation of the explicit expression of the density distribution with applications to confidence interval and hypothesis testing in mind. Our method of derivation is different from those used by other authors: we use the fact that the estimating equation is linear in observations and derive the density by integrating out the multivariate density upon transformation.

For many years, it was believed that the distribution of the nonlinear least squares estimator, even if its unique, cannot be derived in closed form for a general model. We have derived this distribution and shown that it matches the distribution of a special case derived by

Fieller (1932) more than 80 years ago. First, the exact distribution is derived for a nonlinear regression model with one parameter and then extended to models with an arbitrary number of linear parameters and a coefficient (partial linear least squares). Second, the exact density is generalized to the estimating equation approach with fixed sample size.

In short, unlike extensive research on the approximation of the distribution from the asymptotic point of view we derived the distribution *directly* for fixed n. We outline how this derivation can be extended to incorporate the possibilities of nonexistence of the NLS estimate and multiple solutions of the normal equation.

The Gauss-Markov theorem is the landmark result of statistical science (Casella & Berger, 1990; Schervish, 1995). The classic optimal estimation results with fixed n, such as uniformly minimum variance unbiased estimation, hold for the narrow exponential family of distributions (Bickel & Doksum, 2007) and therefore cannot be applied to nonlinear regression. Novel, non-quadratic loss function approaches are needed to expand the optimality theory to more complicated nonexponential distribution statistical models, such as nonlinear regression. As an example, we illustrate the application of the density of the M-estimator by showing that the nonlinear least squares is optimal in the local sense.

The organization of the paper is as follows. The exact density distribution and its comparison with density approximations in the univariate case are presented in Section 2. In that section, we generalize the density to weighted nonlinear least squares and apply the result to a nonlinear regression with a linear part. In Section 3, we generalize our derivation to the estimating equation approach and apply the exact density to a regression with unknown coefficient. Numerical complications arising in nonlinear optimization or the solution of the normal equation and its effect on the density are discussed in Section 4. Section 5 contains results for the multivariate density approximation for nonlinear regression and the estimating equation approach. The formulation of the nonlinear Gauss-Markov theorem and the local optimality of the nonlinear least squares is found in Section 6. In the final section, we outline open problems for exact statistical inference in nonlinear statistical problems with small sample.

2. One intrinsically nonlinear parameter

In this section, we consider nonlinear regression with one intrinsically nonlinear parameter. First, we derive the exact density for the Nonlinear Least Squares (NLS) estimator. Second, we compare the exact density with the near exact density developed previously. Third, we generalize our derivation to the weighted NLS with a known weight matrix. Fourth, we apply the weighted NLS to a regression with one intrinsically nonlinear parameter and an arbitrary number of linear parameters.

With only one nonlinear parameter, β, the nonlinear regression is written as

$$y_i = f_i(\beta) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \qquad (1)$$

where y_i is the ith observation of the dependent variable, $f_i(\beta)$ is the nonlinear regression function, and $\{\varepsilon_i\}$ are iid normally distributed errors with zero mean and variance σ^2. In many applications, the original regression function is written in the form $f(\beta; x_i)$, where x_i is the explanatory (covariate) variable; we prefer the notation, $f_i(\beta) = f(\beta; x_i)$, simply for brevity. Regarding regression functions $\{f_i(\beta)\}$, we assume them to have continuous derivatives up to the second order on a fixed interval, $\beta \in (a, b)$; typically $a = -\infty$ and $b = +\infty$ (we refer to a and b as to the lower and upper domain parameters). Also, to comply with identifiability condition, we assume that two different parameter values cannot produce the same function values. That is, $f_i(b_1) = f_i(b_2)$ for at least one i implies $b_1 = b_2$.

The NLS estimate, $\hat{\beta}=\hat{\beta}_{\text{NLS}}$, minimizes the sum of squared residuals,

$$\sum_{i=1}^{n}(y_i - f_i(b))^2. \quad (2)$$

One more comment regarding the notation: While $\hat{\beta}$ denotes the estimator, b denotes a value it takes (it will also act as a dummy argument of functions). The NLS estimate can be found from the solution of the *normal* equation,

$$\sum_{i=1}^{n}(y_i - f_i(b))f_i(b)=0, \quad (3)$$

where the dot over the function means the derivative, $df_i/d\beta = \dot{f}_i$, because symbol $'$ is reserved for vector/matrix transposition. If the normal equation has a unique solution then the solution is the NLS estimate, otherwise, it may yield spurious solutions. Throughout the paper, we use the vector/matrix notation, $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\mathbf{f}(\beta) = (f_1(\beta), \ldots, f_n(\beta))'$, so that $d\mathbf{f}/d\beta=\dot{\mathbf{f}}$, $d^2\mathbf{f}/d\beta^2=\ddot{\mathbf{f}}$. In this notation, the normal equation takes the form

$$(\mathbf{y} - \mathbf{f}(b))'\ddot{\mathbf{f}}(b)=0. \quad (4)$$

The normal equation may have no solution. To avoid this complication, we may restrict our attention to regression with infinite tails: $\|\mathbf{f}(b)\| \to \infty$ when $|b| \to \infty$. As was shown by Demidenko (2000), this assumption guarantees that equation (4) has at least one solution for each y.

Before formulating the main result, we introduce relevant quantities and functions:

$$\begin{aligned} A &= A(b)=\|\dot{\mathbf{f}}(b)\|^2, \quad B=B(b;\beta)=\ddot{\mathbf{f}}'(b)(\mathbf{f}(b) - \mathbf{f}(\beta)), \\ C &= C(b)=\dot{\mathbf{f}}'(b)\ddot{\mathbf{f}}(b), \quad D=D(b;\beta)=\dot{\mathbf{f}}'(b)(\mathbf{f}(b) - \mathbf{f}(\beta)), \\ E &= E(b)=\|\dot{\mathbf{f}}(b)\|^2, \quad Q=Q(b;\beta,\sigma^2)=\frac{A^2+AB-CD}{\sqrt{\sigma^2 A(AE-C^2)}}. \end{aligned}$$

As is seen from the following theorem, the exact density distribution of the NLS estimator is expressed through the standard normal cumulative distribution function (cdf), Φ and its density ϕ.

Theorem 1—Let the solution of the normal equation, the NLS estimate, uniquely exist for each **y**. Let ε_i be iid $\mathcal{N}(0,\sigma^2)$ and $\dot{f}_i(b)\neq 0$ for at least one i. Then the exact density of the NLS estimator is given by

$$p_{EX}(b;\beta,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\|\dot{\mathbf{f}}(b)\|\exp\left[-\frac{1}{2\sigma^2}\frac{\left(\dot{\mathbf{f}}'(b)(\mathbf{f}(b)-\mathbf{f}(\beta))\right)^2}{\|\dot{\mathbf{f}}(b)\|^2}\right]\times a_{EX}(b;\beta,\sigma) \quad (5)$$

with the adjustment coefficient

$$a_{EX}(b;\beta,\sigma)=\left(1+\frac{AB-CD}{A^2}\right)(2\Phi(Q)-1)+\frac{2\sigma}{A}\sqrt{\frac{AE-C^2}{A}}\phi(Q). \quad (6)$$

The proof is found in the Appendix and proceeds as follows. First, without loss of generality, we can assume that $\dot{f}_1(b)\neq 0$. Expressing the first observation ($i = 1$) through b and the other $n - 1$ observations, a well-known theorem for the density upon nonlinear transformation is applied. Second, we derive the mean absolute value of a linear function of a normally distributed vector in closed form. Third, we express the function value and its derivatives for $i = 1$ through the full component vectors.

We make a few remarks regarding the exact density, p_{EX}. First, we note that $AE-C^2 \geq 0$ due to the Cauchy inequality, so the square root function appearing in the adjustment coefficient is well defined. Second, the adjustment coefficient, $a_{EX}(b; \beta, \sigma) \geq 0$, so that the density cannot be negative. To prove that we rewrite the first term in (6) as

$$\frac{\sqrt{\sigma^2 A(AE-C^2)}}{A^2}x(2\Phi(x)-1), \quad (7)$$

where $x=(A^2+AB-CD)/\sqrt{\sigma^2 A(AE-C^2)}$. Since $x(2\Phi(x) - 1) \geq 0$ for all x, we have $a_{EX}(b; \beta, \sigma) \geq 0$.

2.1. Comparison with the near exact density and other approximations

Early authors, including Pazman (1984), Hougaard (1985), and Skovgaard (1985), developed the so-called Near Exact (NE) density of the NLS estimator. This density coincides with (5) but has a different (simplified) adjustment coefficient,

$$a_{NE}(b;\beta)=1+\frac{AB-CD}{A^2}. \quad (8)$$

A nice feature of the NE density is that the cdf of the NLS estimator can be expressed through the standard normal cdf as

$$F_{NE}(b;\beta,\sigma^2)=\Phi\left(\frac{\dot{\mathbf{f}}'(b)(\mathbf{f}(b)-\mathbf{f}(\beta))}{\sigma\|\dot{\mathbf{f}}(b)\|}\right), \quad (9)$$

which follows from straightforward differentiation. Equivalently, one can write, $\dot{\mathbf{f}}'(\hat{\beta})(\mathbf{f}(\hat{\beta})-\mathbf{f}(\beta))\|\dot{\mathbf{f}}(\hat{\beta})\|^{-1}\sim\mathcal{N}(0,\sigma^2)$, where $\hat{\beta}$ is the NLS estimate. However, in order for (9) to be a real b cdf, three conditions must be met: (1) the argument of Φ must be an increasing function of b for each β, (2) the argument must approach $-\infty$ when b approaches its lower parameter domain and (2) the argument must approach $+\infty$ when b approaches its upper parameter domain.

There are two important distinctions between the EX and NE densities: (1) There is a presence of Φ in the former density, (2) the exact adjustment coefficient contains σ but the NE does not. When $\sigma\to 0$, the argument of Φ in (6) approaches $+\infty$, implying that $\Phi\to 1$. The second adjustment coefficient vanishes when $\sigma\to 0$, so that

$$\lim_{\sigma^2\to 0}\frac{p_{EX}(b;\beta,\sigma^2)}{p_{NE}(b;\beta)}=1.$$

In another extreme situation, when $\sigma\to\infty$, we have $\Phi\to 1/2$, so that

$$\lim_{\sigma^2\to\infty}\frac{p_{EX}(b;\beta,\sigma^2)}{p_{NE}(b;\beta)}=0.$$

This observation suggests that the NE density approximation can be called “a small-variance approximation” because it coincides with the exact one when σ^2 approaches zero.

Skovgaard (1985, formula (3.6)) suggested another expression for the density of the NLS estimator. As in the case of previous authors, the formula differs by the adjustment coefficients. More specifically, it has the same argument Q at functions Φ and ϕ, but different coefficients on these functions:

$$a_{SK}(b;\beta,\sigma)=\left(1+\frac{AB-CD}{A^2}\right)\Phi(Q)+\frac{\sigma}{A}\sqrt{\frac{AE-C^2}{A}}\phi(Q). \quad (10)$$

The difference between this and our a_{EX} (6) is in the coefficient at the first term. Clearly, the two expressions will be close when Q is large. It is possible to prove that the Skovgaard density is always positive because $a_{SK} > 0$.

A nonlinear model that can be reduced to a linear model has the form $\mathbf{f}(\beta) = g(\beta)\mathbf{x}$, where g is a strictly monotonic function. In the terminology of Pazman (1993), this model is an Intrinsically Linear (IL) model. For an IL model, $AB - CD = 0$ and $AE - C^2 = 0$, so we can approximate the density with $a_1 = 1$ and $a_2 = 0$, which leads to the IL approximation of $\hat{\beta}$,

$$p_{IL}(b;\beta,\sigma^2)=\frac{\|\dot{\mathbf{f}}(b)\|}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\frac{\left(\dot{\mathbf{f}}'(b)(\mathbf{f}(b)-\mathbf{f}(\beta))\right)^2}{\|\dot{\mathbf{f}}(b)\|^2}\right]. \quad (11)$$

All three densities, (5), (8), and (11), are the same for an intrinsically linear nonlinear regression.

A further way to simplify the approximation of the density is to replace the derivative, $\dot{\mathbf{f}}(b)$, with the derivative evaluated at the true parameter value β:

$$p_{\mathrm{ILL}}\left(b;\beta,\sigma^2\right)=\frac{\left|\dot{\mathbf{f}}'(\beta)\,\dot{\mathbf{f}}(b)\right|}{\sqrt{2\pi\sigma^2}\|\dot{\mathbf{f}}(\beta)\|}\exp\left[-\frac{1}{2\sigma^2}\frac{\left(\dot{\mathbf{f}}'(\beta)(\mathbf{f}(b)-\mathbf{f}(\beta))\right)^2}{\|\dot{\mathbf{f}}(\beta)\|^2}\right]. \quad (12)$$

This approximation will be called the Intrinsically Linear-Linear (ILL) approximation. This approximation is satisfactory in close proximity to the true value and reduces to the standard normal density upon transformation $z=\dot{\mathbf{f}}'(\beta)(\mathbf{f}(b)-\mathbf{f}(\beta))/\left(\sigma\|\dot{\mathbf{f}}(\beta)\|\right)$ under the assumption that $\dot{\mathbf{f}}'(\beta)(\mathbf{f}(b)-\mathbf{f}(\beta))$ is a monotonic function of b for each β. However, the area under the curve (12) is less than 1 if the range of z does not cover the entire line.

Finally, as follows from the standard asymptotic results (e.g. Gallant 1987), when $n\to\infty$ the distribution of $\hat{\beta}-\beta$ approaches the normal distribution with zero mean and varianc $\sigma^2/\sum_{i=1}^{n}\dot{f}_i^2(\beta)$, which implies the normal approximation,

$$p_N\left(b;\beta,\sigma^2\right)=\frac{\|\dot{\mathbf{f}}(\beta)\|}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(b-\beta)^2\|\dot{\mathbf{f}}(\beta)\|^2\right] \quad (13)$$

routinely used in nonlinear regression analysis.

The problem with approximations (8), (11) and (12) is that the area under the “density” is not 1. The fact that the area under the curve specified by the near exact density is not 1 has been mentioned by other authors, including Hougaard (1988), but perhaps the most discouraging feature of the NE density approximation is that it may become negative. For

example, for an exponential regression $\mathbf{f}(\beta) = e^{\beta \mathbf{x}}$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$, the adjustment coefficient evaluated at $b = 0$ is proportional to

$$\sum x_i^2 + \sum x_i^2 \left(1 - e^{\beta x_i}\right) e^{\beta x_i} - \frac{\sum x_i \left(1 - e^{\beta x_i}\right) \sum x_i^3}{\sum x_i^2}.$$

Assuming that max $x > 0$, some algebra shows that this quantity becomes negative when $\beta \to \infty$. Although negative values of the density usually appear at the tails, it may lead to infinite confidence intervals with large σ; the same problem may arise when the area under the density is greater than one.

Example—Five density approximations, (8), (10), (11), (12) and (13), are compared with the exact density (5) for exponential regression, $f_i(\beta) = e^{\beta x_i}$, where $x_i = i$ with $i = 1, \ldots, n = 6$ and the true value, $\beta = 0$ and $\sigma = 2.5$. The results are depicted in Figure 1. The area under the four curves is evaluated numerically by integration over the interval (−4, 0.5) using function `integrate` in R with 10^7 subdivisions (R Core Team, 2014). As seen from this figure, the problem with the NE density approximation starts when the NLS estimate takes values below −0.5, resulting in a considerable deviation of the area under the "density" from 1. The negative density is due to the negative values of the adjustment coefficient which falls below zero starting from $b = -0.5$. The IL and ILL approximations are close to each other and to the EX/NE densities. From a statistical prospective, the NLS estimator has a long left tail which is less visible in the NE approximation. The NE density deviates from the true one because the curvature of the exponential model is unbounded (Pazman, 1984).

An obvious remedy for possible negative values of the adjustment coefficient (and the density) in the NE approximation is to take the absolute value, $|a_{NE}(b; \beta)|$ or set it zero. This suggestion can be justified by the fact that function $x(2\Phi(x) - 1)$ in (7) can be well approximated by $|x|$. After ignoring the second adjustment coefficient in a_{EX}, we arrive at $|a_{NE}(b; \beta)|$.

2.2. Nonlinear regression with linear parameters

A nonlinear regression with m additive linear parameters takes the form

$$\mathbf{y} = \mathbf{X}\gamma + \mathbf{h}(\beta) + \varepsilon, \quad (14)$$

where it is assumed that the $n \times (m + 1)$ matrix, $[\mathbf{X}, \mathbf{h}(b)]$, has full rank for all b. We can reduce this regression to a univariate linear regression by fixing β and applying least squares for $\boldsymbol{\gamma}$, yielding $\tilde{\mathbf{y}} = \mathbf{W}\mathbf{y}$ and $\mathbf{f}(b) = \mathbf{W}\mathbf{h}(b)$, where $\mathbf{W} = \mathbf{I} - \mathbf{P}$ and $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the projection matrix, so that $\mathbf{W}^{1/2} = \mathbf{W}$. Sometimes, this procedure is called the partially linear least squares, or the Golub-Pereyra algorithm by the name of the original authors, Golub & Pereyra (2003). The reduced sum of squares can be rewritten as

$$(\mathbf{y} - \mathbf{X}\gamma - \mathbf{h}(b))' \mathbf{W} (\mathbf{y} - \mathbf{X}\gamma - \mathbf{h}(b)),$$

where $\boldsymbol{\gamma}$ may be taken as the true parameter because $\mathbf{WX} = \mathbf{0}$. Formally, the nonlinear regression is $\mathbf{f}(b) = \mathbf{X}\boldsymbol{\gamma} - \mathbf{h}(b)$, but since $\dot{\mathbf{f}}(b) = \dot{\mathbf{h}}(b)$ and $\mathbf{WX} = \mathbf{0}$, it is sufficient to use $\mathbf{f}(b) = \mathbf{h}(b)$ with matrix $\mathbf{W} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Consequently, the distribution of $\hat{\beta}$ does not depend on linear parameters $\boldsymbol{\gamma}$. To obtain the joint density, (a) write the density of $\hat{\gamma} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{h}(b))$ conditional on $\hat{\beta} = b$,

$$\hat{\gamma}|b \sim \mathcal{N}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{h}(\beta) - \mathbf{h}(b)), \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right), \quad (15)$$

then (b) the joint density for $\hat{\gamma}$ and $\hat{\beta}$ is the product of p_{EX} and this normal conditional density.

2.3. Possible improvements of Theorem 1

The exact density of the NLS estimator is derived under a stringent assumption on the uniqueness of the solution of the normal equation. If this assumption does not hold the area under p_{EX} may be different from 1, below is a motivating example.

Example (provided by a referee)—Define the circle nonlinear regression as $\mathbf{f}(\beta) = (\cos\beta, \sin\beta)$ on $\beta \in (-\pi, \pi]$, $n = 2$. The normal equation has two solutions, $\hat{\beta}_1 = a\tan(y_2/y_1)$ and $\hat{\beta}_2 = \hat{\beta}_1 - \pi \times sign\left(\hat{\beta}_1\right)$. One solution is the NLS estimate as the minimizer of the residual sum of squares (2) and another is the maximizer, the spurious solution. The conditions of Theorem 1 are violated and it is not difficult to find that the area under p_{EX} density is 2. This example, gives rise to the following improvement of Theorem 1.

Find the density of the NLS estimator under condition that the second derivative of the residual sum of squares is positive: this condition eliminates the spurious solution of the estimating equation corresponding to the maximum of the criterion function.

An advantage of the proof of Theorem 1 is that we can easily incorporate the condition on the positiveness of the second derivatives because it is expressed as a linear function of observations. Indeed, in the notation of the proof from the Appendix, let in addition to the normal equation $(z - g(b))'\dot{\mathbf{f}}(b) = 0$ we define the condition of the positiveness of the second derivative as

$$(\mathbf{z} - \mathbf{g}(b))'\ddot{\mathbf{f}}(b) = u + \sigma^{-1}\|\dot{\mathbf{f}}(b)\|^2,$$

where $\mathbf{z} = \sigma^{-1}(\mathbf{y} - \mathbf{f}(\beta))$ are normalized observations and $\mathbf{g}(b) = \sigma^{-1}(\mathbf{f}(b) - \mathbf{f}(\beta))$, and $u < 0$. The joint density of (b, u) can be derived by integrating out $y_2, \ldots, y_n$ following the line of the proof. Finally, to obtain the marginal distribution of b we condition on $u > 0$ by integration the bivariate density over $(-\infty, 0)$. For example, for the circle regression the density of the NLS (with the spurious solution eliminated) is proportional to

$$-\int_{-\infty}^{0}(2\pi\sigma^2)^{-1}ue^{-\frac{1}{2}\left\{\sigma^{-1}(\cos b-\cos\beta)-(u+\sigma^{-1})\cos b\right\}^2-\frac{1}{2}\left\{\sigma^{-1}(\sin b-\sin\beta)-(u+\sigma^{-1})\sin b\right\}^2}du$$
$$=(2\pi\sigma^2)^{-1}e^{-\frac{1}{2\sigma^2}}\left(1+\sigma^{-1}\cos(b-\beta)\frac{\Phi(\sigma^{-1}\cos(b-\beta))}{\phi(\sigma^{-1}\cos(b-\beta))}\right).$$

This integral has been evaluated in closed form using a technique presented in the proof based on the formula $\int_a^b u\phi(u)\,du=\phi(a)-\phi(b)$. Remarkably, the circle regression gives rise to a new family of distributions on the circle (Mardia & Jupp, 2000). In Figure 2 we depict the empirical (using 100,000 simulations) and theoretical densities of the NLS estimator in the circle regression, wrapped around the unit circle, under two scenarios of true parameters: the densities match perfectly. The density has a peak, the distance from the unit circle, at the true β, and the density with $\sigma = 2$ is more "uniform."

Another possible improvement of Theorem 1 is to incorporate the criterion on the global minimum into the derivation of the density. Indeed, the global criterion is usually formulated as follows: if b is a local minimizer such that $\|\mathbf{y}-\mathbf{f}(b)\|^2 < \|\mathbf{y}-\mathbf{f}(b_0)\|^2$, where b_0 is known, then b is the global minimizer of the sum of squared residuals (Demidenko, 2006, 2008). Express the above inequality as $\|\mathbf{y}-\mathbf{f}(b_0)\|^2-\|\mathbf{y}-\mathbf{f}(b)\|^2 = 2u$ where $u > 0$, or equivalently as $\mathbf{y}'(\mathbf{f}(b)-\mathbf{f}(b_0)) = u-(\|\mathbf{f}(b)\|^2-\|\mathbf{f}(b_0)\|^2)/2$. Since this equation is linear in $\mathbf{y}$ we can derive the joint density of b and u and then find the marginal density of b by integrating out u. The same method can be utilized to incorporate an existence criterion into density derivation which is also expressed as the difference of the sums of squared residuals (Demidenko, 1989, 2000).

3. Extension to the estimating equation approach

Theorem 1 can also be generalized to the case when any function, say $\mathbf{r}$, is used instead of $\dot{\mathbf{f}}$ in the normal equation (4). This leads to the estimating equation (EE) approach:

$$(\mathbf{y}-\mathbf{f}(b))'\mathbf{r}(b)=0, \quad (16)$$

where $\mathbf{r}(b)$ is a nonnull vector function such that $\mathbf{r}'(b)\dot{\mathbf{f}}(b) \neq 0$ for all b and, as before, $\{y_i - f_i(\beta)\}$ are iid $\mathcal{N}(0,\sigma^2)$. The solution of this equation is usually referred to as the M-estimator, or the EE estimator. Here we assume that the solution of (16) exists and is unique for each y (the discussion on the alternative intensity function is deferred to Section 5.2). Without loss of generality, one can assume $\mathbf{r}'(b)\dot{\mathbf{f}}(b) > 0$. It is well known that the EE approach yields a consistent M-estimator of β under mild conditions (Huber, 1981; Schervish, 1995). In particular, the estimating equation approach appears in the framework of the quasi (pseudo) -likelihood and the generalized estimating equation approach (Godambe, 1960; Gong & Samaniego, 1981; Zeger et al., 1988, Pawitan, 2001), robust statistics, and the instrumental variable approach in connection with the measurement error problem (Fuller, 1987).

Theorem 2—The exact density of the M-estimator, $\hat{\beta}_{EE}$, defined by estimating equation (16), is given by

$$p_{EE}\left(b;\beta,\sigma^2\right)=\frac{\|\mathbf{r}\left(b\right)\|}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\frac{\left(\mathbf{r}'\left(b\right)\left(\mathbf{f}\left(b\right)-\mathbf{f}\left(\beta\right)\right)\right)^2}{\|\mathbf{r}\left(b\right)\|^2}\right]\times a_{EX}\left(b;\beta,\sigma\right) \quad (17)$$

with adjustment coefficients computed by the same formulas as in Theorem 1, but

$$A=\|\mathbf{r}(b)\|^2,\ \ D=\mathbf{r}'(b)(\mathbf{f}(b)-\mathbf{f}(\beta)),\ \ B=\dot{\mathbf{r}}'(b)(\mathbf{f}(b)-\mathbf{f}(\beta)),$$
$$C=\mathbf{r}'(b)\dot{\mathbf{r}}(b),\ \ E=\|\dot{\mathbf{r}}(b)\|^2,\ \ G=\left|\mathbf{r}'(b)\dot{\mathbf{f}}(b)\right|.$$

The Near Exact and IL approximations have adjustment coefficients

$$a_{NE}\left(b;\beta\right)=\frac{G}{A}+\frac{AB-CD}{A^2},\ \ a_{IL}\left(b\right)=\frac{G}{A}.$$

The proof is a slight modification of that for Theorem 1 and is found in the Appendix. Basically, we replace $\dot{\mathbf{f}}$ with $\mathbf{r}$ and $\ddot{\mathbf{f}}$ with $\dot{\mathbf{r}}$, and repeat the three steps of the proof outlined in the preceding section. Note that we use the absolute value in G, so $\mathbf{r}'\left(b\right)\dot{\mathbf{f}}\left(b\right)$ may be negative.

As in the case of the NE distribution for the NLS estimator, it is straightforward to check that the EE estimator has the cdf given by

$$F_{EE}\left(b;\beta,\sigma^2\right)=\Phi\left(\frac{\mathbf{r}'\left(b\right)\left(\mathbf{f}\left(b\right)-\mathbf{f}\left(\beta\right)\right)}{\sigma\|\mathbf{r}\left(b\right)\|}\right). \quad (18)$$

Similar to our previous comment, in order for this function to be a cdf, F_{EE} as a function of b must satisfy three conditions.

Linear instrumental variable check: Let $f(\beta)=\beta x$ and $r(\beta)=(r_1, r_2, \ldots, r_n)'$ be a fixed vector such that $r'x>0$, so that the estimating equation takes the form $(\mathbf{y}-\beta\mathbf{x})'\mathbf{r}=\mathbf{0}$. This estimator emerges as the instrumental variable approach (Fuller, 1987). As follows from linear theory, $\hat{\beta}_{EE}=(\mathbf{y}'\mathbf{r})/(\mathbf{r}'\mathbf{x})\sim\mathcal{N}\left(\beta,\sigma^2\|\mathbf{r}\|/(\mathbf{r}'\mathbf{x})^2\right)$. But from Theorem 2, $B=C=E=0$ and $a_1=G/A=(\mathbf{r}'\mathbf{x})\ \|\mathbf{r}\|^2$, $a_2=0$, which yields

$$p_{EE}\left(b;\beta,\sigma^2\right)=\frac{\|\mathbf{r}\|}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\frac{(b-\beta)^2}{\|\mathbf{r}\|^2/(\mathbf{r}'\mathbf{x})^2}\right]\times\frac{(\mathbf{r}'\mathbf{x})}{\|\mathbf{r}\|^2}$$
$$=\frac{1}{\sqrt{2\pi\sigma^2\|\mathbf{r}\|^2/(\mathbf{r}'\mathbf{x})^2}}\exp\left[-\frac{1}{2\sigma^2}\frac{(b-\beta)^2}{\|\mathbf{r}\|^2/(\mathbf{r}'\mathbf{x})^2}\right],$$

the same density as follows from the linear theory. All three densities, NE, IL and ILL, coincide for linear regression.

3.1. Nonlinear regression with an unknown coefficient

In this section, we apply the theory of the estimating equation to regression with one intrinsically nonlinear parameter and an unknown coefficient at the nonlinear function,

$$\mathbf{y}=\nu\mathbf{h}\left(\beta\right)+\varepsilon, \quad (19)$$

assuming that the $n \times 2$ matrix [$\mathbf{h}(b)$, $\mathbf{h}(b)$] has a full rank for all b, and $\nu \neq 0$. After eliminating ν through linear least squares, we arrive at the estimating equation for β,

$$\mathbf{y}'\left(\mathbf{I}-\frac{\mathbf{h}\left(b\right)\mathbf{h}'\left(b\right)}{\left\|\mathbf{h}\left(b\right)\right\|^{2}}\right)\dot{\mathbf{h}}\left(b\right)=0,$$

which is equivalent to

$$\left(\mathbf{y}-\nu\mathbf{h}\left(b\right)\right)'\left(\mathbf{I}-\frac{\mathbf{h}\left(b\right)\mathbf{h}'\left(b\right)}{\left\|\mathbf{h}\left(b\right)\right\|^{2}}\right)\dot{\mathbf{h}}\left(b\right)=0,$$

where ν is the true value. Thus, letting

$$\mathbf{f}=\nu\mathbf{h},\ \ \mathbf{r}=\dot{\mathbf{h}}-\frac{\mathbf{h}'\dot{\mathbf{h}}}{\left\|\mathbf{h}\right\|^{2}}\mathbf{h},$$

Theorem 2 applies (the argument, b, is omitted for brevity). The required condition, $G=\mathbf{r}'\left(b\right)\dot{\mathbf{f}}\left(b\right)>0$ follows from the Cauchy inequality,

$$G=\nu\left(\left\|\dot{\mathbf{h}}\right\|^{2}-\frac{\left(\mathbf{h}'\dot{\mathbf{h}}\right)^{2}}{\left\|\mathbf{h}\right\|}\right)>0$$

since matrix $\left[\mathbf{h},\dot{\mathbf{h}}\right]$ has full rank. The exact and near exact densities require the derivative vector, $\dot{\mathbf{r}}$, which is straightforward to obtain in terms of derivatives of **h**,

$$\dot{\mathbf{r}}=\ddot{\mathbf{h}}-\frac{\mathbf{h}'\dot{\mathbf{h}}}{\left\|\mathbf{h}\right\|^{2}}\dot{\mathbf{h}}-\frac{1}{\left\|\mathbf{h}\right\|^{4}}\left[\left(\left\|\dot{\mathbf{h}}\right\|^{2}+\mathbf{h}'\ddot{\mathbf{h}}\right)\left\|\mathbf{h}\right\|^{2}-2\left(\mathbf{h}'\dot{\mathbf{h}}\right)^{2}\right]\mathbf{h}.$$

Fieller (1932) example. In this example, we test (17) through a distribution derived almost one hundred years in a very special case. Namely, we apply the estimating equation theory to the nonlinear regression

$$\mathbf{y}=\nu\left(1+\beta\mathbf{x}\right)+\varepsilon. \quad (20)$$

In previous notation we have $\mathbf{h}(b) = \mathbf{1}+b\mathbf{x}$, $\mathbf{r}(b) = \mathbf{x} - (\mathbf{1}+b\mathbf{x})'\mathbf{x}(\mathbf{1}+b\mathbf{x}) \; \| \mathbf{1}+\mathbf{bx} \|^{-2}$. For this model the exact distribution of the NLS estimator of β reduces to the distribution of the ratio of two normally distributed random variables. Indeed, the NLS estimator of β is

$$\hat{\beta}=\frac{\widehat{\nu\beta}}{\hat{\nu}}, \quad (21)$$

where $\widehat{\nu\beta}$ and $\hat{\nu}$ are the least squares slope and intercept, respectively. They have a bivariate normal distribution with marginal distributions $\widehat{\nu\beta}\sim\mathcal{N}\left(\nu\beta,\sigma^2/\sum(x_i-\overline{x})^2\right)$ and $\hat{\nu}\sim\mathcal{N}\left(\nu,\sigma^2/\sum x_i^2/\left(n\sum(x_i-\overline{x})^2\right)\right)$, respectively, and correlation coefficient $-\sum x_i/\sqrt{n\sum x_i^2}$. Fieller (1932) initially derived and Hinkley (1969) extended the distribution of the ratio of two normally distributed random variables, Z = X_1/X_2 with means μ_i, variances σ_i^2 $(i=1,2)$, and correlation coefficient ρ. Thus the Fieller result applies with μ_1 = $\nu\beta$, $\sigma_1^2=\sigma^2/\sum(x_i-\overline{x})^2$ and μ_2 = ν, $\sigma_2^2=\sum x_i^2/\left(n\sum(x_i-\overline{x})^2\right.$ and $\rho=-\sum x_i/\sqrt{n\sum x_i^2}$ as the benchmark testing of our EE distribution specified in Theorem 2. After some algebra, one can show that our density p_{EX} and the density of the ratio are identical while other known density approximations, such as NE or Skovgaard do not coincide with the density of the ratio.

3.1.1. Example: Michaelis-Menten model—The Michaelis-Menten model is a popular model in many application fields, especially in chemistry (Seber & Wild, 1989; Hadeler *et al.* 2007). It has a hyperbolic shape and describes the data with the right asymptote,

$$y_i=\frac{\nu x_i}{\beta+x_i}+\varepsilon_i. \quad (22)$$

We use Puromycin data on the velocity of a chemical reaction from the example provided in Bates & Watts (1988, p. 269), n = 12 with the true parameter values β = 0.064 and ν = 212.7 estimated from the data, where x is the substrate concentration and y is the velocity of the enzymatic reaction. The estimated residual standard error was σ = 10.9; we use σ = 40 here to amplify the difference between the densities. Four densities of the NLS estimator $\hat{\beta}$ with the associated 75% density limits are depicted in Figure 3 (we use the 75% confidence level, not 95%, for the illustrative purpose).

The exact and near exact densities practically coincide; that is why only three curves are seen. The IL density deviates only slightly from the exact/NE densities. The normal approximation is adequate only in a close proximity to the true value, but deviates from the exact density elsewhere due to the implied symmetry. Consequently, the normal density limits are narrow and symmetric around β. The Exact/NR and IL limits are wider and shifted to the right due to skewness of the estimates. The same conclusions, on average, can be drawn regarding confidence intervals for β.

In Figure 4, the exact expected value of $\hat{\beta}$ and its MSE are computed using the function `integrate` in `R` based on the exact density as a function of the standard error of residuals. For σ larger than 40, the negative bias becomes considerable (left plot). For $\sigma > 30$ the normal approximation variance, routinely used in existing statistical packages, considerably exceeds the exact value and inflates the p-value as follows from the right plot.

The NLS estimate in the Michaelis-Menten model may not exist especially for large σ; complications associated with nonexistence are discussed below in Section 4. Criteria for the NLS existence for this model are developed in Hadeler *et al.* (2007). A simple criterion based on the concept of the existence level (Demidenko, 2000, 2004), is as follows. Denote $S_1=\sum_{i=2}^{n} y_i^2$ and $S_2=\sum_{i=1}^{n} (y_i - \overline{y})^2$; if there is a starting value for ν and β with the sum of squares less than min(S_1, S_2), the NLS estimate exists.

Now we derive the joint distribution of the estimator for two parameters, the linear coefficient, $\hat{\nu}=\mathbf{y}'\hat{\mathbf{h}}/\hat{\mathbf{h}}'\hat{\mathbf{h}}$ and $\hat{\beta}$, where $\hat{\mathbf{h}}=\mathbf{h}(\hat{\beta})=\mathbf{x}/(\hat{\beta}+\mathbf{x})$. As in (15), we note that the distribution of $\hat{\nu}$ conditional on $\hat{\beta}=b$ is normal with mean $\nu\mathbf{h}'(\beta)\mathbf{h}(b)/\mathbf{h}'(b)\mathbf{h}(b)$ and variance $\sigma^2/\mathbf{h}'(b)\mathbf{h}(b)$. Therefore, the joint distribution is the product of this conditional normal distribution and the marginal density, $p_{EX}(b; \beta, \sigma^2)$.

3.2. Nonlinear regression with an unknown coefficient and linear parameters

The regression with one intrinsically nonlinear parameter, a linear part and an unknown coefficient takes the form

$$\mathbf{y}=\mathbf{X}\gamma+ \nu\mathbf{h}(\beta)+\varepsilon, \quad (23)$$

where $\boldsymbol{\gamma}$ is an m-dimensional parameter vector and ν and β are scalars. It is assumed that the $n \times (m + 2)$ matrix, [**X**, **h**(β), **h**(β)], has full rank for all β. This regression is an obvious combination of (14) and (19) with applications arising from two-compartment modeling as the solution to ordinary differential equations. For example, Gallant (1987) uses model (23) with $\mathbf{h}(\beta) = e^{\beta z}$.

After elimination of the linear parameters, $\boldsymbol{\gamma}$, we come to a one-parameter nonlinear least squares problem with coefficient ν treated as in the above section. Thus, the same formula for the density of $\hat{\beta}$ applies with $\mathbf{f} = \nu\mathbf{h}$ and **h** replaced with $\mathbf{f} = \nu\mathbf{W}\mathbf{h}$ and **Wh**, as in Section 2.2.

4. Numerical complications

The goal of this section is to examine how numerical complications arising when minimizing the sum of squares or solving the normal equation affect the density of the NLS estimator distribution. Apparently, two kinds of complications are possible: the solution does not exist and there are multiple solutions.

4.1. What happens when the NLS estimator does not exist

In the traditional asymptotic approach, nonexistence of the NLS estimate is not an issue because according to the classical maximum likelihood theory this probability vanishes when $n \to \infty$. However, in practice the probability that the NLS estimate does not exist cannot be ignored, see an example below. Only a handful of papers discuss the nonexistence of the NLS estimate either in terms of sufficient criteria Demidenko (1989, 2000, 2008) or necessary and sufficient conditions for specific nonlinear regression models Hadeler *et al.* (2007), Jukíc & Markovic (2010), and Jukíc (2014), to name a few. The purpose of this section is to illustrate how the probability of nonexistence affects the density distribution using simplistic examples of nonlinear regression where this probability is tractable.

The NLS estimate does not exist for some observations $y_1, \ldots, y_n$ when σ^2 is large and the regression curve has finite tails, namely, when $\| \mathbf{f}(b) \|$ does not go to infinity when $|b| \to \infty$. To illustrate, let us consider an exponential regression $f_i(\beta) = e^{\beta x_i}$ with two observations ($n = 2$). For simplicity, we assume that $x_1 = 1$ and $x_2 = 2$, implying that the regression curve is a half parabola in the observation space R^2. The normal equation turns into a cubic polynomial $2\theta^3 + (1 - 2y_2)\theta - y_1 = 0$ where $\theta = e^b$. Although the cubic equation has at least one real root, there may be no positive roots which means that the NLS estimate does not exist.

Using some algebra on the cubic equation similar to Demidenko (2000), one can show that there are no positive roots for θ if $y_1 \le 0$ and $y_2 < 1.2\,|y_1|^{2/3} + 0.5$. These explicit conditions allow computation of the probability that the NLS estimate does not exist for given β and σ expressed as an integral,

$$\Pr\left(\hat{\beta}\,\text{does not exist}\right) = \sigma^{-1}\int_{-\infty}^{0}\phi\left(\frac{y_1 - e^{\beta}}{\sigma}\right)\Phi\left(\frac{1.2(-y_1)^{2/3} + 0.5 - e^{2\beta}}{\sigma}\right)dy_1. \tag{24}$$

These probabilities as functions of σ are depicted in Figure 5 for $\beta = \ln 0.5$ and $\beta = 0$. The closer the true point is on the regression curve to zero or the larger σ, the greater the probability. For example, in regression with the true value $\beta = \ln 0.5 = -0.7$ and the standard deviation, $\sigma = 1$, a quarter of all iterations would diverge to $-\infty$ if standard nonlinear regression software such as `nls` in `R` were used.

Now we investigate the effect of nonexistence on the density of the NLS estimator. Two exact densities with the true value $\beta = 0$ and $\sigma = 0.5$, $\sigma = 1$ computed by formula (5) are depicted in Figure 6. Both densities have a long left tail, especially for $\sigma = 1$. The areas under the densities evaluated by numerical integration are shown in the upper-left corner. While for $\sigma = 0.5$ this area is 1, as it supposed to be, the area under the density for $\sigma = 1$ is less than 1 which reflects the fact that in approximately 20% of cases the NLS estimate does not exist. In general, numerical evaluation of the area under the density is a good test that points out to a nonignorable nonexistence of the NLS estimate.

4.2. What happens when the normal equation has multiple roots

In this section, we analyze the effect of violation of another assumption under which the exact and near exact densities were derived, namely, when the normal equation has multiple roots. Lehmann & Casella (1998, p. 451) and Skovgaard (1990) presented the analysis of multiple solutions in general terms by showing that the maximum likelihood estimator lacks consistency. Here, we illustrate the consequences with an example similar to that from the previous section, the parabolic regression, $f_1(b) = b$ and $f_2(b) = b^2$, $-\infty < b < \infty$ ($n = 2$), see Figure 7. Then the normal equation reduces to a cubic equation which admits a closed-form solution. As was shown in Demidenko (2000), the sum of squares $(y_1 - b)^2 + (y_2 - b^2)$ has two local minima if $y_2 > 3/4^{1/3}\,|y_1|^{2/3} + 1/2$. In the figure, the data point **y** leads to two local minima as the distance to the parabolic curve; the positive value is the true NLS estimate and the negative value is the false NLS estimate. For given β and σ, one can compute the probability of two local minima as the integral

$$\Pr(\text{multiple roots}) = \sigma^{-1}\int_{-\infty}^{\infty}\phi\left(\frac{y_1-\beta}{\sigma}\right)\Phi\left(\frac{\beta^2 - \frac{3}{4^{1/3}}|y_1|^{2/3} - 0.5}{\sigma}\right)dy_1. \quad (25)$$

For example, for $\beta = 0.5$ and $\sigma = 0.5$ the probability that the sum of squares has two local minima is 0.026. In Figure 7, the circle around $(0.5, 0.5^2)$ shows the 95% confidence region of (y_1, y_2). Probability (25) equals the density-weighted area of the intersection of this circle with the shaded area.

Now we turn our attention to how multiple roots (local minima) affect the densities. Although the NLS estimate should yield the absolute minimum of the sum of squares it may not hold in practice where a possibility of local minimum is a reality. In Figure 8, five densities are shown for $\beta = 0.5$ and $\sigma = 0.5$. The first two densities are theoretical and the last three densities are empirical, estimated with the Gaussian kernel based on 100,000 simulations. Three strategies for handling multiple roots are studied through simulations: The “Global minimum” strategy is when the global minimum is taken, which leads to the true NLS estimate. The “False minimum” is when the wrong local minimizer is taken as the false NLS estimate. The “50/50” strategy is when the true or the false estimate is randomly chosen with equal probability. We notice that the NE density has a prominent dip and differs from the EX density in the interval (−0.5, 0.25). The probability of the false NLS estimate computed by formula (25) is 0.026. The left bump around −0.5 in the densities reflects the possibility of the false estimate. In fact, when **y** is close to the center of the multiple minima region ($b = 0$), the roots of the normal equation have the same magnitude but different signs. The existence of multiple roots makes the area under the EX density larger than 1; for this example, it is 1.10. As in the case of the NLS nonexistence, the deviation of the area under the density from 1 is indicative of the nonuniqueness of the NLS estimate.

5. Multivariate case

In multivariate nonlinear regression, parameter $\boldsymbol{\beta}$ is an m-dimensional vector, so we use boldface now,

$$y_i = f_i(\boldsymbol{\beta}) + \varepsilon_i.$$

The NLS estimator, $\hat{\boldsymbol{\beta}}$, is the solution to the vector normal equation

$$\mathbf{F}'(\mathbf{b})(\mathbf{y} - \mathbf{f}(\mathbf{b})) = \mathbf{0}, \quad (26)$$

where $\mathbf{F}(\mathbf{b}) = \partial \mathbf{f}/\partial \mathbf{b}$ is the $n \times m$ derivative (Jacobian) matrix assuming $\det(\mathbf{F}'(\mathbf{b})\mathbf{F}(\mathbf{b})) \neq 0$ for all $\mathbf{b}$ and that $\mathbf{f}(\mathbf{b}_1) = \mathbf{f}(\mathbf{b}_2)$ implies $\mathbf{b}_1 = \mathbf{b}_2$ to ensure the identifiability. As in the univariate case, we assume that the solution of (26) uniquely exists for each $\mathbf{y} \in R^n$. The exact density for a multivariate nonlinear regression model cannot be derived in closed form, so we deal with approximations.

5.1. Density approximations for the NLS estimator

The NE density for the multivariate case was derived by Pazman (1984, 1993) using a geometric approach,

$$p_{NE}(\mathbf{b};\boldsymbol{\beta},\sigma^2) = \frac{\det\left[\mathbf{F}'(\mathbf{b})\mathbf{F}(\mathbf{b}) + \left(\mathbf{I}_m \otimes (\mathbf{f}(\mathbf{b}) - \mathbf{f}(\boldsymbol{\beta}))^{\perp}\right)\mathbf{H}(\mathbf{b})\right]}{(2\pi\sigma^2)^{m/2}\sqrt{\det(\mathbf{F}'(\mathbf{b})\mathbf{F}(\mathbf{b}))}} \times \exp\left[-\frac{1}{2\sigma^2}(\mathbf{f}(\mathbf{b}) - \mathbf{f}(\boldsymbol{\beta}))'\mathbf{P}(\mathbf{b})(\mathbf{f}(\mathbf{b}) - \mathbf{f}(\boldsymbol{\beta}))\right], \quad (27)$$

where $\mathbf{H}(\mathbf{b})$ is a $(nm) \times m$ stack matrix with the ith $m \times m$ block as the Hessian matrix of f_i, namely, $\mathbf{H}_i = \partial^2 f_i/\partial \mathbf{b}^2$, and $\mathbf{P}(\mathbf{b}) = \mathbf{F}(\mathbf{b})(\mathbf{F}'(\mathbf{b})\mathbf{F}(\mathbf{b}))^{-1}\mathbf{F}'(\mathbf{b})$ is the $n \times n$ projection matrix. Notation $^{\perp}$ is used to indicate the row distance vector to the plane spanned by vector columns of matrix $\mathbf{F}(\mathbf{b})$, namely, $(\mathbf{f}(\mathbf{b}) - \mathbf{f}(\boldsymbol{\beta}))^{\perp} = (\mathbf{f}(\mathbf{b}) - \mathbf{f}(\boldsymbol{\beta}))'(\mathbf{I}_n - \mathbf{P}(\mathbf{b}))$. Skovgaard (1985), Hougaard (1985) and Pazman (1999) derived expression (27) using the saddlepoint approximation.

We make several comments on density (27). First, the Kronecker product in (27) can be expressed explicitly as

$$(\mathbf{I}_m \otimes (\mathbf{f}(\mathbf{b}) - \mathbf{f}(\boldsymbol{\beta}))^{\perp})\mathbf{H}(\mathbf{b}) = \sum_{i=1}^{n} (f(\mathbf{b}) - f(\boldsymbol{\beta}))_i^{\perp}\mathbf{H}_i(\mathbf{b}). \quad (28)$$

Second, for $m = 1$, after some algebra, we arrive at the previous expression (5) with adjustment coefficient (8). Third, density (27) is equivalent to saying that

$$(\mathbf{F}'(\hat{\boldsymbol{\beta}})\mathbf{F}(\hat{\boldsymbol{\beta}}))^{-1/2}\mathbf{F}'(\hat{\boldsymbol{\beta}})(\mathbf{f}(\hat{\boldsymbol{\beta}}) - \mathbf{f}(\boldsymbol{\beta})) \sim \mathcal{N}(0, \sigma^2\mathbf{I}). \quad (29)$$

The IL approximation can be used as a simplification of p_{NE} by omitting term (28),

$$p_{IL}\left(\mathbf{b};\boldsymbol{\beta},\sigma^2\right)=\frac{\sqrt{\det\left(\mathbf{F}'\left(\mathbf{b}\right)\mathbf{F}\left(\mathbf{b}\right)\right)}}{\left(2\pi\sigma^2\right)^{m/2}}\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{f}\left(\mathbf{b}\right)-\mathbf{f}\left(\boldsymbol{\beta}\right)\right)'\mathbf{P}\left(\mathbf{b}\right)\left(\mathbf{f}\left(\mathbf{b}\right)-\mathbf{f}\left(\boldsymbol{\beta}\right)\right)\right].$$

Again, one can prove that this density is exact for the intrinsically linear regression model, the nonlinear model that can be reduced to linear after reparametrization. The ILL approximation is a further simplification when matrices are evaluated at the true value:

$$p_{\mathrm{ILL}}\left(\mathbf{b};\boldsymbol{\beta},\sigma^2\right)=\frac{\sqrt{\det\left(\mathbf{F}'\left(\boldsymbol{\beta}\right)\mathbf{F}\left(\boldsymbol{\beta}\right)\right)}}{\left(2\pi\sigma^2\right)^{m/2}}\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{f}\left(\mathbf{b}\right)-\mathbf{f}\left(\boldsymbol{\beta}\right)\right)'\mathbf{P}\left(\boldsymbol{\beta}\right)\left(\mathbf{f}\left(\mathbf{b}\right)-\mathbf{f}\left(\boldsymbol{\beta}\right)\right)\right].$$

The complications outlined in Section 4 apply to these densities as well. Namely, (27) can be negative and the volume under the densities may be not 1. Pazman (1984) expressed the deviation of (27) from the true density in terms of the minimum radius of intrinsic curvature as a part of the *nonoverlapping assumption*. The problem is that, in most nonlinear regression models, this minimum reaches zero if parameter space is not restricted, so that the nonoverlapping assumption does not hold. For example, the radius of intrinsic curvature approaches zero when $\beta \to -\infty$ in the exponential model $f_i(b) = e^{-bx_i}$ with $0 < x_1 < \ldots < x_n$. Of course, one can restrict the parameter space, say, $\beta > 0$, but that restriction would become questionable.

5.2. Density approximations for the estimating equation approach

In the multivariate EE approach, the M-estimator is found from the vector equation

$$\mathbf{R}'\left(\mathbf{b}\right)\left(\mathbf{y}-\mathbf{f}\left(\mathbf{b}\right)\right)=\mathbf{0}, \quad (30)$$

where **R**(**b**) is a $n \times m$ matrix such that matrix $\mathbf{R}'(\mathbf{b})\mathbf{F}(\mathbf{b})$ is positive definite. In a special case when **R** = **F**, the estimating equation approach reduces to NLS. Following the line of the proof in the univariate case, we obtain the NE approximation of the density of the M-estimator,

$$p_{NE}\left(\mathbf{b};\boldsymbol{\beta},\sigma^2\right)=\frac{\det\left[\mathbf{F}'(\mathbf{b})\mathbf{R}(\mathbf{b})+\left(\mathbf{I}_m\otimes(\mathbf{f}(\mathbf{b})-\mathbf{f}(\boldsymbol{\beta}))^{\perp}\right)\mathbf{E}(\mathbf{b})\right]}{\left(2\pi\sigma^2\right)^{m/2}\sqrt{\det(\mathbf{R}'(\mathbf{b})\mathbf{R}(\mathbf{b}))}}$$
$$\times\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{f}\left(\mathbf{b}\right)-\mathbf{f}\left(\boldsymbol{\beta}\right)\right)'\mathbf{P}\left(\mathbf{b}\right)\left(\mathbf{f}\left(\mathbf{b}\right)-\mathbf{f}\left(\boldsymbol{\beta}\right)\right)\right], \quad (31)$$

where **E**(**b**) is a $(nm) \times m$ stack matrix, with the ith $m \times m$ block, $\mathbf{E}_i = \partial \mathbf{r}_i/\partial \mathbf{b}$ and $\mathbf{r}_i = \mathbf{r}_i(\mathbf{b})$, being the ith row vector of matrix **R**(**b**), and the projection matrix is given by

$$\mathbf{P}\left(\mathbf{b}\right)=\mathbf{R}\left(\mathbf{b}\right)\left(\mathbf{R}'\left(\mathbf{b}\right)\mathbf{R}\left(\mathbf{b}\right)\right)^{-1}\mathbf{R}'\left(\mathbf{b}\right).$$

Simple algebra shows that (31) reduces to (17) when $m = 1$. The IL approximation is an obvious simplification,

$$p_{IL}\left(\mathbf{b};\boldsymbol{\beta},\sigma^2\right)=\frac{\det\left(\mathbf{F}'\left(\mathbf{b}\right)\mathbf{R}\left(\mathbf{b}\right)\right)}{\left(2\pi\sigma^2\right)^{m/2}\sqrt{\det\left(\mathbf{R}'\left(\mathbf{b}\right)\mathbf{R}\left(\mathbf{b}\right)\right)}}\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{f}\left(\mathbf{b}\right)-\mathbf{f}\left(\boldsymbol{\beta}\right)\right)'\mathbf{P}\left(\mathbf{b}\right)\left(\mathbf{f}\left(\mathbf{b}\right)-\mathbf{f}\left(\boldsymbol{\beta}\right)\right)\right].$$

The saddlepoint approximation to the density in more general settings expressed in terms of the cumulant generating function were developed by Field (1982) and Strawderman *et al.* (1996), among others (see a nicely written paper by Goutis & Casella (1999) as a general introduction to the saddlepoint approximation). One can show that the saddlepoint approximation yields p_{IL}. There is a bulk of work on the distribution of the M-estimator in a less idealistic situation when there are several solutions of the EE with the density counterpart termed the intensity function derived in a general form. For example, Almudevar *et al.* (2000) provide an explicit solution for the density only for a linear Huber's robust regression.

The multivariate version of the EE densities can be applied to nonlinear regressions with linear part and unknown coefficients, such as $\mathbf{X}\boldsymbol{\gamma}+\nu_1\mathbf{f}_1(\beta_1)+\boldsymbol{\nu}_2\mathbf{f}(\boldsymbol{\beta}_2)$, similar to regression with one intrinsically nonlinear parameter.

5.3. Possible improvements of multivariate density derivation

We have indicated that our derivation of the univariate density may be improved by brushing off spurious solutions or by accounting for nonexistence of the estimate. Many previous derivations of the multivariate density relied on the saddle point approximation. We provide an outline of an alternative derivation of the density as a generalization of the univariate case in the Appendix. As before, we assert that this derivation is flexible enough to incorporate conditions on positive definiteness of the Hessian, or at least positiveness of the diagonal elements, and existence criteria once they are expressed as linear functions of observations.

6. Gauss-Markov theorem for nonlinear regression

The aim of this section is to show how the density results can be used to study methods of optimal parameter estimation. Classic quadratic loss function criteria for optimality do not work in this setting because (a) the moments may not exist, such as in Fieller problem, and therefore the concept of unbiasedeness is invalid, and (b) the distribution is highly asymmetric.

Gauss-Markov theorem is the landmark result of linear regression with normally distributed variables: if $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where components of error vector $\boldsymbol{\varepsilon}$ are iid normally distributed with zero mean and matrix $\mathbf{X}$ has full rank, then the maximum likelihood (least squares) estimators of the beta-coefficients are unbiased and have minimum variance for any finite n. This result holds for the NLS estimator in nonlinear regression in large sample ($n\to\infty$). The paramount question: does it hold for finite n? One can guess that this question was on the mind of statisticians since the Gauss-Markov theorem was proved for linear regression. Two challenging questions should be answered before even formulating the problem:

1. What family of competitive estimators should be considered?

2. How to define the most precise estimator with a known cdf which may have infinite mean or variance like in the Fieller test example when the traditional mean square error loss function does not work?

To answer the first question we suggest to consider the EE approach with the family of estimators defined by equation (30). To motivate our choice we refer to linear regression restricting to linear estimators that reduces to equation $\mathbf{R}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$, where matrix $\mathbf{R}$ has full rank and matrix $\mathbf{R}'\mathbf{X}$ is nonsingular ('linear' means that matrix $\mathbf{R}$ does not depend on $\boldsymbol{\beta}$). The EE estimator $\hat{\boldsymbol{\beta}}_{\mathbf{R}} = \left(\mathbf{R}'\mathbf{X}\right)^{-1}\mathbf{R}'\mathbf{y}$ is unbiased with the covariance matrix $\operatorname{cov}(\hat{\boldsymbol{\beta}}_{\mathbf{R}}) = \sigma^2(\mathbf{R}'\mathbf{X})^{-1}(\mathbf{R}'\mathbf{R})(\mathbf{X}'\mathbf{R})^{-1} \geq \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \operatorname{cov}(\hat{\boldsymbol{\beta}}_{LS})$, as the Gauss-Markov theorem says (the last inequality follows from the fact that matrix $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is nonnegative definite). Restriction on unbiased estimators is crucial—the EE approach defined by (30) can be viewed as the unbiasedeness counterpart.

To answer the second question, we invoke the concept of the confidence interval (CI) termed here as inverse cdf (see Casella & Berger, 1990, pp. 417–418). Hereafter in this section we consider the one-parameter statistical estimation problem. Let the cdf of an estimator be defined as $F(b; \beta)$ such that $F(b; \beta)$ is a strictly decreasing function of β for every fixed b and $\lim_{\beta \to -\infty} F(b; \beta) = 1$, $\lim_{\beta \to \infty} F(b; \beta) = 0$. If α is the significance level, say, $\alpha = 0.05$ and $\hat{\beta}$ is the EE estimator, the lower and the upper limits of the CI for β are the solutions to the equations $F(\hat{\beta}; \beta) = 1 - \alpha/2$ and $F(\hat{\beta}; \beta) = \alpha$, respectively. The solutions exist and unique as follows from the assumptions on cdf F, although the existence of the limiting values for F are not crucial— if the the solution does not exist the CI limit is set to ∞.

Definition 3

Let the cdfs of two estimators of the true β be $F_1(b; \beta)$ and $F_2(b; \beta)$. We say that the first estimator is more precise if its inverse cdf CI is a subinterval of the second for every $\alpha > 0$.

It is easy to see that the two estimators are unbiased and normally distributed with cdfs $\Phi((b - \beta)/\sigma_1)$ and $\Phi((b - \beta)/\sigma_2)$ the first estimator is more precise if and only if $\sigma_1 \leq \sigma_2$.

Now we turn our attention to formulation of the extended Gauss-Markov theorem; as mentioned before one intrinsically nonlinear parameter nonlinear regression with known σ is discussed here. Also we restrict ourselves with near exact approximation of the density distribution since the cdfs for the NLS and EE estimators admit a closed-form expression via Φ defined by (9) and (18). The family of nonlinear regressions has to be constrained as well to comply with the uniqueness of the cdf solutions in inverse cdf CI. We adopt the following definition of the unidirected regression (Demidenko, 2006):

Definition 4

A nonlinear regression is called unidirected if the vector derivatives constitute a sharp angle for any pair of parameter values, i.e. $\dot{\mathbf{f}}'(\beta_1)\dot{\mathbf{f}}(\beta_2) > 0$ for every β_1 *and* β_2.

It is easy to see that for a unidirected regression the inverse cdf CI based on the cdf (9) is unique for any α. Indeed, if $\hat{\beta}$ is the NLS estimator the lower and upper limits of the CI are the solutions to the equation $H.(\beta; \hat{\beta}) = \sigma\Phi^{-1}(1 - \alpha/2)$, where

$$H.(\beta;b) = \frac{\left|\dot{\mathbf{f}}'(b)\left(\dot{\mathbf{f}}'(b) - \mathbf{f}(\beta)\right)\right|}{\|\dot{\mathbf{f}}(b)\|}$$

But for the unidirected function H is a strictly increasing function of β for any fixed b, including $b = \hat{\beta}$ so that the solutions are unique.

Now we turn our attention to the EE estimator as the solution to (16) defined by function **r**. In general terms, the Gauss-Markov theorem for nonlinear regression proves that $\mathbf{r}(b) = \dot{\mathbf{f}}(b)$ yields a more precise EE estimator than any other choice of function $\mathbf{r}(b)$. The expression for cdf (18) is crucial here.

Theorem 5. Local near-exact Gauss-Markov theorem for nonlinear regression

Let (1) be a unidirected nonlinear regression and $\mathbf{r}'(b)\dot{\mathbf{f}}(\beta) > 0$ for every b and β. Denote

$$H_{\mathbf{r}}(\beta;b) = \frac{\left|\mathbf{r}'(b)(\mathbf{f}(b) - \mathbf{f}(\beta))\right|}{\|\mathbf{r}(b)\|}$$

Then (a) the EE estimator with $\mathbf{r} = \mathbf{r}_*$ *is the most precise if* $H_{\mathbf{r}}(\beta; b) \le H_{\mathbf{r}*}(\beta; b)$ *for all* β *and b. (b) The NLS/ML estimator is locally precise meaning that*

$$\lim_{|b-\beta|\to 0} \frac{H_{\mathbf{r}}(\beta;b)}{H.(\beta;b)} \le 1$$

for any function $\mathbf{r} = \mathbf{r}(b)$.

Proof

a. As follows from (18) the inverse cdf (1−α)100% CI can defined as the interval $I_{\mathbf{r}} = \{\beta: H_{\mathbf{r}}(\beta; b) \le \sigma\Phi^{-1}(1-\alpha/2)\}$. If $\mathbf{r}_*$ is such that $H_{\mathbf{r}}(\beta;b) \le H_{\mathbf{r}_*}(\beta;b)$ then the interval $\left\{\beta: H_{\mathbf{r}_*}(\beta;b) \le \sigma\Phi^{-1}(1-\alpha/2)\right\}$ is a subinterval of $I_{\mathbf{r}}$.

b. For the NLS/MLE estimator we have

$$\lim_{|b-\beta|\to 0} \frac{H.(\beta;b)}{|b-\beta|} = \lim_{|b-\beta|\to 0} \frac{\left|\dot{\mathbf{f}}'(b)(\mathbf{f}(b) - \mathbf{f}(\beta))\right|}{\|\dot{\mathbf{f}}(b)\|\,|b-\beta|} = \frac{\left|\dot{\mathbf{f}}'(\beta)\dot{\mathbf{f}}(\beta)\right|}{\|\dot{\mathbf{f}}(\beta)\|} = \|\dot{\mathbf{f}}(\beta)\|.$$

For the EE estimator with any $\mathbf{r} = \mathbf{r}(b)$ we have

$$\lim_{|b-\beta|\to 0}\frac{H_{\mathbf{r}}(\beta;b)}{|b-\beta|}=\frac{\left|\mathbf{r}'(\beta)\dot{\mathbf{f}}(\beta)\right|}{\|\mathbf{r}(\beta)\|}\le\|\dot{\mathbf{f}}(\beta)\|$$

due to Cauchy-Schwarz inequality, end of proof.

We make a few comments regarding this theorem: (a) function H provides a criterion for search of the most precise EE estimator through the vector function **r**(*b*). (b) the Gauss-Markov theorem holds in the local sense, for tight CI.

7. Future work

The present work on the density distribution for the estimating equation approach can be extended in several theoretical research directions. First, the density distribution of the variance estimate, $\hat{\sigma}^2=\sum_{i=1}^{n}(y_i-f_i(\hat{\beta}))^2/(n-m)$ should be derived and the confidence intervals for the parameters should be corrected based on this density. We believe that this knowledge may show the way to find a better estimator of σ^2.

Second, the extension to the estimating equation approach developed in this paper opens the possibility of studying the small-sample properties of the weighted nonlinear regression (Carroll & Ruppert, 1988), and the quasi/pseudo-likelihood approach (Gong & Samaniego, 1981; Zeger *et al.* 1988) with a normally distributed response variable. In the former approach, which is sometimes called generalized nonlinear least squares, the weight matrix is a function of the regression parameters and the estimating equation takes the form

$$\mathbf{F}'(\mathbf{b})\mathbf{W}(\mathbf{b})(\mathbf{y}-\mathbf{f}(\mathbf{b}))=\mathbf{0},$$

where **W**(**b**) is the weight matrix, which turns into estimating equation (30) by letting **R**(**b**) = **W**(**b**)**F**(**b**).

Third, following the line of our density derivation, one may study the small-sample properties of the Generalized Estimating Equation (GEE) approach, often applied to the analysis of longitudinal and cluster data (Fitzmaurice & Molenberghs, 2009), with estimating equation $\sum_{k=1}^{K}\mathbf{X}_k'\mathbf{W}_k(\mathbf{b})(\mathbf{y}_k-\mathbf{X}_k\mathbf{b})=\mathbf{0}$ and the weight matrix $\mathbf{W}_k(\mathbf{b})=(\mathbf{D}_k(\mathbf{b})\mathbf{R}_k\mathbf{D}_k(\mathbf{b}))^{-1}$, where $\mathbf{R}_k$ is the correlation matrix and $\mathbf{D}_k(\mathbf{b})$ is the diagonal matrix of standard deviations.

Fourth, the exact (or improved) statistical inference can be extended to statistical models when variance-covariance matrices are subject to estimation using the estimating equation approach (Paige & Trindade, 2009).

Fifth, the normal approximation density can be improved for multivariate nonlinear regression (Vonesh *et al.*, 2001) and, more generally, nonlinear mixed models (Demidenko, 2013).

Sixth, the near exact density approximation of estimating equation approach allows formulation of the nonlinear Gauss-Markov theorem. We have derived a local version of the theorem; more work has to be done in this direction, intriguing from the theoretical and important from practical perspective. In particular, the question whether NLS/ML estimator remains precise in the global sense is open and awaiting the solution. It is possible that for a concrete regression function one may find a better estimator.

Seventh, the idea finding the exact distribution conditional on the difference of residual sum of squares in Section 2.3 may be applied to adjust for nonexistence of the NLS, frequently forgotten in the literature of higher order asymptotics, because the criteria of existence usually have the form $\|\mathbf{y}-\mathbf{f}(\beta)\|^2-\|\mathbf{y}-\mathbf{f}(\beta_0)\|^2<0$ where β_0 is a fixed parameter value (Demidenko 1989, 2008).

On the practical side, as was mentioned above, the existing commercial statistical packages, such as SAS and STATA, or freely available R rely on the normal approximation which yields symmetric Wald confidence intervals. Arguments against computationally intensive confidence intervals in nonlinear models were adequate several decades ago, but not today. In addition to Wald confidence intervals and *p*-values, software offering nonlinear regression should be augmented with more precise numerical and statistical features, such as testing whether the found minimum of the sum of squares is global and improved asymmetric profile-confidence intervals and associated *p*-values such as implemented in the recent R package nlreg (Brazzale *et al.*, 2007, Brazzale & Davison, 2008).

Acknowledgments

I thank two referees for their helpful and enlightening comments and especially for providing the circle regression example. In part, this work was supported by the following grants from the National Institutes of Health (NIH): R01 LM012012A, 8 P20 GM103534, 1U01CA196386, P30 CA23108-37, and 1UL1TR001086.

Appendix

A. Proof of Theorem 1

The following lemma is an obvious reformulation of a textbook result on the density of the multivariate distribution under a nonlinear transformation.

Lemma 6

Let random vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ have density $p_{\mathbf{y}}(y_1, y_2, \ldots, y_n)$ and random variable b be a unique solution to a nonlinear equation,

$$g(b, y_1, y_2, \ldots, y_n)=0, \quad (32)$$

where g is a nonlinear function such that $\frac{\partial}{\partial y_1} g(b, y_1, y_2, \ldots, y_n) \neq 0$. *Moreover, let* y_1 *be expressed from* (32) *via inverse function,* $y_1 = h(b, y_2, \ldots, y_n)$. *Then the density of b is given by*

$$p(b)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left|\frac{\partial h}{\partial b}\right|p_{\mathbf{y}}(h(b,y_2,\ldots,y_n),y_2,\ldots y_m)dy_2\cdots dy_m.$$

Since the normal equation does not depend on σ, we can assume that σ is fixed and known. Moreover, we can normalize the normal equation by replacing the observations and the regression function with y_i/σ and $f_i(\beta)/\sigma$, respectively, so that observations will have unit variance. Also, without loss of generality, we can assume that $\dot{f}_1(b)\neq 0$ for all b. The proof has three steps.

#**1**. Express y_1 through the NLS estimator b and y_2, …., y_n, as

$$y_1=h(b,y_2,\ldots,y_n)=f_1(b)-\frac{\sum_{i=2}^{n}(y_i-f_i(b))\dot{f}_i(b)}{\dot{f}_1(b)}$$

and apply Lemma 3 to obtain the density of the NLS estimator as an integral:

$$p_{\mathrm{NLS}}(b;\beta)=(2\pi)^{-n/2}\int_{R^{n-1}}\left|\frac{\partial h}{\partial b}\right|e^{-\frac{1}{2}h^2-\frac{1}{2}\sum_{i=2}^{n}(y_i-f_i(b))^2}dy_2\cdots dy_n. \quad (33)$$

The needed derivative is obtained using elementary calculus, $dh/db=R+\sum_{i=2}^{n}z_iq_i$, where $z_i=y_i-f_i(\beta)\sim\mathcal{N}(0,1)$ and

$$R=\frac{1}{\dot{f}_1(b)}\sum_{i=1}^{n}\dot{f}_i^2(b)+\sum_{i=2}^{n}(f_i(\beta)-f_i(b))q_i(b)$$
$$q_i=\frac{\dot{f}_1(b)\dot{f}_i(b)}{\dot{f}_1^2(b)}-\frac{\ddot{f}_i(b)}{\dot{f}_1(b)},\quad i=2,3,\ldots,n.$$

Introducing the $(n-1)\times 1$ vectors $\mathbf{z}$, $\mathbf{q}$ and $\mathbf{g}$ with the ith components z_i, q_i and $g_i=\dot{f}_i(b)/\dot{f}_1(b)$, respectively, rewrite integral (33) in a compact form as

$$p_{\mathrm{NLS}}(b;\beta)=(2\pi)^{-n/2}\int_{\mathbf{z}\in R^{n-1}}\left|R+\mathbf{q}'\mathbf{z}\right|e^{-\frac{1}{2}\left(\mathbf{z}'\mathbf{g}-p\right)^2-\frac{1}{2}\|\mathbf{z}\|^2}d\mathbf{z}, \quad (34)$$

where $p=\left[(\mathbf{f}(b)-\mathbf{f}(\beta))'\ddot{\mathbf{f}}(b)\right]/\dot{f}_1(b)$.

#**2**. Express integral (34) as an expectation of $|R+\mathbf{q}'\mathbf{z}|$ over a normally distributed vector using cdf Φ. Using an elementary fact ($a>0$),

$$\int_{-\infty}^{\infty}|u|e^{-au^2+cu}du=\frac{1}{a}+\frac{c\sqrt{a\pi}}{2a^2}e^{\frac{1}{4a}c^2}\left(2\Phi\left(\frac{c}{\sqrt{2a}}\right)-1\right),$$

which follows from equality $\int_a^b u\phi(u)\,du = \phi(a) - \phi(b)$ where $\boldsymbol{\phi} = \Phi'$. After some algebra, we obtain

$$E_{X\sim\mathcal{N}(\kappa,\tau^2)}(|X|) = \frac{e^{-\frac{1}{2\tau^2}\kappa^2}}{\tau\sqrt{2\pi}}\left[2\tau^2 + \kappa\tau\sqrt{2\pi}e^{\frac{1}{2\tau^2}\kappa^2}\left(2\Phi\left(\frac{\kappa}{\tau}\right) - 1\right)\right]. \quad (35)$$

We apply this formula to (34) letting $X = R + \mathbf{q}'\mathbf{z}$ and expressing the integrand through a density of a $(n-1)$ dimensional normal variable. Using the formula

$$(\mathbf{I} + \mathbf{g}\mathbf{g}')^{-1} = \mathbf{I} - \frac{\mathbf{g}\mathbf{g}'}{1+\|\mathbf{g}\|^2},$$

we represent

$$\|\mathbf{z}\|^2 + \left(\mathbf{z}'\mathbf{g} - p\right)^2 = (\mathbf{z}+\boldsymbol{\mu})'\left(\mathbf{I} + \mathbf{g}\mathbf{g}'\right)(\mathbf{z}+\boldsymbol{\mu}) + \frac{p^2}{1+\|\mathbf{g}\|^2},$$

where $\boldsymbol{\mu} = -p(\mathbf{I} + \mathbf{g}\mathbf{g}')^{-1}\mathbf{g} = -p(1 + \|\mathbf{g}\|^2)^{-1}\mathbf{g}$. Using this result, we can rewrite the exponential part as

$$\begin{aligned}
&(2\pi)^{-(n-1)/2} e^{-\frac{1}{2}\left(\mathbf{z}'\mathbf{g}-p\right)^2 - \frac{1}{2}\|\mathbf{z}\|^2} \\
&= (2\pi)^{-(n-1)/2} e^{-.5c} e^{-\frac{1}{2}(\mathbf{z}+\boldsymbol{\mu})'\left(\mathbf{I} - \frac{\mathbf{g}\mathbf{g}'}{1+\|\mathbf{g}\|^2}\right)^{-1}(\mathbf{z}+\boldsymbol{\mu})} \\
&= \frac{e^{-\frac{p^2}{2(1+\|\mathbf{g}\|^2)}}}{\sqrt{\det\left(\mathbf{I}+\mathbf{g}\mathbf{g}'\right)}}\left[(2\pi)^{-(n-1)/2}\sqrt{\det(\mathbf{I}+\mathbf{g}\mathbf{g}')}\, e^{-\frac{1}{2}(\mathbf{z}+\boldsymbol{\mu})'\left(\mathbf{I} - \frac{\mathbf{g}\mathbf{g}'}{1+\|\mathbf{g}\|^2}\right)^{-1}(\mathbf{z}+\boldsymbol{\mu})}\right].
\end{aligned}$$

But $\det(\mathbf{I} + \mathbf{g}\mathbf{g}') = 1 + \|\mathbf{g}\|^2$, so finally the distribution of $R + \mathbf{q}'\mathbf{z}$ is a constant times the normal distribution, or symbolically,

$$X = R + \mathbf{q}'\mathbf{z} \sim \mathcal{N}\left(R + \frac{p}{1+\|\mathbf{g}\|^2}\mathbf{g}'\mathbf{q}, \|\mathbf{q}\|^2 - \frac{\left(\mathbf{g}'\mathbf{q}\right)^2}{1+\|\mathbf{g}\|^2}\right).$$

Combining this result with formula (35) and letting $X = R + \mathbf{q}'\mathbf{z}$ with

$$\kappa = R + \frac{p}{1+\|\mathbf{g}\|^2}\mathbf{g}'\mathbf{q}, \quad \tau^2 = \|\mathbf{q}\|^2 - \frac{\left(\mathbf{g}'\mathbf{q}\right)^2}{1+\|\mathbf{g}\|^2},$$

we obtain the following result. Let $\mathbf{z}\sim\mathcal{N}(0, \mathbf{I})$, p and R be scalar and $\mathbf{g}$ and $\mathbf{q}$ be vectors. Then

$$(2\pi)^{-n/2}\int_{\mathbf{z}\in R^{n-1}}\left|R+\mathbf{q}'\mathbf{z}\right|e^{-\frac{1}{2}\left(\mathbf{z}'\mathbf{g}-p\right)^2-\frac{1}{2}\|\mathbf{z}\|^2}d\mathbf{z}=\frac{\sqrt{1+\|\mathbf{g}\|^2}}{\sqrt{2\pi}}e^{-\frac{p^2}{2+(1+\|\mathbf{g}\|^2)}}\times(a_1+a_2), \quad (36)$$

where

$$a_1=\frac{R+\frac{p}{1+\|\mathbf{g}\|^2}\mathbf{g}'\mathbf{q}}{1+\|\mathbf{g}\|^2}\left(2\Phi\left(\frac{R+\frac{p}{1+\|\mathbf{g}\|^2}\mathbf{g}'\mathbf{q}}{\sqrt{\|\mathbf{q}\|^2-\frac{\left(\mathbf{g}'\mathbf{q}\right)}{1+\|\mathbf{g}\|^2}}}\right)-1\right), \quad (37)$$

$$a_2=\frac{1}{1+\|\mathbf{g}\|^2}\sqrt{\frac{2}{\pi}}e^{-\frac{\left(R+\frac{p}{1+\|\mathbf{g}\|^2}\mathbf{g}'\mathbf{q}\right)^2}{2\left(\|\mathbf{q}\|^2-\frac{\left(\mathbf{g}'\mathbf{q}\right)^2}{1+\|\mathbf{g}\|^2}\right)}}\sqrt{\|\mathbf{q}\|^2-\frac{(\mathbf{g}'\mathbf{q})^2}{1+\|\mathbf{g}\|^2}}. \quad (38)$$

Note that without the absolute value, we obtain an easier expression,

$$(2\pi)^{-n/2}\int_{\mathbf{z}\in R^{n-1}}\left(R+\mathbf{q}'\mathbf{z}\right)e^{-\frac{1}{2}\left(\mathbf{z}'\mathbf{g}-p\right)^2-\frac{1}{2}\|\mathbf{z}\|^2}d\mathbf{z}=\frac{\sqrt{1+\|\mathbf{g}\|^2}}{\sqrt{2\pi}}e^{-\frac{p^2}{2(1+\|\mathbf{g}\|^2)}}a, \quad (39)$$

where

$$a=\frac{R+\frac{p\mathbf{g}'\mathbf{q}}{1+\|\mathbf{g}\|^2}}{1+\|\mathbf{g}\|^2}. \quad (40)$$

#**3**. Express the density through the *n*-dimensional vector **f** and its derivatives and simplify. Letting

$$a_i=\dot{f}_i(b), b_i=\ddot{f}_i(b), \quad \Delta_i=f_i(\beta)-f_i(b),$$
$$A=\sum_{i=1}^n a_i^2, \quad D=\sum_{i=1}^n a_i\Delta_i, \quad B=\sum_{i=1}^n b_i\Delta_i, \quad C=\sum_{i=1}^n b_i a_i, \quad E=\sum_{i=1}^n b_i^2$$

for brevity, some tedious algebra yields

$$\begin{aligned}
&\tfrac{1}{1+\|\mathbf{g}\|^2}\left(R+\tfrac{p\mathbf{g}'\mathbf{q}}{1+\|\mathbf{g}\|^2}\right)\\
=&\tfrac{1}{A}a_1^2\left[\tfrac{A}{a_1}+\sum_{i=2}^n\Delta_i\left(a_i\frac{b_1}{a_1^2}-b_i\frac{1}{a_1}\right)-\frac{a_1^2}{A}\frac{1}{a_1^2}D\sum_{i=2}^n a_i\left(a_i\frac{b_1}{a_1^2}-b_i\frac{1}{a_1}\right)\right]\\
=&\tfrac{a_1}{A}\left[A+\tfrac{b_1}{a_1}\sum_{i=2}^n\Delta_i a_i-\sum_{i=2}^n\Delta_i b_i-\frac{D}{Aa_1}\left(b_1\sum_{i=2}^n a_1^2-a_1\sum_{i=2}^n a_i b_i\right)\right]\\
=&\tfrac{a_1}{A}\left[A+\tfrac{b_1}{a_1}(D-\Delta_1 a_1)-B+b_1\Delta_1-\tfrac{D}{Aa_1}(b_1(A-a_1^2)-a_1(C-a_1b_1))\right]\\
=&\tfrac{a_1}{A}\left[A+\tfrac{b_1}{a_1}D-b_1\Delta_1-B+b_1\Delta_1-\tfrac{D}{Aa_1}(b_1A-a_1C)\right]\\
=&\tfrac{a_1}{A}\left[A+\tfrac{b_1}{a_1}D-B-D\tfrac{b_1}{a_1}+\tfrac{DC}{A}\right]=\tfrac{a_1}{A}\left[A-B+\tfrac{DC}{A}\right]
\end{aligned}$$

and

$$\begin{aligned}
&\|\mathbf{q}\|^2+\|\mathbf{q}\|^2\|\mathbf{g}\|^2-\left(\mathbf{g}'\mathbf{q}\right)^2\\
=&\sum_{i=2}^n\left(a_i\frac{b_1}{a_1^2}-b_i\frac{1}{a_1}\right)^2+\left[\sum_{i=2}^n\left(a_i\frac{b_1}{a_1^2}-b_i\frac{1}{a_1}\right)^2\right]\sum_{i=2}^n\frac{a_i^2}{a_1^2}-\left(\sum_{i=2}^n\frac{a_i}{a_1}\left(a_i\frac{b_1}{a_1^2}-b_i\frac{1}{a_1}\right)\right)^2\\
=&\tfrac{1}{a_1^4}\sum_{i=2}^n(a_ib_1-a_1b_i)^2+\frac{1}{a_1^6}\sum_{i=2}^n(a_ib_1-a_1b_i)^2\sum_{i=2}^n a_1^2-\frac{1}{a_1^6}\left(\sum_{i=2}^n a_i(a_ib_1-b_ia_1)\right)^2\\
=&\tfrac{b_1^2}{a_1^4}\sum_{i=2}^n a_i^2+\frac{1}{a_1^2}\sum_{i=2}^n b_i^2-\frac{2b_1}{a_1^3}\sum_{i=2}^n a_ib_i+\frac{1}{a_1^6}\left(b_1^2\sum_{i=2}^n a_i^2-2a_1b_1\sum_{i=2}^n a_ib_i+a_1^2\sum_{i=2}^n b_i^2\right)\left(A-a_1^2\right)\\
&-\tfrac{1}{a_1^6}\left(b_1\sum_{i=2}^n a_i^2-a_1\sum_{i=2}^n a_ib_i\right)^2=\tfrac{b_1^2}{a_1^4}(A-a_1^2)\tfrac{1}{a_1^2}(E-b_1^2)-\tfrac{2b_1}{a_1^3}(C-a_1b_1)\\
&+\tfrac{1}{a_1^6}\left[(b_1^2(A-a_1^2)-2a_1b_1(C-a_1b_1)+a_1^2(E-b_1^2)\right](A-a_1^2)-\tfrac{1}{a_1^6}(b_1(A-a_1^2)-a_1(C-a_1b_1))^2\\
=&\tfrac{1}{a_1^4}(C^2-AE).
\end{aligned}$$

Plugging these expressions back into (36), we obtain the exact density of the NLS estimator (5) with the adjustment coefficients, which follow from (37) and (38). Note that the NE density approximation with adjustment coefficient (8) follows from (39). Thus, this approximation is obtained from density (33) by ignoring the absolute value of the derivative/Jacobian.

B. Proof of Theorem 2

This proof follows the proof of Theorem 1 closely. Express y_1 through the NLS estimator b and $y_2, \ldots, y_n$ as

$$y_1=h(b,y_2,\ldots,y_n)=f_1(b)-\frac{\sum_{i=2}^n(y_i-f_i(b))r_i(b)}{r_1(b)}$$

and find the derivative $dh/db=R+\sum_{i=2}^n z_iq_i$, where

$$R=\frac{1}{r_1(b)}\sum_{i=1}^{n}\dot{f}\,(b)\,r_i\,(b)+\sum_{i=2}^{n}\left(f_i\,(\beta)-f_i\,(b)\right)q_i\,(b)$$
$$q_i=\frac{\dot{r}_1(b)r_i(b)}{r_1^2(b)}-\frac{\dot{r}_i(b)}{\dot{r}_1(b)},\quad i=2,3,\ldots,n.$$

Introducing the $(n-1)\times 1$ vectors **z**, **q** and **g** with the ith components z_i, q_i and $g_i = r_i(b)/r_1(b)$, respectively, and $p = [(\mathbf{f}(b)-\mathbf{f}(\boldsymbol{\beta}))'\mathbf{r}(b)]/r_1(b)$. We notice that the derivative $\dot{f}_i$ appears only in the expression of R; in the rest of the expressions, r_i acts as $\dot{f}_i$ and subsequently $\dot{r}_i$ acts as $\ddot{f}_i$. Thus, we need to work on the terms containing R,

$$\begin{aligned}
&\frac{1}{1+\|\mathbf{g}\|^2}\left(R+\frac{p\mathbf{g}'\mathbf{q}}{1+\|\mathbf{g}\|^2}\right)\\
&=\frac{1}{A}a_1^2\left[\frac{G}{a_1A}+\sum_{i=2}^{n}\Delta_i\left(a_i\frac{b_1}{a_1^2}-b_i\frac{1}{a_1}\right)-\frac{a_1^2}{A}\frac{1}{a_1^2}D\sum_{i=2}^{n}a_i\left(a_i\frac{b_1}{a_1^2}-b_i\frac{1}{a_1}\right)\right]\\
&=\frac{a_1}{A}\left[\frac{G}{A}+\frac{b_1}{a_1}\sum_{i=2}^{n}\Delta_i a_i-\sum_{i=2}^{n}\Delta_i b_i-\frac{D}{Aa_1}\left(b_1\sum_{i=2}^{n}a_i^2-a_1\sum_{i=2}^{n}a_ib_i\right)\right]\\
&=\frac{a_1}{A}\left[\frac{G}{A}+\frac{b_1}{a_1}(D-\Delta_1a_1)-B+b_1\Delta_1-\frac{D}{Aa_1}\left(b_1\left(A-a_1^2\right)-a_1\left(C-a_1b_1\right)\right)\right]\\
&=\frac{a_1}{A}\left[\frac{G}{A}+\frac{b_1}{a_1}D-b_1\Delta_1-B+b_1\Delta_1-\frac{D}{Aa_1}\left(b_1A-a_1C\right)\right]\\
&=\frac{a_1}{A}\left[\frac{G}{A}+\frac{b_1}{a_1}D-B-D\frac{b_1}{a_1}+\frac{DC}{A}\right]=\frac{a_1}{A}\left[\frac{G}{A}-B+\frac{DC}{A}\right].
\end{aligned}$$

Thus, we come to the formula for density in Theorem 2 by replacing $\dot{\mathbf{f}}(b)$ with $\mathbf{r}(b)$ and $\ddot{\mathbf{f}}(b)$ with $\dot{\mathbf{r}}(b)$.

C. Outline of the multivariate density derivation

Partition the Jacobian matrix $\mathbf{F}(\mathbf{b})$, the data vector $\mathbf{y}$ and the regression function $\mathbf{f}(\mathbf{b})$ as follows:

$$\mathbf{F}(\mathbf{b})=\begin{bmatrix}\mathbf{F}_1^{m\times m}(\mathbf{b})\\ \mathbf{F}_2^{(n-m)\times m}(\mathbf{b})\end{bmatrix},\quad \mathbf{y}=\begin{bmatrix}\mathbf{y}_1^{m\times 1}\\ \mathbf{y}_2^{(n-m)\times 1}\end{bmatrix},\quad \mathbf{f}(\mathbf{b})=\begin{bmatrix}\mathbf{f}_1^{m\times 1}(\mathbf{b})\\ \mathbf{f}_2^{(n-m)\times 1}(\mathbf{b})\end{bmatrix},$$

assuming that the square matrix $\mathbf{F}_1(\mathbf{b})$ is nonsingular for every **b**. Using the normal equation express $\mathbf{y}_1$ through the NLS estimator **b** and $\mathbf{z}_2 = \mathbf{y}_2 - \mathbf{f}_2(\boldsymbol{\beta})$,

$$\mathbf{y}_1=\mathbf{h}(\mathbf{b},\mathbf{z}_2)\overset{\text{def}}{=}\mathbf{f}_1(\mathbf{b})-\left(\mathbf{F}_1'(\mathbf{b})\right)^{-1}\mathbf{F}_2'(\mathbf{b})(\mathbf{z}_2+(\mathbf{f}_2(\boldsymbol{\beta})-\mathbf{f}_2(\mathbf{b})),$$

and apply the multivariate version of Lemma 6 to obtain the density of the NLS estimator as an integral over R^{n-m},

$$p_{\text{NLS}}(\mathbf{b};\boldsymbol{\beta})=(2\pi)^{-n/2}\int_{R^{n-m}}\left|\det\left(\mathbf{R}+\sum_{i=m+1}^{n}z_{2i}\mathbf{Q}_i\right)\right|e^{-\frac{1}{2}\|\mathbf{h}(\mathbf{b},\mathbf{z}_2)-\mathbf{f}_1(\beta)\|^2-\frac{1}{2}\|\mathbf{z}_2\|^2}d\mathbf{z}_2,$$

$$\mathbf{R}=\mathbf{F}_1(\mathbf{b})+\left(\mathbf{F}_1'(\mathbf{b})\right)^{-1}\mathbf{F}'(\mathbf{b})\mathbf{F}(\mathbf{b})+\sum_{i=2}^{n}(f_i(\beta)-f_i(b))\mathbf{Q}_i,$$

where $\mathbf{Q}_i=-\left(\mathbf{F}_1'(\mathbf{b})\right)^{-1}\mathbf{H}_{2i}$ and $\mathbf{H}_{2i}=\partial^2 f_{i2}/\partial^2\mathbf{b}$.

Approximate

$$\det\left(\mathbf{R}+\sum_{i=m+1}^{n} z_{2i}\mathbf{Q}\right)\simeq|\mathbf{R}|\sum_{i=m+1}^{n}\mathrm{tr}(\mathbf{R}^{-1}\mathbf{Q}_i)z_{2i}.$$

using the Jacobi's formula and proceed as in the univariate case since the right-hand side may be treated as a normally distributed random variable and as such exact density can be derived beyond this point. Note that if the absolute value of the determinant in p_{NLS} expression is approximated as det(**R**) we obtain the NE density.

References

Almudevar A, Field C, Robinson J. The density of multivariate M-estimators. Annals of Statistics. 2000; 28:275–297.

Barndorf-Nielsen OE, Cox DR. Edgeworth and saddlepoint approximations with statistical applications. Journal of the Royal Statistical Society, ser B. 1979; 41:279–312.

Bates DM, Watts DG. Relative curvature measures of nonlinearity (with discussion). Journal of the Royal Statistical Society, ser B. 1980; 42:1–25.

Bates, DM., Watts, DG. Nonlinear regression analysis and its applications. Wiley; New York: 1988.

Brazzale, AR., Davison, AC., Reid, N. Applied asymptotics: Case studies in small-sample statistics. Cambridge University Press; Cambridge, UK: 2007.

Brazzale AR, Davison AC. Accurate parametric inference for small samples. Statistical Science. 2008; 23:465–484.

Beale EML. Confidence regions in nonlinear estimation (with discussion). Journal of the Royal Statistical Society, ser B. 1961; 22:41–88.

Bickel, PJ., Doksum, KA. Mathematical statistics Basic ideas and selected topics. 2nd. Vol. 1. Upper Saddle River; NJ: 2007.

Carroll, RJ., Ruppert, D. Transformation and weighting in regression. Chapman and Hall; New York: 1988.

Casella, G., Berger, RL. Statistical inference. Duxbury Press; Belmont, CA: 1990.

Demidenko, E. Optimization and regression. Nauka (in Russian); Moscow: 1989.

Demidenko E. Is this the least squares estimate? Biometrika. 2000; 87:437–452.

Demidenko, E. Mixed models: Theory and applications with R. 2nd. Wiley, Hoboken; New Jersey: 2013.

Demidenko E. Criteria for global minimum of sum of squares in nonlinear regression. Computational Statistics and Data Analysis. 2006; 53:1739–1753.

Demidenko E. Criteria for unconstrained global optimization. Journal of Optimization Theory and Applications. 2008; 136:375–395.

Field CA. Small sample asymptotic expansions for multivariate *M*-estimates. Annals of Statistics. 1982; 10:672–689.

Fieller EC. The distribution of the index in a normal bivariate population. Biometrika. 1932; 24:428–440.

Fitzmaurice, G., Molenberghs, M. Advances in longitudinal data analysis: An historical perspective. In: Fitzmaurice, G.Davidian, M.Verbeke, G., Molenberghs, G., editors. Longitudinal Data Analysis. CRC Press; Boca Raton, FL: 2009.

Fraser DAS, Reid N, Wu J. A simple general formula for tail probabilities for frequentist and Bayesian inference. Biometrika. 1999; 86:249–264.

Fuller, WA. Measurement error models. Wiley; New York: 1987.

Gallant, AR. Nonlinear statistical models. Wiley; New York: 1987.

Godambe VP. An optimum property of regular maximum likelihood estimation. Annals of Mathematical Statistics. 1960; 31:1208–1211.

Golub GH, Pereyra V. Separable nonlinear least squares: the variable projection method and its applications. Inverse Problems. 2003; 19:1–126.

Gong G, Samaniego FJ. Pseudo maximum likelihood estimation: Theory and applications. Annals of Statistics. 1981; 9:861–869.

Goutis C, Casella G. Explaining the saddlepoint approximation. The American Statistician. 1999; 53:216–224.

Hadeler KP, Jukic D, Sabo K. Least-squares problems for Michaelis–Menten kinetics. Mathematical Methods in Applied Sciences. 2007; 30:1231–1241.

Hinkley DV. On the ratio of two correlated normal random variables. Biometrika. 1969; 56:635–639.

Hougaard P. Saddlepoint approximations for curved exponential families. Statistical and Probability Letters. 1985; 3:161–166.

Huber, PJ. Robust statistics. Wiley; New York: 1981.

Jensen JL, Wood TA. Large deviation and other results for minimum contrast estimators. Ann Inst Statist Math. 1998; 50:673–685.

Jukíc D. A simple proof of the existence of the best estimator in a quasilinear regression model. Journal of Optimization Theory and Applications. 2014; 162:293–302.

Jukíc D, Markovic D. Nonlinear weighted least squares estimation of a three-parameter Weibull density with a nonparametric start. Applied Mathematics and Computations. 2010; 215:3599–3609.

Lehmann, EL., Casella, G. Theory of point estimation. 2nd. Springer; New York: 1998.

Levenberg K. A method for the solution of certain problems in least squares. Quarterly of Applied Mathematics. 1944; 2:164–168.

Lugannani R, Rice S. Saddle point approximation for the distribution of the sum of independent random variables. Advances in Applied Probability. 1980; 12:475–490.

Mardia, KV., Jupp, PE. Directional statistics. Wiley; New York: 2000.

Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. SIAM Journal for Applied Mathematics. 1963; 11:431–441.

Paige RL, Trindade AA. Saddlepoint-based bootstrap inference for quadratic estimating equations. Scandinavian Journal of Statistics. 2009; 36:98–111.

Pawitan, Y. In all likelihood: Statistical modelling and inference using likelihood. Oxford: Clarendon Press; 2001.

Pazman A. Probability distribution of the multivariate nonlinear least squares estimates. Kibernetika. 1984; 20:209–230.

Pazman, A. Nonlinear statistical models. Dordrecht: Kluwer; 1993.

Pazman A. Some properties and improvements of the saddlepoint approximation in nonlinear regression. Annals of the Institute of Statistical Mathematics. 1999; 51:463–478.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2014. URL http://www.R-project.org

Seber, GSF., Wild, CJ. Nonlinear regression. New York: Wiley; 1989.

Skovgaard IM. Large deviation approximations for maximum likelihood estimators. Probability and Mathematical Statistics. 1985; 6:89–107.

Skovgaard IM. On the density of minimum contrast estimators. The Annals of Statistics. 1990; 18:779–789.

Schervish, MJ. Theory of statistics. New York: Springer-Verlag; 1995.

Strawderman RL, Casella G, Wells MT. Practical small-sample asymptotics for regression problems. Journal of American Statistical Association. 1996; 91:643–654.

Vonesh EF, Wag H, Majumdar D. Generalized least squares, Taylor series linearization, and Fisher's scoring in multivariate nonlinear regression. Journal of American Statistical Association. 2001; 96:282–291.

Yule GU. The applications of the method of correlation to social and economic statistics. Journal of the Royal Statistical Society. 1909; 72:721–730.

Zeger SL, Liang KY, Albert PS. Models for longitudinal data: A generalized estimating equation approach. Biometrics. 1988; 44:1049–1060. [PubMed: 3233245]

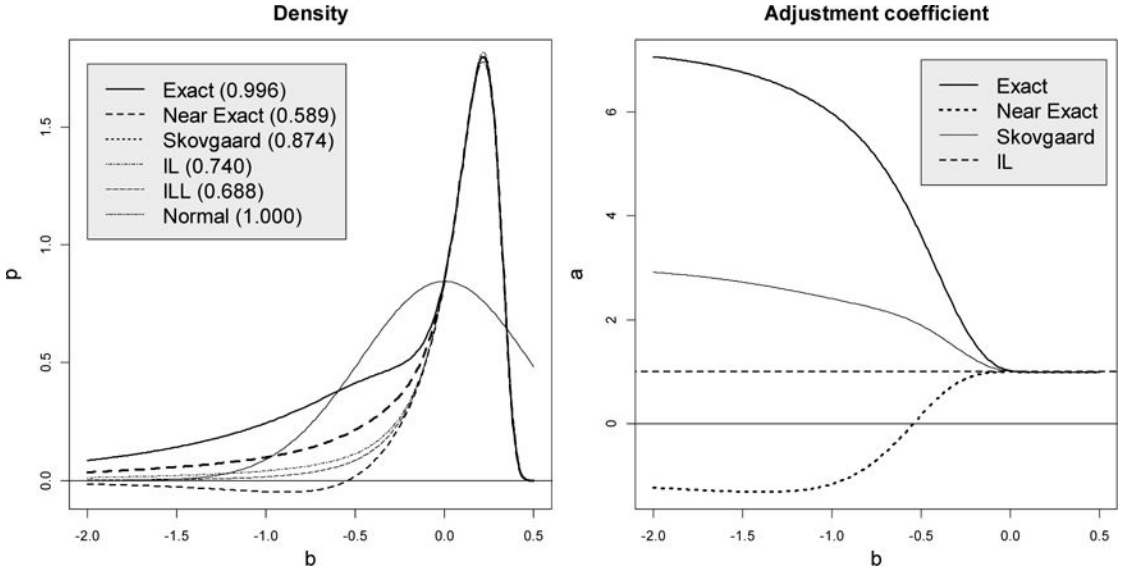

Figure 1.

Exact and four density approximations for the exponential regression $f_i(\beta) = e^{i\beta}$ with $i = 1, 2, \ldots, 6$, $\sigma = 2.5$ and the true $\beta = 0$. The figure in the parentheses shows the area under the curve (must be 1).

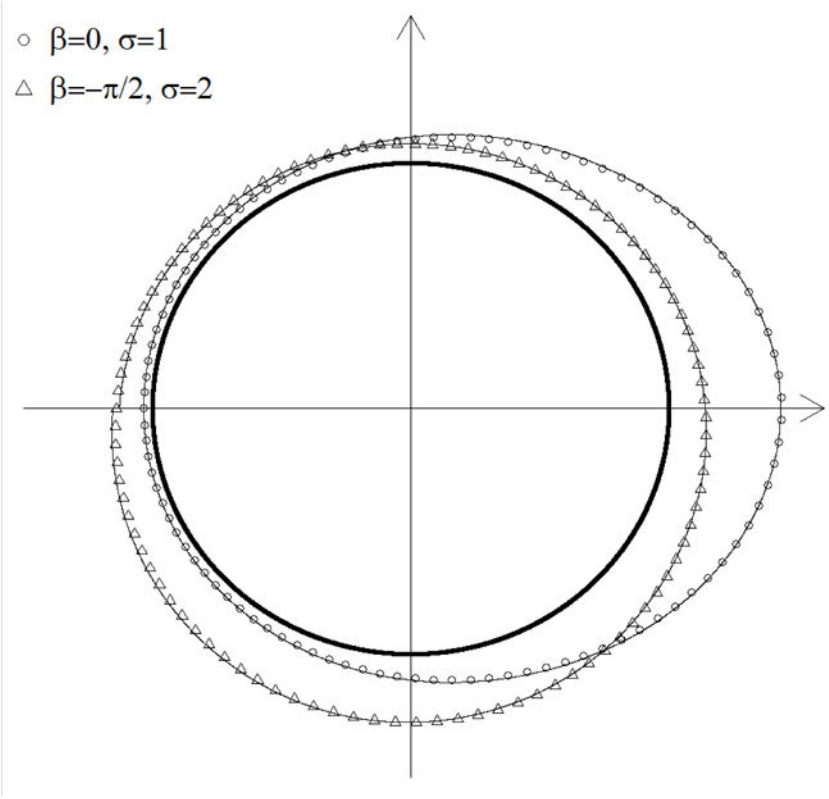

Figure 2.
The theoretical and empirical densities of the NLS estimate in circle regression wrapped around the unit circle under two scenarious (symbols represent simulations and curves represent the analytic density).

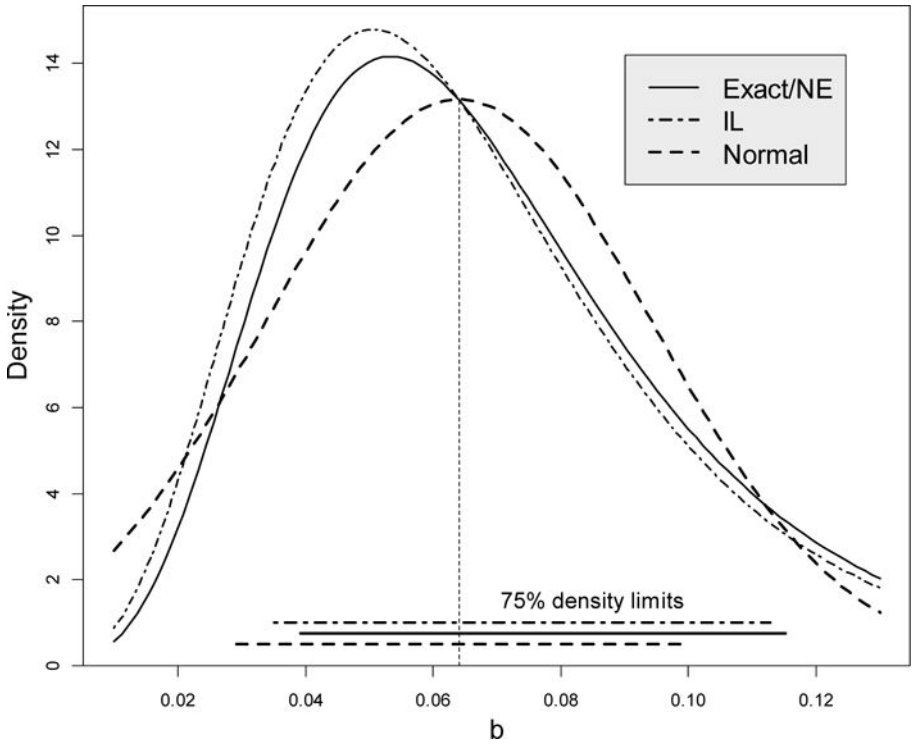

Figure 3.

Four densities of the NLS estimator of $\beta = 0.064$ in the *Puromycin* example ($\sigma = 40$) with the 75% density limits. The exact and near exact densities practically coincide.

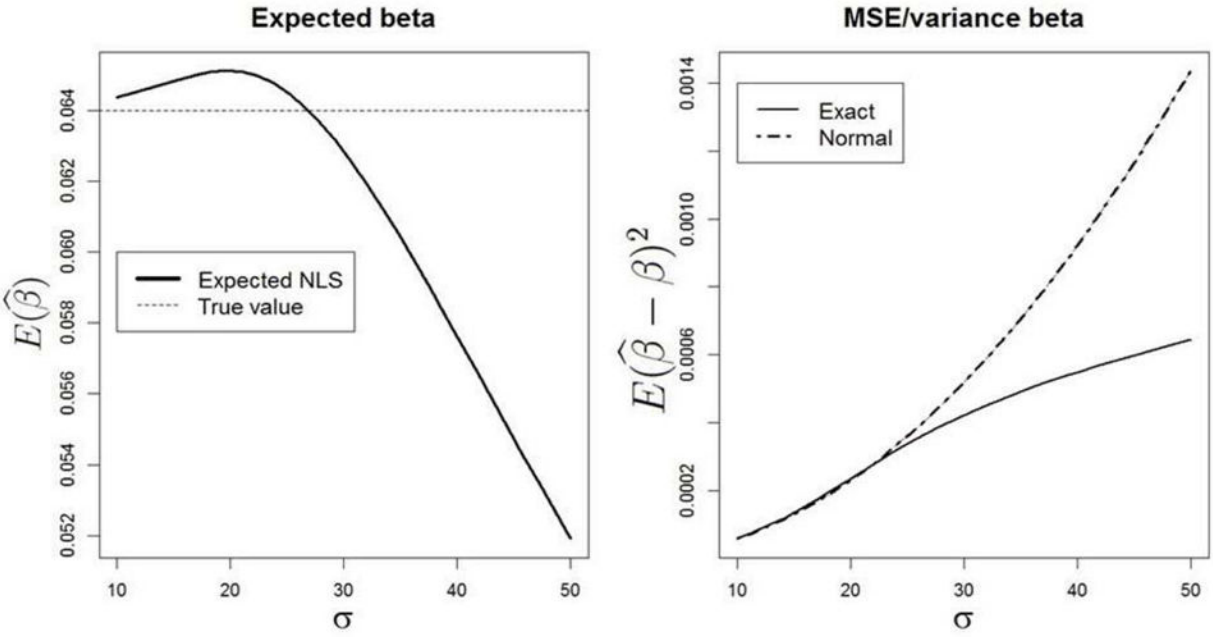

Figure 4.
Computation of the expected value and the MSE of the NLS estimator using the numerical integration of the exact density.

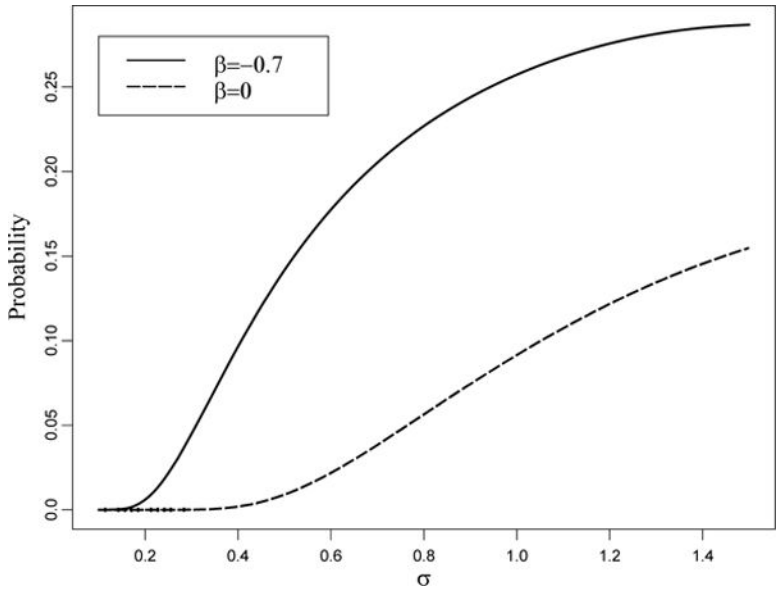

Figure 5.

Probabilities that the NLS estimate does not exist in the exponential regression with two observations ($n = 2$) computed as the integral (24). The smaller the value of β the greater the probability.

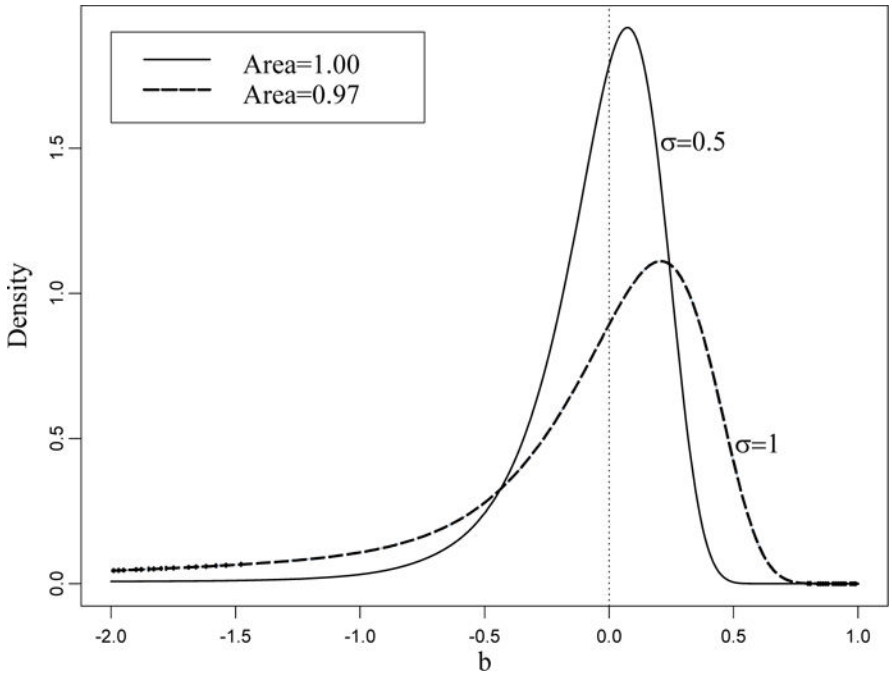

Figure 6.

Two densities for an exponential regression with $\beta = 0$ and different σ. For $\sigma = 0.5$, there is a tiny probability that the NLS estimate does not exist. For $\sigma = 1$ this probability is about 0.1, as follows from the previous figure.

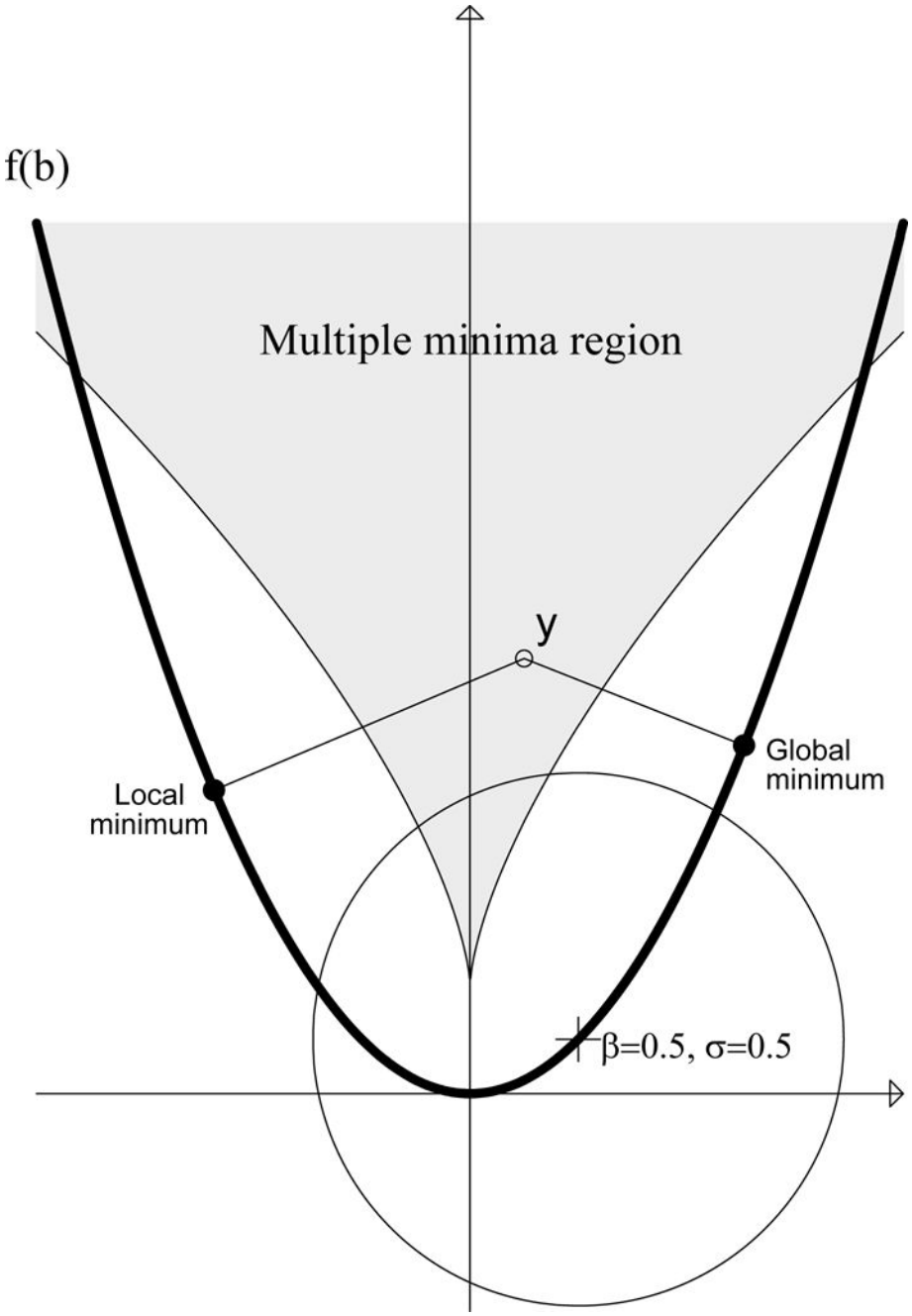

Figure 7.
Parabolic regression on the plane ($n = 2$). If $\mathbf{y} = (y_1, y_2)$ is from the shaded region the sum of squares, $(y_1 - b)^2 + (y_2 - b^2)^2$, has two local minima.

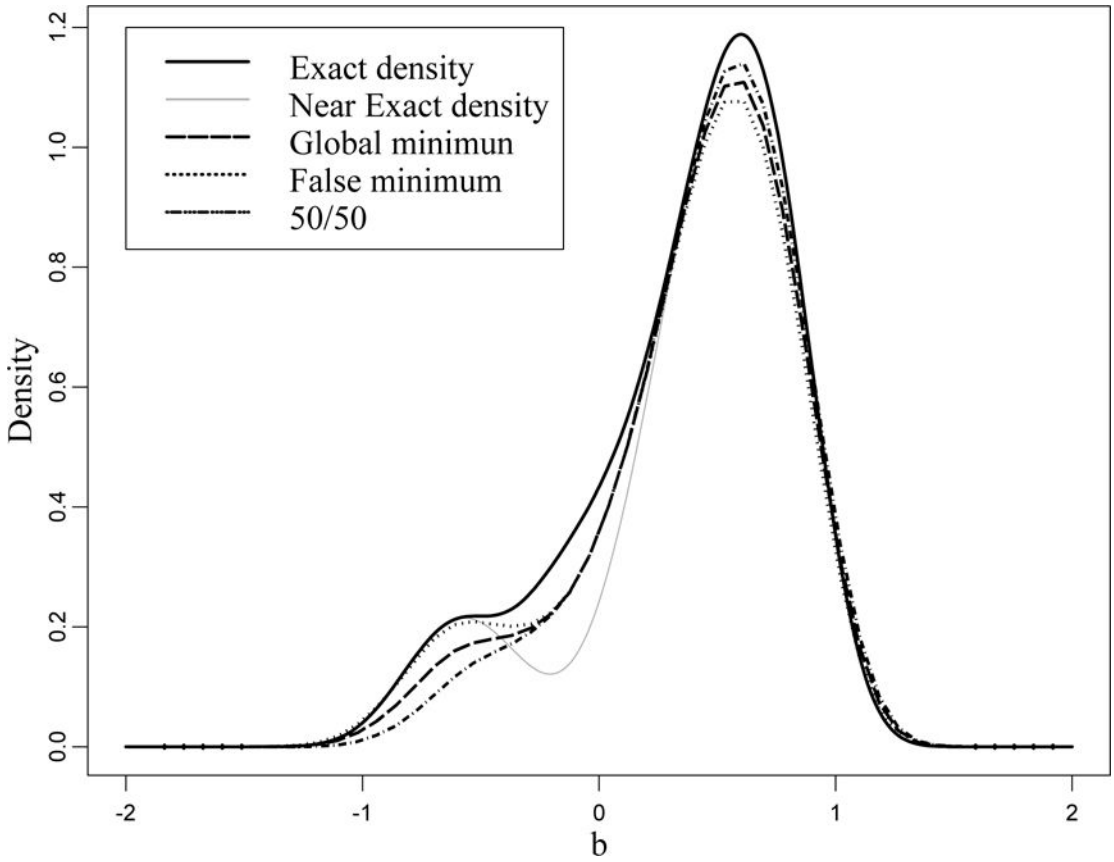

Figure 8.

Two theoretical and three empirical densities from simulations ($N = 100{,}000$) estimated with the Gaussian kernel for parabolic regression, $\beta = 0.5$, $\sigma = 0.5$. The left bump reflects the existence of the false NLS estimate. The probability that the sum of squares has two local minima, computed by formula (25), is 0.026.