Parametrizations of Non-Linear Models

Philip Hougaard

Journal of the Royal Statistical Society. Series B (Methodological), Vol. 44, No. 2 (1982), 244-252.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281982%2944%3A2%3C244%3APONM%3E2.0.CO%3B2-X

Journal of the Royal Statistical Society. Series B (Methodological) is currently published by Royal Statistical Society.

J. R. Statist. Soc. B (1982),
44, *No.2, pp.* 244–252

Parametrizations of Non-linear Models

By PHILIP HOUGAARD

Statistical Research Unit, Danish Medical and Social Science Research Councils, Copenhagen, Denmark

[Received January 1981. Revised August 1981]

SUMMARY

In the literature there have been many suggestions on how to parametrize models. Some properties you can seek are (1) stability of variance of the MLE; (2) normal likelihood; (3) zero asymptotic skewness of the MLE; (4) asymptotic unbiasedness of the MLE.

The parametrizations corresponding to these demands are found in the one-dimensional curved exponential family. They all belong to a general class of transformations, but they are in general not identical. The transformations in this class are characterized by a differential equation. The transformations are identical in the non-linear normal regression model.

Keywords: PARAMETRIZATIONS; CURVED EXPONENTIAL FAMILIES; VARIANCE STABILITY; ASYMPTOTIC SKEWNESS; ASYMPTOTIC BIAS; NORMAL LIKELIHOOD

1. INTRODUCTION

IN statistical models many properties of estimators are not invariant under parameter transformations. Therefore there have been a lot of suggestions on how to transform the parameters or the observations in order to get some nice properties of the parameter or its estimator.

In Section 1 we give a review of the literature. In Section 2 we present the results and in Section 3 the proofs. In Section 4 we consider some examples and in Section 5 we discuss the advantages of the different parametrizations. Special cases are known from the literature, some of them dating back to the 1930s.

Concerning the stability of variance it can easily be shown that if X_n is asymptotically normally distributed $N\{\beta, \sigma^2(\beta)/n\}$, then $f(X_n)$ is asymptotically $N[f(\beta), \sigma^2(\beta)\{f'(\beta)\}^2/n]$ provided f is differentiable and $f'(\beta) \neq 0$. The asymptotic variance is independent of β if and only if $\sigma^2(\beta)\{f'(\beta)\}^2$ is constant. This is equivalent to $f'(\beta) = c/\sigma(\beta)$. For the Poisson distribution Bartlett (1936) found that $\sqrt{X}$ is approximately normal with variance $\frac{1}{4}$.

It is natural to ask: Is there any function f such that the distribution of $f(X_n)$ is especially close to the asymptotic normal distribution? The problem here is to find out, what we mean by "close". The most common approach is to look at the asymptotic skewness or the third central moment. This moment of $f(X_n)$ is usually of order n^{-2} but by a suitable choice of f this term vanishes making the third central moment of order n^{-3}. This was first established for the χ^2-distribution (f known, unknown scale) by Wilson and Hilferty (1931). In this case you should take the third root of the observation. For the Poisson distribution Anscombe (1960) has mentioned the transformation $X^{2/3}$.

These results have later been refined by adding corrections depending on n, which make some terms in the expansions of the variance and third moment vanish. This has been done by Anscombe (1948) and Borges (1970). Borges (1971) considered a more general situation, namely the one-dimensional exponential family.

Anscombe (1964) suggested that we should ask for normality of the likelihood function; by this he meant that the likelihood function should look like the likelihood function of a normal distribution with fixed variance. This means that the logarithm of the likelihood function $l(\beta) = \ln L(\beta)$ should have a third derivative of 0 at the MLE (maximum likelihood estimate) $\hat{\beta}$ making $l(\beta)$ nearly a parabola around $\hat{\beta}$. He found the general solution in the one-dimensional

 0035–9246/82/44244$2.00

exponential family. Later on Sprott (1973) suggested that if this went wrong one should consider the expected likelihood function $E_{\hat{\beta}_0} l(\beta, X)$, which is a function of β for fixed β_0 the true value. The first procedure cannot be used, when $\hat{\beta}$ is not sufficient. Sprott (1973) argued that the distribution of $V_\beta = (\hat{\beta}-\beta)\sqrt{\{ni(\hat{\beta})\}}$ is especially close to its asymptotic standard normal distribution if β has a normal likelihood. Here $i(\beta)$ is the information about β in one observation. There is however a bias term, so the distribution of V_β is close to a normal distribution, not the standard one. Note that V_β is a function of β as well as of $\hat{\beta}$ and therefore it does not correspond to the zero asymptotic skewness transformation, see Example 3 in Section 4. Sprott (1973) did not discuss what was meant by "close to a normal distribution", but referred to Welch and Peers (1963), who found the corrections used. It can be shown that if one accounts for the bias in V_β, the main order term in an expansion of the third moment vanishes, at least in a curved exponential family.

It is fairly common to do unbiased estimation. In many simple models the parameter equals the mean value of the observation and there is no problem, but in general this is not possible in non-linear models. It has been proposed to reduce the asymptotic bias by choosing a function f which removes the first correction term in the bias of the MLE for β. This has for example been done by Box (1971), who relates the asymptotic bias in a non-linear regression model to the difference of two non-linearity measures proposed by Beale (1960).

Wedderburn examined these parametrizations in the one-dimensional exponential family in an unpublished paper. Some of them can be found in Barndorff-Nielsen (1978). Let $X_1, X_2, \ldots$ be i.i.d. with density $e^{\theta t(x)}/\phi(\theta)$ w.r.t. some measure $\mu(dx)$. θ is called the canonical parameter because of the simple way it enters in the exponent. Then Wedderburn found that the parametrizations are given by the formula

$$\psi(\theta_1) = \int_{\theta_0}^{\theta_1} \left\{\frac{d^2}{d\theta^2} \ln \phi(\theta)\right\}^{\delta} d\theta,$$

where δ is a constant. He then showed that different values of δ correspond to different properties of the parametrization as follows:

$\delta = 0$	θ, the canonical parameter,
$\delta = \frac{1}{3}$	normal likelihood,
$\delta = \frac{1}{2}$	stability of variance,
$\delta = \frac{2}{3}$	zero asymptotic skewness,
$\delta = 1$	the mean value parameter $\tau(\theta) = E_\theta\, t(X_1)$.

These results are generalized to a curved exponential family in Section 2.

The curved exponential family is also considered by Efron (1975). He looks at second-order efficiency and especially at those properties which are independent of the parametrization. In an asymptotic expansion of $(\hat{\theta}_n-\theta)\sqrt{\{ni(\theta)\}}$, the second-order efficiency is of order n^{-1}, whereas the asymptotic bias and skewness are of order $n^{-\frac{1}{2}}$ and they are therefore asymptotically of greater importance, when evaluating the approximate distribution of the estimator.

For the case where the parameter space has a dimension p higher than 1 the problem is more difficult, see Hougaard (1980). For the non-linear normal regression, see also Hougaard (1981). See also Example 5 in Section 4.

2. Results

Consider an n-dimensional exponential family with the canonical parameter θ, i.e. a family of distributions with densities $p_\theta(x) = \exp\{\theta' t(x)\}/\phi(\theta)$ w.r.t. some measure $\mu(dx)$. Let P_θ be the measure. $\theta' t(x)$ is the scalar product of the parameter vector θ and the vector of sufficient statistics $t(x)$. Suppose the family is full and regular. That is the parameter set

$$D = \{\theta \mid \int \exp\{\theta' t(x)\}\, \mu(dx) < \infty\}$$

is open. Suppose $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$. We will consider a one-dimensional submodel (a curved family) of this, i.e. a statistical model given by the family of densities $h_\beta(x) = \exp\{\theta(\beta)'t(x)\}/\phi\{\theta(\beta)\}$, $\beta \in B$. Here $B \subseteq \mathbb{R}$, B open, f: $B \to D$ is injective. If θ is sufficiently differentiable, we have a smooth family of order 1.

Let $\chi(\theta) = \ln \phi(\theta)$. Then $d\chi/d\theta$ is an n-vector, $d^2\chi/d\theta^2$ is an $n \times n$-matrix and $d^3\chi/d\theta^3$ is a trilinear form so that if v is an n-vector then $d^3\chi/d\theta^3(v) = \Sigma v_i v_j v_k \partial^3 \chi/\partial\theta_i \partial\theta_j \partial\theta_k$. The problem is to find a transformation g of the parameter β, such that the new parameter $\psi = g(\beta)$ has some specified properties.

Some of the arguments are asymptotic building on m i.i.d. repetitions of the n-dimensional observation. It is well known that a sufficient statistic is $\bar{T}_m(X) = \{t(X_1) + \ldots + t(X_m)\}/m$. The limit used is as $m \to \infty$.

Theorem. Consider the submodel of the exponential family mentioned above: Suppose $\theta(\beta)$, θ: $B \to D$, is twice continuously differentiable and $d\theta/d\beta \neq 0$. Let g: $B \to \mathbb{R}$ be twice differentiable and $dg/d\beta \neq 0$. Consider the differential equation

$$\frac{d^2 g/d\beta^2}{dg/d\beta} = \left\{\delta \frac{d^3\chi}{d\theta^3}\left(\frac{d\theta}{d\beta}\right) + \left(\frac{d^2\theta}{d\beta^2}\right)' \frac{d^2\chi}{d\theta^2}\frac{d\theta}{d\beta}\right\} \Big/ \left\{\left(\frac{d\theta}{d\beta}\right)' \frac{d^2\chi}{d\theta^2}\frac{d\theta}{d\beta}\right\} \tag{2.1}$$

Then:

(1) $\psi = g(\beta)$ has a normal likelihood, that is $E_{\psi_0}(d^3 l/d\psi^3)(\psi_0) = 0$ for all values of ψ_0, where l is the logarithm of the likelihood function, if and only if g is a solution to the differential equation with $\delta = \frac{1}{3}$.

(2) $\psi = g(\beta)$ is a variance stabilizing parameter, that is the information about ψ is independent of ψ if and only if g is a solution to the differential equation with $\delta = \frac{1}{2}$.

(3) $\psi = g(\beta)$ has an asymptotic third central moment for the maximum likelihood estimate $\hat{\psi}$ of order $\sigma(m^{-2})$ if and only if g is a solution to the differential equation with $\delta = \frac{2}{3}$.

(4) $\psi = g(\beta)$ has asymptotic bias for the maximum likelihood estimator $\hat{\psi}$ of order $\sigma(m^{-1})$ if and only if g is a solution to the differential equation with $\delta = 1$.

Remarks. The differential equation can always be solved. The solution of an equation of the form $(d^2 g/d\beta^2)/(dg/d\beta) = h(\beta)$ is

$$g(\beta) = k_1 + k_2 \int_{\beta_0}^{\beta} \exp\left\{\int_{u_0}^{u} h(x)\,dx\right\} du,$$

where β_0, u_0 are chosen in B and k_1, k_2 arbitrary constants ($k_2 \neq 0$). The choice of k_1, k_2 corresponds to making affine transformations of the parameter, which do not change any of the properties.

For $\delta = \frac{1}{2}$ you also have the simpler expression

$$g(\beta) = k_1 + k_2 \int_{\beta_0}^{\beta} \sqrt{i(x)}\,dx,$$

where

$$i(\beta) = \left(\frac{d\theta}{d\beta}\right)' \frac{d^2\chi}{d\theta^2}\frac{d\theta}{d\beta}.$$

Special cases

(i) *Non-linear regression*

The non-linear regression analysis with one parameter and normal errors is a special case of this model. To begin with we assume σ^2 known. The model is $Y \sim N_n(F(\beta), \sigma^2 I)$, F is a known vector function of β. Choose $D = \mathbb{R}^n$, $t(Y) = Y/\sigma^2$, $\theta(\beta) = F(\beta)$. In this case $\chi(\theta) = \frac{1}{2}\sigma^{-2}\Sigma\theta_i^2$ and hence $d^2\chi/d\theta^2 = \sigma^{-2} I$. $d^3\chi/\mathrm{d}\theta^3 = 0$.

Therefore all the differential equations coincide and the four parametrizations are the same. You can use the simplest expression ($\delta = \frac{1}{2}$) for the parameter ψ.

$$\psi = g(\beta_1) = k_1 + k_2 \int_{\beta_0}^{\beta_1} \left\{ \left(\frac{d\theta}{d\beta} \right)' \frac{d\theta}{d\beta} \right\} d\beta.$$

The parametrization does not depend on σ^2 and therefore the results are valid also if σ^2 is unknown varying independently of β. This parametrization will also minimize the curvature as proposed by Beale (1960), that is $N_\theta = N_{\min}$ for all $\beta \in B$, see Hougaard (1981).

(ii) *The one-dimensional exponential family*

If $\theta(B)$ is an affine one-dimensional interval then we have a one-dimensional exponential family and we find stronger results for $\delta = 0$ and $\delta = 1$. For $\delta = 0$ we find the canonical parameter and for $\delta = 1$ we find the mean value parameter, which can of course be estimated exactly unbiased and not just asymptotically unbiased. The results are a generalization of Wedderburn's results mentioned in Section 1. However he formulated the results in terms of the canonical parameter and we derive them for an arbitrary parameter β.

3. Proof of the Theorem

The logarithm of the likelihood function is

$$l(\beta) = -\ln\{\phi(\theta)\} + \theta(\beta)' t(x) = -\chi(\theta) + \theta(\beta)' t(x).$$

$$\frac{dl}{d\beta} = \left(\frac{d\theta}{d\beta} \right)' [t(x) - \tau\{\theta(\beta)\}],$$

where we define $\tau(\theta) = d\chi/d\theta = E_\theta t(x)$

$$\frac{d^2 l}{d\beta^2} = \left(\frac{d^2\theta}{d\beta^2} \right)' [t(x) - \tau\{\theta(\beta)\}] - \left(\frac{d\theta}{d\beta} \right)' \frac{d^2\chi}{d\theta^2} \frac{d\theta}{d\beta}.$$

The information is

$$J = -E \frac{d^2 l}{d\beta^2} = \left(\frac{d\theta}{d\beta} \right)' \frac{d^2\chi}{d\theta^2} \frac{d\theta}{d\beta}.$$

(1) *Normal likelihood*

We want a parameter ψ for which $E(d^3 l/d\psi^3) = 0$ and expressed in $d\psi/d\beta$ and $d^2\psi/d\beta^2$ we find

$$E \frac{d^3 l}{d\psi^3} = \left(\frac{d\psi}{d\beta} \right)^{-3} \left\{ 3 \frac{d^2\psi}{d\beta^2} \left(\frac{d\psi}{d\beta} \right)^{-1} \left(\frac{d\theta}{d\beta} \right)' \frac{d^2\chi}{d\theta^2} \frac{d\theta}{d\beta} - \frac{d^3\chi}{d\theta^3} \left(\frac{d\theta}{d\beta} \right) - 3 \left(\frac{d^2\theta}{d\beta^2} \right)' \frac{d^2\chi}{d\theta^2} \frac{d\theta}{d\beta} \right\}.$$

From $Ed^3 l/d\psi^3 = 0$ we find immediately the desired equation (2.1) with $\delta = \frac{1}{3}$ which is analogous to equation (1) in Anscombe (1964).

(2) *Stability of variance*

It is well known (see Section 1) that the information is constant if

$$\frac{d\psi}{d\beta} = c \cdot \sqrt{J} = c \sqrt{\left\{ \left(\frac{d\theta}{d\beta} \right)' \frac{d^2\chi}{d\theta^2} \frac{d\theta}{d\beta} \right\}}.$$

By differentiation we find

$$\frac{d^2\psi}{d\beta^2} = c \left\{ \frac{d^3\chi}{d\theta^3} \left(\frac{d\theta}{d\beta} \right) + 2 \left(\frac{d^2\theta}{d\beta^2} \right)' \frac{d^2\chi}{d\theta^2} \frac{d\theta}{d\beta} \right\} \Big/ \left[2 \sqrt{\left\{ \left(\frac{d\theta}{d\beta} \right)' \frac{d^2\chi}{d\theta^2} \frac{d\theta}{d\beta} \right\}} \right]$$

giving the desired equation (2.1) with $\delta = \frac{1}{2}$.

(3) *and* (4) *Zero asymptotic skewness and asymptotic unbiasedness*

We expand $\hat{\beta}$ as a function of $\varepsilon = \bar{T}_m(x) - \tau\{\theta(\beta_0)\}$ by implicit differentiation in β_0, the true value of the parameter, i.e. we approximate $\hat{\beta}$ by the estimator $\tilde{\beta} = \beta_0 + L'\varepsilon + M(\varepsilon)$, where L is a vector and M a quadratic form, which will also be written as a matrix $M(\varepsilon) = \varepsilon' M\varepsilon$. L and M are chosen such that $d\hat{\beta}/d\varepsilon = d\tilde{\beta}/d\varepsilon$ and $d^2\hat{\beta}/d\varepsilon^2 = d^2\tilde{\beta}/d\varepsilon^2$ for $\varepsilon = 0$. The expansions are correct under the limit $m \to \infty$, because $\varepsilon \xrightarrow{\text{a.s.}} 0$. We need the first three asymptotic moments of $\hat{\beta}$. Higher moments are of order m^{-3} or less. The moments we derive from the first terms in the Taylor expansion are approximations of the moments of $\tilde{\beta}$ and not necessarily of $\hat{\beta}$. This problem is discussed in Section 5. Define $\beta^* = \tilde{\beta} - \beta_0$.

$$V(\varepsilon) = m^{-1} d^2\chi/d\theta^2, \quad m^{-1}\Sigma, \quad \text{say}$$

The third moment of ε is the trilinear form $m^{-2} d^3\chi/d\theta^3$, $m^{-2}\Lambda$, say.

The fourth moment is, written in co-ordinates,

$$E\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l = m^{-2}\left\{\frac{\partial^2\chi}{\partial\theta_i\partial\theta_j}\cdot\frac{\partial^2\chi}{\partial\theta_k\partial\theta_l}+\frac{\partial^2\chi}{\partial\theta_i\partial\theta_k}\frac{\partial^2\chi}{\partial\theta_j\partial\theta_l}+\frac{\partial^2\chi}{\partial\theta_i\partial\theta_l}\cdot\frac{\partial^2\chi}{\partial\theta_j\partial\theta_k}\right\}+m^{-3}\frac{\partial^4\chi}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l}.$$

Define the following n-vectors, and put in $\beta = \beta_0$.

$$H = d\tau\{\theta(\beta)\}/d\beta = d^2\chi/d\theta^2 \cdot d\theta/d\beta,$$
$$K = d^2\tau\{\theta(\beta)\}/d\beta^2, \quad B = d\theta/d\beta,$$
$$C = d^2\theta/d\beta^2.$$

Up to order ε^2 the likelihood equation is

$$(B + C\beta^*)'(\varepsilon - H\beta^* - \tfrac{1}{2}K\beta^{*2}) = 0.$$

Inserting $\beta^* = L'\varepsilon + M(\varepsilon)$ we get

$$L'\varepsilon = (B'H)^{-1}B'\varepsilon = J^{-1}B'\varepsilon,$$
$$M(\varepsilon) = J^{-1}\{L'\varepsilon C'(I - HL')\varepsilon - \tfrac{1}{2}B'K(L'\varepsilon)^2\}.$$

Thus we can find the asymptotic moments

$$E\tilde{\beta} - \beta_0 = EM(\varepsilon) = m^{-1}\,tr(M\Sigma) = -\tfrac{1}{2}m^{-1}J^{-2}\left\{\frac{d^3\chi}{d\theta^3}\left(\frac{d\theta}{d\beta}\right)+\left(\frac{d^2\theta}{d\beta^2}\right)'\frac{d^2\chi}{d\theta^2}\frac{d\theta}{d\beta}\right\}$$

$$E(\tilde{\beta} - E\tilde{\beta})^2 = m^{-1}J^{-1} + \ldots$$

$$E(\tilde{\beta} - E\tilde{\beta})^3 = E[(L'\varepsilon)^3 + 3(L'\varepsilon)^2\{M(\varepsilon) - EM(\varepsilon)\}]$$
$$= m^{-2}\{\Lambda(L) + 6L'\Sigma M\Sigma L\}$$
$$= m^{-2}J^{-3}\left\{-2\frac{d^3\chi}{d\theta^3}\left(\frac{d\theta}{d\beta}\right)-3\left(\frac{d^2\theta}{d\beta^2}\right)'\frac{d^2\chi}{d\theta^2}\frac{d\theta}{d\beta}\right\}.$$

The asymptotic moments of $\hat{\psi} = g(\hat{\beta})$ are then

$$E\hat{\psi} = \psi + m^{-1}\left[-\tfrac{1}{2}g'(\beta)J^{-2}\left\{\frac{d^3\chi}{d\theta^3}\left(\frac{d\theta}{d\beta}\right)+\left(\frac{d^2\theta}{d\beta^2}\right)'\frac{d^2\chi}{d\theta^2}\frac{d\theta}{d\beta}\right\}+\tfrac{1}{2}g''(\beta)J^{-1}\right]+\ldots$$

$$E(\hat{\psi} - E\hat{\psi})^3 = m^{-2}\{g'(\beta)\}^2$$
$$\left[g'(\beta)J^{-3}\left\{-2\frac{d^3\chi}{d\theta^3}\left(\frac{d\theta}{d\beta}\right)-3\left(\frac{d^2\theta}{d\beta^2}\right)'\frac{d^2\chi}{d\theta^2}\frac{d\theta}{d\beta}\right\}+3g''(\beta)J^{-2}\right]+\ldots$$

The main term in the third central moment will disappear if g is a solution to (2.1) with $\delta = \frac{2}{3}$.

We find the asymptotically unbiased parametrization by looking at the m^{-1}-term in $E\hat{\psi}$. This vanishes if g fulfils (2.1) with $\delta = 1$.

By a similar expansion it can be shown that the m^{-1} order term in $E(V_\beta - EV_\beta)^3$ is zero if and only if β has a normal likelihood. Here $V_\beta = (\hat{\beta}-\beta)\cdot\sqrt{\{mJ(\hat{\beta})\}}$.

4. EXAMPLES

Example 1. The normal distribution $N(\theta, 1)$

This is an example of both special cases.

$$\chi(\theta) = \tfrac{1}{2}\theta^2, \quad \chi'(\theta) = Et(x) = \theta,$$
$$\chi''(\theta) = 1, \quad \chi'''(\theta) = 0.$$

All choices of δ give the same result. By definition θ has normal likelihood and it is well known that $\hat{\theta} = \bar{X}$ is unbiased, has constant variance and has exactly zero skewness.

Example 2. The exponential distribution

If the density is of the form $\beta e^{-\beta x}$, then $t(x) = -x$, $\chi(\beta) = -\ln(\beta)$, $\chi'(\beta) = -\beta^{-1}$, $\chi''(\beta) = \beta^{-2}$.

By Wedderburns equation (see Section 1) $g_\delta(\beta) = \beta^{1-2\delta}$ when $\delta \neq \frac{1}{2}$ choosing k_1, k_2 appropriately and $g_{\frac{1}{2}}(\beta) = \ln(\beta)$. By this

β	is the canonical parameter,
$\beta^{1/3}$	has normal likelihood,
$\ln(\beta)$	is variance stabilizing,
$\beta^{-1/3}$	has zero asymptotic skewness of MLE,
β^{-1}	$(= EX)$ has unbiased MLE.

For $\psi = \beta^\rho$, $V_\psi = (\hat{\psi}-\psi)\sqrt{\{J(\hat{\psi})m\}} = m^{\frac{1}{2}}\rho^{-1}(1-\psi/\hat{\psi})$, which is $\hat{\psi}^{-1}$ apart from an affine transformation. The results are also valid for the Γ-distribution because the properties are invariant under addition of independent repetitions.

Example 3. The Poisson distribution

For the density $P(X = x) = \exp(\beta x - e^\beta)/x!$ we find

$$\chi(\beta) = \exp(\beta), \quad \chi'(\beta) = \chi''(\beta) = \exp(\beta)$$
$$g_\delta(\beta) = \begin{cases} \exp(\delta\beta) = (EX)^\delta & \text{when } \delta \neq 0, \\ \beta & \text{when } \delta = 0. \end{cases}$$

For normal likelihood ($\psi = (EX)^{1/3}$) you have that

$$V_\psi = 3\sqrt{n}\{\bar{X}^{\frac{1}{2}} - \psi\bar{X}^{1/6}\}$$

is close to a normal distribution with variance 1, whereas $\bar{X}^{2/3}$ has zero asymptotic skewness and is therefore close to a normal distribution.

Example 4. Censoring in the exponential distribution

Suppose T is the observed survival time of an individual, and D is an indicator variable telling whether the individual dies ($D = 1$) or is censored ($D = 0$). The individual has a constant hazard of dying, say μ and a constant hazard of being censored λ. This might be the situation in some competing risk study. Suppose λ is known. T and D are independent. T is exponential with hazard $\mu+\lambda$, D is binomial $b(1, p)$, $p = \mu/(\mu+\lambda)$.

The differential equation is

$$\frac{d^2\psi/d\mu^2}{d\psi/d\mu} = \frac{\delta(\mu+\lambda)^{-3}\{\lambda\mu^{-2}(\lambda-\mu)-2\}-\lambda\mu^{-2}(\mu+\lambda)^{-2}}{\mu^{-1}(\mu+\lambda)^{-1}}.$$

The solutions are

$$\begin{aligned}
\delta = \tfrac{1}{3}: &\quad g(\mu) = \mu^{1/3},\\
\delta = \tfrac{1}{2}: &\quad g(\mu) = \ln\{\sqrt{\mu}+\sqrt{(\mu+\lambda)}\},\\
\delta = \tfrac{2}{3}: &\quad g(\mu) = \ln\{1-3\lambda^{1/3}\mu^{1/3}(\mu^{1/3}+\lambda^{1/3})^{-2}\}\\
&\qquad\qquad +6\operatorname{arc\,tg}\{(2\lambda^{-1/3}\mu^{1/3}-1)/\sqrt{3}\},\\
\delta = 1: &\quad g(\mu) = (\mu+\lambda)^{-1}.
\end{aligned}$$

Example 5. The Cox model for survival data

This is a very simplified version of the regression model for survival data, which was proposed by Cox (1972). Suppose (1) $\lambda_0(t)$ is known (without loss of generality $\lambda_0(t) = 1$); (2) there is no censoring; and (3) there is only one covariate z. In the asymptotic arguments the whole experiment is repeated with the same values of z.

Suppose X_i, $i = 1, \ldots, n$, are all independent and X_i is exponentially distributed with intensity $\exp(\beta z_i)$. $z_1, \ldots, z_n$ are known covariates, not all equal to 0. Then

$$\theta_i = -\exp(\beta z_i) \quad \text{and} \quad \chi(\theta) = -\sum \ln(-\theta_i).$$

By the theorem we receive the differential equation

$$\frac{d^2\psi/d\beta^2}{d\psi/d\beta} = (1-2\delta)\sum(z_i^3)/\sum(z_i^2).$$

Let

$$\rho = \sum(z_i^3)/\sum(z_i^2)$$

If $\rho = 0$ or $\delta = \frac{1}{2}$ you find $g(\beta) = \beta$, you should make no transformation.

If $\rho \neq 0$ and $\delta \neq \frac{1}{2}$ you find $g(\beta) = \exp\{(1-2\delta)\rho\beta\}$.

If $\sum z_i^3 = 0$, i.e. the third moment of the covariate around 0 vanishes, then the four parametrizations coincide and you should indeed use the parameter as Cox (1972) proposed. If $\sum z_i^3 \neq 0$ the four parametrizations are different. β is variance stabilizing, the others are powers of the intensity (for any fixed z).

$\exp(\frac{1}{3}\rho\beta)$	gives normal likelihood,
$\exp(-\frac{1}{3}\rho\beta)$	gives zero asymptotic skewness,
$\exp(-\rho\beta)$	gives asymptotic unbiasedness.

This generalizes Example 2.

Some of the results can be generalized to constant but unknown λ_0. In the marginal distribution of $\hat{\beta}$ you find asymptotic skewness 0 and asymptotic unbiased parameters by changing ρ to $\sum(z_i-\bar{z})^3/\sum(z_i-\bar{z})^2$.

β is still variance stabilizing. The marginal third derivative of the likelihood function has no relevance. However if you assume $\bar{z} = 0$, which is possible by changing λ_0, you can transform β such that the mean of the third derivative of the logarithm of the likelihood in this direction vanishes as in the case with λ_0 known.

5. Discussion

We have examined properties of different transformations of a parameter. A natural question arises: Which one shall we use? There is no unique answer to this, but we will discuss the problem.

For the non-linear regression analysis we found that there is a parametrization that has all the properties mentioned and has minimal curvature. It has a very clear geometric interpretation as the curve length in the solution locus. However, it might be very difficult or impossible to calculate the integrals needed to find the transformation. Another conjecture is that it might be a parameter which is difficult to interpret. If you want to use Wald's test statistic $(\hat{\beta}-\beta_0)/\sigma(\hat{\beta})$ or $(\hat{\beta}-\beta_0)/\sigma(\beta_0)$ it is advisable to use the proposed parameter. If the parameter is multidimensional such a parametrization may or may not exist, see Hougaard (1980).

For the curved exponential family the four parametrizations are not identical in general, and you cannot get more than one of the properties. Therefore it is important to discuss the relevance of the parametrizations.

(1) *Normal likelihood*

This assures that the likelihood looks nice. A Bayesian would prefer this parametrization. This implies that an iterative solution to the likelihood equation will converge faster. The distribution of $(\hat{\beta}-\beta)/\sigma(\hat{\beta})$ is close to a normal distribution with variance 1, but with a bias term.

At a first glance you would believe that if you have normal likelihood,

$$-(\beta-\hat{\beta})^2(\partial^2 l/\partial\beta^2)(\hat{\beta})$$

would be a good approximation to the likelihood ratio statistic $2\{l(\hat{\beta})-l(\beta)\}$, which does not depend on the parametrization. This is not correct, because the correction term, corresponding to the fourth derivative is of the same order as that to the third. Therefore you do not gain anything if $(d^3 l/d\beta^3)(\hat{\beta}) = 0$, there is still a correction of order n^{-1}.

(2) *Stability of variance*

This property assures that the information and thereby the asymptotic variance is known. The use of Wald's test is in order as shown by Væth (1981). In the other parametrizations this test can be misleading according to Væth. This transformation has been used on data in order to use techniques from analysis of variance, like fitting Latin squares on square roots of Poisson data. This can give very strange hypotheses in the original model and therefore it should be avoided, one should rather fit multiplicative models directly. This analysis of variance technique is not necessary nowadays. This transformation is usually easier to compute than the others.

(3) *Zero asymptotic skewness*

This implies that the distribution of the maximum likelihood estimator is nice. When distributions of functions of an average is not near the normal the most important reason is usually that the distribution is skew. In this parametrization that does not happen. Probability plots of the MLE will look well. Notice however that there may be some bias.

In testing simple hypothesis $\beta = \beta_0$ with test statistic $(\hat{\beta}-\beta_0)/\sigma(\beta_0)$ this bias may be calculated. This will give a rather good approximation to the P-value.

(4) *Asymptotic unbiasedness*

In this parametrization the MLE will be asymptotically unbiased. However, the asymptotic moments say a lot about the distribution in the central area, while the exact moments are highly dependent on the tails. The arguments for using unbiased estimators (i.e. consistency of the average of independent repetitions) is based on the exact mean value. In exponential families you have a natural scale – the sufficient statistic— in which you calculate averages and then use this average to find the MLE of the desired parametrization.

The properties are all nice to have in applications. No one beats the others. If you make an asymptotic expansion of the distribution of $\hat{\beta}$ there will be two first-order correction terms.

One of these vanishes in case (3) and the other in case (4). Remember, that if you choose one, you miss the others.

ACKNOWLEDGEMENT

I wish to thank Søren Johansen for many discussions and comments on the manuscript.

REFERENCES

ANSCOMBE, F. J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika*, **35**, 246–254.
—— (1960). Notes on sequential sampling plans. *J. R. Statist. Soc.* A, **123**, 297–306. (See p. 301.)
—— (1964). Normal likelihood functions. *Ann. Inst. Stat. Math.*, **16**, 1–19.
BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families.* Wiley.
BARTLETT, M. S. (1936). The square root transformation in analysis of variance. *J. R. Statist. Soc.* , Suppl. 3., 68–78.
BEALE, E. M. L. (1960). Confidence regions in non-linear estimation. *J. R. Statist. Soc.* B, **22**, 41–76.
BORGES, R. (1970). Eine Approximation der Binomialverteilung durch die Normalverteilung der Ordnung $1/n$. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **14**, 189–199.
—— (1971). Derivation of normalising transformations with an error of order $1/n$. *Sankhyā*, A, **33**, 441–460.
BOX, M. J. (1971). Bias in non-linear estimation. *J. R. Statist. Soc.* B. **32**, 171–201.
COX, D. R. (1972). Regression models and life tables. *J. R. Statist. Soc.* B, **34**, 187–220.
EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Stat.*, **3**, 1189–1242.
HOUGAARD, P. (1980). Selected topics from non-linear statistical inference. Unpublished Cand. Stat. thesis (in Danish), Institute of Mathematical Statistics, University of Copenhagen.
—— (1981). The appropriateness of the asymptotic distribution in a non-linear regression model in relation to curvature. Research Report 81/9. Statistical Research Unit, Copenhagen, Denmark.
SPROTT, D. A. (1973). Normal likelihoods and their relation to large sample theory of estimation. *Biometrika*, **60**, 457–465.
VÆTH, M. (1981). On the use of Wald's test for exponential families. Research Report no. 70, University of Aarhus, Institute of Mathematics, Department of Theoretical Statistics.
WELCH, B. L. and PEERS, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. R. Statist. Soc.* B, **25**, 318–329.
WILSON, E. B. and HILFERTY, M. M. (1931). The distribution of chi-square. *Proc. Nat. Acad.*, **17**, 684–688.