Une interprétation de la pseudo-vraisemblance -
An interpretation of the pseudo-likelihood

Monique Graf
Institut de Statistique - Université de Neuchâtel, Switzerland
and Elpacos Statistics, la Neuveville, Switzerland

10e Colloque Francophone sur les Sondages
Université de Lyon
Outline

Introduction

One stage design

Two stage design

Discussion
Introduction

Multilevel models = special case of the generalized mixed model, used for the analysis of survey data with several levels (strata, clusters, units)


In multi-stage surveys, scaling of weights influence the parameter estimates (see e.g. Rabe-Hesketh and Skrondal, 2006 and Asparouhov, 2006).

No theory on the choice of scaling.
Alternatives to the pseudo-likelihood

- Rao et al. (2013) propose a method by estimating functions that have good asymptotic properties.
- Sampling density conditional on the distribution of weights for non-ignorable designs, e.g. Pfeffermann (2011). Bonnéry et al. (2018) establish asymptotic properties of the likelihood obtained with this density.
Goal

- Given a postulated population distribution,
- obtain the pseudo-likelihood,
- find a **proper likelihood**
  - belonging to the same family of distributions as the population distribution
  - as ”close” as possible to the pseudo-likelihood.
- Derive a method for rationally choosing the scaling of weights.
Consider a one stage design. Let

- \( \{y_i, w_i, i = 1, \ldots, n\} \) = sampled units and the corresponding extrapolation weights.
- \( y_i \): realization of a random variable \( Y_i \)
- a model: \( Y_i \) are i.i.d with pdf \( f(., \theta) \) depending on a set of parameters \( \theta \).
Pseudo-log-likelihood

In a one-stage design, the pseudo-log-likelihood given by

\[ \ell_{\text{pseudo}}(\theta; y, w) = \sum_{i=1}^{n} w_i \log f(y_i, \theta) = \sum_{i=1}^{n} \log f(y_i, \theta)^{w_i}. \]

\( \ell_{\text{pseudo}} \) is a proper log-likelihood, if it can be written as a sum of log-densities, up to a constant term not depending on the parameters.

1. Conditions for \( \ell_{\text{pseudo}} \) to be a proper log-likelihood, \( \ell_{\text{proper}} \)?
2. Conditions for pdf \( K_i^{-1} f(y_i, \theta)^{w_i} \) to belong to the same family of distributions as \( f(y_i, \theta) \)?
3. Conditions for the parameters of \( \ell_{\text{pseudo}} \) and \( \ell_{\text{proper}} \) to coincide?
In general,
\[
\int_{-\infty}^{\infty} f(y, \theta)^{xw_i} \, dy = K(xw_i, \theta) = K_i \implies K_i^{-1} f(y, \theta)^{xw_i} \text{ is a pdf.}
\]

Thus
\[
\ell^{\text{proper}} = \sum_{i=1}^{n} \log[K(xw_i, \theta)^{-1} f(y, \theta)^{xw_i}]
\]

Thus
\[
\ell^{\text{pseudo}} = \ell^{\text{proper}} + \sum_i \log[K(xw_i, \theta)] - \sum_i xw_i \log[K(1, \theta)]
\]
\[
= \ell^{\text{proper}} + C(xw, \theta).
\]
Equivalence condition

\[ \ell_{\text{pseudo}} \text{ equivalent to } \ell_{\text{proper}} \]

\[ C(xw, \theta) = C(xw). \]
Sampling pdf

- $K(w_i, \theta)^{-1} f(y, \theta)^{w_i}$ can be interpreted as the sampling pdf of $Y_i$, the random variable associated to the $i$-th sampled unit.
- observations are no longer identically distributed, but still independent (according to the model).
- the sampling pdf depends on the scaling of weights.

How to choose the scaling?
Canonical scaling

- A proper likelihood is the sum of $n$ log-densities where $n$ is the sample size.
- $\tilde{w}_i$, $i = 1, \ldots, n = \text{provided weights.}$
- *Canonical weights*:
  
  $$w_i = n \frac{\tilde{w}_i}{\sum_{k=1}^{n} \tilde{w}_k} = \tilde{w}_i \bar{\tilde{w}} \text{ sum to } n.$$  

- Another scaling can always be defined from the canonical weights.
  
  $x = \text{scaling factor}$
  $xw_i = \text{scaled weight.}$
Normal distribution - One stage design

- \( Y_i \sim N(\mu_i, \sigma^2) \)
- \( X \) = matrix of auxiliary variables;
- \( x_{it} = i\)-th row of \( X \)
- \( \mu_i = \mathbb{E}(Y_i) = x_{it}^T\beta \)
- parameters: \( \theta = (\beta, \sigma) \)
Normal distribution - One stage design

- Population log-likelihood

\[ \ell^{\text{pop}}(\beta, \sigma; y) = \sum_{i=1}^{N} \log \left( \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \frac{(y_i - \mu_i)^2}{\sigma^2} \right) \right). \]

- Pseudo-log-likelihood

\[ \ell^{\text{pseudo}}(\beta, \sigma; y, xw) = \sum_{i=1}^{n} xw_i \log \left[ \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \frac{(y_i - \mu_i)^2}{\sigma^2} \right) \right] \]

- Proper log-likelihood

\[ \ell^{\text{proper}}(\beta, \sigma; y, xw) = \sum_{i=1}^{n} \log \left[ \frac{\sqrt{xw_i}}{(\sigma \sqrt{2\pi})} \exp \left( -\frac{1}{2} \frac{(y_i - \mu_i)^2}{\sigma^2} \right) \right] \]

- Correction term

\[ C(x, w, \theta) = \sum_{i=1}^{n} (xw_i - 1) \log \left( \frac{1}{\sigma \sqrt{2\pi}} \right) - \frac{1}{2} \log \left( \prod_{i=1}^{n} xw_i \right) \]
The correction term can be simplified,

\[ C(x, w, \sigma) = n \left\{ (x - 1) \log \left( \frac{1}{\sigma \sqrt{2\pi}} \right) - \frac{1}{2} [\log(x) + \log(G)] \right\} \]

where \( G \) is the geometric mean of the canonical weights.

- \( C \) does not depend on \( \beta \).
- \( \hat{\beta}^{\text{pseudo}} \) and \( \hat{\sigma}^{\text{pseudo}} \) do not depend on \( x \).
- \( \ell^{\text{proper}}(\beta, \sigma; y, x w) \equiv \ell^{\text{proper}}(\beta, \sigma / \sqrt{x}; y, w) \) thus
  \[ \hat{\sigma}^{\text{pseudo}} = \hat{\sigma}_x^{\text{proper}} \text{ where } \sigma_x = \sigma / \sqrt{x}. \]

- \( C \) does not depend on \( \sigma \) if and only if \( x = 1 \).

With the canonical weights, it is equivalent to estimate the parameters using the pseudo- or the proper log-likelihood.
Exponential distribution - One stage design

\[ g(y; b) = \frac{1}{b} \exp \left( -\frac{y}{b} \right) \quad y > 0; \ b > 0. \]

\[ g^w(y; b) = \left( \frac{1}{b} \right)^w \exp \left( -\frac{wy}{b} \right) = \frac{1}{b/w} \exp \left( -\frac{y}{b/w} \right) \frac{1}{wb^{w-1}} = g(y; b/w) \frac{1}{wb^{w-1}}. \]

The pseudo-log-likelihood is given by

\[ \ell^{\text{pseudo}}(b; y, xw) = \sum_{i=1}^{n} xw_i \log (g(y_i; b)) \]

\[ = \ell^{\text{proper}}(b; y, xw) - n \log(xG) - \sum_{i=1}^{n} (xw_i - 1) \log(b). \]

\[ C(x, w, b) = -n \{ \log(x) + (x - 1) \log(b) + \log(G) \} \]

Same form as before.
Generalized gamma distribution - One stage design

- Probability density of \( Y \sim GG(a, b, p) \):

\[
g(y; a, b, p) = \frac{a}{\Gamma(p)} (y/b)^{ap} \exp\{-(y/b)^a\} \frac{1}{y} \quad a, b, p > 0.
\]

In the applications, \( b = \exp(x^t \beta) \), where \( x \) is a vector of auxiliary variables.

- Change of variable: \( u = \log(y) \); pdf of \( \log(Y) \):

\[
f(u; a, b, p) = \frac{a}{\Gamma(p)} (e^u/b)^{ap} \exp\{-(e^u/b)^a\}
\]
Which pseudo-likelihood?

\[ \ell_{pseudo} \text{ based on } g \neq \ell_{pseudo} \text{ based on } f \]

- \( \ell_{pseudo} \) based on \( g \): weights are applied to \( y \)
- \( \ell_{pseudo} \) based on \( f \): weights are applied to \( \log(y) \)

Weights do not have the same meaning according to the model.

Good reason to choose \( f \):

the sampling density is more similar to the population density.
Generalized gamma distribution - One stage design

The proper log-likelihood is the sum of log-densities, pdf of $GG(a, b/(xw_i)^{1/a}, pxw_i)$. 
Correction term for the pseudo-log-likelihood :

$$C(x, w, a, p) = \sum_i \log \left\{ \left[ \frac{a}{\Gamma(p)} \right]^{xw_i} \frac{\Gamma(pxw_i)}{a} \right\}$$

$$= n(x - 1) \log(a) - nx \log(\Gamma(p)) + \sum_i \log(\Gamma(pxw_i))$$

- $C$ does not depend on $b$
- if $x = 1$, $C$ does not depend on $a$
- if $w_i \neq 1$, the dependence on $p$ remains.

With unequal weights, $\ell^{\text{pseudo}}$ and $\ell^{\text{proper}}$ will give different estimates.
Three approximations of \( \text{digamma}(p) \)

\[ k = 1.80256 \]

\[
\text{digamma}(p) - \log(p) + \frac{1}{p} \\
\text{digamma}(p) - \log(p) + \frac{1}{(k \cdot p)} \\
\text{digamma}(p) - \log(p) + \frac{1}{(2 \cdot p)}
\]
Generalized gamma distribution - One stage design

Set $x = 1$.

$C_1(p) = C(1, w, a, p) = -n \log(\Gamma(p)) + \sum_i \log(\Gamma(p w_i))$.

\[
\frac{\partial}{\partial p} \ell_{pseudo} = \frac{\partial}{\partial p} \ell_{proper} + \frac{d}{dp} C_1(p),
\]

\[
\frac{d}{dp} C_1(p) = -n \psi(p) + \sum_i w_i \psi(p w_i)
\]

\[
\approx -n \left[ \log(p) - \frac{1}{kp} \right] + \sum_i w_i \left[ \log(w_i p) - \frac{1}{kw_i p} \right]
\]

\[
= \sum_i w_i \log(w_i).
\]

\[
\frac{d}{dp} C_1(p) = \sum_i w_i \log(w_i) \pm \frac{1}{2p}.
\]
Noufaily and Jones (2013) unweighted case: the score equation in $p$ is strictly decreasing for given values of $a$ and $b$.

This property extends to the weighted case.

It can be shown that if $n \geq 3$, $\sum_i w_i \log(w_i)$ is always positive.

Thus in general we expect

$$\hat{p}^{\text{pseudo}} > \hat{p}^{\text{proper}}.$$
Densities GG(1,1,p)

- \( p = 3 \)
- \( p = 5 \)

pseudo-lik: One stage design
Two stage design

Primary sampling units (PSU) are selected and within each PSU a sample is selected.
Hypothesis: The model includes an additive random effect that corresponds to the PSU of the design.

Within each PSU \( j \), weights \( \tilde{w}_{ij} \) are provided for the ultimate unit \( i, i \in j \).

- \( n_j \) = sample size in PSU \( j \).
- \( w_{ij} = n_j \frac{\tilde{w}_{ij}}{\sum_{k=1}^{n} \tilde{w}_{kj}} = \frac{\tilde{w}_{ij}}{\tilde{w}_j} \) = canonical weight within primary unit \( j \).
- observations within PSU \( j \) are conditionally independent given random effect \( V_j \),
- \( f_1(y - v; \theta) = \) conditional pdf of \( Y_{ij} \) given the random effect \( V_j = v \).
- within PSU pseudo-log-likelihood = 
  \[ \ell_{j}^{pseudo}(\theta; y_j - v1_{n_j}, xw_j) = \sum_{i=1}^{n_j} xw_{ij} \log[f_1(y_{ij} - v; \theta)] \]
Two stage design

- $V = \text{latent unobserved PSU effect with pdf } f_2(v; \Theta)$.
- $\tilde{W}_j = \text{provided weight of PSU } j, j = 1, \ldots, c$.
- $W_j = \text{canonical weight of PSU } j$,

$$W_j = \sum_{k=1}^{c} n_k \frac{\tilde{W}_j}{\sum_{k=1}^{c} n_k \tilde{W}_k} = \frac{\tilde{W}_j}{\tilde{W}_n}.$$

Total sample size:

$$\sum_j n_j = \sum_j n_j W_j.$$
Two stage design

Total pseudo-log-likelihood =

$$\ell^{\text{pseudo}}(\theta, \Theta; \{y_j, xw_j, j = 1, \ldots, c\}; tW)$$

$$= \sum_{j=1}^{c} tW_j \log \left[ \int_{-\infty}^{\infty} \exp(\ell^{\text{pseudo}}_j(\theta; y_j - v1_{nj}, xw_j)f_2(v; \Theta)dv) \right]$$

$$= \sum_{j=1}^{c} tW_j \log \left[ \int_{-\infty}^{\infty} \exp(\ell^{\text{proper}}_j(\theta; y_j - v1_{nj}, xw_j)f_2(v; \Theta)dv) \right]$$

$$+ \sum_{j=1}^{c} tW_j \left[ C_{1j}(xw_j; \theta) \right]$$

$$= \ell^{\text{proper}}(\theta, \Theta; \{y_j, xw_j, j = 1, \ldots, c\}; tW)$$

$$+ \sum_{j=1}^{c} tW_j \left[ C_{1j}(xw_j; \theta) \right] + C_2(\{xw_j, j = 1, \ldots, c\}, tW; \theta, \Theta)$$
Normal distribution - Two stage design

Population model:

- $\theta = (\beta, \sigma)$
- $Y_{ij} \sim N(x_{ij}^t \beta - \nu, \sigma^2), \ i = 1, \ldots, n_j$ independent observations with pdf $f_1(y - \nu; \theta)$ given random effect $V_j = \nu$.
- $\Theta = (\eta)$
- $V_j \sim N(0, \eta^2), \ j = 1, \ldots, c$ : independent random effects with pdf $f_2(\nu; \Theta) = $.

The model and the within-PSU weighting scheme imply

**Sampling distribution:**

$(Y_1, \ldots, Y_c)$ are independent vectors with

\[ Y_j \sim N(X_j^t \beta, \Gamma_j) \quad \Gamma_j = \frac{\sigma^2}{x} \text{diag}(w_j)^{-1} + \eta^2 11^T. \]

\[ \det(\Gamma_j) = \left( G_j \frac{\sigma^2}{x} \right)^{n_j} \frac{n_j \eta^2 + \sigma^2/x}{\sigma^2/x}. \]

$G_j = \text{geometric mean of weights} \ (w_{ij}, \ i = 1, \ldots, n_j) = w_j.$
Normal distribution - Two stage design

Correction term

$$C(\{xw_j, j = 1, ..., c\}, tW; \sigma, \eta) = \sum_j tW_j C_{1j} + C_2 = C_1 + C_2$$

$$2C_1 = -\sum_j tW_j n_j \{(x - 1) \log (2\pi\sigma^2) + [\log(x) + \log(G_j)]\}$$

$$2C_2 = \sum_j n_j \log(W_j) + (tW_j - 1) \log[\det(\Gamma_j)]$$

$$= \sum_j n_j [\log(tW_j) + (tW_j - 1) \log(G_i)]$$

$$- (\sum_j n_j)(t - 1) \log(\sigma^2/x) + \sum_j (tW_j - 1) \log \left[\frac{n_j \eta^2 + \sigma^2/x}{\sigma^2/x} \right].$$
Normal distribution - Two stage design

- **x = 1** makes $C_1$ independent of $\sigma$ i.e.
  
  $x = 1 \implies \ell_{j}^{\text{pseudo}}$ and $\ell_{j}^{\text{proper}}$ are equivalent for all $j$.

- If moreover $t = 1$, $C_2$ is independent of $\sigma$ and $\eta$ in two instances:
  1. If $n_j = n$, then $\Gamma_j = \Gamma$ and $\sum_j W_j = c$,
     
     $$2C_2 = n \sum_{j=1}^{c} \log(W_j)$$

  2. If $W_j = 1$, 
     
     $$2C_2 = 0.$$

In all other cases, the overall log-likelihoods $\ell_{j}^{\text{pseudo}}$ and $\ell_{j}^{\text{proper}}$ will give different estimates.
Multivariate generalized beta distribution (MGB2)  

Two stage design

MGB2 distribution (Yang et al., 2010):

A set of $n$ random variables $\mathbf{Y} = (Y_1, ..., Y_n)$ conditionally independent given a random scale parameter $\Theta$, with pdf

$$\mathbf{Y}|\{\Theta = \theta\} \sim GG(a, (\theta^{-1/a}b), p)$$

$\Theta \sim invGa(q)$ with pdf

$$g(\theta; q) = \frac{1}{\Gamma(q)} \theta^{-q} e^{-\theta} \frac{1}{\theta}$$

Graf, Marín and Molina (2018) use this setting in the context of small area estimation.

- $\Theta$ : latent area effect
- $\log(b) = X\beta$ : model on scale
- $a$, $p$ and $q$ : shape parameters
MGB2 - two stage

Aim: incorporate weights.
Same setting as in the normal case.

- **PSU** \( j \): sample size \( n_j, j = 1, ..., c \),
  canonical weights \( w_j = (w_{ij}, i = 1, ..., n_j) \)
  \[
  \log(b_{ij}) = x^t_{ij}/\beta
  \]

- **PSU canonical weights**: \( W_j \)
- \( x \) and \( t \) scaling factors.
- \( \ell_j^{\text{proper}} \): sum of log-densities \( GG(a, (\theta xw_{ij})^{-1/a}b_{ij}, pxw_{ij}) \)
- \( \Theta_j \sim invG(tW_jq) \)
- PSU are independent.
MGB2 - two stage

Correction terms

\[ C_{1j}(x) = n_j(x - 1) \log(a) - n_jx \log(\Gamma(p)) + \sum_{i=1}^{n_j} \log(\Gamma(xw_{ij}p)) \]

\[ C_2(t, x) = c(t - 1) \log(a) + \]
\[ \sum_{j=1}^{c} tW_j \log \left[ \frac{\Gamma(xn_jp + q)}{\Gamma(q) \prod_{i=1}^{n_j} \Gamma(xw_{ij}p)} \right] - \]
\[ \sum_{j=1}^{c} \log \left[ \frac{\Gamma(tW_jxn_jp + tW_jq)}{\Gamma(tW_jq) \prod_{i=1}^{n_j} \Gamma(tW_jxw_{ij}p)} \right] \]

\[ C(t, x) = \sum_{j=1}^{c} tW_j C_{1j}(x) + C_2(t, x). \]
MGB2 - two stage

- $C(t, x)$ does not depend on $a$ if and only if $t = 1$ and $x = 1$.
- $C(1, 1)$ does not depend on $q$, if $W_j = 1$.
- $C(1, 1)$ still depends on $p$ and $q$, if $n_j = n$.
- $C(1, 1)$ still depends on $p$ and $q$, if $w_{ij} = 1$ but $W_j \neq 1$.

The estimates based on $\ell^{\text{proper}}$ or $\ell^{\text{pseudo}}$ won’t coincide, except if all the canonical weights are 1.
Discussion

- Design properties of canonical weights.
- Underestimation of between-cluster variance in Gaussian model mentioned by e.g. Rabe-Hesketh and Skrondal (2006) when the expectation of weighted estimates is computed from the population model. It does not occur if the sampling distribution is used.
- Advantage of having a sampling density over a method of moments.
- Simpler than the sampling density based on modeling the weights.