Assessing the precision of estimates of variance components

Douglas Bates

Department of Statistics
University of Wisconsin – Madison
U.S.A.

(bates@stat.wisc.edu)

LMU, Munich
July 16, 2009

Outline

Describing the precision of parameters estimates

- In many ways the purpose of statistical analysis can be considered as quantifying the variability in data and determining how the variability affects the inferences that we draw from it.
- Good statistical practice suggests, therefore, that we not only provide our "best guess", the point estimate of a parameter, but also describe its precision (e.g. interval estimation).
- Some of the time (but not nearly as frequently as widely believed) we also want to check whether a particular parameter value is consistent with the data (i.e.. hypothesis tests and p-values).
- In olden days it was necessary to do some rather coarse approximations such as summarizing precision by the standard error of the estimate or calculating a test statistic and comparing it to a tabulated value to derive a 0/1 response of "significant (or not) at the 5% level".

Modern practice

- Our ability to do statistical computing has changed from the "olden days". Current hardware and software would have been unimaginable when I began my career as a statistician. We can work with huge data sets having complex structure and fit sophisticated models to them quite easily.
- Regrettably, we still frequently quote the results of this sophisticated modeling as point estimates, standard errors and p-values.
- Understandably, the client (and the referees reading the client's paper) would like to have simple, easily understood summaries so they can assess the analysis at a glance. However, the desire for simple summaries of complex analyses is not, by itself, enough to these summaries meaningful.
- We must not only provide sophisticated software for statisticians and other researchers; we must also change their thinking about summaries.

Summaries of mixed-effects models

- Commercial software for fitting mixed-effects models (SAS PROC MIXED, SPSS, MLwin, HLM, Stata) provides estimates of fixed-effects parameters, standard errors, degrees of freedom and p-values. They also provide estimates of variance components and standard errors of these estimates.
- The mixed-effects packages for R that I have written (`nlme` with José Pinheiro and `lme4` with Martin Mächler) do not provide standard errors of variance components. `lme4` doesn't even provide p-values for the fixed effects.
- This is a source of widespread anxiety. Many view it as an indication of incompetence on the part of the developers ("Why can't lmer provide the p-values that I can easily get from SAS?")
- The 2007 book by West, Welch and Galecki shows how to use all of these software packages to fit mixed-effects models on 5 different examples. Every time they provide comparative tables they must add a footnote that `lme` doesn't provide standard errors of variance components.

The Dyestuff data set

- The `Dyestuff`, `Penicillin` and `Pastes` data sets all come from the classic book *Statistical Methods in Research and Production*, edited by O.L. Davies and first published in 1947.
- The `Dyestuff` data are a balanced one-way classification of the `Yield` of dyestuff from samples produced from six `Batches` of an intermediate product. See `?Dyestuff`.

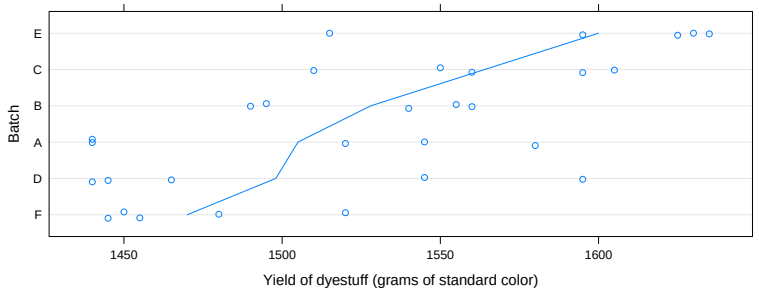
```
> str(Dyestuff)
'data.frame': 30 obs. of  2 variables:
 $ Batch: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ..
 $ Yield: num  1545 1440 1440 1520 1580 ...

> summary(Dyestuff)
 Batch     Yield
 A:5   Min.   :1440
 B:5   1st Qu.:1469
 C:5   Median :1530
 D:5   Mean   :1528
 E:5   3rd Qu.:1575
 F:5   Max.   :1635
```

What does a standard error tell us?

- Typically we use a standard error of a parameter estimate to assess precision (e.g. a 95% confidence interval on μ is roughly $\bar{x} \pm 2\frac{s}{\sqrt{n}}$) or to form a test statistic (e.g. a test of $H_0 : \mu = 0$ versus $H_a : \mu \neq 0$ based on the statistic $\frac{\bar{x}}{s/\sqrt{n}}$).
- Such intervals or test statistics are meaningful when the distribuion of the estimator is more-or-less symmetric.
- We would not, for example, quote a standard error of $\widehat{\sigma^2}$ because we know that the distribution of this estimator, even in the simplest case (the mythical i.i.d. sample from a Gaussian distribution), is not at all symmetric. We use quantiles of the χ^2 distribution to create a confidence interval.
- Why, then, should we believe that when we create a much more complex model the distribution of estimators of variance components will magically become sufficiently symmetric for standard errors to be meaningful?

Dyestuff data plot



- The line joins the mean yields of the six batches, which have been reordered by increasing mean yield.
- The vertical positions are jittered slightly to reduce overplotting. The lowest yield for batch A was observed on two distinct preparations from that batch.

A mixed-effects model for the dyestuff yield

```
> fm1 <- lmer(Yield ~ 1 + (1 | Batch), Dyestuff)
> print(fm1)
```

```
Linear mixed model fit by REML
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
   AIC   BIC logLik deviance REMLdev
 325.7 329.9 -159.8    327.4   319.7
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1764.0   42.00
 Residual             2451.3   49.51
Number of obs: 30, groups: Batch, 6
Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      19.38   78.81
```

- Fitted model `fm1` has one fixed-effect parameter, the mean yield, and one random-effects term, generating a simple, scalar random effect for each level of `Batch`.

Re-fitting the model for ML estimates

```
> (fm1M <- update(fm1, REML = FALSE))
```

```
Linear mixed model fit by maximum likelihood
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
   AIC   BIC logLik deviance REMLdev
 333.3 337.5 -163.7    327.3   319.7
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1388.4   37.261
 Residual             2451.2   49.510
Number of obs: 30, groups: Batch, 6
Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      17.69   86.33
```

(The extra parentheses around the assignment cause the value to be printed. Generally the results of assignments are not printed.)

REML estimates versus ML estimates

- The default parameter estimation criterion for linear mixed models is restricted (or "residual") maximum likelihood (REML).
- Maximum likelihood (ML) estimates (sometimes called "full maximum likelihood") can be requested by specifying `REML = FALSE` in the call to `lmer`.
- Generally REML estimates of variance components are preferred. ML estimates are known to be biased. Although REML estimates are not guaranteed to be unbiased, they are usually less biased than ML estimates.
- Roughly, the difference between REML and ML estimates of variance components is comparable to estimating σ^2 in a fixed-effects regression by $SSR/(n-p)$ versus SSR/n, where SSR is the residual sum of squares.
- For a balanced, one-way classification like the `Dyestuff` data, the REML and ML estimates of the fixed-effects are identical.

Evaluating the deviance function

- The *profiled deviance* function for such a model can be expressed as a function of 1 parameter only, the ratio of the random effects' standard deviation to the residual standard deviation.
- A very brief explanation is based on the n-dimensional response random variation, $\mathcal{Y}$, whose value, $\boldsymbol{y}$, is observed, and the q-dimensional, unobserved random effects variable, $\mathcal{B}$, with distributions

$$(\mathcal{Y}|\mathcal{B}=\boldsymbol{b}) \sim \mathcal{N}\left(\boldsymbol{Z}\boldsymbol{b}+\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n\right), \quad \mathcal{B} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\theta\right),$$

- For our example, $n=30$, $q=6$, $\boldsymbol{X}$ is a 30×1 matrix of 1s, $\boldsymbol{Z}$ is the 30×6 matrix of indicators of the levels of `Batch` and $\boldsymbol{\Sigma}$ is $\sigma_b^2\boldsymbol{I}_6$.
- We never really form $\boldsymbol{\Sigma}_\theta$; we always work with the *relative covariance factor*, $\boldsymbol{\Lambda}_\theta$, defined so that

$$\boldsymbol{\Sigma}_\theta = \sigma^2\boldsymbol{\Lambda}_\theta\boldsymbol{\Lambda}_\theta^\mathsf{T}.$$

In our example $\theta = \frac{\sigma_b}{\sigma}$ and $\boldsymbol{\Lambda}_\theta = \theta\boldsymbol{I}_6$.

Orthogonal or "unit" random effects

- We will define a q-dimensional "spherical" or "unit" random-effects vector, $\mathcal{U}$, such that

$$\mathcal{U} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I_q\right), \ \mathcal{B} = \Lambda_\theta \mathcal{U} \Rightarrow \mathsf{Var}(\mathcal{B}) = \sigma^2 \Lambda_\theta \Lambda_\theta^\intercal = \Sigma_\theta.$$

- The linear predictor expression becomes

$$Zb + X\beta = Z\Lambda_\theta\, u + X\beta = U_\theta\, u + X\beta$$

where $U_\theta = Z\Lambda_\theta$.

- The key to evaluating the log-likelihood is the Cholesky factorization

$$L_\theta L_\theta^\intercal = P\left(U_\theta^\intercal U_\theta + I_q\right) P^\intercal$$

(P is a fixed permutation that has practical importance but can be ignored in theoretical derivations). The sparse, lower-triangular L_θ can be evaluated and updated for new θ even when q is in the millions and the model involves random effects for several factors.

Profiling the deviance with respect to β

- Because the deviance depends on β only through $r^2(\theta, \beta)$ we can obtain the conditional estimate, $\widehat{\beta}_\theta$, by extending the PLS problem to

$$r^2(\theta) = \min_{u,\beta} \left[\|y - X\beta - U_\theta\, u\|^2 + \|u\|^2\right]$$

with the solution satisfying the equations

$$\begin{bmatrix} U_\theta^\intercal U_\theta + I_q & U_\theta^\intercal X \\ X^\intercal U_\theta & X^\intercal X \end{bmatrix} \begin{bmatrix} \tilde{u}_\theta \\ \widehat{\beta}_\theta \end{bmatrix} = \begin{bmatrix} U_\theta^\intercal y \\ X^\intercal y. \end{bmatrix}$$

- The profiled deviance, which is a function of θ only, is

$$-2\tilde{\ell}(\theta) = \log(|L_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r^2(\theta)}{n}\right)\right]$$

The profiled deviance

- The Cholesky factor, L_θ, allows evaluation of the conditional mode $\tilde{u}_{\theta,\beta}$ (also the conditional mean for linear mixed models) from

$$\left(U_\theta^\intercal U_\theta + I_q\right) \tilde{u}_{\theta,\beta} = P^\intercal L_\theta L_\theta^\intercal P \tilde{u}_{\theta,\beta} = U_\theta^\intercal (y - X\beta)$$

Let $r^2(\theta, \beta) = \|y - X\beta - U_\theta\, \tilde{u}_{\theta,\beta}\|^2 + \|\tilde{u}_{\theta,\beta}\|^2$.

- $\ell(\theta, \beta, \sigma|y) = \log L(\theta, \beta, \sigma|y)$ can be written

$$-2\ell(\theta, \beta, \sigma|y) = n\log(2\pi\sigma^2) + \frac{r^2(\theta, \beta)}{\sigma^2} + \log(|L_\theta|^2)$$

- The conditional estimate of σ^2 is

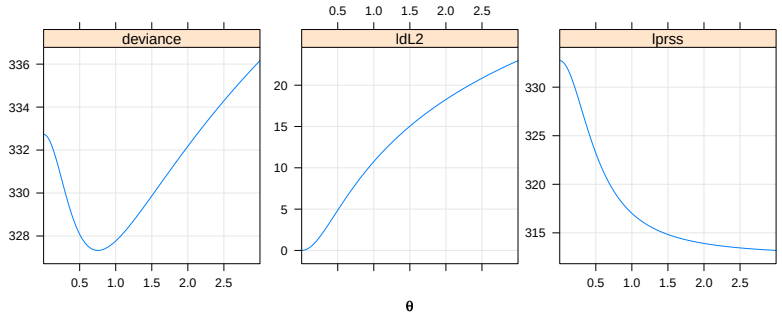
$$\widehat{\sigma^2}(\theta, \beta) = \frac{r^2(\theta, \beta)}{n}$$

producing the *profiled deviance*

$$-2\tilde{\ell}(\theta, \beta|y) = \log(|L_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r^2(\theta, \beta)}{n}\right)\right]$$

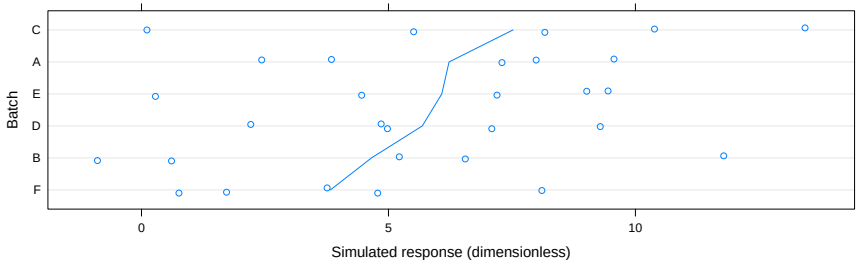
Profiled deviance and its components

- For this simple model we can evaluate and plot the deviance for a range of θ values. We also plot its components, $\log(|L_\theta|^2)$ (`ldL2`) and $n\left[1 + \log\left(\frac{2\pi r^2(\theta)}{n}\right)\right]$ (`lprss`).
- `lprss` measures fidelity to the data. It is bounded above and below. $\log(|L_\theta|^2)$ measures complexity of the model. It is bounded below but not above.

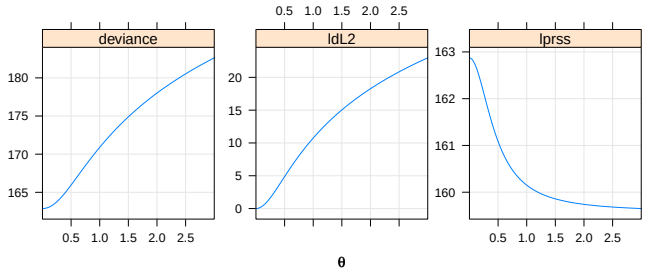


The MLE (or REML estimate) of σ_b^2 can be 0

- For some model/data set combinations the estimate of σ_b^2 is zero. This occurs when the decrease in `lprss` as $\theta \uparrow$ is not sufficient to counteract the increase in the complexity, $\log(|\boldsymbol{L}_\theta|^2)$. The `Dyestuff2` data from Box and Tiao (1973) show this.

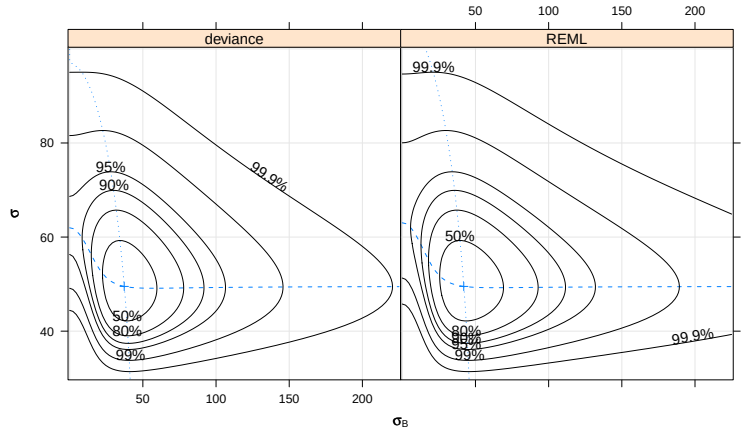


Components of the profiled deviance for `Dyestuff2`



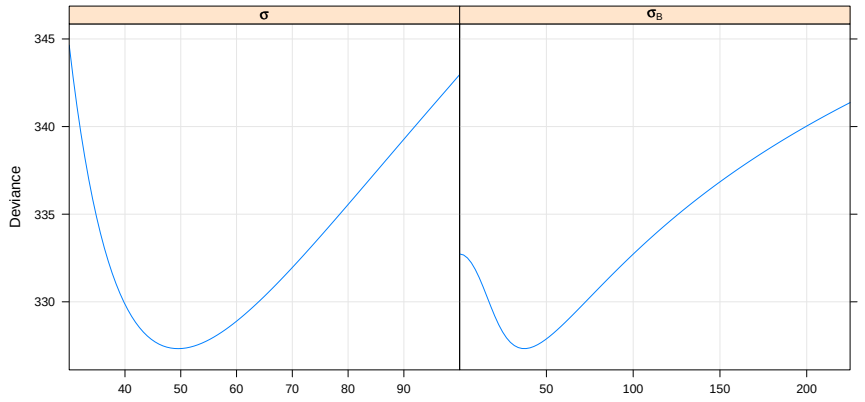
- For this data set the difference in the upper and lower bounds on `lprss` is not sufficient to counteract the increase in complexity of the model, as measured by $\log(|\boldsymbol{L}_\theta|^2)$.
- Software should gracefully handle cases of $\sigma_b^2 = 0$ or, more generally, $\boldsymbol{\Lambda}_\theta$ being singular. This is not done well in the commercial software.
- One of the big differences between inferences for σ_b^2 and those for σ^2 is the need to accomodate to do about values of σ_b^2 that are zero or near zero.

Profiled deviance and REML criterion for σ_b and σ



- The contours correspond to 2-dimensional marginal confidence regions derived from a likelihood-ratio test.
- The dotted and dashed lines are the profile traces.

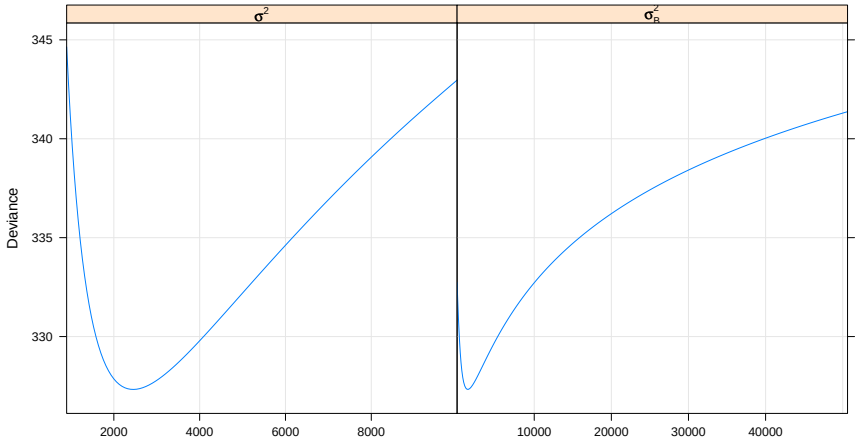
Profiling with respect to each parameter separately



- These curves show the minimal deviance achieveable for a value of one of the parameters, optimizing over all the other parameters.

Profiled deviance of the variance components

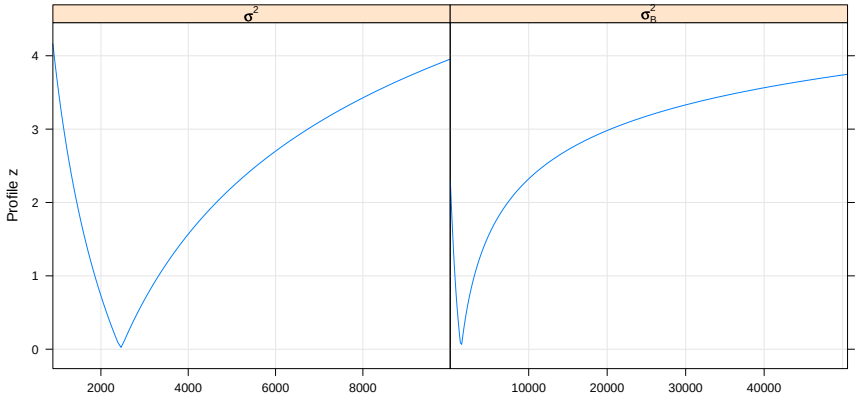
- Recall that we have been working on the scale of the standard deviations, σ_b and σ. On the scale of the variance, things look worse.



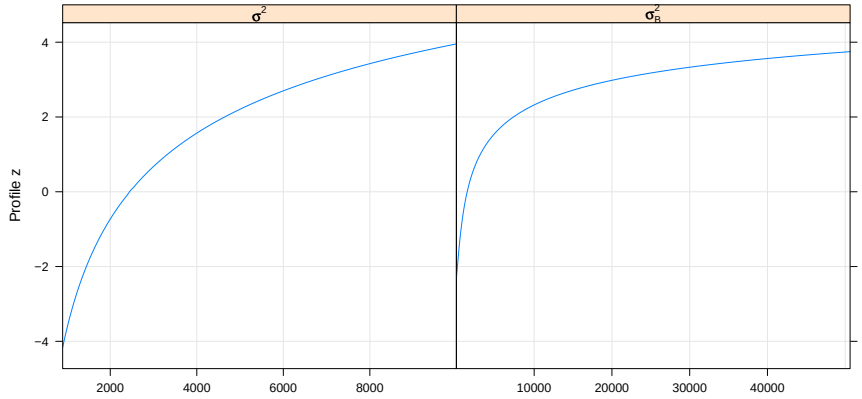
Square root of change in the profiled deviance

- The difference of the profiled deviance at the optimum and at a particular value of σ or σ_b is the likelihood ratio test statistic for that parameter value.
- If the use of a standard error, and the implied symmetric intervals, is appropriate then this function should be quadratic in the parameter and its square root should be like an absolute value function.
- The assumption that the change in the deviance has a χ^2_1 distribution is equivalent to saying that $\sqrt{\text{LRT}}$ is the absolute value of a standard normal.
- If we use the *signed square root* transformation, assigning $-\sqrt{\text{LRT}}$ to parameters to the left of the estimate and $\sqrt{\text{LRT}}$ to parameter values to the right, we should get a straight line on a standard normal scale.

Plot of square root of LRT statistic



Signed square root plot of LRT statistic



Summary

- Summaries based on parameter estimates and standard errors are appropriate when the distribution of the estimator can be assumed to be reasonably symmetric.
- Estimators of variances do not tend to have a symmetric distribution. If anything the scale of the log-variance (which is a multiple of the log-standard deviation) would be the more appropriate scale on which to assume symmetry.
- Estimators of variance components are more problematic because they can take on the value of zero.
- Profiling the deviance and plotting the result can help to visualize the precision of the estimates.