# Assessing the precision of estimates of variance components

Douglas Bates

University of Wisconsin - Madison
and R Development Core Team
<Douglas.Bates@R-project.org>

Max Planck Institute for Ornithology
Seewiesen
July 21, 2009

# Outline

# Outline

# Outline

# Outline

# Outline

# 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.

# Outline

# 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.

# 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 $\mu$ 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 $\chi^2$ 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?

# Outline

# 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 $30 \times 1$ matrix of 1s, $\boldsymbol{Z}$ is the $30 \times 6$ 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, $\boldsymbol{\mathcal{U}}$, such that

$$\boldsymbol{\mathcal{U}} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_q\right), \ \boldsymbol{\mathcal{B}} = \boldsymbol{\Lambda}_\theta \boldsymbol{\mathcal{U}} \Rightarrow \mathsf{Var}(\boldsymbol{\mathcal{B}}) = \sigma^2 \boldsymbol{\Lambda}_\theta \boldsymbol{\Lambda}_\theta^{\mathsf{T}} = \boldsymbol{\Sigma}_\theta.$$

- The linear predictor expression becomes

$$\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\, \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{U}_\theta\, \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta}$$

  where $\boldsymbol{U}_\theta = \boldsymbol{Z}\boldsymbol{\Lambda}_\theta$.

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

$$\boldsymbol{L}_\theta \boldsymbol{L}_\theta^{\mathsf{T}} = \boldsymbol{P}\left(\boldsymbol{U}_\theta^{\mathsf{T}} \boldsymbol{U}_\theta + \boldsymbol{I}_q\right)\boldsymbol{P}^{\mathsf{T}}$$

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

# The profiled deviance

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

$$\left(\boldsymbol{U}_\theta^\intercal \boldsymbol{U}_\theta + \boldsymbol{I}_q\right) \tilde{\boldsymbol{u}}_{\theta,\beta} = \boldsymbol{P}^\intercal \boldsymbol{L}_\theta \boldsymbol{L}_\theta^\intercal \boldsymbol{P} \tilde{\boldsymbol{u}}_{\theta,\beta} = \boldsymbol{U}_\theta^\intercal (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

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

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

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

- The conditional estimate of $\sigma^2$ is

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

  producing the *profiled deviance*

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

# Profiling the deviance with respect to $\boldsymbol{\beta}$

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

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

  with the solution satisfying the equations

$$\begin{bmatrix} \boldsymbol{U}_\theta^\intercal \boldsymbol{U}_\theta + \boldsymbol{I}_q & \boldsymbol{U}_\theta^\intercal \boldsymbol{X} \\ \boldsymbol{X}^\intercal \boldsymbol{U}_\theta & \boldsymbol{X}^\intercal \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_\theta \\ \widehat{\boldsymbol{\beta}}_\theta \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_\theta^\intercal \boldsymbol{y} \\ \boldsymbol{X}^\intercal \boldsymbol{y}. \end{bmatrix}$$

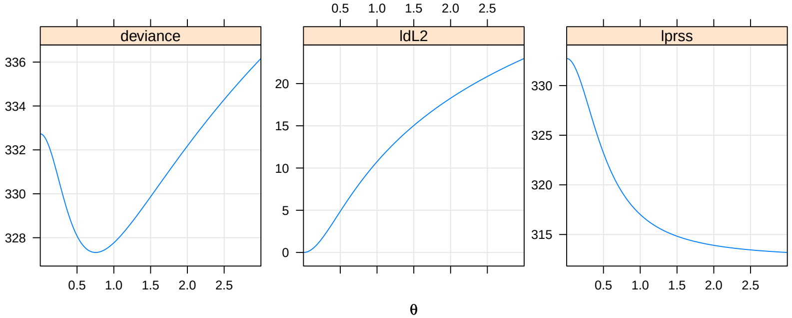
- The profiled deviance, which is a function of $\boldsymbol{\theta}$ only, is

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

# Outline

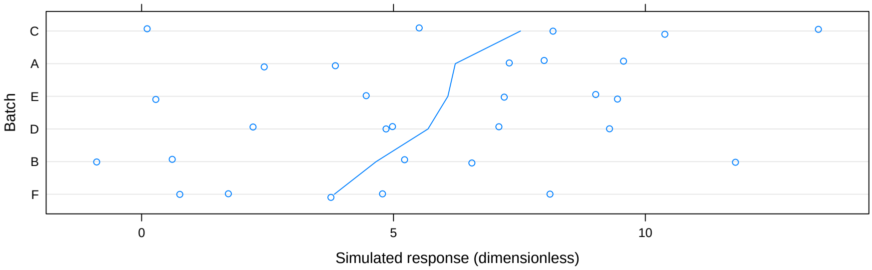
# Profiled deviance and its components

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

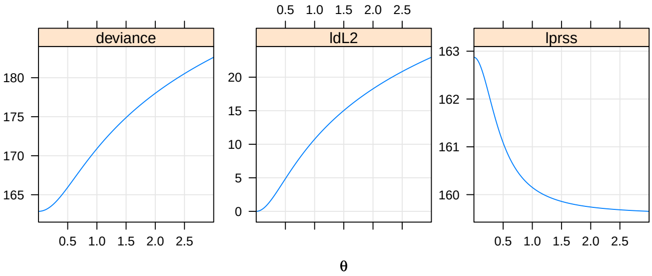


# The MLE (or REML estimate) of $\sigma_b^2$ can be $0$

- For some model/data set combinations the estimate of $\sigma_b^2$ 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.

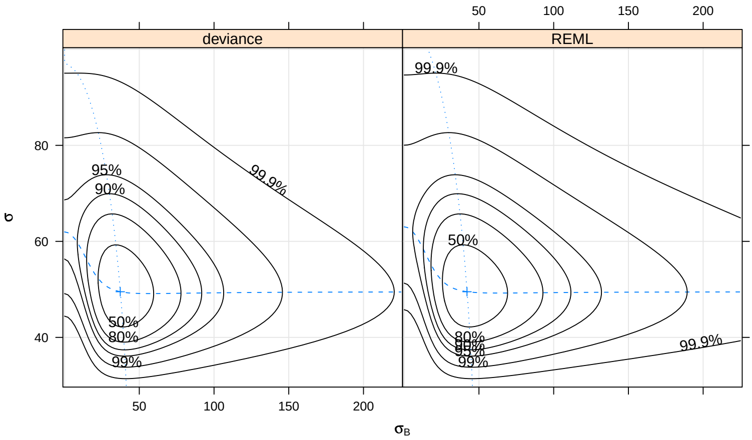


# Components of the profiled deviance for Dyestuff2



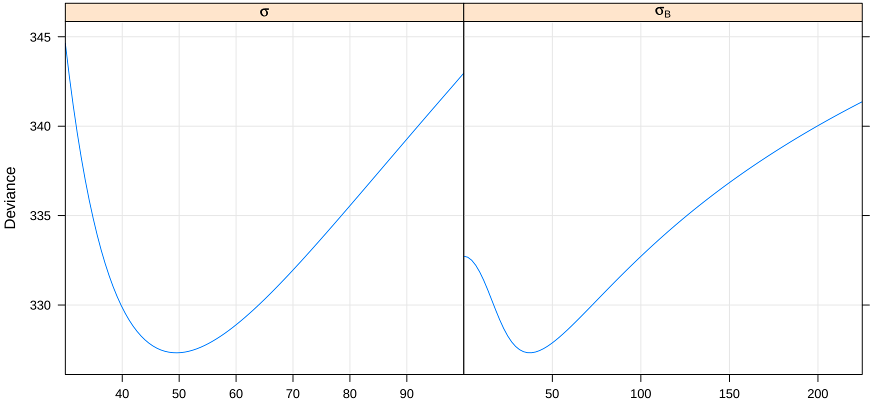

- For this data set the difference in the upper and lower bounds on `lprss` is not sufficient to counteract the increase in complexity of the model, as measured by $\log(|\boldsymbol{L}_\theta|^2)$.
- Software should gracefully handle cases of $\sigma_b^2 = 0$ or, more generally, $\boldsymbol{\Lambda}_\theta$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

# Profiled deviance and REML criterion for $\sigma_b$ and $\sigma$




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

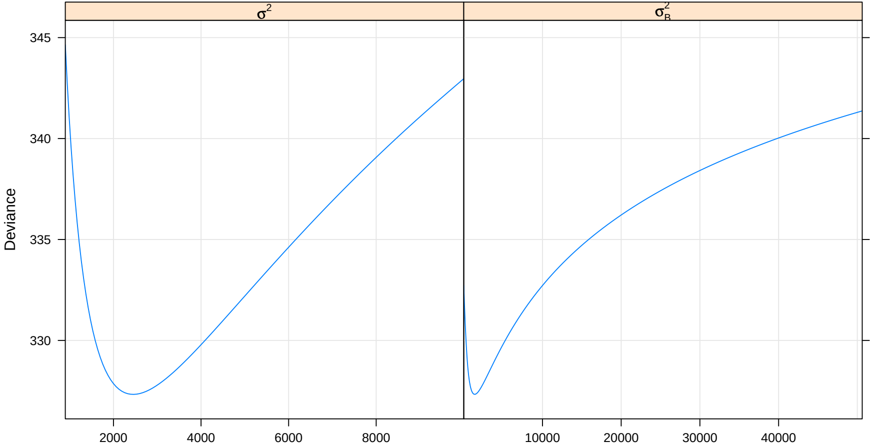
# Profiling with respect to each parameter separately




- 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

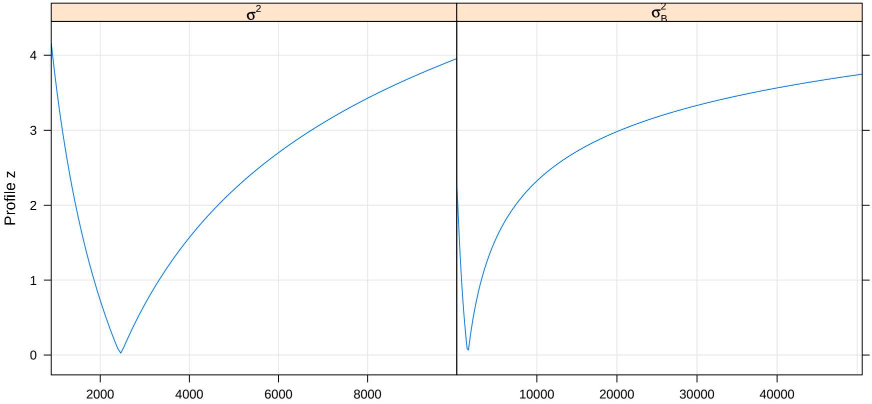
- Recall that we have been working on the scale of the standard deviations, $\sigma_b$ and $\sigma$. On the scale of the variance, things look worse.



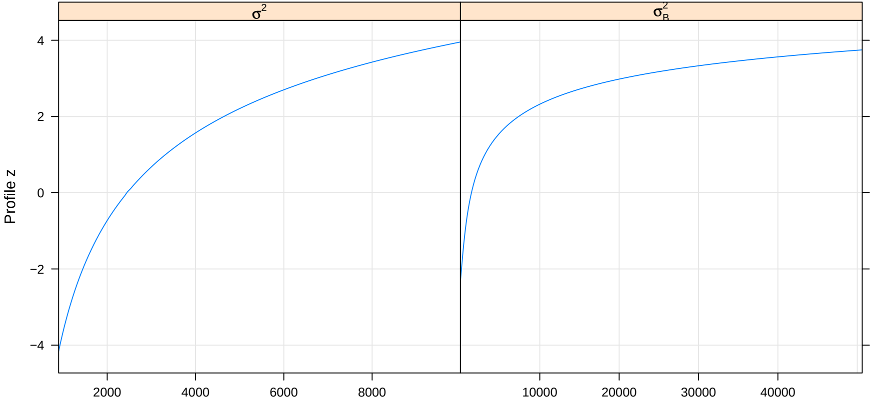
# Square root of change in the profiled deviance

- The difference of the profiled deviance at the optimum and at a particular value of $\sigma$ or $\sigma_b$ 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 $\chi^2_1$ 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



# Outline

# 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.