Web based supplementary materials for "Prediction of random effects in linear and generalized linear models under model misspecification" by Charles E. McCulloch and John M. Neuhaus

February 11, 2010

1 Web Appendix A: Best predicted values under different assumed distributions

1.1 Assuming b_i a mixture of normals

Assume that b_i follows a scaled, two-component, mixture distribution. If b_i is sampled from the first component (indicated as $c = 1$) then its conditional distribution is $\mathcal{N}(-\sigma_b \delta(1-p), \sigma_b^2 \tau^2)$, where $\tau^2 = 1 - \delta^2 p[1-p]$. When $c = 2$ (second component) and using the notation $\sigma_{bw}^2 = \sigma_b^2 \tau^2$ (for the variance of b within a component) the conditional distribution is $\mathcal{N}(\sigma_b \delta p, \sigma_{bw}^2)$. As in Verbecke and Lesaffre (1996) it is straightforward to derive the best predicted values in a linear mixed model:

$$\begin{aligned}
\tilde{b}_i &= \mathrm{E}[b_i|\bar{Y}_{i\cdot}] \\
&= \mathrm{E}\left[\mathrm{E}[b_i|\bar{Y}_{i\cdot}, c]\right], \qquad (1)\\
&= \pi(1)\mathrm{E}[b_i|\bar{Y}_{i\cdot}, c=1] + [1-\pi(1)]\mathrm{E}[b_i|\bar{Y}_{i\cdot}, c=2] \\
&= \pi(1)\left\{-\sigma_b\delta(1-p) + \frac{\sigma_{bw}^2}{\sigma_{bw}^2+\sigma_\epsilon^2/n}[\bar{Y}_{i\cdot} - \bar{x}_{i\cdot}'\boldsymbol{\beta} + \sigma_b\delta(1-p)]\right\} \\
&\quad +[1-\pi(1)]\left\{\sigma_b\delta p + \frac{\sigma_{bw}^2}{\sigma_{bw}^2+\sigma_\epsilon^2/n}[\bar{Y}_{i\cdot} - \bar{x}_{i\cdot}'\boldsymbol{\beta} - \sigma_b\delta p]\right\} \\
&= \frac{\sigma_{bw}^2}{\sigma_{bw}^2+\sigma_\epsilon^2/n}(\bar{Y}_{i\cdot} - \bar{x}_{i\cdot}'\boldsymbol{\beta}) + \frac{\sigma_\epsilon^2/n}{\sigma_{bw}^2+\sigma_\epsilon^2/n}\left(-\pi(1)\sigma_b\delta(1-p) + [1-\pi(1)]\sigma_b\delta p\right)
\end{aligned}$$

where $\pi(1)$ is the conditional probability of being from component 1 given $\bar{Y}_{i\cdot}$ and is given by

$$\pi(1) = \frac{p\exp\left\{-\frac{(\bar{Y}_{i\cdot}-\bar{x}_{i\cdot}'\beta+\delta\sigma_b[1-p])^2}{2(\sigma_{bw}^2+\sigma_\epsilon^2/n)}\right\}}{p\exp\left\{-\frac{(\bar{Y}_{i\cdot}-\bar{x}_{i\cdot}'\beta+\delta\sigma_b[1-p])^2}{2(\sigma_{bw}^2+\sigma_\epsilon^2/n)}\right\} + (1-p)\exp\left\{-\frac{(\bar{Y}_{i\cdot}-\bar{x}_{i\cdot}'\beta-\delta\sigma_b p)^2}{2(\sigma_{bw}^2+\sigma_\epsilon^2/n)}\right\}}. \qquad (2)$$

1.2 Assuming b_i exponential

When the distribution of the random effects is assumed to be exponential, it is also possible to explicitly calculate the BPs. Rewriting the linear mixed model in terms of random variables, a_i with standard exponential distributions ($b_i = \sigma_b[a_i - 1]$), we have

$$\begin{aligned}
Y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} - \sigma_b + \sigma_b a_i + \epsilon_{it} \\
a_i &\sim \text{i.i.d. } \mathcal{E}(1) \text{ independent of} \qquad (3)\\
\epsilon_{it} &\sim \text{i.i.d. } \mathcal{N}(0, \sigma_\epsilon^2).
\end{aligned}$$

We proceed by first calculating the conditional distribution of a_i given $\mathbf{Y}$, which is proportional to the joint distribution of a_i and $\mathbf{Y}$. We suppress the index i and write μ_t for $\mathbf{x}'_{it}\boldsymbol{\beta}$ to simplify the presentation.

$$
\begin{aligned}
f_{a|Y} &\propto \exp\left\{\sum_t \frac{-1}{2\sigma_\epsilon^2}(Y_t - \mu_t + \sigma_b - \sigma_b a)^2 - a\right\} I_{\{a>0\}} \\
&\propto \exp\left\{\frac{-n\sigma_b^2}{2\sigma_\epsilon^2}a^2 - \frac{2\sigma_\epsilon^2}{2\sigma_\epsilon^2}a + \frac{2\sigma_b a}{2\sigma_\epsilon^2}(\Sigma_t Y_t - \Sigma_t \mu_t + n\sigma_b)\right\} I_{\{a>0\}} \\
&\propto \exp\left\{\frac{-n\sigma_b^2}{2\sigma_\epsilon^2}\left(a - \left[\frac{1}{\sigma_b}(\bar{Y} - \bar{\mu} + \sigma_b) - \frac{\sigma_\epsilon^2}{n\sigma_b^2}\right]\right)^2\right\} I_{\{a>0\}}. \qquad (4)
\end{aligned}
$$

This shows that the conditional distribution of a_i given $\mathbf{Y}$ is a truncated Gaussian, truncated above 0. The standardized variable, $a^* = \frac{\sqrt{n}\sigma_b}{\sigma_\epsilon}a$, follows a truncated Gaussian distribution with mean

$$
\Delta = \frac{\sqrt{n}}{\sigma_\epsilon}(\bar{Y} - \bar{\mu} + \sigma_b) - \frac{\sigma_\epsilon}{\sqrt{n}\sigma_b}, \qquad (5)
$$

variance 1 and again truncated above 0. The expected value of a^* given $\mathbf{Y}$ is

$$
\mathrm{E}[a^*|\mathbf{Y}] = \Delta + \frac{\phi(\Delta)}{\Phi(\Delta)}, \qquad (6)
$$

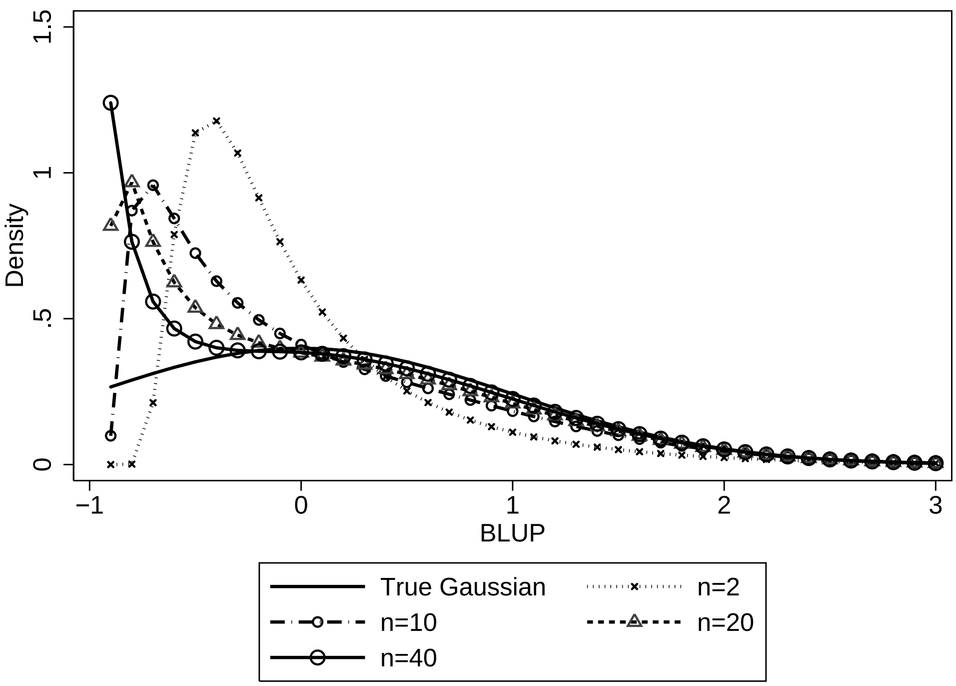
where $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the standard Gaussian p.d.f. and c.d.f. Finally the BP of b_i is given by

$$
\begin{aligned}
\tilde{b}_i &= \mathrm{E}[\sigma_b(a_i - 1)|\mathbf{Y}] \\
&= \mathrm{E}\left[\sigma_b\left(\frac{\sigma_\epsilon}{\sqrt{n}\sigma_b}a_i^* - 1\right)\middle|\,\mathbf{Y}\right] \\
&= \sigma_b\left(\frac{\sigma_\epsilon}{\sqrt{n}\sigma_b}\Delta + \frac{\sigma_\epsilon}{\sqrt{n}\sigma_b}\frac{\phi(\Delta)}{\Phi(\Delta)} - 1\right) \\
&= \bar{Y}_{i\cdot} - \bar{\mathbf{x}}'_{i\cdot}\boldsymbol{\beta} - \frac{\sigma_\epsilon^2}{n_i\sigma_b} + \frac{\phi(\Delta_i)\sigma_\epsilon}{\Phi(\Delta_i)\sqrt{n_i}}. \qquad (7)
\end{aligned}
$$

where, in the last line, we have returned the subscript i, $\bar{\mathbf{x}}_{i\cdot} = \sum_t \bar{\mathbf{x}}_{it}$, and $\Delta_i = \sqrt{n_i}(\bar{Y}_{i\cdot} - \bar{\mathbf{x}}'_{i\cdot}\boldsymbol{\beta} + \sigma_b)/\sigma_\epsilon - \sigma_\epsilon/(\sqrt{n_i}\sigma_b)$.

Using numerical methods, it is straightforward to calculate and plot the distribution of the BPs. Figure 1 shows the density for various cluster sizes, an assumed exponential distribution, but a true Gaussian distribution and again using $\sigma_\epsilon^2 = 3$ and $\sigma_b^2 = 1$.

Figure 1: Plot of best predictor density for various cluster sizes with an assumed exponential but true Gaussian density



1.2.1 MSE of prediction under an assumed exponential distribution

We next evaluate the mean square error of prediction under a model with an assumed exponential distribution for the random effects. Rewriting the best predicted value, (7), using $\bar{Y}_{i\cdot} - \bar{\mathbf{x}}'_{i\cdot}\boldsymbol{\beta} = b_i + \bar{\epsilon}_{i\cdot}$ we obtain

$$\tilde{b}_i = b_i + \bar{\epsilon}_{i\cdot} - \frac{\sigma_\epsilon^2}{n_i\sigma_b} + \frac{\phi\left(\frac{\sqrt{n_i}}{\sigma_\epsilon}(b_i + \bar{\epsilon}_{i\cdot} + \sigma_b) - \frac{\sigma_\epsilon}{\sqrt{n_i}\sigma_b}\right)\sigma_\epsilon}{\Phi\left(\frac{\sqrt{n_i}}{\sigma_\epsilon}(b_i + \bar{\epsilon}_{i\cdot} + \sigma_b) - \frac{\sigma_\epsilon}{\sqrt{n_i}\sigma_b}\right)\sqrt{n_i}}. \tag{8}$$

Using the notation $\overset{D}{=}$ to represent "equal in distribution", the conditional distribution of $\tilde{b}_i - b_i$ given b_i has the following equivalency

$$\tilde{b}_i - b_i | b_i \overset{D}{=} Z\frac{\sigma_\epsilon}{\sqrt{n_i}} - \frac{\sigma_\epsilon^2}{n_i\sigma_b} + \frac{\phi\left(\frac{\sqrt{n_i}}{\sigma_\epsilon}(b_i + \sigma_b) + Z - \frac{\sigma_\epsilon}{\sqrt{n_i}\sigma_b}\right)\sigma_\epsilon}{\Phi\left(\frac{\sqrt{n_i}}{\sigma_\epsilon}(b_i + \sigma_b) + Z - \frac{\sigma_\epsilon}{\sqrt{n_i}\sigma_b}\right)\sqrt{n_i}}, \tag{9}$$

where $Z \sim \mathcal{N}(0,1)$. This construction makes it straightforward to numerically evaluate the mean square error of prediction under different true distributions since it is equal to the double integral of the square of (9) with respect to a standard Gaussian distribution and the true distribution of b_i.

1.3 MSE of prediction under an assumed mixture distribution

The mean square error of prediction under a model with an assumed mixture distribution for the random effects can be calculated as in the previous section. Using (1) we can rewrite $\tilde{b}_i$ as follows

$$\tilde{b}_i \overset{D}{=} \frac{\sigma_{bw}^2}{\sigma_{bw}^2 + \sigma_\epsilon^2/n}(b_i + \sigma_\epsilon Z/\sqrt{n}) + \frac{\sigma_\epsilon^2/n}{\sigma_{bw}^2 + \sigma_\epsilon^2/n}\left(-\pi(1)\sigma_b\delta(1-p) + [1-\pi(1)]\sigma_b\delta p\right) \tag{10}$$

with

$$\pi(1) \overset{D}{=} \frac{p\exp\left\{-\frac{(b_i + \sigma_\epsilon Z/\sqrt{n} + \delta\sigma_b[1-p])^2}{2(\sigma_{bw}^2 + \sigma_\epsilon^2/n)}\right\}}{p\exp\left\{-\frac{(b_i + \sigma_\epsilon Z/\sqrt{n} + \delta\sigma_b[1-p])^2}{2(\sigma_{bw}^2 + \sigma_\epsilon^2/n)}\right\} + (1-p)\exp\left\{-\frac{(b_i + \sigma_\epsilon Z/\sqrt{n} - \delta\sigma_b p)^2}{2(\sigma_{bw}^2 + \sigma_\epsilon^2/n)}\right\}}. \tag{11}$$

Again, this construction makes it relatively easy to numerically evaluate the mean square error of prediction under different true distributions.

Table 1: Percentiles of a Gaussian and standardized Tukey(0.5,0.1) distribution

Percentile	Gaussian(0,1)	Standardized Tukey(0.5,0.1)
.1%	-3.09	-1.89
1%	-2.33	-1.40
2.5%	-1.96	-1.21
5%	-1.64	-1.06
10%	-1.28	-0.89
50%	0	-0.21
90%	1.28	1.09
95%	1.64	1.73
97.5%	1.96	2.47
99%	2.33	3.62
99.9%	3.09	7.68

2 Web Appendix B: Additional simulations

2.1 Random Intercepts

We performed additional simulation studies to evaluate the performance of the BPs under different true and assumed distributions in the more realistic situation in which all the parameters were estimated. We tested two distributions for the random intercepts: a Gaussian distribution and a Tukey(g, h) distribution. The Tukey distribution was chosen to evaluate the performance under a distribution representing an extreme departure from the Gaussian. Depending on the values of g and h the Tukey distribution can be quite skewed and/or heavy-tailed. See He and Raghunathan (2006) for a recent reference. We chose $g = 0.5$ and $h = 0.1$, which gives a mean of 0.31, variance of 2.27, skewness of 3.41 and a kurtosis of 44.24. This was then standardized to have mean zero and variance one. Table 1 gives some percentiles for this standardized distribution, illustrating its extreme right tail.

Using these two random effects distribution, we simulated eight different scenarios. For a continuous outcome, linear mixed model with Gaussian errors and for a binary outcome, logistic regression, we simulated the four combinations of assumed and true distributions (Gaussian and Tukey). The simulations used two covariates: one within cluster (x_w) and one between cluster covariate (x_b). The within cluster covariate was equally spaced between 0 and 1. The between cluster covariate was binary with a 25%/75% division. The parameter values were set as follows: $\beta_0 = -2, \beta_{between} = 1, \beta_{within} = 1, \sigma_b = 1$, and, for continuous outcomes, $\sigma_\epsilon = 1$. The number of clusters, m, was set to 100 and a variety of cluster sizes used ($n = 2, 4, 6, 10, 20$, and 40). The linear predictor for subject i, observation t was given by

$$\beta_0 + b_{0i} + \beta_{between} x_{b,i} + \beta_{within} x_{w,it}. \tag{12}$$

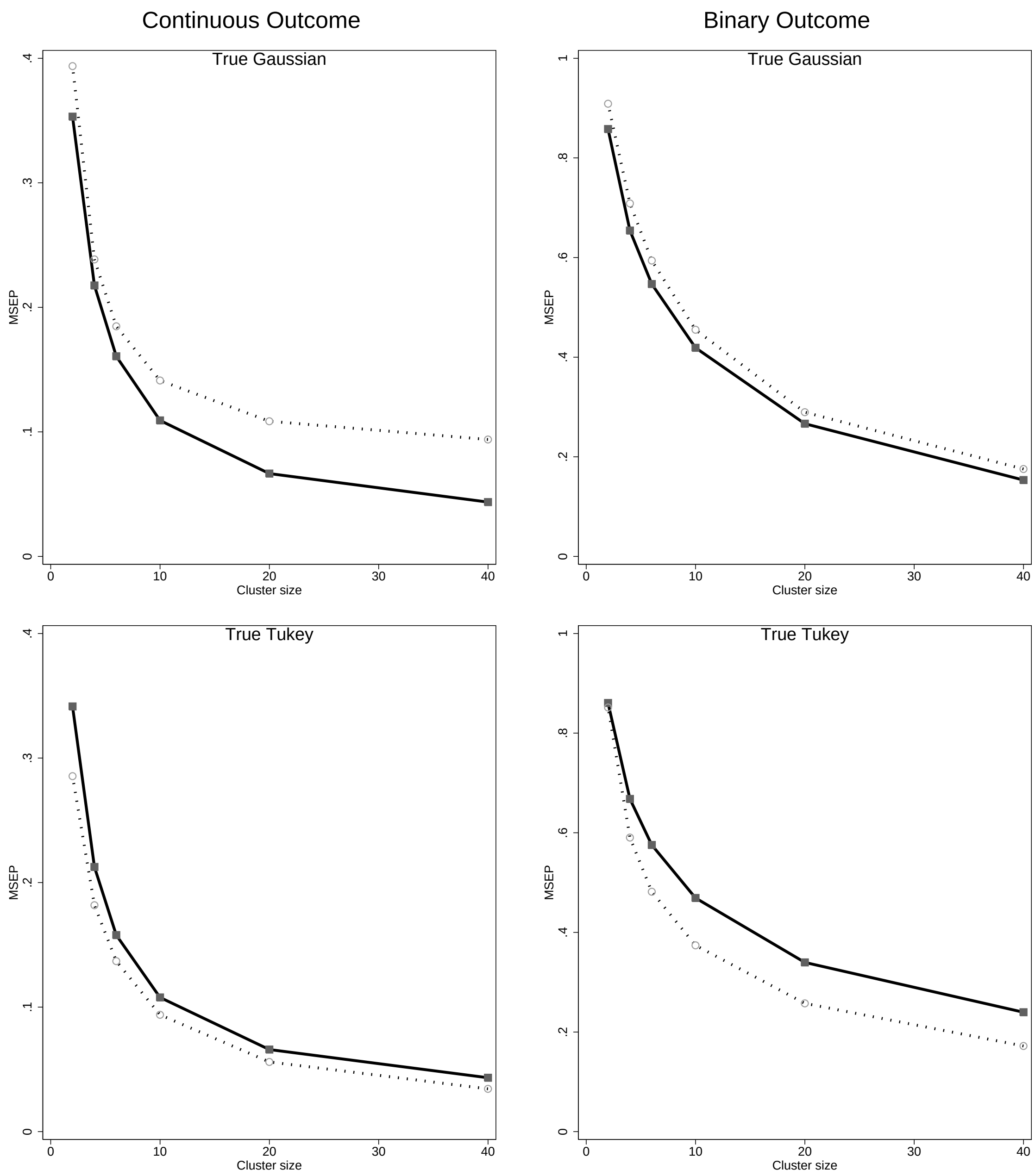
To each simulated data set we fit two GLMMs with either an identity or logistic link. One model assumed that the random effects were standard Gaussian while the other assumed the random effects followed a standardized Tukey(0.5, 0.1) distribution. To assess the effect of misspecification, we did not allow the program to estimate the parameters of the Tukey distribution, but rather fixed them at g=0.5 and h=0.1. Otherwise the comparison confounds the issue of misspecification of distributional shape with estimation of two additional shape parameters.

Figure 2 gives the results; the lefthand panels displays the results for continuous outcomes and the righthand panels binary outcome. The rows correspond to the true distributions (first row true Gaussian and second row true Tukey). Each panel plots the mean square error of prediction versus cluster size for each of the two assumed distributions.

The main message is that the primary determinant of the MSEP is the cluster size. In each case, using the incorrect distribution causes only a slight degradation in the MSEP, especially for smaller cluster sizes (i.e., less than 20). So even in this case, where the two distributions are quite different, only a modest impact is seen in getting the distributional assumptions wrong.

We investigated more closely the reasons for the modest discrepancies (continuous outcome, true Gaussian and binary outcome, true Tukey). In the case of the continuous outcome most of the discrepancy was due to the fact that the mean of the predicted random effects under the assumed Tukey was not zero. Simply normalizing the values to have mean zero generated much more accurate predictions. In the case of the binary outcome a very small percentage of extremely large random effects (less than 0.5%) generated under the Tukey distribution were poorly predicted under the assumed Gaussian distribution. When those were excluded, the prediction error for the Gaussian distribution was almost the same as under the true Tukey distribution.

Figure 2: MSE of Prediction for continuous and binary outcomes under assumed and true Gaussian and Tukey distributions



Assumed distributions: Solid line/square=Gaussian, dotted line/circle=Tukey

Even though there were modest discrepancies in MSEP, the rank correlations between random effects calculated under different assumed distributions was uniformly high. In each of the four cases displayed in Figure 2 the Spearman rank correlation between pairs of predictions was greater than 0.95.

2.2 Random intercepts and slopes

We also simulated data from a random intercepts and slopes model for both normally distributed and exponentially distributed random effects. This was similar to the random intercepts models but with a linear predictor given by

$$\beta_0 + b_{0i} + \beta_{between} x_{b,i} + (\beta_{within} + b_{1i}) x_{w,it}. \tag{13}$$

We used SAS Proc NLMIXED to fit the models. Since NLMIXED is restricted to normally distributed random effects, for the exponentially distributed random effects we rewrote each random term as a function of a normal distribution:

$$b_{ki} = -\log(1 - \Phi(Z_k)) - 1, \tag{14}$$

where Z_k were distributed with means 0, variances 1 and correlation ρ. As with the random intercepts we simulated data sets with 100 clusters of size $n = 2, 4, 6, 10, 20,$ and 40 and used the same parameter values as above except that $\text{var}(b_{ki}) = 0.5 = \rho$.

Figure 3 reports the results for predicting both the random slopes and intercepts. As before, there is a modest efficiency loss due to having the incorrect distribution for estimating the intercepts. For estimating the slopes the results are much more comparable, with little efficiency loss.

We also simulated a more challenging situation with $\text{var}(b_{ki}) = 2$ and $\rho = 0.5$. Table 2 gives the results for estimating the random intercepts for continuous (linear link) and binary outcomes (logistic link) when the true distribution is exponential. We see again the similar pattern: under the incorrect Gaussian assumption there is a modest loss of efficiency. This is somewhat worse for the binary outcome and worsens with increasing cluster size.

Figure 3: MSE of prediction for continuous outcomes under assumed and true Gaussian and exponential distributions using a random intercepts and slopes model

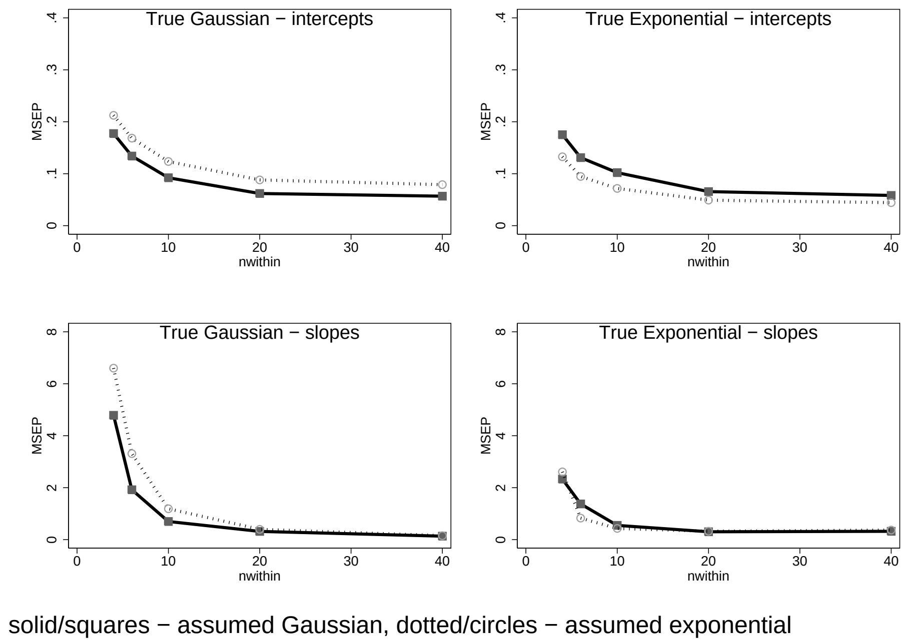
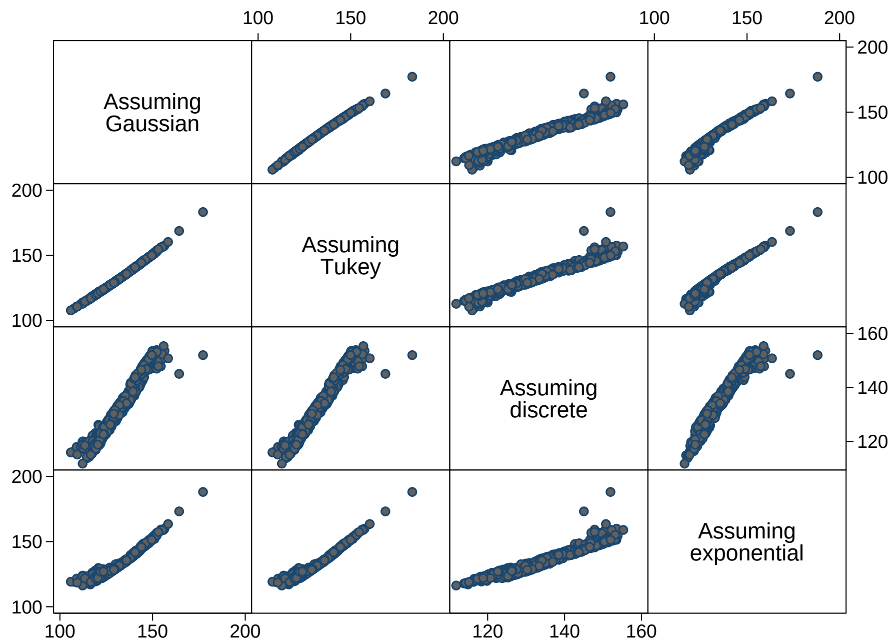


Table 2: Mean square error of prediction of random intercepts under a true exponential distribution and assumed Gaussian or exponential distributions mixed effects model with random intercepts and slopes

	Continuous		Binary	
Cluster size	Gaussian	Exponential	Gaussian	Exponential
4	0.26	0.22	1.06	0.98
6	0.19	0.17	0.86	0.82
10	0.14	0.14	0.67	0.62
20	0.13	0.11	0.50	0.45
40	0.11	0.09	0.43	0.32

Figure 4: Matrix scatter plot of predictions from the four random intercept models for the HERS example



3 Web Appendix C - Plot of best predicted values for the HERS example

Figure 4 gives a plot of the predicted values derived from the four different random effects distributions (Gaussian, Tukey, discrete and exponential) fitted to the random intercepts model. The Tukey and Gaussian predictions are very similar, indicative of the fact that the estimated Tukey distribution was fairly close to Gaussian. There is good agreement between all the distributions, with the main differences at the extremes. The predicted values under the exponential assumption reflect the truncation in the left tail.

4 Web Appendix D - Numerical details for main manuscript

Integration to generate Figures 1 through 5 in Section 3 was performed using Maple Version 12 (Maplesoft: Waterloo, Ontario, Canada). Figure 5 was double checked by Monte Carlo integration and simulation using Matlab Version 6.5 (Mathworks: Natick MA). The simulations in Section 5 were conducted using SAS Version 9.1 (SAS Institute: Cary NC) with the models fit using NLMIXED. The fits to the example in Section 6 were performed using Proc MIXED and NLMIXED in SAS Version 9.1, excepting the discrete distribution fit which used the GLLAMM module (www.gllamm.org) in Stata Version 11.0 (Statcorp: College Station TX).

NLMIXED estimates random effects using posterior modes rather than the actual best predicted value, defined as a conditional or "posterior" mean of the random effect given the data. These agree for a linear, Gaussian, mixed model but may not otherwise. We conducted limited simulation studies to compare the performance of posterior modes versus means using NLMIXED and Stata's GLLAMM macro (which does estimate posterior means) under binary outcome, logit link models with random effects variances of 1 and 4 and a true Gaussian distribution. These studies indicated that predictions based on posterior modes closely corresponded to those based on posterior means. Correlations of calculated MSEP for posterior means and modes were about 0.95 across replicate data sets. Absolute values of the MSEP were slightly higher (about 10%) for the posterior modes (not surprising since the posterior mean is optimized for MSEP). We did not explore the impact of using posterior modes under other true random effects distributions.

References

He, Y. and Raghunathan, T. (2006). Tukey's gh distribution for multiple imputation. *The American Statistician* **60,** 251–256.

Verbecke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91,** 217–221.