Notes for 751-752
Sections 22

Martin Lindquist*

December 7, 2017

*Thanks to Brian Caffo and Ingo Ruczinski for supplying their notes, upon which much of this document is based.

22 Mixed Models

It is often the case that parameters of interest in linear models are naturally thought of as being random rather than fixed. The rational for this can come about for many reasons. The first occurs when the natural asymptotics have the number of parameters tending to infinity with the sample size. As an example, consider the `Rail` dataset in `nlme`. The measurements are echo times for sound traveling along railroad lines (a measure of health of the line). Multiple (3) measurements are collected for each rail. Consider a model of the form

$$y_{ij} = \mu + U_i + \epsilon_{ij},$$

where i is rail and j is measurement within rail. Treating the u_i as fixed effects results in a circumstance where the number of parameters goes to infinity with the rails, which can lead to inconsistent parameter estimates.

A solution to this problem is to put a distribution on the U_i, say $U_i \sim_{iid} N(0, \sigma_u^2)$. This is highly related to ridge regression (from the penalization chapter). However, unlike penalization, this problem allows for thinking about the random effect distribution as a population distribution (the population of rails in our example).

Another way to think about random effects is to consider a fixed effect treatment of the U_i terms. Since we included an intercept, we would need to add one linear constraint on the U_i for identifiability. Consider the constraint, $\sum_{i=1}^n U_i = 0$. Then, μ would be interpreted as and overal mean and the U_i terms would be interpreted as the rail-specific deviation around that mean. The random effect model simply specifies that the U_i are iid $N(0, \sigma_u^2)$ and mutually independent from ϵ_{ij}. The mean of the distribution on the U_i has to be 0 (or fixed at a number), since it would not be identified from μ otherwise.

A perhaps preferable way to specify the model hierarchically, $y_{ij} \mid U_i \sim N(\mu, \sigma^2)$ and $U_i \mid \ \sim N(0, \sigma_U^2)$. Consider the implications of this model. First, note that

$$\begin{aligned}
\text{Cov}(y_{ij}, y_{i'j'}) &= \text{Cov}(U_i + \epsilon_{ij}, U_{i'} + \epsilon_{i'j'}) && (1)\\
&= \text{Cov}(U_i, U_{i'}) + \text{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) && (2)\\
&= \begin{cases} \sigma^2 + \sigma_U^2 & \text{if } i = i'1 \text{ and } j = j' \\ \sigma^2 & \text{if } i = i' \text{ and } j \neq j' \\ 0 & \text{Otherwise} \end{cases} && (3)
\end{aligned}$$

And thus the correlation between observations in the same cluster is $\sigma_u^2/(\sigma_u^2 + \sigma^2)$. This is the ratio between the between subject variability, σ_u^2, and the total variability, $\sigma_u^2 + \sigma^2$.

Notice that the marginal model for $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ is normally distributed with mean $\mu \times \mathbf{J}_{n_i}$ and variance $\sigma^2 \mathbf{I}_{n_i} + \mathbf{J}_{n_i} \mathbf{J}'_{n_i} \sigma_u^2$. It is by maximizing this (marginal) likelihood that we obtain the ML estimates for $\mu, \sigma^2, \sigma_U^2$.

We can predict the U_i by considering the estimate $E[U_i \mid \mathbf{Y}]$. To derive this, note that the density for $U_i \mid \mathbf{Y}$ is equal to the density of $U_i \mid \mathbf{y}_i$, since U_i is independent of every $y_{i'j}$ for $i \neq i'$. Then further note that the

density for $U_i \mid \mathbf{y}_i$ is propotional to the joint density of y_i, U_i, which is equal to the density of $y_i \mid U_i$ times the density for U_i. Omitting anything that is not proportional in U_i, and taking twice the natural logarithm of the the densities, we obtain:

$$||\mathbf{y}_i - \mu \mathbf{J}_{n_i} - U_i||^2/\sigma^2 + U_i/\sigma_U^2.$$

Expanding the square, and discarding terms that are constant in U_i, we obtain that U_i is normally distributed with mean

$$\frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma^2}{n}}(\bar{y}_i - \mu).$$

Thus, if $\hat{\mu} = \bar{Y}$, our estimate of U_i is the estimate that we would typically use shrunken toward zero. The idea of shrinking estimates when simultaneously estimating several quantities is generally a good one. This has similarities with James/Stein estimation.

Shrinkage estimation works by trading bias for lower variance. In our example, the shrinkage factor is $\sigma_u^2/(\sigma_u^2 + \sigma^2/n)$. Thus, the better estimated the mean for that group is (σ^2/n is small), or the more variable the group is (σ_u^2 is large), the less shrinkage we have. On the other hand, the fewer observations that we have, the larger the residual variation or the smaller the inter-subject variation, the more shrinkage we have. In this way the estimation is optimally calibrated to weigh the contribution of the individual versus the contribution of the group to the estimate regarding this specific individual.

22.1 General case

The general linear mixed model can be written as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{U}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots m.$$

Here: $\mathbf{y}_i$ is an $n_i \times 1$ vector of observations; $\mathbf{X}_i$ is an $n_i \times p$ design matrix for the fixed effects; $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects; $\mathbf{Z}_i$ is an $n_i \times k$ design matrix for the random effects; $\mathbf{U}_i$ is a $k \times 1$ vector of random effects; $\boldsymbol{\varepsilon}_j$ is an $n_i \times 1$ vector of error terms. In addition, we assume that the error terms and random effects are uncorrelated across groups, i.e., when $i \neq i'$:

$$\operatorname{cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'}) = \mathbf{0}, \quad \operatorname{cov}(\mathbf{U}_i, \mathbf{U}_{i'}) = \mathbf{0}, \quad \operatorname{cov}(\boldsymbol{\varepsilon}_i, \mathbf{U}_{i'}) = \mathbf{0}.$$

When $i = i'$, we have that

$$\operatorname{cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i) = \Sigma_\epsilon, \quad \operatorname{cov}(\mathbf{U}_i, \mathbf{U}_i) = \Sigma_u, \quad \operatorname{cov}(\boldsymbol{\varepsilon}_i, \mathbf{U}_i) = \mathbf{0}.$$

We can reformulate the model as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{pmatrix} \quad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{Z}_m \end{pmatrix}$$

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_m \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{pmatrix}.$$

We can write the model as follows: $\mathbf{y} \mid \mathbf{U} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U}, \sigma^2\mathbf{I})$ and $\mathbf{U} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_U)$. This is marginally equivalent to specifying

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\epsilon}.$$

Here, the marginal likelihood for $\mathbf{y}$ is normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance $\mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}' + \sigma^2\mathbf{I}$. Maximum likelihood estimates maximize the marginal likelihood via direct numerical maximization or the EM algorithm. Notice, for fixed variance components, the estimate of $\boldsymbol{\beta}$ is a weighted least squares estimate.

It should be noted the distinction between a mixed effect model and simply specifying a marginal variance structure. The same marginal likelihood could be obtained via the model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}' + \sigma^2\mathbf{I})$. However, some differences tend to arise. Often, the natural specification of a marginal variance structure doesn't impose positivity constraints that random effects do. For example, in the previous section, we saw that the covariance between measurements in the same cluster was $\sigma_u^2/(\sigma_u^2 + \sigma^2)$, which is guaranteed to be positive. However, if fitting a general marginal covariance structure, one would typically simply parameterize the covariance structure as either positive or negative.

Another difference lies in the hierarchical model itself. We can actually estimate the random effects if we specify them, unlike marginal models. This is a key (perhaps "the" key) defining attribute of mixed models. The Best Linear Unbiased Predictor (BLUP) is given by

$$E[\mathbf{U} \mid \mathbf{Y}]$$

As a homework exercise, derive the general form of the BLUPs.

22.2 Estimation

Next, we focus on estimation of model parameters. For a given $\mathbf{V}$ we can estimate $\boldsymbol{\beta}$ using:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

However, in practice, $\mathbf{V}$ is unknown and needs to be estimated.

Consider re-expressing the model as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{(1)}\mathbf{u}_1 + \cdots + \mathbf{Z}_{(k)}\mathbf{u}_k + \boldsymbol{\varepsilon}$$

Here we assume each $\mathbf{Z}_{(h)}$ is an $n \times r_h$ matrix that specifies membership in the various clusters or subgroups, and the $\mathbf{u}_h$ are different random effects. Let $E(\mathbf{u}) = E(\boldsymbol{\varepsilon}) = \mathbf{0}$, and let us assume

$$\text{var}(\boldsymbol{\varepsilon}) = \sigma_\epsilon^2 \mathbf{I}_N$$

and

$$\text{var}(\mathbf{u}_l) = \sigma_l^2 \mathbf{I}_{r_h}.$$

Here r_h represents the number of elements in $\mathbf{u}_h$. In addition, $\text{cov}(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{0}\ \forall i \neq j$, and $\text{cov}(\mathbf{u}, \boldsymbol{\varepsilon}) = \mathbf{0}$.

We summarize the model as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{h=1}^{k} \mathbf{Z}_{(h)}\mathbf{u}_h + \boldsymbol{\varepsilon}$$

where

$$\mathbf{V} \equiv \text{var}(\mathbf{y}) = \sum_{h=1}^{k} \sigma_h^2 \mathbf{Z}_{(h)}\mathbf{Z}'_{(h)} + \sigma_\epsilon^2 \mathbf{I}_N.$$

A useful extension is to make the following definition:

$$\mathbf{u}_0 \equiv \boldsymbol{\varepsilon} \qquad \mathbf{Z}_{(0)} \equiv \mathbf{I}_N \qquad \sigma_0^2 \equiv \sigma_\epsilon^2$$

We can now write the model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{h=0}^{k} \mathbf{Z}_{(h)}\mathbf{u}_h$$

with

$$\mathbf{V} = \sum_{h=0}^{k} \sigma_h^2 \mathbf{Z}_{(h)}\mathbf{Z}'_{(h)}.$$

22.2.1 Maximum Likelihood

After specifying the appropriate model, the next task is to estimate the fixed effects and variance components. The most common approaches towards estimating the parameters in the covariance matrices is to use maximum likelihood (ML) or restricted maximum likelihood (REML). The MLE of $\mathbf{V}$ is based on using the marginal model

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

The log-likelihood of $\mathbf{y}$ is given by:

$$\ell(\boldsymbol{\beta}, \mathbf{V}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

By first optimizing over $\boldsymbol{\beta}$ for fixed $\mathbf{V}$ we obtain the familiar estimate:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

Next, by plugging this value into $\ell(\boldsymbol{\beta}, \mathbf{V})$ we obtain the profile log-likelihood for $\mathbf{V}$:

$$\begin{aligned}\ell_P(\mathbf{V}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &\propto -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}).\end{aligned}$$

The MLE for $\mathbf{V}$ can be found by maximizing this function.

22.2.2 Restricted Maximum Likelihood

Note the MLE estimator for linear mixed effect model is biased and the error on the random effects covariance may be large. In contrast, REML estimates tend to be less biased than the ML estimates. For example, if $y_i \sim_{iid} N(\mu, \sigma^2)$, REML yields the unbiased variance estimate (divided by $n-1$) rather than the biased variance estimate obtained via ML. REML estimates are often the default for linear mixed effect model programs.

The key idea of REML is to perform maximum likelihood estimation for $\mathbf{Ky}$ rather than $\mathbf{y}$, where $\mathbf{K}$ is chosen so that the distribution of $\mathbf{Ky}$ involves only the variance components, not $\boldsymbol{\beta}$. For this to occur, we require $\mathbf{K}$ to be a full-rank matrix such that $\mathbf{KX} = \mathbf{0}$. In this setting, $E(\mathbf{Ky}) = \mathbf{KX}\boldsymbol{\beta} = \mathbf{0}$. It turns out that $\mathbf{K}$ must be of the form

$$\begin{aligned}\mathbf{K} &= \mathbf{C}(\mathbf{I} - \mathbf{H}) \\ &= \mathbf{C}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\end{aligned}$$

where $\mathbf{C}$ specifies a full-rank transformation of the rows of the projection matrix $\mathbf{I} - \mathbf{H}$. There are an infinite number of such $\mathbf{K}$'s, and it does not matter which is used.

Consider the marginal model where

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

with

$$\mathbf{V} = \sum_{h=0}^{k} \sigma_h^2 \mathbf{Z}_{(h)}\mathbf{Z}'_{(h)}.$$

Let $\mathbf{K}$ be specified as above. Then,

$$\mathbf{Ky} \sim N\Big(\mathbf{0}, \mathbf{K}\Big(\sum_{h=0}^{k} \sigma_h^2 \mathbf{Z}_{(h)}\mathbf{Z}'_{(h)}\Big)\mathbf{K}'\Big).$$

Thus the distribution of the transformed data $\mathbf{Ky}$ involves only the $k+1$ variance components as unknown parameters. In order to estimate the variance components, the next step in REML is to maximize the likelihood of $\mathbf{Ky}$ with respect to these variance components.

We now develop a set of estimating equations by taking partial derivatives of the log likelihood with respect to the variance components, and setting them equal to zero. The log-likelihood can be expressed as follows:

$$\begin{aligned}
\ell(\sigma_0^2, \quad \ldots \quad , \sigma_k^2) & \\
&= -\frac{n-r}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{KVK}'| - \frac{1}{2}\mathbf{y}'\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{Ky} \\
&= -\frac{n-r}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}\Big(\sum_{h=0}^{k}\sigma_h^2\mathbf{Z}_{(h)}\mathbf{Z}'_{(h)}\Big)\mathbf{K}'| \\
&\quad -\frac{1}{2}\mathbf{y}'\mathbf{K}'\Big[\mathbf{K}\Big(\sum_{h=0}^{k}\sigma_h^2\mathbf{Z}_{(h)}\mathbf{Z}'_{(h)}\Big)\mathbf{K}'\Big]^{-1}\mathbf{Ky}
\end{aligned}$$

Using these results, we take the partial derivative of $\ell(\sigma_0^2, \ldots, \sigma_k^2)$ with respect to each of the σ_h^2:

$$\begin{aligned}
&\frac{\partial}{\partial\sigma_h^2}\ell(\sigma_0^2, \ldots, \sigma_k^2) \\
&= -\frac{1}{2}\mathrm{tr}\Big((\mathbf{KVK}')^{-1}\Big[\frac{\partial}{\partial\sigma_h^2}(\mathbf{KVK}')\Big]\Big) \\
&\quad +\frac{1}{2}\mathbf{y}'\mathbf{K}'(\mathbf{KVK}')^{-1}\Big[\frac{\partial}{\partial\sigma_h^2}(\mathbf{KVK}')\Big](\mathbf{KVK}')^{-1}\mathbf{Ky} \\
&= -\frac{1}{2}\mathrm{tr}\Big((\mathbf{KVK}')^{-1}\mathbf{KZ}_{(h)}\mathbf{Z}'_{(h)}\mathbf{K}')\Big) \\
&\quad +\frac{1}{2}\mathbf{y}'\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{KZ}_{(h)}\mathbf{Z}'_{(h)}\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{Ky}
\end{aligned}$$

Setting this result equal to zero, we obtain a set of $k+1$ estimating equations for $\sigma_0^2, \ldots \sigma_k^2$ given by

$$\begin{aligned}
&\mathrm{tr}\Big((\mathbf{KVK}')^{-1}\mathbf{KZ}_{(h)}\mathbf{Z}'_{(h)}\mathbf{K}'\Big) \\
&= \mathbf{y}'\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{KZ}_{(h)}\mathbf{Z}'_{(h)}\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{Ky}
\end{aligned}$$

Note, the expected value of the quadratic form on the right side is given by the left side of the equation. In certain special cases these equations can be simplified to yield closed-form estimating equations. However, in most cases, numerical methods are required to solve the equations.

An alternative way to derive the REML estimates is using a Bayesian formulation. Consider a model where

$$\mathbf{y} \mid \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\Sigma_u\mathbf{Z}' + \sigma^2\mathbf{I})$$

and $\boldsymbol{\beta} \sim N(0, \theta\mathbf{I})$. Calculating the mode for Σ_u and σ^2 after integrating out $\boldsymbol{\beta}$ as $\theta \to \infty$ results in the REML estimates. While this is not terribly useful for general linear mixed effect modeling, it helps us think about REML as it relates to Bayesian analysis and it allows us to extend REML in settings where residuals are less well defined, like generalized linear mixed models.

22.2.3 Inference for β

Estimates of the variance components can be inserted into $\mathbf{V}$ to obtain:

$$\hat{\mathbf{V}} = \sum_{h=0}^{k} \hat{\sigma}_h^2 \mathbf{Z}_{(h)}\mathbf{Z}'_{(h)}.$$

We can use this to estimate:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}.$$

This is sometimes called the estimated generalized least-squares. An approximate estimate of the variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is

$$\text{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}.$$

Note this ignores the variability due to the estimation of $\mathbf{V}$. For large samples this extra variability is negligible, but it can be substantial for smaller samples. If the sample size is small it can be preferable to use the parametric bootstrap methods to approximate the distribution of the test statistics.

22.2.4 Likelihood Ratio Tests

We can compare two nested models m_0 and m_1 using the likelihood ratio test statistic:

$$-2\log(LR(\mathbf{y})) = -2(\ell(\hat{\boldsymbol{\beta}}_0, \hat{\mathbf{V}}_0|\mathbf{y}) - \ell(\hat{\boldsymbol{\beta}}_1, \hat{\mathbf{V}}_1|\mathbf{y})),$$

where $\hat{\boldsymbol{\beta}}_0$, $\hat{\mathbf{V}}_0$ and $\hat{\boldsymbol{\beta}}_1$, $\hat{\mathbf{V}}_1$ are the estimates for the parameters under the two models. This test statistics is asymptotically distributed as a χ^2 with degrees of freedom equal to the difference in number of parameters between the two models. If the models differ in their fixed effects, it is not possible to use REML estimates in the likelihood ratio statistics. REML estimates the random effects by considering linear combinations of the data that remove the fixed effects and therefore the two likelihood functions are not comparable. Note

this asymptotic result is based on some technical assumptions that are not always satisfied in practice. In particular, the parameters under the null model cannot lie on the boundary of the parameter space. This is a problem for testing variance components, which are constrained to be positive (or positive definite), when the null hypothesis is they are equal to zero.

22.2.5 Testing Variance Components

Consider the model:

$$y_{ij} = \beta_0 + u_i + \beta_1 x_{ij} + \epsilon_{ij}.$$

where $u_i \sim_{iid} N(0, \sigma_u^2)$ and $\epsilon_{ij} \sim_{iid} N(0, \sigma_\epsilon^2)$. Suppose we want to test whether the intercepts of the individuals i are significantly different from one another. This is equivalent to testing

$$H_0 : \sigma_u^2 = 0 \quad versus \quad H_a : \sigma_u^2 > 0.$$

Under certain independence assumptions, the asymptotic distribution when H_0 is true implies that there is a 50% chance that $\hat{\sigma}_u^2 = 0$. This leads to the following approximate result:

$$-2\log(LR(\mathbf{y})) \sim 0.5\chi_0^2 + 0.5\chi_1^2$$

where χ_0^2 is a point mass at 0. This results assumes that $\mathbf{y}$ can be partitioned into subvectors that are independent. This assumption does not necessarily hold for all mixed models. The asymptotic distribution for general mixed models is more difficult.

22.3 Prediction

Consider generally trying to predict U from observed data Y. Let f_{uy}, f_u, f_y, $f_{u|y}$ and $f_{y|u}$ be the joint, marginal and conditional densities respectively. Let $\theta(Y)$ be our estimator of U. Consider evaluating the prediction error via the expected squared loss

$$E[(U - \theta(Y))^2]$$

We now show that this is minimized at $\theta(Y) = E[U \mid Y]$. Note that

$$\begin{aligned}
&E[(u - \theta(y))^2] \\
&\quad = E[(u - E[u|y] + E[u|y] - \theta(y))^2] \\
&\quad = E[(u - E[u|y])^2] - 2E[(u - E[u|y])(E[u|y] - \theta(y))] \\
&\qquad + E[(E[u|y] - \theta(y))^2] \\
&\quad = E[(u - E[u|y])^2] + E[(E[u|y] - \theta(y))^2] \\
&\quad \geq E[(u - E[u|y])^2].
\end{aligned}$$

Here note that $E(E[(u - E[u|y])(E[u|y] - \theta(y))]|y) = 0$.

This argument should seem familiar. (In fact, Hilbert space results generalize these kinds of arguments into one theorem.) Therefore, $E[U \mid Y]$ is the best predictor. Note, that it is always the best predictor, regardless of the settting. Furthermore, in the context of linear models, this predictor is both linear (in $\mathbf{Y}$) and unbiased. We mean unbiased in the sense of:

$$E[U - E[U \mid Y]] = 0.$$

Therefore, even in the more restricted class of linear estimators, in the case of mixed models, $E[U \mid Y]$ remains best.

A complication arises in that we do not know the variance components. As that is the case, we must plug in the estimates (either REML or ML). The BLUPs lose their optimality properties then and are thus often called EBLUPs (for empirical BLUPs).

Prediction of this sort relates to so-called empirical Bayesian prediction and shrinkage estimation. In your more advanced classes on decision theory, you'll learn about loss functions and uniform desirability of shrinkage estimators over the straightforward estimators. (In our case the straightforward estimator is the one that treats the random effects as if fixed.) This line of thinking yields yet another use for random effect models, where we might apply them merely for the benefits of shrinkage, but don't actually think of our random effects as if random. Consider settings like genomics. The genes being studied are exactly the quantities of interest, not a random sample from a population of genes. However, it remains useful to treat effects associated with genes as if random to obtain the benefits of shrinkage.

22.4 Regression splines

The application to splines has been a very successful, relatively new, use of mixed models. To discuss the methodology, we need to introduce splines briefly. We will only overview this area and focus on regression splines, while acknowledging that other spline bases may be preferable.

Let $(a)_+ = a$ if $a > 0$ and 0 otherwise. Let ξ be a known knot location. Now consider the model:

$$E[y_i] = \beta_0 + \beta_1 x_i + \gamma_1 (x_i - \xi)_+.$$

For x_i values less than or equal to ξ, we have

$$E[y_i] = \beta_0 + \beta_1 x_i$$

and for x_i values above ξ we have

$$E[y_i] = (\beta_0 + \gamma_1 xi) + (\beta_1 + \gamma_1) x_i.$$

Thus, the response function $f(x) = \beta_0 = \beta_1 x + \gamma_1 (x - \xi)_+$ is continuous at ξ and is a line before and after. This allows us to create "hockey stick" models, with a line below ξ and a line with a different slope

afterwards. Furthermore, we could expand the model as

$$E[y_i] = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} \gamma_k (x_i - \xi_k)_+.$$

where ξ_k are knot points. Not the model is a spiky, but flexible, function that is linear between the knots and meets at the knots.

To make the fit less spiky, we want continuous differentiability at the knot points. First note that the function $(x)_+^p$ has $p-1$ continuous derivatives at 0. To see this, take the limit to zero of the derivatives from the right and the left. Thus, the function

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{k=1}^{K} \gamma_k (x_i - \xi_k)_+^2$$

will consist of parabolas between the knot points and will have one continuous derivative at the knot points. This will fit a smooth function that can accommodate a wide variety of data shapes.

22.5 Coding example

```
> library(SemiPar)
> library(lme4)
> library(ggplot2)
> data(pig.weights)

> ggplot(pig.weights, aes(x = num.weeks, y = weight, group = id.num)) +
  geom_point() + geom_path() +
  labs(x = "Week", y = "Weight")

> pig.mixed = lmer(weight ~ (1 | id.num) + num.weeks, data = pig.weights)
> summary(pig.mixed)

> pig.weights$pig.mixed.fit = fitted(pig.mixed)
> ggplot(pig.weights, aes(x = num.weeks, y = weight, group = id.num)) +
  geom_point() + geom_path(alpha = .2) +
  labs(x = "Week", y = "Weight") +
  geom_line(aes(y = pig.mixed.fit), color = "blue", alpha = .5)
```