SB1a Applied Statistics
Lectures 15-16

Dr Geoff Nicholls

Week 8 MT14
- Hierarchical models - concepts

- Example and modeling considerations (intercept)

- Fitting and Model selection

- Example and modeling considerations (slope)

- Goodness of Fit

- Conclusions
Hierarchical Models

Generalize GLM’s (and NLM’s). We consider here the extension of NLM’s.

A First generalization:

Suppose data come in $J$ groups of $n_j, j = 1, 2, ..., J$ observations. The offset for the $j^{th}$ group is $b_j$. Let $y_{ij}$ be the response for the $i^{th}$ observation in the $j^{th}$ group.

Suppose we have one covariate $x_{ij}$. The simple NLM is

$$ y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij} $$

$$ \epsilon_{ij} \sim N(0, \sigma_y^2) \text{ jointly independent.} $$
Our first generalization allows the intercept to vary randomly between groups

\[ y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij} \]

\[ \epsilon_{ij} \sim N(0, \sigma_y^2) \]

\[ \alpha_j \sim N(\alpha, \sigma_\alpha^2) \]

We assume the \( \alpha_j \) and \( \epsilon_{ij}, j = 1..J, i = 1..n_j \) are jointly independent.

*Example*: Gelman and Hill give radon measurements for US homes. Following that text I have extracted the \( n = 919 \) measurements from Minnesota and taken logs. The measurements come in groups by county. There are \( J = 85 \) counties. Denote by

\[ y_{ij}, i = 1..n_j, j = 1..J \]

the response in the \( i'th \) home in the \( j'th \) county.
Measurements $y_{ij}$ can be taken in the basement or on the ground floor. Covariate $x_{ij} \in \{0, 1\}$ records where the measurement was taken.

What is the baseline (log) radon level in each county?

The background radon levels vary from county to county, depending on local geology. It is natural to suppose that this background determines a baseline level $\alpha_j$ and each house varies around that local mean. Baseline levels are not unrelated - they are assumed to vary around the Minnesota-mean level $\alpha$.

This leads us to the Hierarchical model given above, with a random effect on the intercept.

[There is another covariate, $u_j$ giving the (log of the) soil uranium yield in county $j$. We will ignore this for the moment.]
Individuals in group $j$

Population

$J$ Groups

$85$ Counties

State

$n_j$ houses in county $j$

$\alpha$

$\alpha_j$

$y_{ij}$

$n_j$ Individuals in group $j$
Alternative notations

There are other ways to write this exact same model. For example, we can stack the group data in a big column vector $y = (y_1, .. y_n)$ with $n = \sum_j n_j$ entries, mapping $(i, j)$ to $k$ using $k = (j-1)n_j + i$ and create a categorical group variable $j_k \in \{1, 2, ..., J\}$. In this indexing

$$
\begin{align*}
y_k &= \alpha_{jk} + \beta x_k + \epsilon_k \quad k = 1, .., n \\
\epsilon_k &\sim N(0, \sigma_y^2) \quad k = 1, ..., n \\
\alpha_j &\sim N(\alpha, \sigma_\alpha^2), j = 1, ..., J
\end{align*}
$$

Apart from the last line, this is just the same as treating the group variable as a categorical variable with $J$ levels, and using indicator variables. After all,

$$
\alpha_{jk} = \sum_j \alpha_j \mathbb{I}_{j_k = j}
$$

The novelty here is that the effects $\alpha_j$ are treated as random variables: they are random effects.
R notation works with zero-mean offsets, setting

\[ \alpha_{jk} = \alpha + b_{jk} \]

\[ y_k = \alpha + b_{jk} + \beta x_k + \epsilon_k \]

\[ b_j \sim N(0, \sigma_\alpha^2) \]
We can actually write the model without any random effects, actually integrating them out. In the $k$-indexing the model above has covariances

$$\text{cov}(y_k, y_{k'}) = \begin{cases} 
\sigma^2_y + \sigma^2_\alpha & \text{if } k = k' \\
\sigma^2_\alpha & j_k = j_{k'} \text{ - same groups} \\
0 & j_k \neq j_{k'} \text{ - different groups}
\end{cases}$$

Let $\Sigma$ be $n \times n$ with $\Sigma_{k,k'} = \text{cov}(y_k, y_{k'})$ as above. Let

$$\eta_k = \alpha + \beta x_k, \ k = 1, ..., n.$$  

Since everything is normal, our NLM is

$$y \sim N(\eta, \Sigma).$$

We are not assuming we know $\sigma^2_y$ or $\sigma^2_\alpha$. This form makes it clearer that although we introduced $n$ random effects, the predictive model for $y$ still has just $p = 2$ explanatory variables and two unknown variances.
If we ignore the group structure we ignore this correlation in the errors. Correlation reduces the effective sample size, yielding falsely small estimates of variance. This leads to falsely inflated significance levels (falsely small $p$-values).
In order to make transparent the consequence of taking random effects, consider two extreme, simple, models - one with a single common mean, and one with a separate mean for each group. Let $\epsilon_k \sim N(0, \sigma_y^2) \; k = 1, \ldots, n$ throughout.

Pooled: $y_k = \alpha + \epsilon_k$; MLE $\hat{\alpha} = \bar{y}$.

Un-pooled: $y_k = \alpha_j I_{j_k=j} + \epsilon_k$; MLE $\hat{\alpha}_j = \bar{y}_j$.

The corresponding HM is

$y_k = \alpha_j \mathbb{I}_{j_k=j} + \epsilon_k \; k = 1, \ldots, n$

$\alpha_j \sim N(\alpha, \sigma_{\alpha}^2), \; j = 1, \ldots, J$
Exercise: Write down the likelihood for $\alpha_j$, $j = 1, ..., J$ in the HM and show (by completing the square) that the MLE’s are

$$\hat{\alpha}_j = \frac{n_j \bar{y}_j}{\sigma_y^2} + \frac{\bar{y}}{\sigma_\alpha^2}$$

$$= \bar{y}_j - \frac{\bar{y}_j - \bar{y}}{1 + \frac{n_j \sigma_\alpha^2}{\sigma_y^2}}$$

We see that the MLE’s vary between the pooled and un-pooled estimates:

pooled $0 \longrightarrow \sigma_\alpha^2 \longrightarrow \infty$ un-pooled.

$$\bar{y} \longrightarrow \hat{\alpha}_j \longrightarrow \bar{y}_j$$

This is related to the idea of shrinkage. By assuming a common underlying distribution for the $\alpha_j$’s, we shrink them towards the pooled mean. How much we shrink depends on the ratio $n_j \sigma_\alpha^2 / \sigma_y^2$. 
Example (cont.)

```r
> rad.lme1<-lme(y~x,random=~1|county)
> summary(rad.lme1)
...
Random effects:
Formula: ~1 | county
     (Intercept) Residual
StdDev:  0.3282224  0.7555892

Fixed effects: y ~ x
              Value Std.Err DF  t-value p-value
  (Intercept) 1.461  0.0516 833  28.34       0
  x           -0.693  0.0704 833  -9.84       0
...
Number of Observations: 919
Number of Groups: 85
```
> head(random.effects(rad.lme1))
  (Intercept)
 1  -0.27009749
 2  -0.53395107
 3   0.01761646
 ...
> coef(rad.lme1)[2,]
  (Intercept)  x
 2  0.9276468 -0.6929937

In terms of the model

\[ y_k = \alpha + b_{jk} + \beta x_k + \epsilon_k \]

\[ b_j \sim N(0, \sigma_{\alpha}^2) \]

Our estimates are \( \hat{\sigma}_y = 0.756 \), \( \hat{\sigma}_\alpha = 0.328 \)

Our estimates for the fixed effects \( \hat{\alpha} = 1.461 \) and \( \hat{\beta} = -0.693 \)
Our estimates for the \( J = 85 \) random effects \( \hat{b}_j \) are given group by group in the output from \texttt{random.effects()}.

Our estimates for the fitted values \( \hat{y}_k \) are

\[
\hat{y}_k = 1.461 + \hat{\alpha}_{jk} - 0.693x_k
\]

For example the 5th house \( k = 5 \) is in county \( j = 2 \). The measurement was in the basement so \( x_k = 0 \). We estimate

\[
\hat{y}_k = 1.461 - 0.534 = 0.93
\]

for the log-radon level under the conditions the observations was made. Compare \texttt{coef(rad.lme1)[2,1]} above.
REML, LRT’s and model selection

The `lme()` function in `library(nlme)` fits using REML (Restricted Maximum Likelihood) by default.

REML is a two stage process which estimates the variances $\sigma^2_y$ and $\sigma^2_\alpha$ using a first regression, and then computes a 'profile' MLE for all other parameters conditioned on these estimates. It can be shown to have lower bias than the MLE, which tends to underestimate. The details of this algorithm are not part of the course, though we do use REML estimators. See Pinheiro and Bates Section 2.2.5 for more info.

Although REML gives good estimates, it cannot be used for LRT’s (the REML value isn’t a ML value). We need to tell `lme()` to do the fit by seeking the MLE’s if we want to use the results in a LRT, for model selection.
> rad.lme2<-lme(y~x,random=~1|county,method="ML")
> summary(rad.lme2)

... 

Random effects:
  Formula: ~1 | county
        (Intercept) Residual
  StdDev:   0.324464 0.7551596

Fixed effects: y ~ x

  Value  Std.Err  DF  t-value  p-value
(Intercept)  1.461 0.0513 833  28.48     0
x         -0.693 0.0704 833  -9.83     0

Compare these numbers with the previous results.

We have a large-ish sample, with lots of groups so both variances are well estimated by both methods, and there little difference REML/ML here.
We can use `rad.lme2` in a LRT. Compare the model with random effects by county with the model with offsets for county without random effects:

```r
> rad.lm2 <- lm(y ~ x + as.factor(county))
> anova(rad.lme2, rad.lm2)

Model df  AIC   BIC logLik Test L.Ratio p-value
rad.lme2 1   4  2171.7 2190.9 -1081.8     
rad.lm2   2  87  2178.6 2598.2 -1002.3 1 vs 2   159.03  <.0001
```

The AIC favors the model with random effects. The LRT test rejects the ('simpler') random effects model. Which model you go with will depend on how you intended to use the model, and wider physical considerations.
Random effects, blocks and treatments

In an experimental design with blocks and treatments, the block variables often come from a larger population.

We have seen blocks variables index patches in a field, litters of pigs, technicians. In this context it is natural to shrink the block effects using a mixture model.

Example: Piglet diet data from L5

<table>
<thead>
<tr>
<th>Litter</th>
<th>Diet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>I</td>
<td>89</td>
</tr>
<tr>
<td>II</td>
<td>78</td>
</tr>
<tr>
<td>III</td>
<td>114</td>
</tr>
<tr>
<td>IV</td>
<td>79</td>
</tr>
</tbody>
</table>

Let $X$ be the matrix of treatment treatment variables and $Z$ be the matrix of block variables.
Here $\beta$ will be the fixed effects and we may model $\gamma_j, j = 1\ldots J$ as random effects (with $J$ the number of blocks).

If we use random effects (\texttt{lme()} we are adding $\gamma_j \sim N(0, \sigma^2_\gamma)$ to the model.

See R-example.
**Random effects on slopes**

If some variables are drawn from a population then the same principals which apply to the intercepts may apply to the slopes. We are sticking with the case where we just have one explanatory variable $x$, so our basic regressions would be $y_k = \alpha + \beta x_k + \epsilon_k$. We have observation errors $\epsilon_k \sim N(0, \sigma^2_y)$ iid as usual and

\begin{align*}
  y_k &= \alpha_{j_k} + \beta_{j_k} x_k + \epsilon_k
\end{align*}

with

\[
\begin{pmatrix}
  \alpha_j \\
  \beta_j
\end{pmatrix} \sim N\left[
\begin{pmatrix}
  \mu_\alpha \\
  \mu_\beta
\end{pmatrix},
\begin{pmatrix}
  \sigma^2_\alpha & \rho \sigma_\alpha \sigma_\beta \\
  \rho \sigma_\alpha \sigma_\beta & \sigma^2_\beta
\end{pmatrix}
\right]
\]

for $j = 1, 2, ..., J$. Here $\rho$ is a parameter allowing correlation of the random effects in slope and intercept.
We looked at the radon data. Let us see if there is anything to gain from allowing a random effect on slope. The model we are fitting is the one above.

Refer accompanying R code.

Discussion of Goodness of Fit in connection with example: very close to NLM GOF.
Conclusions

Hierarchical models and more general models with random effects are very widely used to accommodate group structure in explanatory variables.

If the number of groups is large then $\sigma_\alpha$ is well estimated. If there are just a few groups, this is poorly estimated and the unpooled or poole models could be used.

Gelman and Hill: "There is little risk from applying a multilevel model, assuming we are willing to put in the effort to set up the model and interpret the resulting inferences."

This try-it-and-see approach makes sense - we should use mixed effects models fairly routinely for data sets of any reasonable size.