Multilevel modelling: Beyond the basic applications

Daniel B. Wright\(^1\)* and Kamala London\(^2\)

\(^1\)Department of Psychology, Florida International University, Miami, Florida, USA
\(^2\)University of Toledo, Toledo, Ohio, USA

Over the last 30 years statistical algorithms have been developed to analyse datasets that have a hierarchical/multilevel structure. Particularly within developmental and educational psychology these techniques have become common where the sample has an obvious hierarchical structure, like pupils nested within a classroom. We describe two areas beyond the basic applications of multilevel modelling that are important to psychology: modelling the covariance structure in longitudinal designs and using generalized linear multilevel modelling as an alternative to methods from signal detection theory (SDT). Detailed code for all analyses is described using packages for the freeware R.

1. Introduction

There has been a large increase in the use of multilevel models, in some form and by different names, within the social and medical sciences over the past decade. The number of computer programs that have specialized multilevel modules has also increased from a few specialist programs in the early 1990s, like ML2 (a precursor to MLwiN) and HLM, to being included in general packages like SAS, SYSTAT, and SPSS.

Different authors use different notations. We will use two notations in this paper. First, we follow Goldstein’s (2003) notation, which we feel is the simplest for presenting multilevel models when describing the statistics. The second notation is that used in R, which we use to describe the computations for the examples. Using the first notation, the standard linear multilevel model with a single predictor variable is:

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \]

where \( j \) is the subscript for the different groups and \( i \) is the subscript for the different individuals. For this model, the intercept for group \( j \) is estimated by \( \beta_0 + u_j \), an
estimate of the population intercept \( (\beta_0) \) and variation around it for the individual groups \((u_j)\). The \( e_{ij} \) are the individual level residuals. In this equation all the groups are assumed to have the same slope \( \beta_1 \). It is usually assumed that the \( u_j \) and the \( e_{ij} \) are normally distributed around 0 with unknown standard deviations, \( \sigma_u \) and \( \sigma_e \). The equal slope assumption can be relaxed by letting the \( \beta_1 \) vary such that \( \beta_1j = \beta_1 + u_1j \), where \( u_1j \) is a random variable usually also assumed to be normally distributed. Similarly heteroscedasticity as a function of \( x_{ij} \) in the level-1 residuals can be modelled by including additional level 1 random variation that is function of \( x_{ij} \) (for example, \( x_{ij} \cdot e_{1ij} \)). Different assumptions can be made and tested for correlations among the variables. More predictor variables can be added for multiple multilevel regressions, and if \( x_{ij} \) is replaced with a series of dummy variables, then multilevel ANOVAs can be analysed. These, in a sense, are the standard multilevel models and there are several introductions to these for psychologists (e.g. Hoffman & Rovine, 2007; Wright, 1998).

The purpose of this paper is to go beyond the basic multi-level model and to describe some of the techniques that can be particularly beneficial for psychologists. This paper is written for psychologists with good quantitative skills and not for statisticians. Where there is some statistical debate about, for example, different algorithms, we direct readers to where they can read more about the debate. More detailed reviews, in ascending order of the statistical knowledge they assume, are: Kreft and de Leeuw (1998); Hox (2002); Singer and Willett (2009); Bryk and Raudenbush (2002); Pinhiero and Bates (2000); and Goldstein (2003). We take the view that most quantitative psychologists are more interested in how to run a particular model, than specifics about the algorithms. Because of this we focus more on computer implementation than mathematics.

Each multi-level package has a different interface and different capabilities; therefore the choice of which to use is important. To make this paper useable by the largest audience we chose to present these models using the freeware R (R Development Core Team, 2008), which can be downloaded from the web (see Appendix). While general packages like SAS, SPSS, S-Plus, and SYSTAT are widely available to most academics, they are expensive without a site licence. Other packages like aML, MIXOR (which is now superseded by the commercial SuperMix), and WinBUGS are free, but are more specialized than R so are less useful for other analyses. MLwiN (Rasbash, Steele, Browne, & Prosser, 2005) was another contender both because it is in our opinion the state-of-the-art multi-level package and also that it is currently free for UK users (http://www.cmm.bristol.ac.uk/MLwiN/ordering/ac-uk.shtml). However, given the international readership of this journal we opted for R. Crawley (2005, 2007) has written an introduction to R that is suitable for the same audience as this paper and a detailed manual. For more advanced coverage see Chambers (2008) and for an introductory book see Venables, Smith, and the R Development Core Team (2008). In an ascending order of statistical knowledge assumed, Wright and London (2009), Faraway (2004, 2006), and Fox (2002) provide tutorials on running regressions using R. There are different packages for R that estimate multi-level models. We will use \texttt{nlme} (Pinheiro & Bates, 2000; Pinheiro, Bates, DebRoy, & Sarkar, 2008) and \texttt{lme4} (Bates, 2007). These two packages have similar syntax and can do similar things. \texttt{nlme} has in-built correlation structures which makes it better suited for our first example and \texttt{lme4} allows generalized linear models (GLM) which makes it better suited for our second example.
This is a pedagogical piece. This is why R code and output are included in the text. No new statistical procedures are described. We look at two situations, common in psychology, and show how extensions of the standard multi-level model presented above can be used to analyse the data. These are longitudinal designs and using multi-level logistic regression as an alternative to methods from SDT.

1. Example #1 – Longitudinal designs: Child birth

To download data:

```r
ayers <- read.table("http://www.sussex.ac.uk/Users/danw/MLM//ayers.dat")
or
install.packages("sdtalt")
library(sdtalt)
then
attach(ayers)
```

Ayers’ (1999) longitudinal study of women’s mental health during and after pregnancy is used to illustrate examining possible variance-covariance structures. There were four time points (during pregnancy, 1 week, 6 weeks, and 6 months after birth) and 287 women in total. We will look at their anxiety scores, which range from 0 to 21, at these four time points: anx1; anx2; anx3; and anx4. Here are the values for the first few cases (the dep (depression) variables are not considered here):

<table>
<thead>
<tr>
<th>partno</th>
<th>dep1</th>
<th>dep2</th>
<th>dep3</th>
<th>dep4</th>
<th>anx1</th>
<th>anx2</th>
<th>anx3</th>
<th>anx4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>2</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

Some of the values are missing and are labelled NA in the data file (e.g. participant #22 for the 3rd and 4th sessions). There are a variety of different methods for dealing with missing values (Little & Rubin, 2002), often estimating some value for each missing value. These can be computationally cumbersome with traditional within-subject
approaches to data analysis. Because of this, in the past researchers would go to great lengths to include all participants at each testing phase, make sure that the testing phases were all at the same time, and often would throw out participants with incomplete data. The multi-level approach solves some of these computational difficulties because the individual measurements are treated as randomly sampled elements nested within the individual and a variable for time can be used in the models. Conceptual difficulties still remain with the multi-level approach if the missing values are not missing at random, but the computational means for including incomplete cases are addressed.

The data were positively skewed for each anxiety variable (0.66, 0.90, 1.10, and 0.91, for anx1 to anx4). The models in this example assume that the residuals are normally distributed. While having normally distributed response variables does not imply that the residuals will be normally distributed (nor vice versa), it was the case here and therefore we transformed the variables. Several transformations were tried and the square root of the variable plus .5 was used. Let anx1 = √(anx1 + .5) or in R:

\[
\text{anx1} \leftarrow \text{sqrt}(\text{anx1} + .5); \text{anx2} \leftarrow \text{sqrt}(\text{anx2} + .5)
\]
\[
\text{anx3} \leftarrow \text{sqrt}(\text{anx3} + .5); \text{anx4} \leftarrow \text{sqrt}(\text{anx4} + .5)
\]

The new skewness values are: −0.12, 0.17, 0.23, and 0.10.

The numbers of women completing each wave were 251, 244, 219, and 201, and the means at these times were: 2.59; 2.19; 2.09; and 2.14, respectively. Complete data are available for only 177 of the 287 women. As mentioned above, with the traditional approach missing values must be dealt with. The two simplest approaches for dealing with missing data are: exclude incomplete cases and calculate measures for all pairs with data. The next step in analysing these data should be to look at the relationships between pairs of these variables, both graphically and numerically. Figure 1 shows the scatterplots, the histograms, and in the upper triangle the correlations and their 95% confidence intervals. The code for Figure 1 follows the example for pairs on the R help facility. From Figure 1, it is clear that the anxiety scores are all positively related.

Table 1 shows the correlations between data at the four time points in the lower triangle, the covariances in the upper triangle, and the variances along the diagonal. The left side of the table shows the values only for participants with complete data and the right side for all pairwise complete. The relationship between covariances, correlations, and variances is simple and is important for understanding R output. The covariance between variables x and y is: \( \text{cov}_{xy} = \text{sd}_x \text{sd}_y \text{cor}_{xy} \). In R the correlation and covariance tables can be combined into a single table, like Table 1, with:

\[
\text{anxvars} \leftarrow \text{cbind}(\text{anx1,anx2,anx3,anx4})
\]
\[
\text{coranx} \leftarrow \text{cor}(\text{anxvars,use} = \text{"complete.obs")}
\]
\[
\text{covanx} \leftarrow \text{cov}(\text{anxvars,use} = \text{"complete.obs")}
\]
\[
\text{cmat1} \leftarrow -\text{upper.tri}(\text{covanx,diag} = \text{T})*\text{covanx} + \text{lower.tri} (\text{coranx})*\text{coranx}
\]
\[
\text{coranx2} \leftarrow -\text{cor}(\text{anxvars,use} = \text{"pairwise.complete.obs")}
\]
\[
\text{covanx2} \leftarrow -\text{cov}(\text{anxvars,use} = \text{"pairwise.complete.obs")}
\]
\[
\text{cmat2} \leftarrow -\text{upper.tri}(\text{covanx2,diag} = \text{T})*\text{covanx2} + \text{lower.tri} (\text{coranx2})*\text{coranx2}
\]
\[
\text{print}(\text{cbind}(\text{cmat1,cmat2}),\text{digits} = \text{2})
\]
There are a few things to note from Table 1. First, the correlations are all substantial and positive. Second, the correlations tend to be slightly higher when the variables are nearer in time (for the pairwise complete data: .47, .54, .61, for one step away compared with .45, .49, for two steps away; and .43 for three steps away). In many correlation matrices these differences are more pronounced, but it is still worth examining if this pattern should be taken into account here.

**Table 1.** The correlations (lower triangle), variances (diagonal), and covariance (upper triangle) for the four waves of data for anxiety from Ayers (1999)

<table>
<thead>
<tr>
<th></th>
<th>Excludes incomplete</th>
<th>All pairwise data included</th>
</tr>
</thead>
<tbody>
<tr>
<td>anx1</td>
<td>.87</td>
<td>.87</td>
</tr>
<tr>
<td>anx2</td>
<td>.46</td>
<td>.47</td>
</tr>
<tr>
<td>anx3</td>
<td>.43</td>
<td>.45</td>
</tr>
<tr>
<td>anx4</td>
<td>.44</td>
<td>.43</td>
</tr>
</tbody>
</table>
Although this is not a multi-level dataset per se, researchers now often conceptualize repeated measures data as multi-level, and use multi-level algorithms to circumvent some of the problems of the standard repeated measures ANOVA (Singer & Willett, 2003). The first step is restructuring the dataset so that there is a single response variable, anx, a session variable called session, and a participant number (called partno2).

```
anx <- c(anx1, anx2, anx3, anx4)
session <- rep(1:4, each = length(anx1))
partno2 <- rep(partno, 4)
```

The missing values were also removed. In R this is done by:

```
detach(ayers)
ayers <- na.omit(data.frame(partno2, session, anx))
rm(partno2); rm(session); rm(anx)
attach(ayers)
```

Here are the first few lines for these data:

```
ayers[1:3,]
partno2 session anx
 1   67     1 1.870829
 2   22     1 1.581139
 3   40     1 2.549510
```

There are other ways to restructure the data. In some packages when creating a new multi-level structure the original variables would no longer be active. Here they are, so anx1 still exists. This is why partno2 was used rather than partno, and why the remove function (rm) was used to clean-up the working environment (other variables could also be removed). Most R statistics functions allow for the inclusion or exclusion of missing values, but here all NAs were removed with the na.omit function. The new data file ayers has 915 lines for the 915 anxiety measurements (i.e. 251 + 244 + 219 + 201).

The main substantive question in Ayers (1999) was about differences in anxiety at the different points in time and how these relate to other factors. Here we concentrate on finding a good correlation structure. Sometimes researchers are specifically interested in the correlations among variables, but an obvious question is whether choosing a good correlation structure makes a difference for estimating the fixed effects. The short answer is that it usually does not greatly affect the estimates of the fixed effects, but it often affects the precision of these estimates. Therefore, it is important to explore if confidence intervals or p values are to be reported.

```
gls – generalized least squares from nlme library (Pinheiro & Bates, 2000)
gls(respvar ~ covariates, weight = variance function,
correlation = correlation structure)
Example variance and correlation functions (also available for lme):
varIdent(form = ~ 1 | repeated measure)
corAR1(form = ~ 1 | level)
```
It is necessary to examine both the variances of the individual measures and the correlations among the measures. We will use the gls function from nlme (Pinheiro & Bates, 2000; Pinheiro et al., 2008), and when using this function these two facets (variances and correlations) are considered separately using the weight and the correlation options. The nlme package has several built-in variance and correlation structures. For the variance terms, they can all be equal (assume homogeneity), all be different with no particular relationships among them (unstructured heterogeneity), be a function of the time series (e.g. variances closer in time more similar), or be a function of other variables (e.g. variances a function of predicted values).

Similarly, the variables can be assumed to be uncorrelated as in a between-subjects design, have equal correlations, all have different correlations without any particular pattern, or a variety of correlation structures. Most of the structures in which psychologists would be interested come with nlme, but the user can construct their own (Pinheiro et al., 2008). The most common pattern for longitudinal studies is first order autoregressive (AR1) where the correlations steadily decrease with distance from the main diagonal. Table 2 shows the patterns of standard deviations and correlations for some of these models (assumed covariances can be calculated using \( \text{cov}_{xy} = s_{x} s_{y} \text{cor}_{xy} \)).

The default contrasts for an ordered variable (session) are polynomials (linear, quadratic, etc.) and these will be used for the fixed part of the model. The default estimation procedure for these models is restricted maximum likelihood (REML) and this will be used for this example. The main alternative is maximum likelihood (set method = "ML"). Here is the code for these 10 models. The model, \( mij \), corresponds to the \( i \)th row and \( j \)th column of Table 2. The only difference between the first five and the last five models is the standard deviations are allowed to differ in the later models.

\[
\begin{align*}
m_{11} &< - \text{gls}(\text{anx} \sim \text{as.ordered(session)}) \\
m_{12} &< - \text{gls}(\text{anx} \sim \text{as.ordered(session)}, \\
& \quad \text{correlation} = \text{corCompSymm}(\text{form} = \sim 1 | \text{partno2})) \\
m_{13} &< - \text{gls}(\text{anx} \sim \text{as.ordered(session)}, \\
& \quad \text{correlation} = \text{corAR1}(\text{form} = \sim 1 | \text{partno2})) \\
m_{14} &< - \text{gls}(\text{anx} \sim \text{as.ordered(session)}, \\
& \quad \text{correlation} = \text{corARMA}(\text{form} = \sim 1 | \text{partno2}, p = 3, q = 0)) \\
m_{15} &< - \text{gls}(\text{anx} \sim \text{as.ordered(session)}, \\
& \quad \text{correlation} = \text{corSymm}(\text{form} = \sim 1 | \text{partno2})) \\
m_{21} &< - \text{gls}(\text{anx} \sim \text{as.ordered(session)}, \\
& \quad \text{weights} = \text{varIdent}(\text{form} = \sim 1 | \text{session})) \\
m_{22} &< - \text{gls}(\text{anx} \sim \text{as.ordered(session)}, \\
& \quad \text{weights} = \text{varIdent}(\text{form} = \sim 1 | \text{session}), \\
& \quad \text{correlation} = \text{corCompSymm}(\text{form} = \sim 1 | \text{partno2})) \\
m_{23} &< - \text{gls}(\text{anx} \sim \text{as.ordered(session)}, \\
& \quad \text{weights} = \text{varIdent}(\text{form} = \sim 1 | \text{session}), \\
& \quad \text{correlation} = \text{corAR1}(\text{form} = \sim 1 | \text{partno2})) \\
m_{24} &< - \text{gls}(\text{anx} \sim \text{as.ordered(session)}, \\
& \quad \text{weights} = \text{varIdent}(\text{form} = \sim 1 | \text{session}), \\
& \quad \text{correlation} = \text{corARMA}(\text{form} = \sim 1 | \text{partno2}, p = 3, q = 0)) \\
m_{25} &< - \text{gls}(\text{anx} \sim \text{as.ordered(session)}, \\
& \quad \text{weights} = \text{varIdent}(\text{form} = \sim 1 | \text{session}), \\
& \quad \text{correlation} = \text{corSymm}(\text{form} = \sim 1 | \text{partno2}))
\end{align*}
\]
Table 2. Some possible variance and correlation structures for Ayers (1999) for the anxiety measures across the four time points

<table>
<thead>
<tr>
<th></th>
<th>Autoregressive 0 (AR0)</th>
<th>Autoregressive 1 (AR1)</th>
<th>Autoregressive 2 (AR2)</th>
<th>Autoregressive 3 (AR3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncorrelated (between-subjects)</td>
<td>Homogeneous $\sigma = \sigma_i$</td>
<td>$\sigma$</td>
<td>$\sigma$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>$d_f = d_f$</td>
<td>$2.50486, 2.528957$</td>
<td>$2.26584, 2.30434$</td>
<td>$2.26584, 2.30434$</td>
<td>$2.26584, 2.30434$</td>
</tr>
<tr>
<td>$AIC, BIC$</td>
<td>$2,504.86$, $2,528.93$</td>
<td>$2,266.17$, $2,318.82$</td>
<td>$2,266.17$, $2,318.82$</td>
<td>$2,266.17$, $2,318.82$</td>
</tr>
<tr>
<td>$2 \times LL$</td>
<td>$1,247.43$</td>
<td>$1,127.12$</td>
<td>$1,127.12$</td>
<td>$1,127.12$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Uncorrelated (within-subject)</th>
<th>Autoregressive 0 (AR0)</th>
<th>Autoregressive 1 (AR1)</th>
<th>Autoregressive 2 (AR2)</th>
<th>Autoregressive 3 (AR3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homogeneous $\sigma = \sigma_i$</td>
<td>$\sigma$</td>
<td>$\sigma$</td>
<td>$\sigma$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>$d_f = d_f$</td>
<td>$2.50486, 2.528957$</td>
<td>$2.26584, 2.30434$</td>
<td>$2.26584, 2.30434$</td>
<td>$2.26584, 2.30434$</td>
</tr>
<tr>
<td>$AIC, BIC$</td>
<td>$2,504.86$, $2,528.93$</td>
<td>$2,266.17$, $2,318.82$</td>
<td>$2,266.17$, $2,318.82$</td>
<td>$2,266.17$, $2,318.82$</td>
</tr>
<tr>
<td>$2 \times LL$</td>
<td>$1,247.12$</td>
<td>$1,127.12$</td>
<td>$1,127.12$</td>
<td>$1,127.12$</td>
</tr>
</tbody>
</table>

| Homogeneous $\sigma = \sigma_i$ | $\sigma$ | $\sigma$ | $\sigma$ | $\sigma$ |
| $d_f = d_f$ | $2.50486, 2.528957$ | $2.26584, 2.30434$ | $2.26584, 2.30434$ | $2.26584, 2.30434$ |
| $AIC, BIC$ | $2,504.86$, $2,528.93$ | $2,266.17$, $2,318.82$ | $2,266.17$, $2,318.82$ | $2,266.17$, $2,318.82$ |
| $2 \times LL$ | $1,247.43$ | $1,127.12$ | $1,127.12$ | $1,127.12$ |

| Homogeneous $\sigma = \sigma_i$ | $\sigma$ | $\sigma$ | $\sigma$ | $\sigma$ |
| $d_f = d_f$ | $2.50486, 2.528957$ | $2.26584, 2.30434$ | $2.26584, 2.30434$ | $2.26584, 2.30434$ |
| $AIC, BIC$ | $2,504.86$, $2,528.93$ | $2,266.17$, $2,318.82$ | $2,266.17$, $2,318.82$ | $2,266.17$, $2,318.82$ |
| $2 \times LL$ | $1,247.43$ | $1,127.12$ | $1,127.12$ | $1,127.12$ |

Note: The lowest AIC and BIC are in italic and underlined. Other models were tested and are discussed on the website. Calculating the values for $\sigma$ and $\rho$ for these matrices is somewhat cumbersome. Methods are shown on the web page for doing this.
These models can be compared with the `anova` function:

```r
anova(m11,m12,m13,m14,m15,m21,m22,m23,m24,m25)
```

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Test</th>
<th>L.Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>m11</td>
<td>1</td>
<td>2504.864</td>
<td>2528.937</td>
<td>-1247.432</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m12</td>
<td>2</td>
<td>2266.170</td>
<td>2295.057</td>
<td>-1127.085</td>
<td>1 vs 2</td>
<td>240.69381</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>m13</td>
<td>3</td>
<td>2294.733</td>
<td>2323.621</td>
<td>-1141.367</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m14</td>
<td>4</td>
<td>2265.838</td>
<td>2304.354</td>
<td>-1124.919</td>
<td>3 vs 4</td>
<td>32.89553</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>m15</td>
<td>5</td>
<td>2265.855</td>
<td>2318.815</td>
<td>-1121.928</td>
<td>4 vs 5</td>
<td>5.98259</td>
<td>.1125</td>
</tr>
<tr>
<td>m21</td>
<td>6</td>
<td>2510.231</td>
<td>2548.747</td>
<td>-1247.116</td>
<td>5 vs 6</td>
<td>250.37579</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>m22</td>
<td>7</td>
<td>2271.064</td>
<td>2314.395</td>
<td>-1126.532</td>
<td>6 vs 7</td>
<td>241.16657</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>m23</td>
<td>8</td>
<td>2298.414</td>
<td>2341.745</td>
<td>-1140.207</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m24</td>
<td>9</td>
<td>2270.274</td>
<td>2323.234</td>
<td>-1124.137</td>
<td>8 vs 9</td>
<td>32.14057</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>m25</td>
<td>10</td>
<td>2271.213</td>
<td>2338.616</td>
<td>-1121.606</td>
<td>9 vs 10</td>
<td>5.06100</td>
<td>.1674</td>
</tr>
</tbody>
</table>

The \( \chi^2 \) value is two times the difference in their log-likelihoods: \( 2(1124.919 - 1124.919) \).

2 The R package `sdtalt` calculates these and 13 other common statistics (e.g. \( A' \), odds ratio, eta, and weighted kappa) from signal detection theory, along with the confidence intervals for the sample means of these statistics (Wright, Horry, & Skagerberg, in press).
DeCarlo (1998) described how this is equivalent to a GLM, the probit regression, for each participant (see Faraway, 2006; Wright & London, 2009, for introduction to GLMs in R). Thus, the following function produces the same result as \texttt{sdt1} for a single set of values using the \texttt{glm} function. The first two lines inside the function reorganized the data into two variables.

\begin{verbatim}
sdt2 <- function(hits,fas,misses,cr)
  {sayold <- c(rep(1,sum(hits,fas)),rep(0,sum(misses,cr)))
   old <- c(rep(.5,hits),rep(-.5,fas),rep(.5,misses),rep(-.5,cr))
   model <- glm(sayold ~ old,family = binomial(link = "probit"))
   d <- model$coef[2]; csdt <- -model$coef[1]
   return(c(d,csdt))}
\end{verbatim}

Both \texttt{sdt1(65,6,34,75)} and \texttt{sdt2(65,6,34,75)} produce $d' = 1.85$ and $C = 0.52$. Pictorially, Figure 2 shows these values.

These two methods are equivalent for calculating $d'$ and $C$, so if these statistics are all that is needed then either can be used. Using \texttt{sdt2} allows access to other information, so if \texttt{summary(model)} was embedded in \texttt{sdt2} then it would show that, for example, the standard error of $d'$ for these data is 0.47. The flexibility of using the \texttt{glm} function means that other methods can also be used. In medicine, the logistic model is more common than the Gaussian/normal model (Zho, Obuchowski, & McClish, 2002). This can be found by changing \texttt{sdt2} so that it uses the default \texttt{family = binomial} link function, the logit. The coefficient of interest for the logistic model is the log of the odds ratio, ln(OR). Conceptually, ln(OR) can be thought of in a very similar way to $d'$. While $d' = z(\text{hit rate}) - z(\text{false alarm rate})$, ln(OR) = ln(odds of a hit) − ln(odds of a false alarm). An approximate relationship between them is ln(OR) $\sim 1.6d'$, and this holds except when $d'$ is either very large or very small (i.e. $|d'| > 4$). For the example above, ln(OR) = 3.17 which is about 1.6 times the observed $d'$. We will use the logistic model because that is the natural link function for the binomial distribution (Hoffmann, 2004).

Another way in which the GLM approach is flexible is that predictor variables which vary trial-by-trial, like confidence or response time, can be included within the model.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{distribution.png}
\caption{A pictorial representation of $d' = 1.85$ and $C = 0.52$ for the equal variance normal signal detection theory model.}
\end{figure}
This is difficult to do with the standard approach to SDT. The norm within most memory recognition research is to calculate $d'$ (or other measures) for each participant, sometimes separately for different within-subject conditions, and compare these aggregate measures. One of the main arguments for multi-level modelling is to move away from analysing aggregate level data like these.

In this example the function \texttt{lmer} is used rather than \texttt{lme} or \texttt{gls} because \texttt{lmer} allows generalized linear multi-level modelling. \texttt{lmer} was not used in the last example because it does not have built-in covariance structures. For \texttt{lmer} the random effects are placed within the model. Thus, \texttt{m12} from the first example could be written as:

\begin{verbatim}
  lmer(anx ~ as.ordered(session) + (1|partno2))
\end{verbatim}

The output is slightly different because different estimations procedures are used, but the basic models are equivalent.

With \texttt{lmer} for GLMs, maximum-likelihood estimates of the parameters are found with an iterative procedure. The \texttt{lme4} manual lists three approximation methods (Bates, 2007). In order of their level of accuracy and their computation time, these are: penalized quasi-likelihood (PQL); Laplacian approximation (Laplace); and adaptive Gaussian quadrature approximation (AGQ). Bates recommends using Laplace, so it is used here. Details of \texttt{lme4} are in Bates (2007, \url{http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf}).

We begin with a simple example where traditional SDT could also be used: showing that white people have better recognition memory for white faces than for black faces. We then add in a continuous predictor variable that varies by trial, the log of the response time. The data are from the white English participants of Wright, Boyd, and Tredoux (2003).

To download data:
\begin{verbatim}
  memrec <- read.table("http://www.sussex.ac.uk//Users//danw//MLM//memrec.dat", header = T)
\end{verbatim}

or
\begin{verbatim}
  library(sdtalt)
  then
  attach(memrec)
\end{verbatim}
The data are in memrec. The variables are: face; saysold; faceold; facewhite; intime; and partno. Figure 3 shows the frequency of old responses in the different conditions. The proportion of hits is about the same for white and black faces, but the number of false alarms is much greater for black faces. This is consistent with the memory literature (Horry & Wright, 2008).

The model to test for what is called the own race bias compares model1 and model2 below. model1 includes the main effects for whether the face was previous shown (faceold, which measures accuracy), how many of these were white faces (facewhite), and it allows the intercept to vary for people. In SDT terminology this corresponds to people having different response criteria. In statistical notation it is:

$$\logit(p[saysold]) = \beta_0 + \beta_1 \text{facewhite} + \beta_2 \text{faceold} + u_j + e_{ij}$$

where the $e_{ij}$ are assumed to be binomially distributed and the $u_j$ are assumed to be normally distributed. In R this is:

```r
model1 <- lmer(saysold ~ facewhite + faceold + (1|partno), family = binomial, method = "Laplace")
```

Writing `summary(model1)` produces statistics for the overall fit of this model and the individual coefficient estimates: $\beta_0 = -0.99; \beta_1 = -0.51; \beta_2 = 2.02; \text{ and var}(u_j) = 0.12$. We can conclude that people say ‘old’ more to black faces and to items that are old. Both of these effects are evident from Figure 3.

model2 adds the interaction between the two predictor variables so it tests whether accuracy depends on whether the face was white or black. The `update` function is convenient for incrementally building models.

```r
model2 <- update(model1, . ~ . + facewhite:faceold)
```

ANOVA:

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>model1</td>
<td>4</td>
<td>3455.4</td>
<td>3479.5</td>
<td>-1723.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>model2</td>
<td>5</td>
<td>3422.2</td>
<td>3452.3</td>
<td>-1706.1</td>
<td>35.225</td>
<td>1</td>
<td>2.937e-09***</td>
</tr>
</tbody>
</table>

Figure 3. The probability of an ‘old’ response for the data from white English participants in Wright et al. (2003).
The own race bias is observed, $\chi^2(1) = 35.23$, $p < .001$. Participants were more accurate responding to white faces. The coefficient for whether something was previously shown (here faceold) measures discriminability (a measure of memory) and interactions between this and other variables show whether these other variables moderate accuracy. The fixed effects for model2 can be found either by typing `model2` (which produces the fixed effects and a lot of other output) or by `fixef(model2)`.

The estimated parameter for faceold is 1.57. This estimates ln(OR)$_{\text{black}}$. The interaction was 1.02, so the estimate of ln(OR)$_{\text{white}}$ is $1.57 + 1.02 = 2.59$. The probit model can also be used and the multi-level estimates for $d'_\text{black}$ and $d'_\text{white}$ by using the probit link.

The estimated value for $d'_\text{black}$ is 0.97 and for $d'_\text{white}$ is $0.97 + 0.59 = 1.56$. Notice that the ln($\text{OR}$) $< 1.6$ approximation holds.

If we had calculated ln(OR) and $d'$ separately for each individual for each race, the means of these would be ln(OR)$_{\text{black}} = 1.69$, ln(OR)$_{\text{white}} = 5.47$, $d'_\text{black} = 1.02$, and $d'_\text{white} = 2.52$. Notice that these values are very different from those found with the multi-level model for white faces and that the ln(OR) $\approx 1.6 d'$ approximation does not hold for the white faces ($5.47/2.31 = 2.37$). This is because some individuals had extremely high values for these (many $d' > 4$) and that the mean is not a robust statistic. The 20% trim means are: ln(OR)$_{\text{black}} = 1.74$; ln(OR)$_{\text{white}} = 3.02$; $d'_\text{black} = 1.06$; and $d'_\text{white} = 1.77$. The approximation now holds and these values are closer to those found with the multi-level approach.

It is worth exploring if participants' accuracy varies. This can be done by changing the random part of the model to (faceold|partno). This adds both a term for the variance of accuracy and the covariance between accuracy and responding old, and therefore there is an increase of two degrees of freedom in the model. Including this term increases the fit of the model, $\chi^2(2) = 12.59$, $p = .002$. It can be written with the following and compared with model12. The difference is statistically significant, $\chi^2(2) = 12.59$, $p = .002$, although the BIC value is increased.

Up to this point traditional SDT could have been used to reach the same basic conclusion, providing care was taken to use robust estimators. The next step involves adding a continuous variable which varies by trials: the log of the response time.
The theoretical question is about the relationship between response time and accuracy for the different faces. Much research shows quicker responses tend to be more accurate, but the exact relationship is unclear (Weber, Brewer, Wells, Semmler, & Keast, 2004). Our question is whether the relationship is similar for old and new faces, and white and black faces.

The main effect of time was non-significant, $\chi^2(1) = 0.28, p = .60$, but is retained as interactions including this term are added to the model. The best model in terms of AIC, BIC, and significance tests, includes only adding the interaction between \textit{lntime} and \textit{faceold}: improvement $\chi^2(1) = 52.43, p < .001$.

The coefficients for this model are:

```
model3 <- update(model2b, . ~ . + \textit{lntime})
anova(model2b, model3)
```

```
df  AIC       BIC   logLik  Chisq  Chi df  Pr(>Chisq)
model2b 7 3413.6  3455.7 -1699.8       
model3 8 3415.4  3463.4 -1699.7  0.2781 1  0.5979
```

The coefficient associated with the \textit{faceold:lnetime} interaction is negative and therefore accuracy decreases with increased response time. The lack of other interactions improving the model shows that the relationship between time and accuracy is similar for white and black faces.

In Figure 4 the probability of a correct response by whether the face was new (where it is a probability of a correct rejection) or old (where it is a probability of a hit) and the race of the face is plotted with response time. We calculated the predicted probabilities using the estimates above and transformed the predicted values with $\frac{e^x}{1 + e^x}$, the inverse of the logit transformation. We then changed these predicted values to 1 minus themselves for new faces, so that the probabilities were for correct responses.

```
mod4 <- ~ -7.51 + faceold*13.31 + facewhite* -1.07 + lnetime
*0.86 + faceold*facewhite*.99 + faceold*lnetime* -1.50
predprob <- exp (mod4) / (1 + exp (mod4))
rightprob <- - predprob*faceold + (1 - faceold)*(1 - predprob)
```

This shows that after controlling for response time the probability of correct response is highest for new white faces.
There are several other extensions to the multi-level GLMs that can be explored, including tests of the variance/covariance matrix at each level of the model, as discussed with the first example. It is also often worth examining the size of $\text{var}(e_{ij})$ to see if there is more or less variation than predicted from the binomial distribution (Browne, Subramanian, Jones, & Goldstein, 2005; Wright, 1997). This is reported with the summary information and for these data it is very similar to that predicted by the binomial distribution. Another extension is to model more flexible relationships using multi-level generalized additive models (see Wood, 2006, for details, and also Ng, Carpenter, Goldstein, & Rasbash, 2006). Wood (2006) has written a specialist package within R to run these, but it can also be done with the \texttt{bs} function from the \texttt{splines} library. \texttt{bs} is a B-spline which is a set of polynomials linked together smoothly at knots (see Wood, 2006, for more details). The following model allows the relationship between response time and responding old to be modelled in a more flexible way which increases the chances of detecting other effects. This is a multi-level gamcova (Wright & London, 2009).

```
model8 <- lmer(saysold ~ faceold*facewhite + bs(lntime,df = 4) + faceold + lntime:faceold + (faceold|partno), family = binomial, method = "Laplace")
anova(model4,model8)
```

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>model4</td>
<td>9</td>
<td>3364.9</td>
<td>3419.0</td>
<td>-1673.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>model8</td>
<td>12</td>
<td>3359.2</td>
<td>3431.2</td>
<td>-1667.6</td>
<td>11.751</td>
<td>3</td>
<td>0.008288**</td>
</tr>
</tbody>
</table>

Finally, the usual practice in memory recognition research is not to worry about differences among the stimuli within any category. Within linguistics there is a long tradition of taking into account the differences among stimuli (Clark, 1973). Baayen, Davidson, and Bates (2008) show how to include a random variable in their models using \texttt{lmer}. This creates what is called a cross-classified model. In this data set there is a variable, face, for the face number.

\textbf{Figure 4.} The probability of a correct response for previously seen and previously unseen white and black faces, with the raw response times, for the white English participants from Wright et al. (2003).
As can be seen, this greatly improves the model. This shows that there were more ‘old’ responses for some faces than for others.

2.1. Summary – Memory recognition

While traditional methods from SDT are often used for memory recognition studies, they present two difficulties. First, they are usually done by calculating measures (like $d'$) for each individual and then using these aggregate measures in analysis. Outliers, particularly those based on few cases, can have a large impact. This is one of the main reasons for multi-level modelling. Second, if interested in a covariate which can take many possible values, like response time, the standard SDT methods are difficult to implement since the covariate has to be split into bins and SDT measures created for each person for each bin. This can create lots of problems particularly when there are few or unequal numbers of observations per person per bin because the estimates for the individual bins can be unstable.

The multi-level modelling approach is well suited to overcome these problems. Because the standard SDT approach is equivalent to a GLM (DeCarlo, 1998), we use GLMs (in particular logistic regression, although link = "probit" could be used throughout this section) for analyses, treating trials as nested within participants. Although multi-level GLMs have been used with memory recognition data for many years (e.g. Wright & McDaid, 1996), they are still not very common. While SDT has a long history within memory research, and has been very useful in theory construction, it is likely that the flexibility of the approach used here will mean multi-level models become the norm for recognition data in years to come. These have the methodological advantage that researchers can use more variables that differ by trial within their designs. In this example response time was included which given the number of studies now conducted via computer is a variable that could be included in most studies.

3. Summary

Multilevel modelling is one of the hot statistical methods in several areas of science, including psychology. While the traditional example has been with people nested within larger clusters (e.g. pupils nested within classrooms), because of the great amount of medical research with multiple measurements per person, multi-level models with the person as the higher order level are now common (perhaps more common). Harvey Goldstein, one of the pioneers of this approach, talks about how there are hierarchies everywhere. Multilevel modelling is now one of the tools expected for social and psychological scientists.

We end with a caveat. While multi-level models are now expected to be used in areas where the hierarchical structure is obvious, more research is necessary to see how useful they are when the levels are not such clean structures and where the components at different levels cannot be viewed as some random sample of those at that level. This was Cohen’s (1976) main criticism of Clark’s (1973) language as a fixed effect fallacy.

```r
model9 <- lmer(saysold ~ faceold*facewhite + Intime*faceold + (1|face) + (faceold|partno), family = binomial, method = "Laplace")
anova(model4,model9)
```

<table>
<thead>
<tr>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>3364.9</td>
<td>3419.0</td>
<td>21673.5</td>
<td></td>
<td></td>
<td>&lt;2.2e-16***</td>
</tr>
<tr>
<td>10</td>
<td>2896.6</td>
<td>2956.7</td>
<td>1438.3</td>
<td>470.27</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

As can be seen, this greatly improves the model. This shows that there were more ‘old’ responses for some faces than for others.
Perhaps some of the resampling techniques (and local causal inference) will be applied to these situations. As with all statistical procedures, it is critical to examine the data carefully and consider the alternatives before running any statistical test. No amount of statistical sophistication can fix a bad design.

References

Ng, E. S. W., Carpenter, J. R., Goldstein, H., & Rasbash, J. (2006). Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. Statistical Modelling, 6, 23–42.
R was used for these analyses. To download R go to http://cran.r-project.org/ and following instructions for Linux, MacOS X, or Windows.

The data and detailed code for R 2.7 are available on http://www.sussex.ac.uk/Users/danw/MLM.htm. The output is annotated. The code and pages will be updated as needed. You may need to access libraries including nlme and lme4. To install and to load these use the install.packages and library functions. For example:

```r
install.packages("lme4")
library(lme4)
```

The data are both on the paper's website and part of the sdtalt package which can be installed and loaded.