Abstract

In the following review article, I address the fitting of multi-level models for the analysis of hierarchical data in laboratory animal medicine. Using an example of paternal dietary effects on the weight of offspring in a mouse model, this review outlines the reasons and benefits of using a multi-level modeling approach. To start, the concept of clustered/autocorrelated data is introduced, and the implications of ignoring the effects of clustered data on measures of association/model coefficients and their statistical significance are discussed. The limitations of other methods compared with multi-level modeling for analyzing clustered data are addressed in terms of statistical power, control of potential confounding effects associated with group membership, proper estimation of associations and their statistical significance, and adjusting for multiple levels of clustering. In addition, the benefits of being able to estimate variance partition coefficients and intra-class correlation coefficients from multi-level models is described, and the concepts of more complex correlation structures and various methods for fitting multi-level models are introduced. The current state of learning materials including textbooks, websites, and software for the nonstatistician is outlined to describe the accessibility of multi-level modeling approaches for laboratory animal researchers.

Key Words: autocorrelated data; clustered data; correlation structure; hierarchical data; multi-level modeling; random effects; intra-class correlation coefficient; variance partition coefficient

Introduction

In animal health research, whether conducted in a laboratory or field setting, it is relatively common to study groups of animals. For instance, studies may be conducted using litters of mice, pens of pigs, or even troops of primates, and the outcomes being evaluated could include measures of growth, behavior, and/or the development of disease (Bauer et al. 2012; Sadagurski et al. 2014; Tucker et al. 2011; Watkins and Sinclair 2014; Zhao et al. 2012). It is important to consider in these situations whether the outcomes experienced by the animals within a group are actually independent of each other. This concept of independence is a fundamental assumption of most statistical tests and ignoring it has serious implications, including underestimating standard errors (resulting in p values and 95% confidence intervals appearing smaller than their true values) and, in some situations, measuring incorrectly the size and/or direction of the effect of a treatment or exposure on the outcome of interest. Consequently, a great deal of statistical research, textbooks, and review articles have focused on the analysis of clustered or auto-correlated data (i.e., data where the outcomes among observations are not independent of each other) with particular emphasis on the application of multi-level models, also known as hierarchical or mixed models (Diez Roux and Aiello 2005; Dohoo et al. 2009, 2012; Goldstein 2003; Rabe-Hesketh et al. 2002; Rabe-Hesketh and Skrondal 2008; Schukken et al. 2003; Skrondal and Rabe-Hesketh 2003).

However, multi-level modeling offers more to the researcher than just proper estimation of regression model coefficients (i.e., the size of effects) or their p values. In situations where the researcher has variables measured at several hierarchical levels, they offer an opportunity to measure the effect of variables measured at multiple levels and in many situations understand how much of the variation in the outcome of interest is explained at each hierarchical level. For instance, recently there has been interest in the impact of paternal diet on metabolic function and cardiovascular function in offspring, and many of these studies have involved mouse models (Watkins and Sinclair 2014). Inherently, these studies could have a hierarchical structure with males, the different litters sired by these males, and the pups from these litters forming a 3-level hierarchy (Figure 1). A multi-level model would allow the researchers to investigate the effects of independent/explanatory variables measured at each hierarchical level on the dependent/outcome variable of interest. In addition, the variance components from these models would provide information on how much of the variation in a particular outcome was explained at the offspring, litter, and sire levels (Table 1). Consequently, multi-level modeling approaches

David L. Pearl

Making the Most of Clustered Data in Laboratory Animal Research Using Multi-Level Models

© The Author 2014. Published by Oxford University Press on behalf of the Institute for Laboratory Animal Research. All rights reserved.

For permissions, please email: journals.permissions@oup.com
not only allow for improved estimation of effects in the face of clustered data, but can also offer a researcher an opportunity to investigate biological processes in a more holistic manner.

In the following review, I will focus on the application of multi-level models and address the following topics using the influence of paternal diet on offspring weight as an example that includes hypothetical study results:

(i) Advantages of multi-level models over other potential approaches for analyzing clustered/autocorrelated data;
(ii) Basic theory behind the use of multi-level models with particular emphasis on the concept of random intercepts;
(iii) Information that can be gathered from variance components from multi-level models, and how they can be interpreted and used in designing future studies;
(iv) Correlation structures, random slopes, and methods developed for fitting multi-level models; and
(v) Software, websites, and textbooks that would be helpful for scientists interested in applying these methods to their research.

Advantages of Multi-level Models over Other Potential Approaches for Analyzing Clustered Data

Multi-level models are an extension of linear and generalized linear regression models. Consequently, they are flexible in estimating the association between independent/explanatory variables (both continuous and categorical) and dependent variables/outcomes using a variety of distributions, including normal, binomial, and Poisson. In addition, multi-level models can account for potential autocorrelation among observations because of membership in a group, repeated measures in time, or proximity in space. However, before exploring
multi-level models, it is important to recognize that researchers have and continue to use a variety of other tools to analyze clustered data. These methods or approaches include the following: performing statistical analyses on a group-level statistic (e.g., mean weight of each group); including an independent categorical variable in a regression model to account for group (i.e., fixed effects approach); the use of robust standard errors to make estimated $p$ values more conservative; and the use of generalized estimating equations (GEEs) to account for the correlation structure among observations. Although many of these methods under certain circumstances can adequately deal with clustered data, they each have some of the following limitations: loss of statistical power, inability to account for confounding by group, limited to 2-level modeling situations, and inability to examine the distribution of variance in the outcome among hierarchical levels.

Loss of Power

For instance, to evaluate the effect of a paternal dietary treatment on the weight of their offspring by the first month of life, a researcher could simply take the average weight of each litter and examine the association between the sire’s diet and these average litter weights using simple linear regression (assuming there is one litter per sire). However, by using the litter as the unit of analysis, the researcher would lose a great deal of power due to the reduced sample size and could not statistically control for potential confounding variables at the pup level (e.g., sex) or explore interaction effects between variables measured at different hierarchical levels (e.g., diet measured at the sire level and sex measured at the pup level).

Inability to Control for Group-level Confounding

Many statistical packages will allow for the reporting of robust standard errors instead of model-based standard errors (also referred to as Huber-White or sandwich estimators). The standard errors typically reported in regression models and used to determine the significance of a coefficient in a model (i.e., the significance of an explanatory variable) are based on the assumption that the distributional assumptions of the model are true and the observations are independent. In contrast, robust standard errors are less sensitive to model assumptions and estimate the expected variability in a coefficient if repeated samples of the same size as the dataset were drawn from the original population (Dohoo et al. 2009, 2012; Rabe-Hesketh and Skrondal 2008). Furthermore, they can be estimated to allow intragroup correlation, thus relaxing the assumption of independence. Unfortunately, this approach cannot deal with models that have multiple hierarchical levels and does not adjust the coefficients from the models for potential confounding effects by group.

Limited to Two-Level Models

Some researchers working within a regression modeling framework will simply add a categorical variable to account for group membership. This approach is again relatively simple, but if the number of groups is large, particularly for regression models that use maximum likelihood estimation, the model will need to estimate the effect of a large number of variables and estimation problems can occur. Furthermore, the statistical significance of variables measured at the group level cannot be assessed, because they are accounted for once the group variable has been modeled as a fixed effect. The use of GEEs can overcome this issue by estimating average group effects while accounting for the correlation structure among observations within clusters; however, like the fixed effects approach, it is generally limited to dealing with 2-level models. Consequently, if a researcher were investigating a situation in which individual animal, litter, and sire effects needed to be considered, these approaches would be inadequate (Figure 1). However, for binary outcomes, there are some statistical software packages that have alternative algorithms that allow models fit using the GEE method to account for 2 levels of clustering (i.e., 3-level models; Dohoo et al. 2009).

Examining the Distribution of Variance in the Outcome Among Hierarchical Levels

Aside from multi-level models, the approaches described above either cannot (e.g., fixed effects approach) or are limited (i.e., GEE) in their ability to explore the variance explained in the outcome at different hierarchical levels. This type of information can offer important insight into where most of the variation is being explained, and has been used in studies of human and animal disease to target potential interventions (Diez Roux 2005; Dohoo, et al. 2001; Leonard et al. 2012).

Basic Theory behind the Use of Multi-level Models

Multi-level models differ from linear regression models and generalized linear models in that they can include both fixed and random effects. Fixed effects are the mean effects that we typically see estimated in regression models to indicate the magnitude and direction of the effect of a risk factor or exposure on the outcome of interest. In the paternal diet effect example, these fixed effects could include the sex of the pup, the size of the litter, and the sire’s diet (e.g., low protein vs. normal), with the coefficient for diet representing the effect of this diet on the offspring’s weight while controlling for other variables in the model (Table 1). With random effects, we assume that some of the terms in the model are values from a probability distribution rather than having constant values (Brown and Prescott 2006). In other words, with random effects, the focus shifts from the individual group to variability in the population of groups (Dohoo et al. 2009, 2012). In the case of random intercepts, these random effects are typically used to reflect membership within a group within the population of groups of a hierarchical level. In our paternal health effects example, random intercepts would be used to account for...
for the autocorrelation among offspring that are part of a common litter and among litters that were fathered by a common sire (Figure 1; Table 1). The random effects in the model result in the creation of separate variance components for each level for multi-level models for normal data, and there are methods to estimate these variance components for the lowest level of the hierarchy (e.g., pup level) for models involving discrete outcomes (Brown and Prescott 2006; Dohoo et al. 2009, 2012; Goldstein 2003; Rabe-Hesketh et al. 2002; Rabe-Hesketh and Skrondal 2008). Typically, these random effects are assumed to have been selected from a random process or at least be representative of the population and have a normal distribution and constant variance. These models can then be expressed as a function of their fixed and random effects. For instance, if our linear outcome were weight and our fixed and random effects scenario would contain coefficients for the diet (e.g., low-protein diet vs. normal diet); \( X_1 \) is the value of the paternal diet for pup\(_{ijk} \) represented by indicator variables (e.g., \( 1 = \) low protein and \( 0 = \) normal diet); \( u \) and \( v \) represent the random intercepts for sire and litter; \( e \) represents the residual error term for each observation; and \( N(0, \sigma^2) \) indicates a normal distribution with a mean of zero and a constant variance (\( \sigma^2 \)). It should be noted that the estimates of these random effects, also known as best linear unbiased predictors, are used to assess the distributional assumptions of these models (i.e., normality and constant variance).

Similarly, for a discrete outcome like the presence/absence of hypertension, the model could be expressed as follows:

\[
y_{ijk} = \beta_0 + \beta_1 X_{ijk} + u_{(sire(k))} + v_{(litter(jk))} + e_{ijk}
\]

where \( y_{ijk} \) is the model prediction; \( \beta_0 \) is the model intercept; \( \beta_1 \) is the coefficient for the diet (e.g., low-protein diet vs. normal diet); \( X_1 \) is the value of the paternal diet for pup\(_{ijk} \) represented by indicator variables (e.g., \( 1 = \) low protein and \( 0 = \) normal diet); \( u \) and \( v \) represent the random intercepts for sire and litter; \( e \) represents the residual error term for each observation; and \( N(0, \sigma^2) \) indicates a normal distribution with a mean of zero and a constant variance (\( \sigma^2 \)).

Interpreting Variance Components

The variance of the individual observations in the model is the sum of all variance components at each hierarchical level, or in our example, the sum of variance components at the pup, litter, and sire levels after accounting for the fixed effects in the model:

\[
\sigma^2_{total} = \sigma^2_{sire} + \sigma^2_{litter} + \sigma^2_{pup}.
\]

The intra-class correlation coefficient (ICC) and the variance partition coefficient (VPC) can be calculated from these variance components. Depending on the researcher, these measures may be estimated using the variance components from an intercept-only model (i.e., no fixed effects) and/or the final model that includes both fixed and random effects (Dohoo et al. 2001; Leonard et al. 2012; Williams et al. 2013); the ICCs and VPCs estimated from the full model account for the fixed effects in the model, while ICCs and VPCs from an intercept-only model explain the distribution of variance components among hierarchical levels without accounting for these effects. Many researchers find it useful to see how much these measures change after accounting for the fixed effects in the model, and consequently report these estimates using full and intercept-only models.

VPC

Calculating a VPC is relatively straightforward regardless of the number of hierarchical levels. In our hypothetical example of paternal diet effects (Table 1), the VPC for each hierarchical level would be calculated as follows (please note that the total variance in the example below is less than 100% due to rounding errors):

\[
i. \quad VPC_{sire} = \frac{\sigma^2_{sire}}{\sigma^2_{sire} + \sigma^2_{litter} + \sigma^2_{offspring}} = \frac{0.152}{0.152 + 0.256 + 0.325} = 20.7\% \\
ii. \quad VPC_{litter} = \frac{\sigma^2_{litter}}{\sigma^2_{sire} + \sigma^2_{litter} + \sigma^2_{offspring}} = \frac{0.256}{0.152 + 0.256 + 0.325} = 34.9\% \\
iii. \quad VPC_{offspring} = \frac{\sigma^2_{offspring}}{\sigma^2_{sire} + \sigma^2_{litter} + \sigma^2_{offspring}} = \frac{0.325}{0.152 + 0.256 + 0.325} = 44.3\%
\]

In the above example, most of the variation in weight is explained at the level of the individual offspring, but substantial amounts of variation are explained at the litter and sire levels. The multi-level modeling approach has allowed a relatively simple approach to measure these effects.
The ICC, sometimes called rho (\(\rho\)), uses these variance components to estimate the correlation among individuals within a group. In our 3-level modeling scenario, we could estimate the correlation among offspring within the same litter (ICC_{litter}) and the correlation among offspring from different litters that have the same sire (ICC_{sire}) as follows:

\[
\begin{align*}
\text{i. ICC}_{\text{litter}} &= \frac{\sigma^2_{\text{sire}} + \sigma^2_{\text{litter}}}{\sigma^2_{\text{sire}} + \sigma^2_{\text{litter}} + \sigma^2_{\text{offspring}}} \\
&= \frac{0.152 + 0.256}{0.152 + 0.256 + 0.325} = 55.7% \\
\text{ii. ICC}_{\text{sire}} &= \frac{\sigma^2_{\text{sire}}}{\sigma^2_{\text{sire}} + \sigma^2_{\text{litter}} + \sigma^2_{\text{offspring}}} \\
&= \frac{0.152}{0.152 + 0.256 + 0.325} = 20.7%
\end{align*}
\]

The information concerning the ICC can be used to assess design effects (i.e., the effects of the clustered design on standard errors for fixed effects in the model) and to adjust sample size requirements to account for a lack of independence among observations (Dohoo et al. 2009, 2012; Goldstein 2003); in 2-level models, the following formula can be used to adjust a sample size to account for clustering, but more complex hierarchical models, like the 3-level model described, require simulation methods to estimate sample sizes:

\[n' = n (1 + \rho [mn - 1]),\]

where \(n'\) is the adjusted sample size, \(n\) is the original sample size calculated using a formula that does not account for clustering, \(\rho\) is the ICC, and \(m\) is the average cluster size.

Interestingly, for discrete outcomes, the estimation of VPCs and ICCs is more complex, because the different levels of the model are not on the same scale. For instance, for dichotomous outcomes (e.g., disease vs. no disease), the estimation of ICCs or VPCs is complicated, because the errors in the model are from the Bernoulli distribution rather than the normal distribution. A variety of simulation approaches have been developed to address this issue, because for discrete data, there is no unique value of the variance at the lowest level (e.g., pup level in our example), because the variance is a function of the predicted mean and is not constant (Goldstein 2003). However, for dichotomous outcomes, it has become common to use a latent variable approach where the error term is assumed to follow a logistic distribution (similar to a normal distribution in shape) with a mean of zero and a variance of \(\pi^2/3 = 3.29\) (Dohoo et al. 2009, 2012; Rabe-Hesketh and Skrondal 2008); similar approximations do not exist for other discrete outcomes. Furthermore, it is important to note that the values of the variance components will vary depending on the approach used to estimate these components (i.e., maximum likelihood or restricted maximum likelihood). In addition, ICCs and VPCs are intended for models with hierarchical structures that include only random intercepts, so these measures cannot be meaningfully estimated in models with random slopes and more complex correlation structures (Brown and Prescott 2006; Dohoo et al. 2001, 2009, 2012; Goldstein 2003; Rabe-Hesketh and Skrondal 2008).

**Correlation Structures, Random Slopes, and Methods Developed for Fitting Multi-level Models**

**Correlation Structures and Random Slopes**

This review has largely focused on the most common application of multi-level modeling in dealing with hierarchical data structures in which an equal correlation is expected among all members within a group (i.e., exchangeable correlation structure). However, the multi-level modeling approach is much more flexible. For instance, random slopes (also referred to as random coefficients) can be introduced into models to allow for the coefficients of fixed effects in models to vary by group; this approach has been used to account for repeated measurements in time from an individual where time is included as both a fixed effect and a random slope/coefficient (Brown and Prescott 2006; Cernicchiaro et al. 2010). Repeated measures in time or proximity in space can be modeled using covariance pattern models that allow the fitting of a variety of covariance/correlation patterns to account for the relationships between observations in time or space. In other words, these covariance structures account for observations that are closer in time and/or space being more strongly correlated, and different correlation structures can allow this effect to decay more quickly or slowly as time and/or space increase (Brown and Prescott 2006; Lawson et al. 2003). In our paternal diet effects example, if multiple measures were taken from the offspring over time, the researcher might consider adding an additional random intercept for the individual animal and accounting for autocorrelation in time by using a random slope or covariance pattern model. In addition to the classic hierarchical we used in our paternal diet effect example (Figure 1), it could be possible that following birth, offspring would be regrouped into cages based on body size, in which case a cross-classified correlation structure might be required to account for these housing effects after the littermates were regrouped (Rabe-Hesketh and Skrondal 2008; Rasbash et al. 2004).

**Methods Developed for Multi-level Modeling**

It is also important for researchers who are thinking of using these models to recognize that software is available to fit these models in both frequentist and Bayesian frameworks, the former being the basis of most standard statistical approaches while Bayesian analysis can make use of prior information and has proven very effective in dealing with spatial autocorrelation and imputation problems related to missing data. However, the above issues that especially benefit from a Bayesian approach are more likely to occur in epidemiological observational studies rather than laboratory-based animal
Software, Websites, and Textbooks about Multi-level Modeling

The purpose of this section is not to endorse any particular book, software package, or website, but to identify some materials that are suitable for nonstatisticians. For instance, the Centre for Multilevel Modeling at Bristol University (http://www.bristol.ac.uk/cmm) maintains a website that supports several of their specialized multi-level modeling packages (e.g., MLwiN; Centre for Multilevel Modeling, University of Bristol, UK) and also offers free introductory on-line courses and helpful reviews concerning various software packages from a variety of companies and shareware development groups. In terms of software, many general statistical software packages can support multi-level modeling, including STATA (StataCorp, College Station, TX), SAS (SAS Institute, Cary, NC), R (R-Project, http://www.r-project.org), SPSS (IBM, Armonk, NY), and SYSTAT (SYSTAT Software, San Jose, CA), whereas other packages are specific for multi-level/mixed models such as OpenBUGS (MRC Biostatistics Research Unit, Cambridge, UK), MLwiN, and HLM (SSI, Skokie, IL). Some of the packages are quite flexible in offering the ability to use frequentist and Bayesian approaches with several different frequentist approaches for discrete data (e.g., SAS), whereas other packages specialize in a particular approach (e.g., OpenBUGS for Bayesian analysis). There are also programs that combine the functionality of some general statistical software with specialized multi-level modeling software as seen with a “runmlwin” command that allows MLwiN to fit models in STATA (Leckie and Charlton 2013).

Textbooks are available for nonstatisticians that provide relatively applied information to fit these models, and some also provide specific coding or guidance using the graphical user interface for certain statistical packages in their text or accompanying website (Brown and Prescott 2006; Dohoo et al. 2009, 2012; Lawson et al. 2003). Besides the user’s guides for specific packages, on-line sites, such as that maintained by the Institute for Digital Research and Education at the University of California, Los Angeles (https://idre.ucla.edu/stats), offer on-line tutorials for multi-level statistical modeling using a variety of software packages including SAS, SPSS, STATA, and R. Researchers with an applied knowledge of regular linear regression should be able to quickly develop an applied understanding of multi-level modeling and use this analytical approach for their research. Many of the textbooks, cited above, provide details concerning special issues associated with testing model assumptions and fit for multi-level models. Researchers are strongly encouraged to conduct and report these diagnostics as part of standard research practice.

Conclusions

Researchers in laboratory animal research and other fields need to account for clustering or autocorrelation in their data. Multi-level modeling offers a flexible framework to account for clustering in estimating the association between exposures and an outcome in a manner that results in correct statistical significance and the magnitude and direction of the effect. These models can be fitted to deal with both continuous and discrete exposures and outcomes, and can apply a variety of correlation structures.

The variance components estimated from models with random intercepts can be useful for estimating ICCs and VPCs that can provide insight on the hierarchical levels that most contribute to variability in an outcome and at which level to apply interventions. The increasing availability of less technical textbooks, software, and supporting internet materials has made these methods increasingly accessible to laboratory animal researchers.

References


Dohoo IR, Martin SW, Stryhn H. 2009, 2012; Lawson et al. 2003). With regard to discrete data, a variety of frequentist approaches exist, including quadrature-based methods and pseudo-likelihood–based methods that vary in their ability to converge (i.e., estimate the model parameters) and provide unbiased estimates of model coefficients (Goldstein 2003; Rabe-Hesketh and Skrondal 2008). The decision by most researchers in selecting an analytical approach for fitting a multi-level model is often based on their familiarity with a statistical package or the ability of the model to converge. Some researchers have recommended that multiple approaches should be investigated when dealing with discrete data, and manuscripts showing the comparative results of different methodologies using common datasets have been published (Dohoo et al. 2001; Pearl et al. 2008).


