13.2.2 GLMMs and Complex Sample Survey Data

As was mentioned above, GLMMs are specifically designed to address non-independence or "dependency" of observations. In HLMs, the dependency arises because observational units are hierarchically clustered, such as students within classrooms and classrooms with schools. In repeated measures or longitudinal models, the dependency arises because observations are clustered within individuals—for example, daily diaries of food intake for NHANES respondents, or longitudinal measures of household assets for HRS panel respondents (see Chapter 11).

This problem of lack of independence for observations should sound familiar. The survey data analysis techniques that have been the core subject of this book are designed to address the intraclass correlation among observations in sampled clusters. It is natural to draw the analogy between these two estimation problems. Take for example two parallel sets of data. The first data set records scores on a test administered to students who are clustered within classrooms, schools, and school districts. The second data set includes test scores for the same standardized test administered to a probability sample of students in their homes, where households were selected within area segments and primary sampling units (PSUs) of a multi-stage national sample design. In both cases, there is a hierarchical ordering of units—districts, schools, classes, and students in the first case, and PSUs, area segments, households, and students in the second case.

If the survey analyst was only interested in inferences concerning national student performance on the standardized test, robust inferences could be obtained using standard design-based

estimates of population parameters (such as the mean test performance) and their standard errors (see, for example, Chapter 5). School districts would define the ultimate clusters in the first data set, and PSUs would form the ultimate clusters of the second. However, the survey analyst using the first data set may have broader analytic goals. Specifically, he/she may be interested in estimating the proportion of variance in test scores that is attributable to the student, the class, the school, and the district. The analyst working with the second data set may not be especially interested in the **components of variance** associated with PSUs, secondary sampling units (SSUs) within PSUs, households, and individuals. The first analyst will likely pursue a **model-based analysis**, specifying an HLM form of the GLMM.

Following this approach to the analysis of complex sample survey data, the first analyst would specify the best possible probability model for the data set in hand (possibly using a Bayesian approach), which includes the specification of an appropriate random effects structure to reflect the multi-stage cluster sampling and estimate the variance components of interest (Kim et al., 2022). The variance of the estimated parameters in this case would be estimated with respect to the specified model, and if the weights and stratum codes were associated with the dependent variable of interest, these would also need to be accounted for in the model specification. The second analyst will be satisfied with a standard **design-based analysis**, in which weighted estimates of the parameters of interest are computed, and Taylor Series Linearization or replication estimates of the overall variance of the sampling distribution (computed with respect to the sample design) are used to develop confidence intervals and test statistics. Interested readers can refer to Hansen et al. (1983) for more discussion of the differences between model-based and design-based approaches.

Consider another problem in which a cluster sample of individuals is asked whether they have experienced any hay fever symptoms in the past week. The dependent variable for each individual is an indicator of whether they experienced hay fever symptoms ($y_{ij} = 1$) or did not ($y_{ij} = 0$), where i indexes individuals and j indexes ultimate clusters. The independent variables in the analysis might include age, gender, and an indicator of a previous allergy diagnosis. To estimate the relationship of hay fever symptoms with age, the survey analyst could choose among three approaches that all use variants of logistic regression modeling:

1) A cluster-specific model-based analysis using a GLMM, in which the logit[$P(y_{ij} = 1)$] is modeled as a function of **fixed effects** that include the constant effects of age, gender, and previous allergy diagnosis, and **random effects** of the randomly sampled ultimate clusters (enabling estimation of between-cluster variance);

2) A marginal model-based analysis using a GEE model, in which the logit model relating symptoms to the covariates is estimated using GEE methods and the covariance matrix for the model coefficients is separately estimated using the robust Huber-White **sandwich estimator** (Diggle et al., 2002); or

3) A marginal design-based analysis using a program / command / procedure such as Stata's `svy: logit` command, with a single record (row) in the data set for each individual, multiple individuals (rows in the data set) per ultimate cluster, and the ultimate clusters identified as the "clusters" for variance estimation purposes (see, for example, Chapter 8).

The GLMM analysis is a cluster-specific analysis, explicitly controlling for the random cluster effects, and would yield estimates of the fixed effects (or regression parameters) associated with the covariates *in addition to* an estimate of the variance of the random cluster effects. The GEE and `svy: logit` approaches are population-averaged (or marginal) modeling techniques and would provide comparable estimates of robust standard errors for the estimated logistic regression coefficients for the covariates. The GEE and design-based population-averaged approaches would *not* separately estimate the variance of the random cluster effects or their contributions to the total sampling variability. This is the key distinction between these alternative approaches to analyzing clustered or longitudinal data: GLMMs enable analysts to make inferences about between-subject (or between-cluster) variance, based on the variances of the subject-specific (or cluster-specific) random effects explicitly included in the models, while GEE and design-based modeling approaches are only concerned with overall estimates of parameters and their total sampling variance. See Hsu et al. (2017) for an in-depth simulation study that clarifies this distinction.

It is increasingly common to see survey samples designed to be optimal for analysis under a GLMM-type model. For example, a multi-stage national probability sample of school districts, schools, classrooms, and students could be consistent with a GLMM model that would enable the education researcher to study the influence of each level in this hierarchy on student outcomes (see Kim et al., 2022). To achieve data collection efficiency, the sampling statistician designing the sample of individuals for the hay fever study could select a primary stage sample of U.S. counties and then a sample of individuals within each primary stage county. The result would be a two-level data set with individuals nested within the sampled counties.

However, even when care is taken to design a probability sample to support multilevel or longitudinal analysis using GLMM-type models, there are several theoretical and practical issues to be addressed when fitting these models to complex sample survey data:

1. Stratification. If the stratification used in the sample selection (regions, urban/rural classification, population size, etc.) is **informative**, that is, associated with the survey variables of interest (which is typically the case in practice), fixed effects of the stratum identifiers, or at a minimum the major variables used to form the strata, will need to be included in the model.

2. Cluster Sampling. In the general sample design context, clustering of population elements serves to reduce the costs of data collection. The increase in sampling variance attributable to the intraclass correlation of characteristics within the cluster groupings is considered a "nuisance," inflating standard errors of estimates to no analytical benefit. In the context of a specific analysis involving a hierarchical linear (or multilevel) model, the clustering of observations must be reflected (or incorporated) to obtain stable estimates of the effects and variance components at each level of the hierarchy. Ideally, the sample design clusters may be integrated into the natural hierarchy of the GLMM and the cluster effects on outcomes can be directly modeled using additional levels of random effects—say nesting students within classrooms, classrooms within schools, and schools within county PSUs.

3. Weighting. Conceptually, one of the more difficult problems in the application of GLMMs to complex sample survey data is how to handle the survey weights. Theoretically, if the weights

were **noninformative**, they could be safely omitted from the model estimation. If the sample design and associated weights are **informative** for the analysis, fixed effects of the variables used to build the weights or appropriate functional forms of the weight values themselves could be included in the model (DuMouchel and Duncan, 1983; Little, 1991; Korn and Graubard, 1999, Section 4.5; Fuller, 2009, Ch. 6). However, there are other complications with weighting in GLMMs, as we discussed in Chapter 11. Attrition adjustments that are often included in longitudinal survey weights can yield different weight values for each measurement time point. Clusters in a multistage sample design may not enter the sample with equal probability. Even in a national equal-probability multi-stage sample of students, the most efficient samples require that counties and school units are selected with probability proportionate to size (PPS). Conditional on a given stage of sampling, the observed units would enter with varying probability. What role should this information play in the estimation of GLMMs?

4. Subpopulations. As was discussed in Chapter 4, *unconditional* subpopulation analyses are important when conducting design-based analyses of complex sample survey data and focusing inference on subclasses of interest. Do the same issues still apply when fitting GLMMs? For example, if one wanted to make inference about variance components due to classrooms and schools in lower-income school districts exclusively, and district income was not an explicit stratification variable in the original sample design, how should one approach the analysis? Koziol (2019) addressed this question directly via simulation and found that the answer depends on the sizes of the subpopulation samples in the clusters (assuming that the sample design is informative, and weights should be used). When the sample sizes are small (e.g., less than 15 per cluster), standard design-based approaches to the subpopulation analysis ("single-level"

approaches) are recommended (e.g., the regression models described in Chapters 7-9). When the sample sizes are larger, multilevel approaches or single-level approaches can be used.

In this section, we focus on approaches for fitting GLMMs to cross-sectional complex sample survey data that use the weights associated with different randomly sampled units to define the (pseudo) likelihood function for the observed data. These approaches are identical to the approaches outlined in Chapter 11 for weighted multilevel modeling, with the difference being that Level 1 units are generally survey respondents (rather than different time points), and the Level 2 units are generally ultimate clusters (rather than individuals in a longitudinal survey):

- Survey weights for the Level 1 respondents need to reflect probabilities of selection *conditional* on the higher-level ultimate cluster to which the respondent belongs being selected into the sample (including possible adjustments for nonresponse);
- The (adjusted) conditional weights for Level 1 respondents need to be appropriately *scaled* within Level 2 ultimate clusters, with the "normalizing" approach (which ensures that the sum of the scaled weights within a given ultimate cluster is equal to the sample size for that cluster) generally serving as a good multi-purpose approach for estimation (Pfeffermann et al., 1998; Rabe-Hesketh and Skrondal, 2006; Carle, 2009);
- Survey weights for the sampled Level 2 ultimate clusters, reflecting the probabilities of selection for the clusters, need to be accounted for in the (pseudo) likelihood function as well (much like the "baseline" weights for sampled individuals at Level 2 in Chapter 11); see Pfeffermann et al. (1998), Rabe-Hesketh and Skrondal (2006), Rao et al. (2013), or Kim et al. (2022) for the technical details underlying this requirement.

We use the weights associated with different levels of the data hierarchy to compute unbiased estimates of both the fixed effects and the variance components in a GLMM (see West et al., 2015, for a case study). Ideally, such estimates will be unbiased with respect to both the sample design and the model being specified (Pfeffermann, 1993). A failure to use the weights associated with Level 2 ultimate clusters, for example, could lead to biased estimates of the variance components describing between-cluster variability in the larger population. This makes it essential for these weights to be included in a survey data set if analysts of those data will be interested in fitting GLMMs. The use of the weights in estimation also provides some protection against model misspecification, as discussed in Chapter 7; estimates of the fixed effects and variance components in a GLMM will still be unbiased with respect to the sample design if a model has been misspecified.

Importantly, the weights associated with Level 1 respondents need to be *conditional* and *scaled*, as indicated above. A failure to use these conditional, scaled weights (e.g., if an analyst were to use the *overall* survey weights reflecting all stages of sample selection that are typically provided in public-use survey data sets) could lead to bias in the estimates of multilevel model parameters (Pfeffermann et al., 1998). If a data producer is willing to provide both Level 1 and Level 2 weights, *in addition to* joint probabilities of inclusion for Level 1 and Level 2 units, then inferences can be improved even further (Rao et al., 2013).

When using the weights associated with all randomly sampled units to define the pseudo-likelihood function used to estimate the parameters in a GLMM, the variances of the maximum likelihood estimates can be computed using a variation of Taylor Series Linearization developed

by Binder (1981, 1983) for generalized linear models fitted to complex sample survey data (see Chapter 8). This variance estimation approach can explicitly account for multistage stratified cluster sampling, and is sometimes referred to as a robust, sandwich-type approach to computing standard errors in the software that implements these approaches. Pfeffermann et al. (1998) and Rabe-Hesketh and Skrondal (2006) provide technical details regarding these variance estimation approaches.

Recent years have seen additional developments related to optimal estimation approaches for analysts interested in fitting these types of GLMMs to complex sample survey data. These have included Bayesian approaches for implementing the methods described above. do N. Silva and da S. Moura (2022) and Savitsky and Williams (2022) describe Bayesian approaches to the survey-weighted estimation of GLMMs and demonstrate via simulation and an empirical evaluation that this approach has improved inferential properties relative to the other approaches introduced above. The authors provide data and code demonstrating how to implement this approach using the `rstan` package in the R software, and we have included this code on the ASDA website.

13.2.3 Alternative Approaches to Fitting GLMMs to Survey Data: The PISA Example

We now consider an example of fitting a GLMM to cross-sectional survey data arising from a complex sample design. These data come from the 2000 Programme for International Student Assessment (PISA) and were also analyzed by Rabe-Hesketh and Skrondal (2006) using the `gllamm` command that they developed for Stata (see www.gllamm.org). The dependent variable of interest in this case (ISEI) is continuous and measures the socio-economic status (SES) of a

given student. The predictor variable of interest in this case is COLLEGE (an indicator of whether the highest level of education for either parent is college). Also included in the data set (freely available online for Stata users) is an ID code for the sampled school (the PSU in the case of this sample design), the overall final student weight (not conditional!), or W_FSTUWT, and a final school weight (WNRSCHBW). The data can be loaded into Stata's working memory using the following command:

```
use http://www.stata-press.com/data/r13/pisa2000
```

We desire to fit the following GLMM to these data:

$$ISEI_{ij} = \beta_0 + \beta_1 COLLEGE_{ij} + u_{0j} + u_{1j} COLLEGE_{ij} + e_{ij}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix} \equiv D \right)$$

$$e_{ij} \sim N(0, \sigma^2)$$

We note that this model includes two random effects (u_{0j} and u_{1j}), allowing each school (denoted by j) to have a unique intercept and a unique relationship of COLLEGE with ISEI. The random effects are assumed to follow a bivariate normal distribution, with mean vector 0, unique variance components for each random effect, and zero covariance. The random errors are assumed to follow a normal distribution with constant variance. We desire to compute weighted estimates of the two fixed effects (β_0 and β_1) and the two variance components (σ_0^2 and σ_1^2) for the PISA target population. The intercept represents the expected ISEI for students who do not have any college-educated parents, the fixed effect for COLLEGE represents the change in the

expected ISEI for students with college-education parents, and the variance components describe variation among the schools in the intercepts and the college effects, respectively.

We first use the `mixed` command in Stata (appropriate for the approximately normal dependent variable) to fit a standard GLMM for comparison purposes, completely ignoring the weights in the estimation:

```
mixed isei college || id_school: college, ///
   covariance(independent) variance
```

We note in this syntax that the dependent variable (ISEI) is listed first, followed by the predictor variable. The two vertical bars indicate the first level of clustering, which is defined by the values of ID_SCHOOL. The inclusion of COLLEGE after a colon following the cluster identifier means that the coefficient for this predictor is allowed to randomly vary across clusters (schools), and a random intercept will be included in the model by default. The covariance structure for the two random effects is independent (meaning that they have zero covariance), and estimates of variance components are desired in the output (instead of the default standard deviations). The estimates of the parameters in this model are included in Table 13.1.

Table 13.1 Parameter estimates (and standard errors in parentheses) from the different multilevel modelling approaches for the 2000 PISA data

Parameter	No Weights	Weights as a Covariate	Scaling Method 1: Effective	Scaling Method 2: Size

FIXED EFFECTS				
Intercept	38.79 (0.62)	36.82 (1.14)	35.89 (0.91)	35.89 (0.91)
COLLEGE	12.65 (0.90)	12.60 (0.90)	14.28 (1.42)	14.28 (1.42)
WEIGHT		<0.01 (<0.01)		
VARIANCE COMPONENTS				
Var(Intercepts)	16.14 (5.45)	13.94 (5.22)	17.74 (6.43)	17.79 (6.43)
Var(College)	43.12 (10.85)	42.26 (10.58)	41.03 (13.74)	41.06 (13.73)
Var(Residuals)	219.33 (7.20)	219.92 (7.22)	214.96 (12.82)	214.92 (12.84)
Pseudo Log(L)	-8611.88	-8609.91	-1439307.8	-1443258.0

Next, we fit a model including a fixed effect of the final overall student weight, but not using the weights to estimate the parameters of interest:

```
mixed isei college w_fstuwt || id_school: college, ///
   covariance(independent) variance
```

Estimates of the parameters in this model are included in Table 13.1 as well.

We now fit the model of interest using the weights to define the pseudo-likelihood function (and thus computing weighted estimates of the parameters of interest). We first compute the conditional student weight by dividing the final student weight (accounting for all probabilities of selection) by the final school weight. (In fact, one can show that this step is not necessary if the Level-1 weights are normalized within clusters; one could obtain the same results in this case

by simply inputting the overall respondent weight into the command, *provided that the appropriate option for normalizing these weights is included as well*.) We then fit the model of interest, specifying 1) the conditional student weights at Level 1, via the `[pweight = …]` option; 2) the final school weights at Level 2, via the `pweight()` option after the cluster identifier; and 3) the weight scaling method, which in this case is "`size`" for the "normalizing" method. We also fitted the model using the "`effective`" weight scaling method ("Method 1" in Pfeffermann et al., 1998), and the results were virtually identical (as suggested by Carle, 2009).

```
gen conwt = w_fstuwt / wnrschbw
```

```
mixed isei college [pw = conwt] || id_school: college, ///
covariance(independent) variance pweight(wnrschbw) pwscale(size)
```

The results from fitting this model are also included in Table 13.1. Examining the results from the different approaches, we can make the following conclusions:

- Parental college education has a strong positive relationship with mean SES, regardless of the method used;
- The weighted estimates of the fixed effects are different from the unweighted estimates, especially so for the intercept (i.e., the mean for students with non-college educated parents);
- Including the final student-level weight as a covariate changes interpretation of the parameters: the intercept now corresponds to the mean for someone with weight equal to

zero, which is meaningless (using weight deciles, for example, may be a more effective approach);

- The weighted estimates of the variance components also differ:
    - There is more evidence of variability across schools in the means for students with non-college educated parents (i.e., the random intercepts) when computing weighted estimates; and
    - There is less variability in college vs. non-college gaps (i.e., the random coefficients) across the sampled schools.
- Different weight scaling methods do not result in different estimates or conclusions; in this case, one would use the "size" method (Method 2), per Carle (2009); and
- The robust, sandwich-type standard errors for the weighted estimates are generally larger, as expected, but this does not change the overall inferences.

We provide code for fitting the same model using other software on the ASDA website. At the time that this third edition is being written, LMMs for continuous dependent variables like the model fitted above (dropping the 'G') can also be fitted to complex sample survey data using procedures available in R. The contributed `svylme` package, developed by Thomas Lumley, and the contributed `pwlmm` package, developed by Veiga et al. (2014), currently only fit LMMs. Mplus, SAS (PROC GLIMMIX), and the HLM and MLwiN software packages are each capable of fitting GLMMs to data from complex samples.

This worked example has illustrated the importance of examining the sensitivity of GLMM results to alternative estimation approaches. In this example, our inferences were generally quite

similar whether the weights were used or not, suggesting that these weights were not especially informative about the parameters in this model. This will not always be the case for different surveys (e.g., West et al., 2015) or different analyses of the same data set. For example, a series of simulation studies by Zheng and Yang (2016) in an **item response theory (IRT)** context demonstrates the importance of accounting for survey weights when using multilevel models for IRT investigations. Existing software makes it very straightforward to perform these kinds of analyses examining the sensitivity of estimated parameters in multilevel models to the use of survey weights (at each level) for estimation.

In addition to assessing changes in parameter estimates due to the use of weights, assessment of changes in estimated standard errors when fully accounting for the complex sampling features is also important. In this regard, Stapleton and Kang (2018) evaluate design effects for weighted parameter estimates in multilevel models fitted to educational survey data. These authors found that ignoring complex sampling features (i.e., stratification and cluster sampling) in variance estimation *above and beyond* the levels being modeled (e.g., fitting a weighted multilevel model with students nested within schools, but ignoring higher-level primary sampling units, such as school districts or counties; see Rabe-Hesketh and Skrondal, 2006) generally has a minor impact on inferences, and that accounting for the fixed effects of stratum identifiers at Level 2 of a given multilevel model (for higher-level clustering units, such as schools) is a reasonable strategy for inferential purposes.