11.3.3.1 Example: Weighted Multilevel Modeling with Error Covariance Structures

We now consider an illustration of fitting a multilevel model to these data using the alternative approach described by Veiga et al. (2014). Doing this requires downloading, modifying, and running a Mata program from GitHub, and then running this program in Stata.[1] Before we do this, we start by opening the original HRS data set in wide format (with one row per individual), creating four new measures of log-transformed income in preparation for the generation of a vertical data file, and creating an indicator variable identifying cases with complete data on income:

```
gen ln_inc1 = ln(H8ITOT + 1)
gen ln_inc2 = ln(H9ITOT + 1)
gen ln_inc3 = ln(H10ITOT + 1)
gen ln_inc4 = ln(H11ITOT + 1)

* Identify cases with complete data.
gen comp = (ln_inc1 != . & ln_inc2 != . & ln_inc3 != . /// &
ln_inc4 != .)
```

Next, we fit a response propensity model to these data, modeling the probability of having complete data across the four waves as a function of covariates in the baseline wave (2006) and accounting for the complex sampling features:

[1] https://github.com/AlinneVeiga/MATA-code

```
* Fit a response propensity model predicting the
* probability of providing income data at all 4 waves.

* Modal imputation of missing covariate values.
replace selfrhealth_06 = 3 if selfrhealth_06 == .
replace marcat_06 = 2 if marcat_06 == .
replace diabetes_06 = 0 if diabetes_06 == .
replace arthritis_06 = 1 if arthritis_06 == .
replace racecat = 2 if racecat == .
replace edcat = 2 if edcat == .

svyset secu [pweight = kwgtr], strata(stratum)
svy: logit comp ln_inc1 i.selfrhealth_06 age_06 ///
   i.marcat_06 diabetes_06 arthritis_06 i.racecat i.edcat
predict phat, p
```

As an alternative to the response propensity modeling approach, one could also consider the calibration approach described in Section 11.2.3.5, which is designed to make cases with complete data across the waves representative of the target population.

Next, we adjust the baseline survey weights in 2006 for cases with complete data by the inverse of the predicted probability of having complete data across the four waves (accounting for the cases that did not have complete data). We then create a new unique cluster ID variable that combines the HRS stratum and cluster codes, create a unique ID variable for each respondent,

and reshape the wide data set of cases with complete data into a vertical format after keeping only the variables of interest. We note that the cluster-specific weights required by the Veiga et al. approach are set to 1 in this case, given that weights for each of the HRS clusters are not provided in the public-use data files:

```
* Adjust baseline sampling weights in 2006 by inverse of
* predicted probability of complete data.
gen indwgt = kwgtr * (1 / phat)

* Set cluster-specific weight equal to 1 (HRS SECU weights * not
provided in the public-use data).
egen newclust = group(stratum secu)
gen clustwgt = 1

* Create unique person ID.
egen newid = concat(hhid pn)
destring newid, gen(newid_num)

* Only keep cases with complete data.
keep if comp == 1

* Only keep variables of interest.
keep newid_num newclust gender ln_inc1-ln_inc4 indwgt ///
clustwgt
```

```
* Reshape data to vertical format.
reshape long ln_inc, i(newid_num) j(wave)
```

Now, in the vertical data set, we create unique indicator variables for each wave (required by the Veiga et al. approach), in addition to an indicator of being a male respondent and a constant variable equal to 1 for all cases:

```
* Create unique indicators for each wave.
gen wave1 = (wave == 1)
gen wave2 = (wave == 2)
gen wave3 = (wave == 3)
gen wave4 = (wave == 4)

* Compute constant variable for Veiga et al. macro.
gen cons = 1

* Recode gender.
gen male = (gender == 1)
```

We are now ready to modify the Mata program provided by Veiga et al. (2014) so that it recognizes the number of HRS waves (4) and the covariance structure for the repeated income measures that we wish to estimate in our model (Toeplitz). Following this approach, we will be estimating the variance of the random cluster intercepts, in addition to the constant variance and

three covariance parameters defining the Toeplitz covariance structure for the four waves of data (constant variance among individuals over time, and common covariances for all measures one, two, and three waves apart). In total, there are therefore *s* = 5 variance-covariance parameters that we wish to estimate, and we need to modify two lines of the Veiga et al. Mata program that define *s* (near the beginning and near the end; see the program pwigls_genlin_adcv_modAV1.do on the ASDA website):

```
s = 5
```

The next modification that we need to make is the definition of the five "delta" matrices that will be combined to define the Toeplitz covariance structure, as described in the Veiga et al. (2014) paper. There are actually five symmetric 4 by 4 matrices (for the four waves) that need to be defined, but the first matrix is all zeroes:

```
delta_matrix=J(s,16,0)

delta_matrix[2,]=
(1,0,0,0,
 0,1,0,0,
 0,0,1,0,
 0,0,0,1)

delta_matrix[3,]=
(0,1,0,0,
```

```
 1,0,1,0,
 0,1,0,1,
 0,0,1,0)

delta_matrix[4,]=
(0,0,1,0,
 0,0,0,1,
 1,0,0,0,
 0,1,0,0)

delta_matrix[5,]=
 (0,0,0,1,
 0,0,0,0,
 0,0,0,0,
 1,0,0,0)

rowshape(delta_matrix[1,],4)
rowshape(delta_matrix[2,],4)
rowshape(delta_matrix[3,],4)
rowshape(delta_matrix[4,],4)
rowshape(delta_matrix[5,],4)

name1 = tokens(varlist)
```

```
name3 = ("Sigma_u_2","Genlin(1)","Genlin(2)", ///
"Genlin(3)","Genlin(4)","Genlin(5)")
```

Following the notation of this program, “Sigma_u_2” is the variance of the random cluster intercepts, and the remaining “Genlin” parameters (1 through 4) are the variance and covariance parameters describing the Toeplitz structure (5 is not relevant here).

Finally, we need to modify each line of the Mata program defining the Toeplitz covariance structure, given the number of waves. We note that we refer to the four non-zero “delta” matrices defined above in these expressions:

```
theta_genlin=theta[2,]*rowshape(delta_matrix[2,],4)+
theta[3,]*rowshape(delta_matrix[3,],4)+
theta[4,]*rowshape(delta_matrix[4,],4)+
theta[5,]*rowshape(delta_matrix[5,],4)
```

```
theta0_genlin=theta0[2,]*rowshape(delta_matrix[2,],4)+
theta0[3,]*rowshape(delta_matrix[3,],4)+
theta0[4,]*rowshape(delta_matrix[4,],4)+
theta0[5,]*rowshape(delta_matrix[5,],4)
```

These modifications are needed in 5 places in the Mata program, and these are clearly commented in the file pwigls_genlin_adcv_modAV1.do on the ASDA website.

Once these modifications have been made and the updated Mata program has been saved, we can run the Mata program in Stata (assuming that the .do file has been saved in the current working directory) to define the function that will be used to fit the model, and sort the data set by cluster ID, individual ID, and wave:

```
/*** Running the .do file with the mata function ***/
do "pwigls_genlin_adcv_modAV1.do"

sort newclust newid_num wave
```

Next, we declare two global macro variables, one ("wav") containing the four wave indicators and a second ("x") containing the covariates of interest in the model (the four wave indicators and the male indicator). Recall that we are fitting a model with no intercept, and estimating the means for each of the four waves:

```
gl wav wave1 wave2 wave3 wave4
gl x wave1 wave2 wave3 wave4 male
```

Finally, we fit the model of interest using the function defined by the Mata program. For illustration purposes, we fit the model to a random subset of HRS financial reporters with numeric IDs less than 20,000,000. (Fitting the model to the entire data set simply takes longer for convergence.) In order, the arguments to the function are: the covariates, the dependent variable, the wave indicators, the constant term, the cluster ID, the cluster weights, and the individual weights (the program will automatically scale the individual weights so that they are normalized within clusters):

```
keep if newid_num < 20000000
```

```
mata: pwigls_genlin_adcvw4toep1("$x", "ln_inc", "$wav", "cons",
"newclust", "clustwgt", "indwgt")
```

The output generated by the model fitting function is shown below:

```
------------------------------------------------------------------------------
           Probability Weighted Iterative Generalized Least Squares
------------------------------------------------------------------------------
General Information

Response Variable =              ln_inc
Weight at Level 2 =            clustwgt
Weight at Level 1 =              indwgt

Start running on 27 Mar 2024 at  22:59:41
Number of Iterations    =   7
Number of Time points   =   4
Number of Level 1 units = 666
Number of Level 2 units =  67

------------------------------------------------------------------------------
       Fixed Effects|   Coef.    Std.Err.    z     P>|z|   [95%Conf.Interval]  Init.Val.
--------------------+---------------------------------------------------------
              wave1 |  10.2636   .064375  159.43   0.000   10.1374   10.3898   10.2732
              wave2 |  10.2596    .06362  161.26   0.000   10.1349   10.3843   10.2693
              wave3 |  10.1561   .068067  149.21   0.000   10.0226   10.2895   10.1657
              wave4 |  9.95694   .067421  147.68   0.000   9.82479   10.0891   9.96659
               male |  .427748    .08468    5.05   0.000   .261778   .593717   .425256
------------------------------------------------------------------------------

------------------------------------------------------------------------------
 Variance Components|   Coef.    Std.Err.    z     P>|z|   [95%Conf.Interval]  Init.Val.
--------------------+---------------------------------------------------------
          Sigma_u_2 |  .059156   .025703    2.30   0.021   .016878   .101435        .5
          Genlin(1) |  1.36033   .217148    6.26   0.000   1.00315    1.7175   1.30586
          Genlin(2) |   .70801   .174523    4.06   0.000   .420944   .995075        .5
          Genlin(3) |  .710341   .175359    4.05   0.000   .421902    .99878        .5
          Genlin(4) |  .629109    .17227    3.65   0.000    .34575   .912469        .5
------------------------------------------------------------------------------


 General Linear Matrix
[symmetric]
                1             2             3             4
    +-------------------------------------------------------+
  1 |  1.360328393                                          |
  2 |  .7080096706   1.360328393                            |
  3 |  .7103411129   .7080096706   1.360328393              |
  4 |  .6291093613   .7103411129   .7080096706   1.360328393  |
    +-------------------------------------------------------+


 Total Variance
[symmetric]
                1             2             3             4
    +-------------------------------------------------------+
  1 |  1.419484785                                          |
  2 |  .7671660627   1.419484785                            |
  3 |   .769497505   .7671660627   1.419484785              |
  4 |  .6882657534    .769497505   .7671660627   1.419484785  |
    +-------------------------------------------------------+
```

```
------------------------------------------------------------------------------
Note: Robust Standard Errors
```

In the output above, we see the number of sampled financial reporters (666), the number of HRS sampling error computation units, or clusters (67), the cluster and individual weight variables, the dependent variable, and then weighted population estimates of the fixed effects and the covariance parameters. In addition, we see robust standard errors for the weighted estimates that are estimated with respect to both the model and the sample design (as described earlier). Examining the weighted estimates of the fixed effects and the corresponding 95% confidence intervals, we see evidence of similar mean income in the first three waves, followed by a more substantial decline in the fourth wave (2012). The gender gap is once again apparent, with males tending to have significantly higher mean income.

We also note weak evidence of unexplained variance between HRS sampling error computation units (or clusters), denoted by Sigma_u_2. However, recall that we assumed that each cluster had an equal probability of selection (which may be resulting in a biased estimate of this variance component). In other applications where cluster-specific sampling weights are available, these weights should certainly be used in estimation. Finally, we note the weighted estimates of the between-individual variance at each wave (1.36) and the covariances of the individual effects at each wave (e.g., adjacent observations have an estimated covariance of 0.71). These covariances seem to be similar regardless of the time between a given pair of waves. Other covariance structures could certainly be considered, as discussed by Veiga et al. (2014); for example, the between-individual variance could be allowed to vary depending on the wave.

The weighted multilevel modeling approach with error covariance structure modeling included that is illustrated in this example is also possible in R. In 2022, Veiga and colleagues published the `pwlmm` contributed package[2] for R, which enables the exact same type of modeling shown above using the Mata code (in addition to more standard two-level weighted multilevel modeling). At the time that the third edition is being written, this package only enables this type of modeling for continuous outcomes (the "linear" case). We will provide any updates along these lines on the ASDA website.

[2] https://cran.r-project.org/web/packages/pwlmm/index.html