Multilevel Health Disparities Data Analysis: Part 1 – Continuous Outcomes

Tor Neilands, PhD

Jennifer Toman, PhD

Analysis Core

Center for Aging in Diverse Communities (CADC)

November 13, 2024

Goals of Today's Talk

- Introduce considerations for inferential analyses prompted by outcome data with a clustered/multilevel structure.
- Introduce three commonly used estimation methods for conducting multilevel inferential analyses:
  - 1) Cluster-adjusted robust standard errors (SE)
  - 2) Generalized estimating equations (GEE)
  - 3) Multilevel models (MLM) (also known as mixed effects, random effects, and hierarchical linear models [HLM])
- Demonstrate the application of these methods to a CADC scientist's health disparities research dataset.

Challenges in Health Disparities Research

- Jeffries et al (2019, p. S28) lay out four broad and complementary types of methods they recommend for health disparities research to optimize research findings' rigor and impact:
  1. study design and analytical methods that maximize the ability to draw causal inferences from observational data,
  2. modeling techniques that account for the multilevel nature of health disparity causes,
  3. complex systems and simulation methods for modeling dynamic relations, and
  4. qualitative and mixed methods that allow a better understanding of relationships that cannot be achieved using quantitative methods alone.

Multilevel Data

- Health disparities research is inherently complex (Jeffries et al. 2019).
- *Multilevel data* structures are one way in which the complexity of health disparities research is manifested.
- *Multilevel* or *clustered data* refers to data in which observations, represented by rows in a dataset, are grouped or organized together in a meaningful way that **includes** correlation across observations/rows in the dataset.
- Examples:
  - Education: children within classrooms or schools
  - Medicine: patients within providers, hospitals, wards, or clinics
  - Public Health: survey respondents with neighborhoods or venues

Descriptive Analysis of Multilevel Data

- **Descriptive analyses** (e.g., frequencies, means) usually don't require adjustments. Give thought to which units of analysis you want to analyze (e.g., clinic level, patient level, or both).
- One exception: If you are working with data originating from a probability or complex sampling approach, you may need to account for one or more of the following:
  - Weighting
  - Primary (e.g., clinics) and secondary (e.g., patients) sampling units
  - Stratification

  Fortunately, Stata's -svy- prefix commands typically make this process fairly painless. You first use -svyset- to inform Stata of the weight, sampling unit, and/or stratification variables and then prefix analysis commands using the -svy- prefix. For instance, -tab race- would become -svy: tab race-.

Inferential Analysis of Multilevel Data

- **Inferential analysis** of multilevel data presents challenges because most inferential analysis approaches assume rows in the dataset originate from independent subjects.
- Concrete example: standard linear regression models assume residuals from observations have a constant variance and are *uncorrelated*.
- However, observations from multilevel data structures are correlated, so their residuals are also often *correlated*.
- Depending on the research setting, the multilevel structure of the data may be quite complex. For example, patients might be nested within multiple providers at different hospitals. For simplicity, in this presentation, we'll assume a simple hierarchical nesting structure (e.g., a single clinic per patient).

General Linear Models (GLM) Review

- To understand why correlation of observations across rows of data in a dataset is a problem for traditional analysis, methods, let's take **linear regression** as an example.
- Linear regression is a specific case of the general linear model (GLM). GLMs express an outcome (dependent) variable **y** as a function of **Xb** + **ε**, where **X** is a design matrix of fixed effects, **b** are estimated regression coefficients, and **ε** is a vector of residuals.
- **ε** is assumed to be normally distributed with a mean of zero and variance σ^2.

GLM Model Structure

- GLM formulation:

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon}$$

where:

- **y** is a vector containing the outcome scores
- **X** is a design matrix containing the design (i.e., predictor or independent) variables
- **b** is a vector containing the parameter estimates
- **ε** is a vector containing the residual values
- The values of X are *fixed* by the design of the study, so we refer to the estimates contained in the vector **b** as the *estimates of fixed effects.*

GLM Assumptions

- GLMs typically have three main assumptions:
  - Residuals are normally distributed
  - Residuals' variances are constant across levels of the outcome (regression) or groups (ANOVA)
  - **Residuals are independent of each other**
- The independence assumption is problematic for multilevel data because measures from participants (e.g., patients) in the same cluster (e.g., clinic) are (usually) not independent.
- If we are willing to assume any dependence among observations within clusters is linear, it can be represented using either *covariances* (unstandardized metric) or *correlations* (standardized metric).

Structure of **ε** in a GLM

- If we had five observations within a cluster and fitted a GLM, what would the structure of **ε** residuals be? In matrix form, it would look like this:

Observations	Data	Model
Independent (not clustered)	$\sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$	$\sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
Correlated (clustered)	$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 & \sigma_{45} \\ \sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_5^2 \end{bmatrix}$	$\sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

- **Main point:** The GLM residuals structure is correctly specified for independent data but not correctly specified for clustered data!

Consequences of Correlation

What can go wrong if we analyze correlated clustered data using the wrong model (e.g., ANOVA, regression, GLM)?

- If observations within clusters are positively correlated, the variances for between-cluster effects in our example will be *underestimated* (Dunlop, 1994). Standard errors will be *too small*.
- On the other hand, for within-cluster effects, the variances will be *overestimated* (Dunlop, 1994). Standard errors will be *too large*.
- These inaccurate variance estimates will lead to increased Type I (false positive) and Type II (false negative) errors, respectively.
- We might erroneously conclude that a predictor had a significant effect when it actually didn't, or vice versa, **leading us to draw incorrect substantive conclusions.**

Addressing Correlated Observations from Multilevel Data: Three Approaches

- If standard regression based on the GLM yields untrustworthy standard errors and thus tests of significance, **what methods can we use instead to obtain accurate inferences when analyzing multilevel data?**
- We will describe three popular approaches:
  - (1) Cluster-adjusted robust standard errors (SE) and test statistics (a.k.a. independent estimating equations [IEE])
  - (2) Generalized estimating equations (GEE)
  - (3) Multilevel models (MLM)

(1) Cluster-Adjusted Robust SE (IEE)

- In a linear regression model, a general ("sandwich") formula to determine the variance-covariance matrix of the parameter estimates, **b**, is:

$$\mathbf{s}_{\mathbf{b}}^2 = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{\Phi}\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

  where $\mathbf{\Phi} = \mathbf{e}\mathbf{e}^T$ is the variance-covariance matrix of the vector of *n* residuals between the predictions from the regression equation and the observations
- When the residuals meet the constant variance assumption, the variance in the residuals at each value of *x* is estimated by the same value, diag[$s^2_{b_i b_j}$].
- However, when the residuals' variance is not constant (heteroskedasticity), a **robust estimator** is available where **Φ** is computed as diag[e^2_i], which is a diagonal matrix of squared empirical residuals, which is then adjusted for degrees of freedom by multiplying it by $n/(n-p-1)$ for *p* predictors.
- There is a **cluster-adjusted version of this estimator** that computes the residuals in e^2_i **based on cluster-level residuals instead of individual-level residuals**.

(1) Cluster-Adjusted Robust SE Properties

- The robust and cluster-adjusted variance estimators are *statistically consistent*, meaning that their estimates approach the true population parameters as the sample size increases.
- Therefore, this estimator works best when the number of clusters is large (e.g., 50 or larger).
  - Modified versions have been proposed for smaller numbers of clusters. See Hayes & Cai (2007) for details, including formulas. Stata has implemented some of these for linear regression models (e.g., HC3).
- When using a robust variance estimator, the parameter estimates will be the same as those from an analysis using the default model-based estimator. **Only the standard errors, Z-tests, p-values, and confidence intervals will change**.

(1) Cluster-Adjusted Robust SE

- Analyses using cluster-adjusted robust variance estimators will tend to have few convergence challenges and converge quickly, making them especially well-suited to analyses involving large numbers of clusters and observations where other methods will not converge and/or exhibit prohibitively long run times.
- Robust and cluster-adjusted robust SEs can also be used in combination with more sophisticated statistical modeling approaches that don't assume the observations within clusters are uncorrelated (e.g., GEE and multilevel models, described next).
- They may work especially well when cluster sizes are informative (i.e., when the size of the clusters influences the outcome).

(2) Generalized Estimating Equations (GEE)

- The cluster-adjusted robust variance estimation approach is widely available and computationally fast. However, it ignores information in the *correlation of residuals* that could help to minimize the variance of the parameter estimates (i.e., improve *statistical efficiency*).
- *Generalized estimating equations* (GEE) addresses this limitation by using methods similar to ordinary least squares paired with an iterative weighting approach to **estimate a working correlation matrix *R* among the residuals**.
- GEE was originally developed for longitudinal data in which multiple repeatedly measured observations are clustered within persons. However, it can also be used to analyze multilevel data.
- See Hanley et al. (2003) for a gentle introduction to GEE.

(2) More about GEE

- Estimates from GEE are statistically consistent even if *R* is misspecified.
- The closer *R* is to the true underlying correlation structure among observations, the less variance the estimates will have.
- A typical choice for *R* in a GEE analysis of multilevel data is *exchangeable*, which assumes each observation is **equally correlated** with every other observation within the same cluster.
- While model-based standard errors are available from GEE, research has shown that cluster-adjusted standard errors will tend to perform better than model-based standard errors.
- Robust SE versions that work better in samples with smaller numbers of clusters are available to Stata users via the -xtgeebcv- community-contributed command. See Gallis et al. (2020).

(3) Multilevel Models (MLM)

- The cluster-adjusted robust variance method ignores the correlation among observations within the same cluster, except to compute empirical standard errors based on cluster-level residuals.
- GEE improves on the cluster-adjusted robust variance approach by estimating the working correlation matrix *R*, though *R* is still treated as a nuisance.
- **What if we wanted our model estimates to explicitly estimate cluster-level variability?** What if we also wanted our model to estimate the fixed effects conditional on cluster-level variability?
- MLMs enable the estimation of cluster-level variability and estimate fixed effects in the presence of cluster-level variability. **For continuous outcomes, we can fit MLMs via *linear mixed models* (LMMs).**

GLM Recapitulation

- GLM formulation:

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon}$$

  where:

  - **y** is a vector containing the outcome scores
  - **X** is a design matrix containing the design (i.e., predictor or independent) variables
  - **b** is a vector containing the parameter estimates
  - **ε** is a vector containing the residual values
  - The values of X are *fixed* by the design of the study, so we refer to the estimates contained in the vector **b** as the *estimates of fixed effects.*

Fixed vs. Random Effects

- The predicted value of **y** = **Xb** + **ε** is **y** = **Xb.** Therefore, **ε** = **y** – **Xb.**
- **X** is the same for every participant and so is **b**, so the predicted value is the same for everyone in each **X** and **b** combination. But **ε** will be different for each participant on a participant-to-participant basis.
- Since the value of **ε** differs for each person, we can build a *distribution* for it. GLM assumes the distribution of **ε** is normal with variance σ^2.
- So, now we have two sources of variance in outcome scores:
  - Variance due to the fixed effects under our control from **Xb**
  - Variance in the deviations **ε** = **y** – **Xb** due to participant-to-participant variability, which we assume is random
- Thus, we treat the participant-to-participant variability as a *random* effect. This is an important concept for mixed models, which are covered next.

Linear Mixed Model (LMM) Formulation

- The LMM in matrix notation is:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}$$

where:

- **y** is a vector containing the outcome scores.
- **X** is a design matrix containing the dummy or other (e.g., effect) coded fixed effect variables. **b** is a vector containing the parameter estimates for the fixed effects.
- **Z** is a design matrix containing *random effect* variables (e.g., dummy variables for each cluster). **u** is a vector containing the parameter estimates for the random effects.
- **ε** contains the within-cluster residual values.

LMM Formulation Benefits

- Why is the LMM formulation better than the GLM formulation (i.e., structure) for analyses of clustered data?
- Recall earlier we noted that in an analysis with *independent* observations, there are two sources of variability in outcome scores: variance explained by fixed effects and participant-to-participant variability, which is randomly distributed with a mean of zero and variance σ_e^2.
- In an analysis of clustered data, however, we have *three* sources of variability: variance explained by fixed effects, within-cluster variance, and between-cluster variance.
- The **between-cluster variance source is new**.

Accounting for Between-Cluster Variance

- We could account for between-cluster variance by including a dummy variable for each cluster in **X** along with the dummy variables for the other fixed effects of interest.
- Conceptually, that is what a fixed effects analysis does even though we don't see it because software programs take care of creating the dummy variables.
- A limitation with this approach is that both the cluster dummy variables and the dummy variables for the fixed effects of interest are stored in the same design matrix **X**.
- But they need to be treated differently because the effects of interest are fixed effects whereas the cluster dummy variables comprise one or more random effects due to the (theoretical) sampling of the clusters from a broader population.

Modeling Between-Cluster Variance

- LMMs overcome this limitation by placing random effects in **Z**, not **X**.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}$$

- Because the between-cluster dummy variables are in the separate design matrix **Z**, the LMM can treat them differently from the fixed effects of interest stored in the design matrix **X**.
- Since we assume clusters are randomly selected from some broader population, we will treat their effects similarly to how we treated **ε** in the GLM: we will assume random effects have normal distributions with means of zero and a variance σ_j^2.
- These random effects are modeled as variances and covariances stored in a matrix **G**. Residuals can also be correlated and modeled in a residuals correlation matrix R (not covered further).

Random Intercepts

- The simplest LMM adds a random intercept to the usual linear regression model:
  - Standard linear regression: $y_i = b_0 + b_1x1 + b_2x2 + \ldots + e_i$
  - Mixed-effects: $y_{ij} = b_0 + b_1x1 + b_2x2 + \ldots + u_j + e_{ij}$
  - In the standard linear regression specification, *i* indexes participants in both the standard and mixed models. In the mixed-effects specification, *j* indexes clusters.
  - Estimates of u_j typically produced by statistical software represent the cluster-specific deviation from the overall intercept b_0. $b_0 + u_j$ yields the cluster-specific intercept estimates for cluster *j*.
  - See model **3a** in the Stata -mixed- command documentation for an example application of this model.

Intracluster Correlation (ICC)

- For the model shown in the previous slide, the LMM will assume u_j has a normal distribution with a zero mean and a variance σ^2_{int}.
- σ^2_{int} represents the **between-cluster variance**.
- It is possible to quantify the proportion of the total variance in the outcome attributable to cluster membership as:

$$ICC = \sigma^2_{int} / (\sigma^2_{int} + \sigma^2_e)$$

where σ^2_e is the within-cluster residual variance estimate.

- The ICC can be computed from a model containing only the fixed intercept and random intercept (*unconditional* ICC) or from a model containing other fixed effects predictors (*conditional* ICC).

LMM Assumptions

- LMMs share the following assumptions with the GLM described earlier:
  - Residuals are normally distributed.
  - Residuals' variances are constant across levels of the outcome (regression) or groups (ANOVA).
- Additional assumptions specific to the LMM:
  - The correlation structure among the residuals is correctly specified within clusters.
  - Clusters are statistically independent (e.g., uncorrelated), just as individual observations are in the GLM/regression model.
  - The random intercepts u_j are uncorrelated with the observation-level residuals $e_{j.}$

LMM Computing

- SAS was one of the first software programs to offer LMMs through its MIXED procedure (PROC MIXED) beginning in the early 1990s. Today PROC MIXED has many features.
- Stata and SPSS added LMM functionality to their programs as well. R also has LMM commands.
- LMMs can also be fitted in specialty programs like HLM and M*plus*.
- Features differ across the programs and new features and enhancements are added in new releases.
- See West & Galecki (2012) and McCoach et al (2018) for article reviews of LMM software. West et al. (2015) is a readable textbook devoted to illustrating how to fit mixed models in multiple software programs using numerous examples.

Motivating Example: IDEAS Cohort Analysis

- CADC Scientist Dr. Charles Windon investigated correlates associated with amyloid plaque deposits in the brains of older adults.
- For illustrative purposes, we consider a continuous outcome, *the centiloid score*, for 10,361 patients with centiloid values sampled from 500 clinics located across the United States.
  - Centiloid scores range from 0 (no plaque detected) to 100 (maximum possible plaque detected).
- Patients were nested within clinics. Because patients from the same clinic may share similar characteristics and life experiences, centiloid observations from different patients from the same clinic may be correlated.
- To account for this possibility, we analyzed **9,470** non-missing observations from **490** clinics using the three methods previously described, using **Stata**. These methods are available to varying degrees in other programs (e.g., SAS, R, M*plus*).

IDEAS Analysis: Explanatory Variables

Patient-level explanatory variables:

- Age (continuous years)
- Gender (female vs. male)
- Race/ethnicity (White as reference; Hispanic, Black non-Hispanic, and Asian non-Hispanic)
- Education (some college or more vs. high school or less)
- English fluency (fluent vs. not fluent)
- Impairment level (Dementia vs. MCI)
- Living with a child (yes vs. no)

IDEAS Analysis: No Clustering Adjustment

regress centiloid_value age i.e_gender_r i.race_ethnicity_r i.education_r i.fluent_english i.impairment_level lives_child

Source	SS	df	MS		
Model	762023.662	9	84669.2958	**Number of obs**	**= 9,470**
Residual	22005839.7	9,460	2326.1987	F(9, 9460)	= 36.40
Total	22767863.3	9,469	2404.46334	Prob > F	= 0.0000
				R-squared	= 0.0335
				Adj R-squared	= 0.0325
				Root MSE	= 48.231

centiloid_value	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.5836976	.079334	7.36	0.000	.428186	.7392093
e_gender_r						
Female	5.841138	1.017498	5.74	0.000	3.846624	7.835652
race_ethnicity_r						
Hispanic	-10.96587	2.696902	-4.07	0.000	-16.25237	-5.67936
Black, Non-Hispanic	**-5.92834**	**2.864649**	**-2.07**	**0.039**	**-11.54367**	**-.313013**
Asian, Non-Hispanic	-23.08462	3.623497	-6.37	0.000	-30.18745	-15.98179
education_r						
Some college or more	**2.385898**	**1.115335**	**2.14**	**0.032**	**.1996014**	**4.572195**
fluent_english						
Fluent in English	13.65044	5.139161	2.66	0.008	3.576576	23.7243
impairment_level						
Dementia	13.30795	1.054124	12.62	0.000	11.24164	15.37426
lives_child	**-3.747984**	**1.852102**	**-2.02**	**0.043**	**-7.378502**	**-.1174662**
_cons	-20.19985	8.027671	-2.52	0.012	-35.93581	-4.463896

IDEAS Analysis: (1) Cluster-Adjusted Robust SE

```
regress centiloid_value age i.e_gender_r i.race_ethnicity_r i.education_r i.fluent_english
i.impairment_level lives_child, cluster(practice_id)

Linear regression                               Number of obs     =      9,470
                                                F(9, 489)         =      39.79
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0335
                                                Root MSE          =     48.231

                                    (Std. Err. adjusted for 490 clusters in practice_id)
------------------------------------------------------------------------------
                             |               Robust
               centiloid_value |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------------------+------------------------------------------------
                         age |    .5836976   .0889602     6.56   0.000     .4089062    .7584891
                             |
                   e_gender_r |
                      Female |    5.841138   1.163874     5.02   0.000     3.554328    8.127948
                             |
                race_ethnicity_r |
                    Hispanic |   -10.96587    3.02816    -3.62   0.000    -16.91568   -5.016057
           Black, Non-Hispanic |    -5.92834   3.121939    -1.90   0.058    -12.06241    .2057307
        Asian, Non-Hispanic |   -23.08462   3.322703    -6.95   0.000    -29.61315   -16.55608
                             |
                 education_r |
           Some college or more |    2.385898   1.344664     1.77   0.077    -.2561343    5.02793
                             |
                fluent_english |
          Fluent in English |    13.65044   4.852918     2.81   0.005      4.11529    23.18558
                             |
             impairment_level |
                    Dementia |    13.30795   1.273069    10.45   0.000     10.80658    15.80931
                  lives_child |   -3.747984   1.951433    -1.92   0.055    -7.582212    .0862438
                        _cons |   -20.19985   8.446361    -2.39   0.017    -36.79549   -3.604215
------------------------------------------------------------------------------
```

IDEAS Analysis: (2) GEE

```
xtset practice_id
       panel variable:  practice_id (unbalanced)

xtgee centiloid_value  age i.e_gender_r i.race_ethnicity_r i.education_r  i.fluent_english
i.impairment_level lives_child, robust

GEE population-averaged model                   Number of obs      =       9,470
Group variable:                practice_id      Number of groups   =         490
Link:                             identity      Obs per group:
Family:                           Gaussian                    min =           1
Correlation:                  exchangeable                    avg =        19.3
                                                              max =         195
                                                Wald chi2(9)       =      373.67
Scale parameter:                  2328.178      Prob > chi2        =      0.0000

                                    (Std. Err. adjusted for clustering on practice_id)
------------------------------------------------------------------------------
                             |               Robust
      centiloid_value        |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------------+------------------------------------------------------------------
                         age |    .6984801   .0855118     8.17   0.000     .5308801    .8660801
                             |
                  e_gender_r |
                      Female |    6.087034   1.128038     5.40   0.000     3.876121    8.297947
                             |
             race_ethnicity_r |
                    Hispanic |   -8.316242   2.719911    -3.06   0.002    -13.64717   -2.985313
          Black, Non-Hispanic |   -5.584241   3.101088    -1.80   0.072    -11.66226    .4937793
       Asian, Non-Hispanic   |   -21.25838   3.222373    -6.60   0.000    -27.57411   -14.94264
                             |
               education_r   |
Some college or more         |    1.200759   1.220883     0.98   0.325    -1.192127    3.593645
                             |
             fluent_english  |
   Fluent in English         |     13.1081   5.326468     2.46   0.014     2.668416    23.54779
                             |
           impairment_level  |
                    Dementia |    13.30725   1.216738    10.94   0.000     10.92249    15.69201
                 lives_child |   -3.319833   1.910693    -1.74   0.082    -7.064723    .4250561
                       _cons |   -26.17546   8.370517    -3.13   0.002    -42.58137   -9.769548
------------------------------------------------------------------------------
*estat wcorr = .0506997   (exchangeable correlation)
```

IDEAS Analysis: (3a) Multilevel Model

```
mixed centiloid_value age i.e_gender_r i.race_ethnicity_r i.education_r i.fluent_english
i.impairment_level lives_child  || practice_id:

Mixed-effects ML regression                     Number of obs      =      9,470
Group variable: practice_id                     Number of groups   =        490

                                                Obs per group:
                                                              min =          1
                                                              avg =       19.3
                                                              max =        195

                                                Wald chi2(9)       =     330.36
Log likelihood = -50045.809                     Prob > chi2        =     0.0000
```

centiloid_value	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.6928632	.0803622	8.62	0.000	.5353562	.8503702
e_gender_r						
Female	6.083673	1.007187	6.04	0.000	4.109622	8.057724
race_ethnicity_r						
Hispanic	-8.47419	2.697449	-3.14	0.002	-13.76109	-3.187287
Black, Non-Hispanic	**-5.634006**	**2.869926**	**-1.96**	**0.050**	**-11.25896**	**-.0090537**
Asian, Non-Hispanic	-21.40225	3.655176	-5.86	0.000	-28.56626	-14.23824
education_r						
Some college or more	**1.257155**	**1.160741**	**1.08**	**0.279**	**-1.017855**	**3.532165**
fluent_english						
Fluent in English	13.2943	5.71153	2.33	0.020	2.099909	24.48869
impairment_level						
Dementia	13.29192	1.093616	12.15	0.000	11.14847	15.43536
lives_child	**-3.352901**	**1.83777**	**-1.82**	**0.068**	**-6.954864**	**.249062**
_cons	-26.05054	8.430791	-3.09	0.002	-42.57459	-9.526497

IDEAS Analysis: (3a) Multilevel Model

```
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
practice_id: Identity        |
                  var(_cons) |   100.9563   15.27483      75.04905    135.8067
-----------------------------+------------------------------------------------
               var(Residual) |    2220.61   32.81977      2157.207    2285.876
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 184.42        Prob >= chibar2 = 0.0000
```

```
. estat icc
```

Residual intraclass correlation

```
------------------------------------------------------------------------------
                       Level |        ICC   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
                  practice_id |   .0434863   .0063763      .0325685    .0578451
------------------------------------------------------------------------------
```

IDEAS Analysis: (3b) MLM with Robust SE

```
mixed centiloid_value age i.e_gender_r i.race_ethnicity_r i.education_r i.fluent_english
i.impairment_level  lives_child || practice_id:, vce(cluster practice_id)

Mixed-effects regression                        Number of obs      =      9,470
Group variable: practice_id                     Number of groups   =        490

                                                Obs per group:
                                                              min =          1
                                                              avg =       19.3
                                                              max =        195

                                                Wald chi2(9)       =     375.23
Log pseudolikelihood = -50045.809               Prob > chi2        =     0.0000

                                  (Std. Err. adjusted for 490 clusters in practice_id)
------------------------------------------------------------------------------
                             |               Robust
               centiloid_value |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------------+------------------------------------------------
                         age |    .6928632   .0854942     8.10   0.000     .5252976    .8604287
                             |
                   e_gender_r |
                      Female |    6.083673   1.128049     5.39   0.000     3.872737    8.294609
                             |
              race_ethnicity_r |
                    Hispanic |   -8.47419    2.735712    -3.10   0.002    -13.83609   -3.112293
          Black, Non-Hispanic |   -5.634006   3.096821    -1.82   0.069    -11.70366    .4356508
       Asian, Non-Hispanic |   -21.40225    3.215393    -6.66   0.000     -27.7043   -15.1002
                             |
                 education_r |
      Some college or more |    1.257155   1.220365     1.03   0.303    -1.134716    3.649026
                             |
                fluent_english |
           Fluent in English |     13.2943   5.304452     2.51   0.012     2.897767    23.69084
                             |
              impairment_level |
                    Dementia |    13.29192   1.214844    10.94   0.000     10.91087    15.67297
                  lives_child |   -3.352901   1.910036    -1.76   0.079    -7.096502    .3907001
                        _cons |   -26.05054   8.361571    -3.12   0.002    -42.43892   -9.662165
------------------------------------------------------------------------------
```

IDEAS Analysis: Results Comparison

- Results from regression analysis that did not adjust for clustering should be viewed as suspect.
- Remaining analyses **(1-3b)** adjusted for clustering in different ways:
  - 1) Linear regression with cluster-adjusted robust SE approach is computationally fast and easy to specify but does not utilize correlation information among patients within clinics to inform estimates.
  - 2) GEE approach improves on the cluster-adjusted SE method by utilizing correlation information among patients within clinics.
  - 3a) MLM approach directly models clinic-to-clinic variability while relying on a correct specification of the model.
  - 3b) MLM with robust SEs option may help to mitigate some of the effects of assumption violations.

Conclusion

- Multilevel data are prevalent in health disparities and aging research and reflect the complexity of the health disparities research landscape.
- To avoid drawing incorrect conclusions about the statistical significance of explanatory variables when analyzing multilevel data, it is important to be conversant with analysis methods developed for clustered data structures.
- We covered and illustrated three popular methods for analyzing clustered data: 1) cluster-adjusted robust standard errors, 2) GEE, and 3) MLMs. Each has strengths and limitations: MLMs can yield valuable information about cluster-level variability but make stronger assumptions about the correctness of the underlying model than cluster-adjusted standard errors or GEE.
- In general, any should be superior to ignoring clustering and using naïve estimators that assume observations are uncorrelated.

Acknowledgements

- Slide review and comments:
  - Estie Hudes, PhD
  - Anita Stewart, PhD

Thank you!

References

- Dunlop, DD. Regression for Longitudinal Data: A Bridge from Least Squares Regression, *The American Statistician*, 48:299–303, 1994.
- Gallis JA, Li F, Turner EL. xtgeebcv: A command for bias-corrected sandwich variance estimation for GEE analyses of cluster randomized trials. *Stata J.* 2020;20(2):363-381.
- Glantz SA, Slinker BK, Neilands TB. *Primer of Applied Regression and Analysis of Variance.* 3rd ed. Columbus, Ohio: McGraw Hill Education; 2016.
- Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol.* 2003;157(4):364-375.
- Hardin J, Hilbe J. *Generalized Estimating Equations.* New York: Chapman & Hall/CRC; 2003.
- Jeffries N, Zaslavsky AM, Diez Roux AV, et al. Methodological Approaches to Understanding Causes of Health Disparities. *Am J Public Health.* 2019;109(S1):S28-S33. PMCID: PMC635612
- McCoach DB, Rifenbark GG, Newton SD, et al. Does the Package Matter? A Comparison of Five Common Multilevel Modeling Software Packages. *Journal of Educational and Behavioral Statistics.* 2018;43(5):594-627.
- West BT, Galecki AT. An Overview of Current Software Procedures for Fitting Linear Mixed Models. *Am Stat.* 2012;65(4):274-282.
- West BT, Welch KB, Galecki AT. *Linear mixed models: a practical guide using statistical software.* Boca Raton, FL: CRC Press; 2015.