Biostatistics 140.656
Lab 1

Topics:

- Two-stage normal-normal model
- Visual assessment of the within vs. between cluster variance
- Factors that affect estimates of the level-2 random intercept

Learning Objectives:

Students who successfully complete this lab will be able to:

- Fit and interpret all the parameters from the two-stage normal-normal model
- Describe the estimation of cluster-specific means and random intercepts within a two-stage normal-normal model
- Describe characteristics of the level-2 unit (most importantly sample size) and the model that influence the estimation of the cluster-specific means and random intercepts

Scientific Background:

In Homework 1, you will be analyzing a cross-sectional study of high school mathematics achievement from the High School and Beyond (HS&B) study conducted within the National Education Longitudinal Studies (NELS) program of the National Center for Education Statistics (NCES). The NELS was established to study the educational, vocational, and personal development of young people beginning with their elementary or high school years, and following them over time as they begin to take on adult roles and responsibilities. Thus far, the NELS program consists of five major studies: the National Longitudinal Study of the High School Class of 1972 (NLS-72), High School and Beyond (HS&B), the National Education Longitudinal Study of 1988(NELS:88), the Education Longitudinal Study of 2002 (ELS:2002), and the High School Longitudinal Study of 2009 (HSLS:09).

The HS&B survey included two cohorts: the 1980 senior class, and the 1980 sophomore class. Both cohorts were surveyed every two years through 1986, and the 1980 sophomore class was also surveyed again in 1992.

We have available data from one of the assessments for 7042 students within 156 schools.

The study variables include:

Level 1: student

mathach: a measure of mathematics achievement (MA)
minority: dummy variable for student being non-white
female: dummy variable for student being female
ses: socioeconomic status (SES) based on parental education, occupation and income (z-score)

Level 2: school

schoolid: school identified
sector: dummy variable for a school being Catholic
pracad: proportion of students in the academic track
disclaim: scale measuring disciplinary climate
himinty: dummy variable for more than 40% minority enrollment
size: number of students enrolled at the school
newid: rescaled school identifier, counts 1 to 156 (we created this for you)

Two-stage Normal-Normal Model:

In this lab exercise, we will focus on estimation of the school-specific mean mathematics achievement (MA) scores. **To keep things simple we will consider data from the first 25 schools (newid ≤ 25).**

Assume the data are generated as follows where Y_{ij} is the MA score for student *j, j = 1, ..., n_i,* from school *i*, *i =1, ..., 25*.

Student-level model: $Y_{ij} = \theta_i + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2)$

School-level model: $\theta_i = \theta + b_i, b_i \sim N(0, \tau^2)$

Note that in this model we are defining two models; one for the student's MA score and one for the school's mean MA score, and two "residuals": ε_{ij} and b_i, both which are assumed to follow a normal distribution. This is why this model is referred to as the two-stage normal-normal model.

Lab Exercise:

1. Model interpretation: In your group, write down the definitions of the model parameters below within the context of the HS&B data:

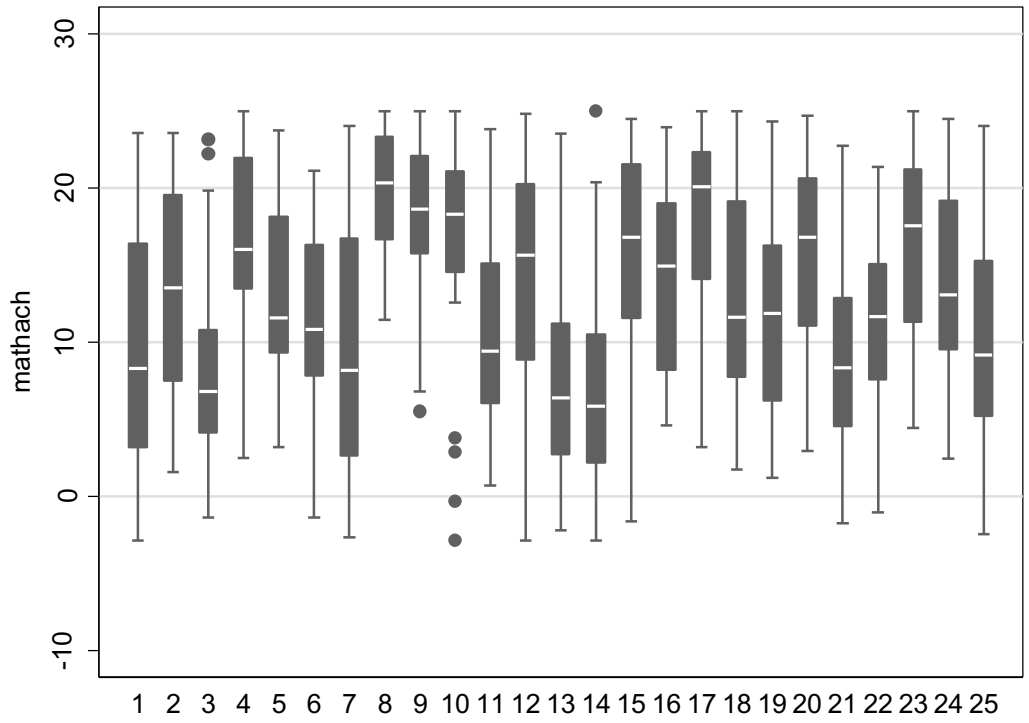
   a. θ

   b. θ_i

   c. b_i

   d. ε_{ij}

2. Model visualization and fit: The two-stage normal-normal model partitions information in Y_{ij} into within level-2 units and between-level-2 units. To visualize this partitioning, one can construct a graph looking at side-by-side boxplots of Y_{ij} for each school.



Without any other information, what percentage of the total variance in Y_{ij}, $Var(Y_{ij}) = \tau^2 + \sigma^2$, do you think is attributable to differences between the school-mean MA scores? Recall that this defines the intraclass correlation coefficient; ICC = $\tau^2/(\tau^2 + \sigma^2)$.

Below, is the output from fitting the two-stage normal-normal model to these data. From the model fit, estimate the percentage of the total variance in Y_{ij} that is attributable to differences between the school-mean MA scores.

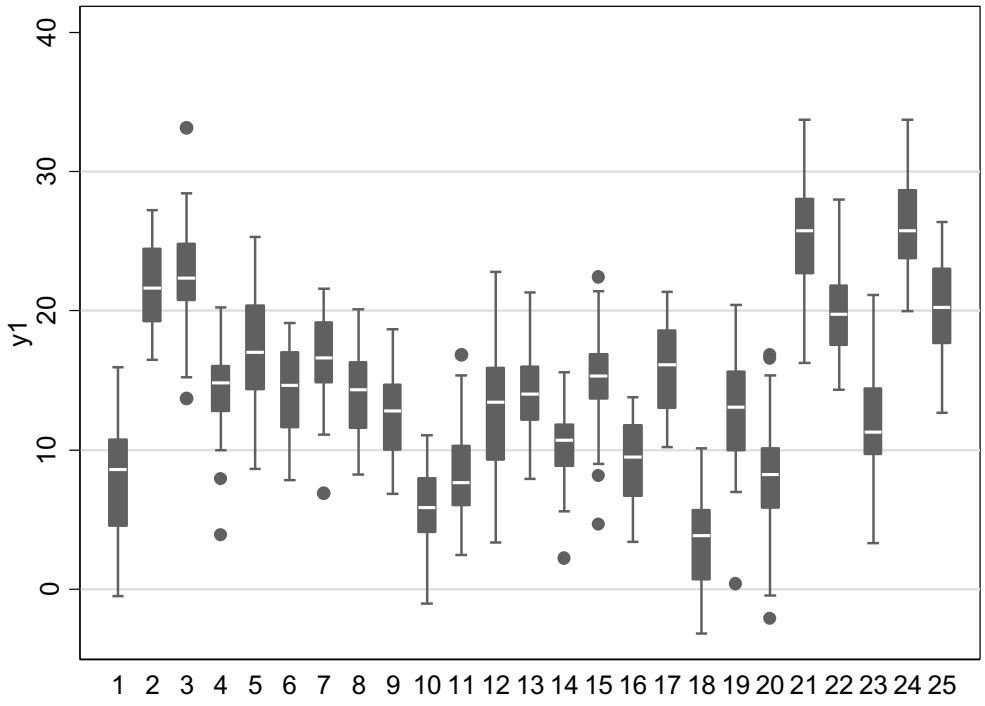
```
Mixed-effects ML regression                     Number of obs      =      1097
Group variable: newid                           Number of groups   =        25

                                                Obs per group: min =        20
                                                               avg =      43.9
                                                               max =        67


                                                Wald chi2(0)       =         .
Log likelihood =  -3589.991                     Prob > chi2        =         .

------------------------------------------------------------------------------
     mathach |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   12.99362   .6954502    18.68   0.000     11.63056    14.35668
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
newid: Identity              |
                  var(_cons) |   11.12486   3.432375      6.076751    20.36655
-----------------------------+------------------------------------------------
               var(Residual) |   38.41937   1.659601       35.3005    41.81379
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   199.04 Prob >= chibar2 = 0.0000
```

3. Test your skill:

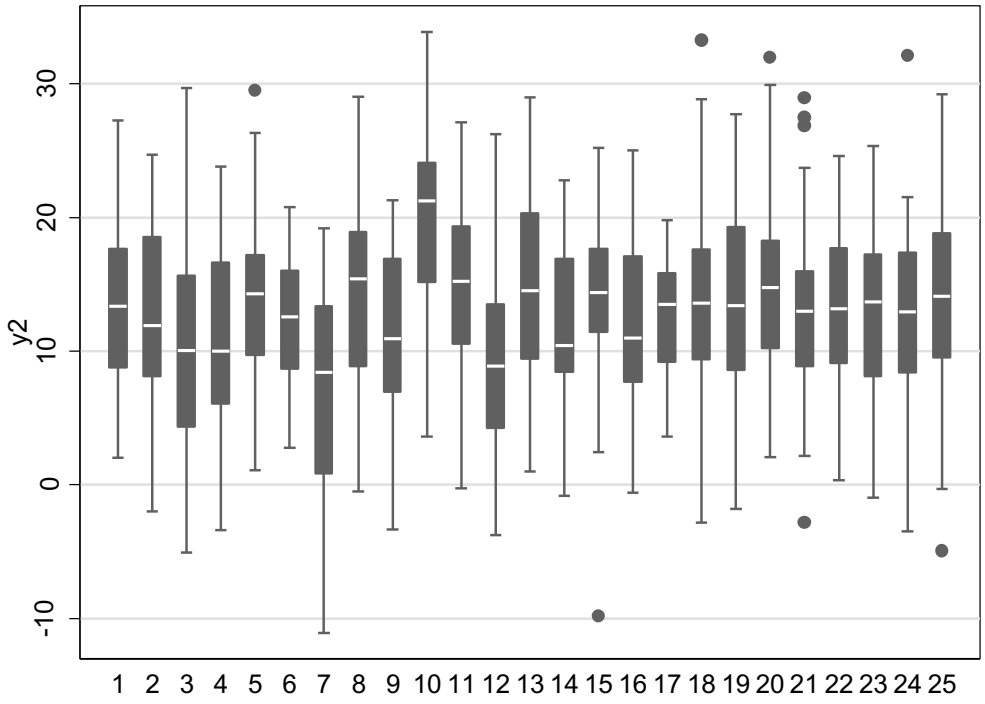
I simulated three datasets. Each of the three datasets has the same structure as our subset of the HS&B data we have been considering, i.e. $i = 1, \ldots, 25$ and $j = 1, \ldots, n_i$. In each dataset, the total variance in Y_{ij} is fixed at 50 (roughly the total variance in Y_{ij} that was observed) and the population mean MA score is 13. For each of the three datasets, I constructed side-by-side boxplots displaying the variation in Y_{ij} partitioned into between and within school variation. For each of the three datasets, guess what fraction of the total variance is attributable to differences across the school-mean MA scores.

Simulation	1	2	3
ICC			

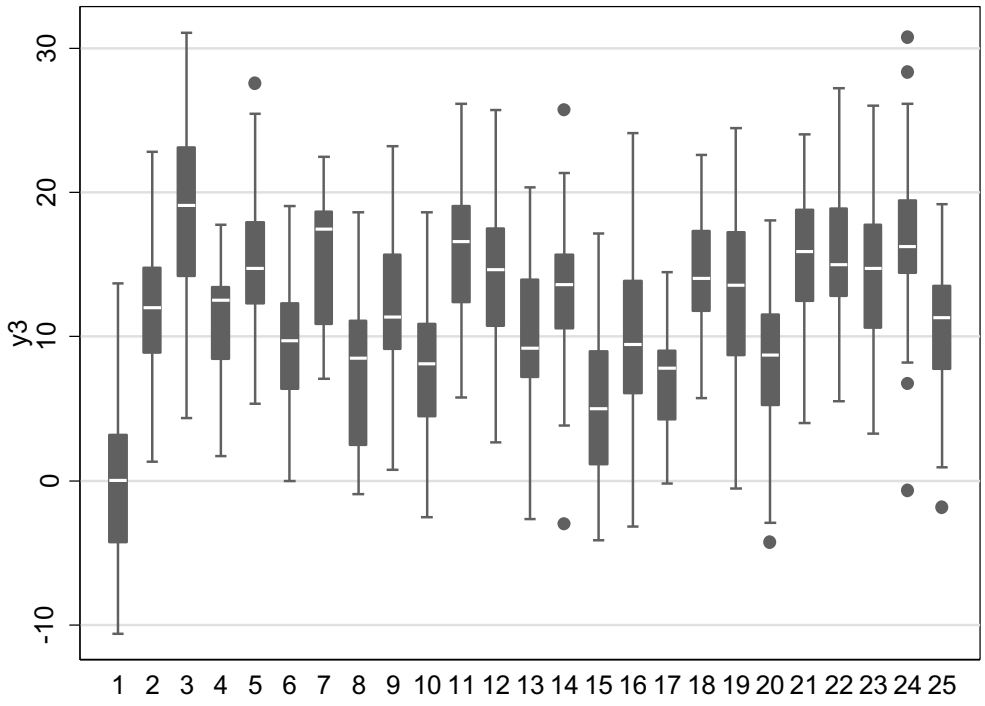
Simulation 1:



Simulation 2:



Simulation 3:



4. Estimated school-mean MA scores OR Estimation of the school-level random intercept

   From the fit of the two-stage normal-normal model, we can obtain estimates for the school-mean MA scores, θ_i. Recall that the model for the school-mean MA scores is $\theta_i = \theta + b_i$. Therefore, estimation of the school-mean MA scores requires us to estimate b_i, the random intercept.

   The estimated school-mean MA scores is: $\hat{\theta}_i = \hat{\theta} + \hat{b}_i$ where $\hat{b}_i = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2/n_i}(\bar{Y}_i - \hat{\theta})$

   When $\hat{b}_i = 0$, then we refer to the estimate of $\hat{\theta}_i$ as a "total shrinkage" estimate since for this particular school, we are using the population average estimate as the estimate for the school-specific mean. NOTE: $\hat{b}_i = 0$ would occur only when $\tau^2 = 0$, i.e. there is information in the clustering of students within schools; this is unlikely, but you could see $\hat{b}_i$ close to 0.

   When $\hat{b}_i = \bar{Y}_i - \hat{\theta}$, then we refer to the estimate of $\hat{\theta}_i$ as a "no shrinkage" estimate since we are using the sample mean from school *i* as our estimate for the school-specific mean.

   In your group, explore the behavior of $\hat{b}_i$ as a function of τ^2 and σ^2 for three selected schools.

   Specifically, do the following:

   a. Consider the sample mean MA scores and numbers of students for schools "newid" = 3, 4 and 22. Confirm the data listed below for these three schools is accurate:

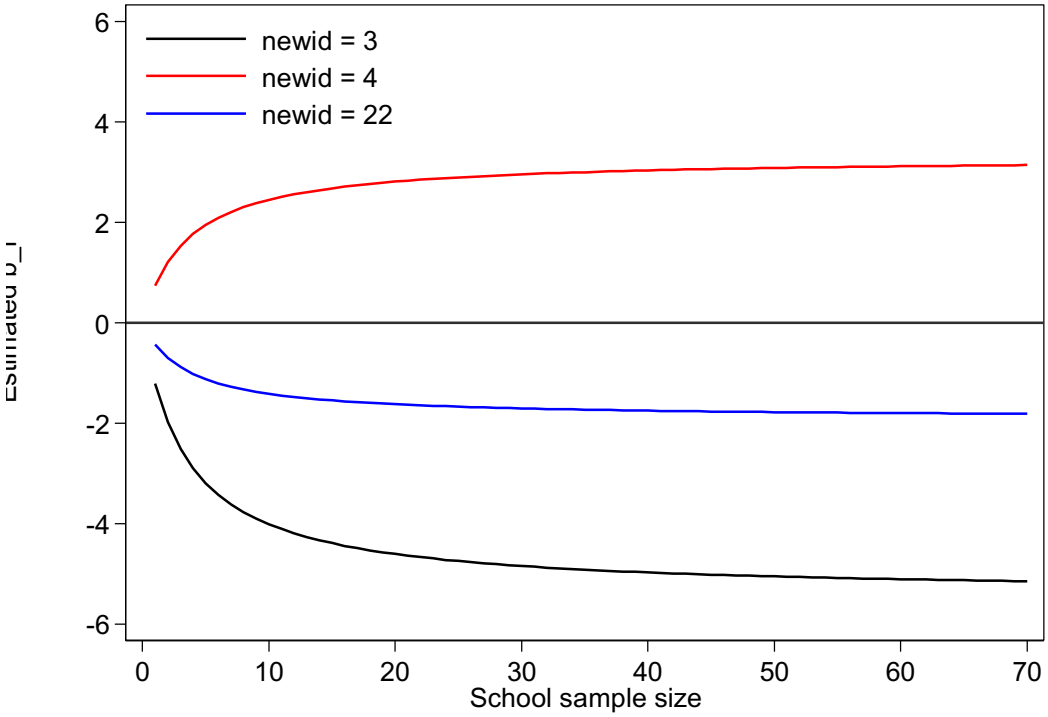
	newid =3	newid =4	newid =22

$\bar{Y}_i$	7.6	16.3	11.1
n_i	48	20	67

a. Fix $\bar{Y}_i$, $\hat{\theta}$, $\hat{\tau}^2$, and $\hat{\sigma}^2$ at the estimated values. On the same graph, plot how $\hat{b}_i$ changes for each school as a function of n_i.

See lab1.do file for details on how to generate the figure below.

The number of students represented in each school in the HS&B data ranges from 20 to 67. We expanded that range from 0 to 70.



You should note that as the sample size goes to 0, $\hat{b}_i$ approaches 0 indicating "total shrinkage" and the estimated school-mean MA score is shrunk towards the estimate population mean MA score; $\hat{\theta}_i \rightarrow \hat{\theta}$, as $n_i \rightarrow 0$

You should also note that as the sample size gets larger, $\hat{b}_i$ approaches $\bar{Y}_i - \hat{\theta}$ and there is little to no shrinkage. E.g. for newid = 3, the school sample mean is 7.6 and the estimated population mean is 13, $\hat{b}_i$ is roughly 7.6 – 13 = -5.4 when the sample size for this school gets large.

b. Fix $\bar{Y}_i, n_i$, $\hat{\theta}$ and $\hat{\sigma}^2$ at the estimated values. On the same graph, plot how $\hat{b}_i$ changes for each school as a function of $\hat{\tau}^2$

c. Fix $\bar{Y}_i, n_i, \hat{\theta}$ and $\hat{\tau}^2$ at the estimated values. On the same graph, plot how $\hat{b}_i$ changes for each school as a function of $\hat{\sigma}^2$

d. Fix $\bar{Y}_i, n_i, \hat{\theta}$ and $\hat{\tau}^2 + \hat{\sigma}^2$ at the estimated values. On the same graph, plot how $\hat{b}_i$ changes for each school as a function of the ICC.

Be prepared to describe the patterns that you observe in your graphs.

