**BST5220 Multilevel HW2 Due 2/25 Monday**

**Author: Miao Cai** [miao.cai@slu.edu](mailto:miao.cai@slu.edu)

The ECLS-K is a national survey of children enrolled in kindergarten in 1998-1999, representative of all kindergarteners attending school that year. The survey is administered by the National Center for Education Statistics, U.S. Department of Education (NCES), and uses a multistage probability sample design (counties, schools, then randomly selected children). The baseline (fall and spring of kindergarten) waves of the ECLS-K, conducted in 1998-1999, included about 20,000 children, with sizable over-samples of at-risk children, including poor and minority children. Children and families were followed in the fall and spring of first grade (1999-2000), spring of third grade (2002), and spring of fifth grade (2004). The entire longitudinal data can be found from this link: <http://nces.ed.gov/ecls/dataproducts.asp>

**The data for this homework is a subset of the above data, which is the baseline data. The name of the data set is “DataForHW2.sas7bdat”. Using Proc freq to find the code for the categorical variables in the data set.**

In this homework, we will evaluate the relationship between child’s **age- and sex-specific BMI percentile** (bmipct) and variables at both school and child levels at the baseline (spring of 1999), by fitting a multilevel model to nest children (level 1) within schools (level 2). The multilevel approach allows us to model variations in BMIPCT within and between schools. The analysis is based on part of the study in the reference paper, “The role of local food availability in explaining obesity risk among young”, which can be found in the Blackboard.

Child, school, and neighborhood measures are described as follows.

**Child-level outcome variable**:

* **Outcome**: **the age- and sex-specific BMI percentile** (bmipct) for each child was calculated using the SAS program (SAS Institute, Inc., Cary, North Carolina) developed by the Centers for Disease Control and Prevention (CDC) based on the updated 2000 Growth Charts.Height and weight were measured by ECLS-K trained assessors using a Shorr Board (Shorr Productions, Olney, MD) and a digital bathroom scale (Seca Model 840, Hanover, MD), which reduced concerns regarding the reliability of parental reports and BMI was calculated from measured height and weight. Children with unreasonable or implausible values were also dropped. These included cases where: BMI values were less than 10 or greater than 50 in the kindergarten. These were likely data entry, coding, or reporting errors. Those with a BMI percentile of 95 or higher were classified as obese.

**Child-level outcome independent variables**:

The ECLS-K attempted to collect parental information at each wave (generally from the child’s mother). Using data from the parent interviews, we consider measures of the child’s racial/ ethnic background, gender, age, household income, educational attainment of parents, parental health (self-reported), and family structure (e.g., single parent). These socio-demographic factors are important correlates of childhood obesity and may also be correlated with residential location and other neighborhood characteristics of interest. More proximate behavioral factors potentially related to BMI include parental reports of how many days per week the child engaged in 20 min or more of vigorous activity or exercise (where the child’s heart rate is consistently elevated) outside the school context, and parental reports of hours the child spends watching television and videos. These two measures serve as indicators of how sedentary children’s lives are outside the school environment.

* Gender: gender of the focal child: 1=male, 2=female
* Childgender: gender of the focal child: “1=female”, “2=male”
* ChildAge: Age of child in months (Continuous)
* Childbmi: child composite bmi
* Meducation: A composite variable of the mother’s highest attained education levels.
* ChildRace: Race and ethnicity of the focal child
* FamilyStructure: Classification of the focal child’s parents who reside in the household
* HouseIncome: Household income (Continuous)
* Phealth: Child’s parent health status
* ExerciseFreeTime : Child’s physically active free time(Continuous)
* TV: Number of hours child spent on watching TV (Continuous).

**School-level variables**

Several school-based factors (public or private, urbanicity, and percent of students who are eligible for school meal programs e a proxy for school-level disadvantage) may be indirectly related to obesity risk and tap into school-level dynamics.

* Region (Census region): Indicates the geographic region in which the child lives
* Urban: School urbanicity designation by the Census Bureau
* Schooltype: School type from the school administrator questionnaire
* PctMinority: School Percentage of minority students in school(Continuous)
* PCTFreeLunch: School Percentage of students eligible for free lunch in school(Continuous)

**Questions:**

Run proc mixed using data set DataForHW2 on the outcome variable BMIPCT along with the above independent variables. Use model building strategies to build the final model. Interpret significant effects. Find the variances at different levels.

**My solutions**:

Since the outcome bmipct is the age- and sex- specific BMI percentile, we do not need to add age or sex in our statistical model. The child composite BMI is also not included since the outcome is recoded using this variable. All potential variables to be selected are:

* Child level: meducation ChildRace FamilyStructure HouseIncome pHealth ExerciseFreeTime TV
* School level: region urban schooltype PctMinority PCTFreeLunch

Recode variables:

* Child race was classified into white, black, and others.
* Annual household income was categorized into three categories: low (<25,000), middle (25,000-75,000), and high (>75,000), which equally distributed about 30% of respondents into each category.
* PCTMinority was recoded as 1, 2 (including 2 and 3), and 3 (including 4 and 5).
* Exercise at free time was recoded as frequent and infrequent (including 1 and 2).

SAS Code:

**DATA** HW2.DATHW2clean;

KEEP school\_ID bmipct

meducation ChildRace FamilyStructure HouseIncome pHealth ExerciseFreeTime TV

region urban schooltype PctMinority PCTFreeLunch;

SET HW2.DataForHW2;

IF ChildRace='5=white' THEN ChildRace='White';

IF ChildRace='4=black' THEN ChildRace='Black';

IF ChildRace NOT IN ('White','Black') THEN ChildRace='Other';

IF HouseIncome IN ('1<15000', '2:15000-25000') THEN HouseIncome='Low';

IF HouseIncome IN ('3:25000-35000', '4:35000-50000', '5:50000-75000') THEN HouseIncome='Middle';

IF HouseIncome IN ('6:75000-100000', '7:>100000') THEN HouseIncome='High';

IF meducation IN ('3=some college', '4=>=bachelor') THEN meducation='3>=college';

IF PctMinority = **1** THEN PctMinority=**1**;

IF PctMinority IN (**2**, **3**) THEN PctMinority=**2**;

IF PctMinority IN (**4**, **5**) THEN PctMinority=**3**;

IF ExerciseFreeTime IN (**1**, **2**) THEN ExerciseFreeTime=**1**;

IF ExerciseFreeTime = **3** THEN ExerciseFreeTime=**2**;

\*if nmiss(of \_numeric\_) > 0 then delete;

**RUN**;

# Model 1: Random intercept only model

The statistical model:

**Level 1** : bmipctij = b0j + eij  (1)

**Level 2 (intercept):** b0j = g00 + u0j  (2)

**Combined model:** bmipctij = g00 + (u0j + eij ) (3)

SAS code:

**PROC** **MIXED** DATA=HW2.DATHW2clean noclprint covtest noitprint method=reml;

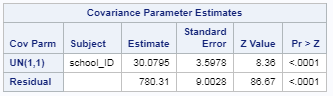
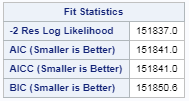
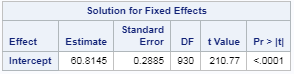
CLASS school\_ID;

MODEL bmipct = / solution ddfm = bw;

RANDOM intercept / subject=school\_ID TYPE=UN;

**RUN**; \*-2loglikelihood: 151837.0;

SAS output:

Since the random intercept and between-cluster variance are statistically siginificant, the random intercept model is good as a starting point.

# Model 2: Random intercept + child level variables, delete insignificant variables (FamilyStructure and ExerciseFreeTime)

Then, I add all child level variables to the random intercept only model. Since some of the variables at the child level were not significant, I excluded them (FamilyStructure and ExerciseFreeTime) and fit the model.

The statistical model:

**Level 1**: bmipctij = b0j + b1 meducationij + b2 ChildRaceij + b3 HouseIncomeij + b4 pHealthij + b5 TVij + eij  (1)

**Level 2 (intercept):** b0j = g00 + u0j  (2)

**Combined model**: bmipctij = g00 + b1 meducationij + b2 ChildRaceij + b3 HouseIncomeij + b4 pHealthij + b5 TVij + (u0j + eij ) (3)

SAS code:

/\* 02 Random Intercept model with level 1 variables - delete insignificant: FamilyStructure ExerciseFreeTime\*/

**PROC** **MIXED** DATA=HW2.DATHW2clean noclprint covtest noitprint method=reml;

CLASS school\_ID meducation ChildRace HouseIncome pHealth;

MODEL bmipct = meducation ChildRace HouseIncome pHealth TV/ solution ddfm = bw;

RANDOM intercept / subject=school\_ID TYPE=UN;

**RUN**; \*-2loglikelihood: 126244.7;

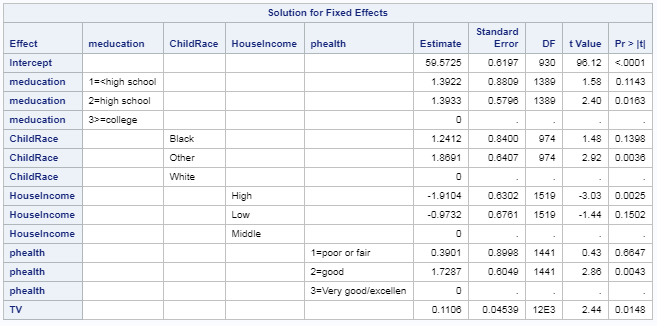
**DATA** pvalue;

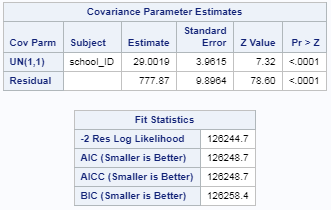
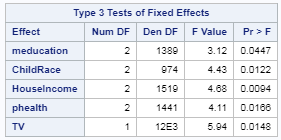
DF = **2**; CHISQ = **151837** – **126244.7**;

PVALUE = **1** - PROBCHI(CHISQ, DF);

**RUN**;

SAS ouputs:





All child level variables were significant in this model. It was significantly better than the intercept only model by deviance test.

# Model 3: Random intercept + child level variables + school level variables (delete insignificant schooltype)

After all child level variables were significant, I started to add all school level variables. School type was deleted since it was not significant.

The statistical model:

**Level 1** : bmipctij = b0j + b1 meducationij + b2 ChildRaceij + b3 HouseIncomeij + b4 pHealthij + b5 TVij + eij  (1)

**Level 2**: b0j = g00 + g01Regionj + g02Urbanj + g02PctMinorityj + g01PCTFreeLunchj + u0j (2)

**Combined model**: bmipctij = (g00 + b1 meducationij + b2 ChildRaceij + b3 HouseIncomeij + b4 pHealthij + b5 TVij g01Regionj + g02Urbanj + g02PctMinorityj + g01PCTFreeLunchj) + (u0j + eij ) (3)

SAS code:

/\* 12.5 Random Intercept with both level variables - delete Schooltype\*/

**PROC** **MIXED** DATA=HW2.DATHW2clean noclprint covtest noitprint method=reml;

CLASS school\_ID meducation ChildRace HouseIncome pHealth Region Urban;

MODEL bmipct = meducation ChildRace HouseIncome pHealth TV Region Urban PctMinority/ solution ddfm = bw;

RANDOM intercept/ subject=school\_ID TYPE=UN;

**RUN**; \*-2loglikelihood: 123083.9;

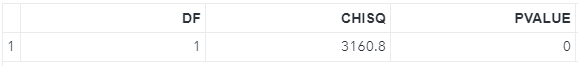
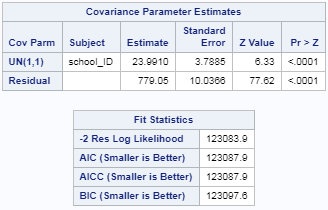
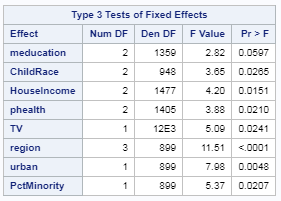
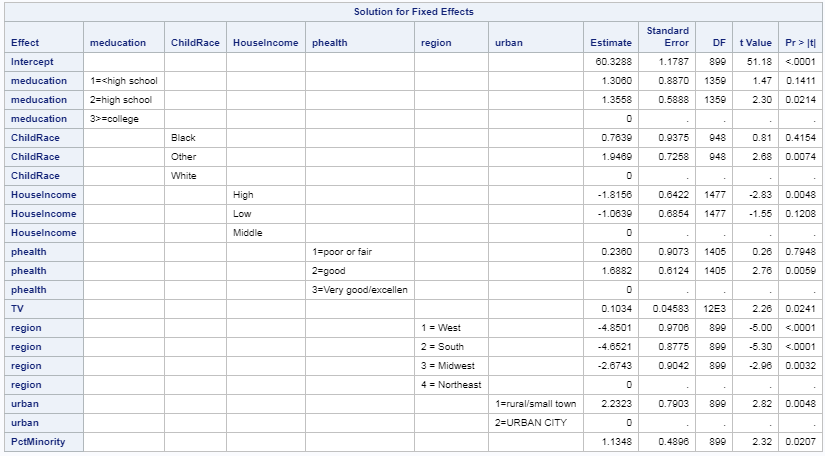
**DATA** pvalue;

DF = **1**; CHISQ = **126244.7** - **123083.9**;

PVALUE = **1** - PROBCHI(CHISQ, DF);

**RUN**;

SAS output:



All school level variables were significant in this model. Deviance test shows that this model is significantly better than the previous model.

# Model 4: Random slope model

I converted all child level variables as numeric and added them to random slope model one-by-one.

The statistical model:

**Level 1** : bmipctij = b0j + b1 meducationij + b2 ChildRaceij + b3 HouseIncomeij + b4 pHealthij + b5 TVij + eij

**Level 2 (intercept)**:b0j = g00 + g01Regionj + g02Urbanj + g02PctMinorityj + g01PCTFreeLunchj + u0j

**Level 2 (slope)**: bkj = g10 + u1j (k = 1, 2, …, 2)

**Combined model**: bmipctij = (g00 + g10 meducationij + b2 ChildRaceij + b3 HouseIncomeij + b4 pHealthij + b5 TVij g01Regionj + g02Urbanj + g02PctMinorityj + g01PCTFreeLunchj) + (u0j + u1j \* meducationij + eij ) (in the example of random intercept for meducationij)

SAS code:

/\* convert string to numeric\*/

**DATA** HW2.DATHW2RSLOPE;

SET HW2.DATHW2clean;

IF ChildRace='White' THEN ChildRace=**1**;

IF ChildRace='Black' THEN ChildRace=**2**;

IF ChildRace='Other' THEN ChildRace=**3**;

IF HouseIncome='Low' THEN HouseIncome=**1**;

IF HouseIncome='Middle' THEN HouseIncome=**2**;

IF HouseIncome='High' THEN HouseIncome=**3**;

IF meducation='1=<high school' THEN meducation=**1**;

IF meducation='2=high school' THEN meducation=**2**;

IF meducation='3>=college' THEN meducation=**3**;

IF pHealth='1=poor or fair' THEN pHealth=**1**;

IF pHealth='2=good' THEN pHealth=**2**;

IF pHealth='3=Very good/excellen' THEN pHealth=**3**;

\*if nmiss(of \_numeric\_) > 0 then delete;

**RUN**;

**DATA** HW2.DATHW2RSLOPE;

SET HW2.DATHW2RSLOPE;

ChildRace1 = INPUT(ChildRace, **20.**);

HouseIncome1 = INPUT(HouseIncome, **20.**);

meducation1 = INPUT(meducation, **20.**);

pHealth1 = INPUT(pHealth, **20.**);

**RUN**;

/\* 21 Random Intercept and random slope model - add meducation\*/

**PROC** **MIXED** DATA=HW2.DATHW2RSLOPE noclprint covtest noitprint method=reml;

CLASS school\_ID ChildRace HouseIncome pHealth Region Urban;

MODEL bmipct = meducation1 ChildRace HouseIncome pHealth TV Region Urban PctMinority/ solution ddfm = bw;

RANDOM intercept meducation1/ subject=school\_ID TYPE=UN;

**RUN**; \*UN(2, 2) NOT SIGNIFICANT;

/\* 22 Random Intercept and random slope model - add ChildRace\*/

**PROC** **MIXED** DATA=HW2.DATHW2RSLOPE noclprint covtest noitprint method=reml;

CLASS school\_ID meducation HouseIncome pHealth Region Urban;

MODEL bmipct = meducation ChildRace1 HouseIncome pHealth TV Region Urban PctMinority/ solution ddfm = bw;

RANDOM intercept ChildRace1/ subject=school\_ID TYPE=UN;

**RUN**; \*Estimated G matrix is not positive definite;

/\* 23 Random Intercept and random slope model - add HouseIncome\*/

**PROC** **MIXED** DATA=HW2.DATHW2RSLOPE noclprint covtest noitprint method=reml;

CLASS school\_ID meducation ChildRace pHealth Region Urban;

MODEL bmipct = meducation ChildRace HouseIncome1 pHealth TV Region Urban PctMinority/ solution ddfm = bw;

RANDOM intercept HouseIncome1/ subject=school\_ID TYPE=UN;

**RUN**; \*-2loglikelihood: 123079.6, this one works!;

**DATA** pvalue;

DF = **1**; CHISQ = **123083.9** - **123079.6**;

PVALUE = **1** - PROBCHI(CHISQ, DF);

**RUN**;

/\* 24 Random Intercept and random slope model - add pHealth\*/

**PROC** **MIXED** DATA=HW2.DATHW2RSLOPE noclprint covtest noitprint method=reml;

CLASS school\_ID meducation ChildRace HouseIncome Region Urban pHealth;

MODEL bmipct = meducation ChildRace HouseIncome pHealth1 TV Region Urban PctMinority/ solution ddfm = bw;

RANDOM intercept pHealth1/ subject=school\_ID TYPE=UN;

**RUN**; \*Estimated G matrix is not positive definite.;

/\* 25 Random Intercept and random slope model - add pHealth\*/

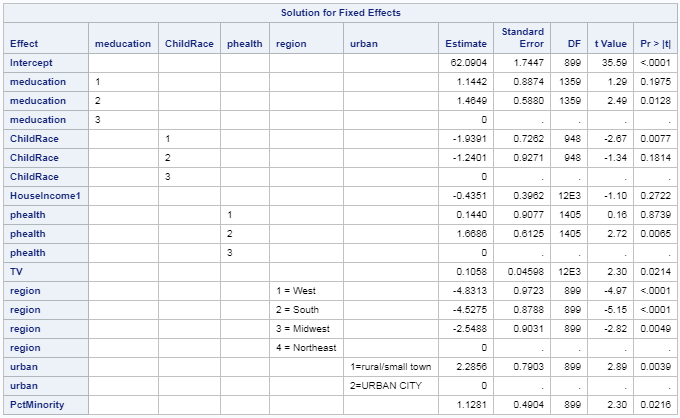
**PROC** **MIXED** DATA=HW2.DATHW2RSLOPE noclprint covtest noitprint method=reml;

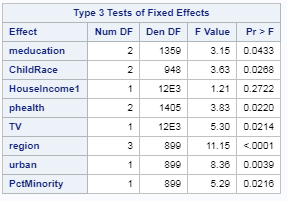
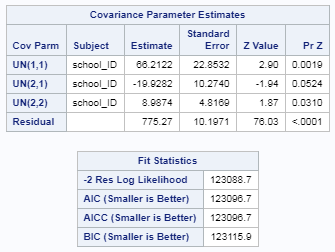
CLASS school\_ID meducation ChildRace HouseIncome pHealth Region Urban;

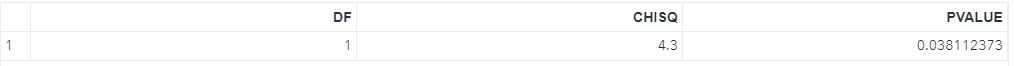
MODEL bmipct = meducation ChildRace HouseIncome pHealth TV Region Urban PctMinority/ solution ddfm = bw;

RANDOM intercept TV/ subject=school\_ID TYPE=UN;

**RUN**; \*Model NOT Working;





All fixed effects were significant except for numeric houseincome1. The variance of the random slopes were significant, and the deviance test suggests that this model is significantly better than the previous one. So this random slope model is significantly better than the previous one. (THIS TURNS OUT TO BE THE BEST FIT MODEL.)

# Model 5: Cross level interactions

I added all interactions between the random slope household income and school level variables.

The statistical model:

**Level 1** : bmipctij = b0j + b1 meducationij + b2 ChildRaceij + b3 HouseIncomeij + b4 pHealthij + b5 TVij + eij

**Level 2 (intercept)**:b0j = g00 + g01Regionj + g02Urbanj + g02PctMinorityj + g01PCTFreeLunchj + u0j

**Level 2 (slope)**: b3j = g30 + g31Regionj + g32Urbanj + g32PctMinorityj + u3j

**Combined model**: bmipctij = (g00 + b1 meducationij + b2 ChildRaceij + g30 HouseIncomeij + b4 pHealthij + b5 TVij g01Regionj + g02Urbanj + g02PctMinorityj + g31Regionj\* HouseIncomeij + g32Urbanj\* HouseIncomeij + g32PctMinorityj\* HouseIncomeij) + (u0j + u1j \* meducationij + eij )

SAS code:

/\* 31 cross level interaction - add pHealth\*/

**PROC** **MIXED** DATA=HW2.DATHW2RSLOPE noclprint covtest noitprint method=reml;

CLASS school\_ID meducation ChildRace pHealth Region Urban;

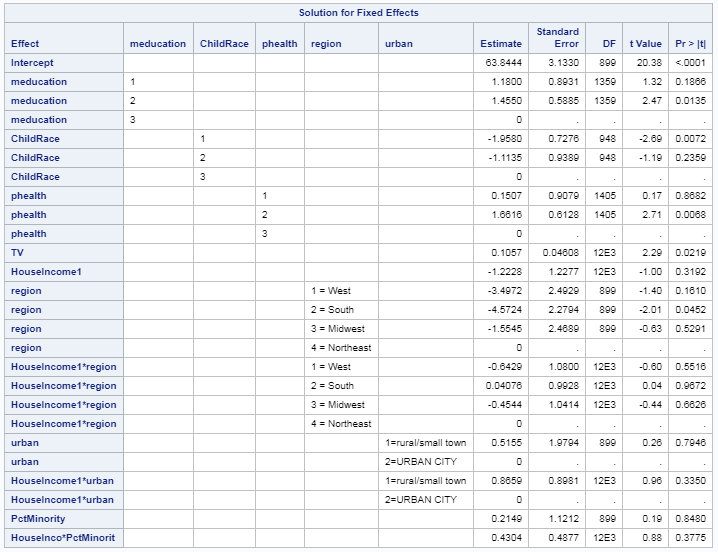
MODEL bmipct = meducation ChildRace pHealth TV HouseIncome1|Region HouseIncome1|Urban HouseIncome1|PctMinority/ solution ddfm = bw;

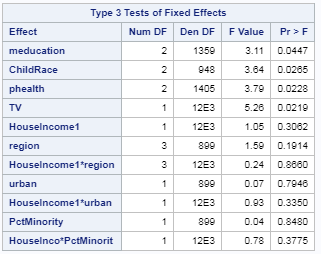
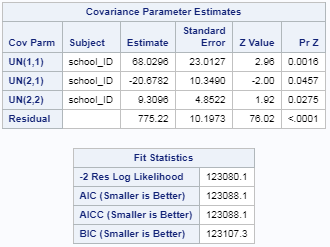
RANDOM intercept HouseIncome1/ subject=school\_ID TYPE=UN;

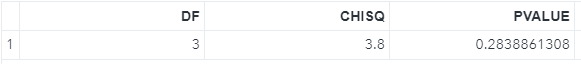
**RUN**; \*-2loglikelihood: 123080.1;

According to the SAS outputs, all the interactions were not significant per the type 3 test. The model fit is not significantly better than the previous no interaction model, as suggested by the deviance test. Therefore, the previous one was the optimal model.

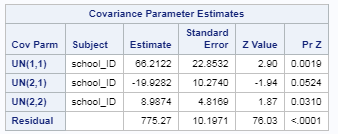
SAS outputs:





# 6: Interpretation



**Interpretation on the variance-covariance matrix:**

The variance/covariance matrix suggests that the variance of the random intercept UN(1, 1) and random slope UN(2, 2) were both significant across schools. But the covariance UN(2, 1) between them were not significant.

**Interpretation on the coefficients:**

The coefficients table below suggests that high mother’s education, poor the parents’ health, residence in rural areas, and studying in high percent of minority students were associated with higher age and sex adjusted BMI. On the other hand, being white or black, low household income, and living in the West, South or Midwest compared to Northeast, were associated with higher age and sex adjusted BMI.

