Problem Set 2: Introduction to Multilevel Modeling

Due: Sunday 9/24 (at 11:59pm)

Overview

This problem set has a few questions focused on thinking about, fitting, and understanding multilevel models. The goal of this assignment is to solidify your foundational understanding of what multilevel models are; after this assignment we are going to give you some data analysis questions, where you will analyze an actual dataset using these tools. For this problem set, each problem is an independent set of small questions, and you can do them in any order.

What to turn in

Turn in a report as a PDF file, ideally made with R Markdown.[1] Before submitting, review the Assignment Guidelines.[2]

[1] You can, if you must, print to PDF from, e.g., MS Word. See references on PSet 0, or the solution for PSet 0 for more pointers on report generation.

[2] Assignment Guidelines

Before you start

Before you start, please be sure you have looked at the course readings for Packets 2.1 and 2.2.[3] Some concepts in these homework problems may not have been explicitly discussed in lecture.

[3] Course readings are listed on Canvas. See packet pages for an annotated reading list.

The Problems

1. Neighborhood-effects data[4]

[4] This problem is a variant of 3.1, pg 172, from RH&S

This problem uses the data from last problem set (and Section 1), from the study where Garner and Raudenbush (1991), Raudenbush and Bryk (2002), and Raudenbush et al. (2004) considered neighborhood effects on educational attainment for young people who left school between 1984 and 1986 in one education authority in Scotland.

The dataset `neighborhood.dta` has the following variables:

Level 1 (students)

- `attain`: a measure of end-of-school educational attainment capturing both attainment and length of schooling (based on the number of O-grades and Higher SCE awards at the A—C levels)
- `p7vrq`: verbal—reasoning quotient (test at age 11—12 in primary school)
- `p7read`: reading test score (test at age 11—12 in primary school)

- `dadocc`: father's occupation scaled on the Hope—Goldthorpe scale in conjunction with the Registrar General's social-class index (Willms 1986)
- `dadunemp`: dummy variable for father being unemployed (1: unemployed; 0: not unemployed)
- `daded`: dummy variable for father's schooling being past the age of 15
- `momed`: dummy variable for mother's schooling being past the age of 15
- `male`: dummy variable for student being male

Level 2 (neighborhoods)

- `neighid`: neighborhood identifier
- `deprive`: social—deprivation score derived from poverty concentration, health, and housing stock of local community.

Your tasks:

a) Fit a random intercepts model with `attain` as the response variable and without any covariates. What are the estimated variance components between and within neighborhoods?

b) Calculate the estimated intraclass correlation (ICC) and give a sentence of interpretation.

c) Include the covariate `deprive` in the model and interpret the estimates. Discuss the changes in the estimated standard deviations of the random intercept and level-1 residual.

d) Extend the `lmer()` command fitting your model from above to include all the student-level covariates.

e) Comment on *whether* and *why* all the estimated standard deviations have changed from (c) to (d). You do *not* need to interpret the student coefficients.

f) Obtain the overall coefficient of determinations R^2 for the models in part (c) and (d).

Using all your above results, briefly discuss whether our covariates are explaining much. (The R^2 values you just calculated can be interpreted just like the R^2 values you get from simple OLS.)

We can estimate the R^2 value by comparing the total unexplained (left over) variance of a model without covariates to a model with covariates. In either case, the *total variation* is going to be $\sigma^2_{total} = \sigma^2_\alpha + \sigma^2_y$. We then have the following for R^2, if we have σ^2_{null} be the total variance of our "null" model with no covariates, and σ^2_{full} the total variance of our model with all covariates:

$$R^2 = 1 - \sigma^2_{full}/\sigma^2_{null}.$$

2. Model building and method selection[5]

[5] This problem does not involve coding or actual data. Often model equations or pseudocode can help with answering the question clearly, but this is not required.

You want to describe the difference in posted starting salaries of teachers depending on the characteristics of the hiring school (in particular, the median household income of the neighborhood the school is in). For a collection of job postings spanning about a decade, you have a sample of schools nested in around 50 different districts. Within a given district, you generally have 4 to 7 schools, mostly with different values of median neighborhood income (you only have a single measure of neighborhood income, not measures across time). You also have the district-level median household income.

a) You are interested in understanding whether districts tend to offer more pay to teachers in lower-income schools as an incentive. Describe a model (linear regression, fixed effects, multilevel, or something else) that would allow you to investigate this question. Be sure to specify what kind of standard errors you would use.

b) Now you are interested in whether richer districts are paying more than poorer districts. What model would you use for this question?

c) Come up with and describe an alternate modeling approach for either (a) or (b), and identify a pro or a con with this alternate approach.

If you wish to simplify this problem you can elect to dichotomize household income into “high” and “low”.