PS 2: Jen Xin and Marshae

2023-09-24

```
data <- read.dta13('./PS 1/neighborhood.dta')
```

1. Neighborhood effects data

a) Random-intercept model

The estimated between group neighborhood variance is 0.202 and the within neighborhood variance is 0.804.

```
m1 <- lmer(attain ~ 1 + (1 | neighid), data = data)

summary(m1, digits = 3)
```

```
## Warning in summary.merMod(m1, digits = 3): additional arguments ignored
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: attain ~ 1 + (1 | neighid)
##    Data: data
##
## REML criterion at convergence: 6421.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.33301 -0.65534  0.01499  0.58163  2.96234
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  neighid  (Intercept) 0.2024   0.4498
##  Residual             0.8044   0.8969
## Number of obs: 2310, groups:  neighid, 524
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.08201    0.02847   2.881
```

```
display(m1, digits = 3) #sd for the between and within. Thinking that attain is a measure of educationa
```

```
## lmer(formula = attain ~ 1 + (1 | neighid), data = data)
## coef.est  coef.se
##    0.082    0.028
##
## Error terms:
```

```
##  Groups   Name        Std.Dev.
##  neighid  (Intercept) 0.450
##  Residual             0.897
## ---
## number of obs: 2310, groups: neighid, 524
## AIC = 6427.3, DIC = 6410.7
## deviance = 6416.0
```

b) ICC

The ICC is the the proportion of the variance explained by the grouping structure (neighborhood) in the population (i.e how much variance is explained by the neighborhood random effect). An ICC of 0.201 indicates that there is fairly little variation between the neighborhoods, and most of the variation is at the student level. In other words, 20.1% of the total variation is due to neighborhood effects.

```
## By hand. Between group variance / (within group variance + between group variance).
sigma.alpha = sigma.hat(m1)$sigma$neighid

sigma.y = sigma(m1)

ICC = sigma.alpha^2/(sigma.alpha^2 + sigma.y^2)

ICC
```

```
## (Intercept)
##   0.2010082
```

```
##using sjstats package
performance::icc(m1)
```

```
## # Intraclass Correlation Coefficient
##
##     Adjusted ICC: 0.201
##   Unadjusted ICC: 0.201
```

c) Adding a covariate

Compared with the model in part (a), the standard deviation of the random intercept is 0.295 (after adding deprive) vs 0.450 (before adding deprive); a decrease in the standard deviation of the random intercept once we have added the fixed effect for deprive. **The neighborhood-level variation has declined after adding a level 2 covariate deprive**; deprive can explain some of the between neighborhood variation, which makes sense for a level 2 covariate.

The standard deviation of the level-1 residual did not change much; 0.897 before adding the covariate and 0.901 after adding the covariate.

The estimated effect estimate for deprive is -0.521. On average, for every unit increase in neighborhood social deprivation score, students at the neighborhood tend to have an estimated 0.521 unit decrease in end of school educational attainment, accounting for clustering by neighborhood.

```
m2 <- lmer(attain ~ deprive + (1 | neighid), data = data)
summary(m2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: attain ~ deprive + (1 | neighid)
##    Data: data
## 
## REML criterion at convergence: 6275.9
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -2.2880 -0.6518  0.0074  0.5833  3.4927 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  neighid  (Intercept) 0.08707  0.2951  
##  Residual             0.81210  0.9012  
## Number of obs: 2310, groups:  neighid, 524
## 
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.09991    0.02373    4.21
## deprive     -0.52068    0.03834  -13.58
## 
## Correlation of Fixed Effects:
##         (Intr)
## deprive -0.045
```

```
display(m2, digits = 3)
```

```
## lmer(formula = attain ~ deprive + (1 | neighid), data = data)
##             coef.est coef.se
## (Intercept)  0.100    0.024 
## deprive     -0.521    0.038 
## 
## Error terms:
##  Groups   Name        Std.Dev.
##  neighid  (Intercept) 0.295   
##  Residual             0.901   
## ---
## number of obs: 2310, groups: neighid, 524
## AIC = 6283.9, DIC = 6255.2
## deviance = 6265.6
```

d) Adding more covariates

```
m3 <- lmer(attain ~ deprive  + p7vrq + p7read +
             + dadocc + dadunemp + daded + momed + male + (1 | neighid), data = data)
summary(m3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: attain ~ deprive + p7vrq + p7read + +dadocc + dadunemp + daded +  
##     momed + male + (1 | neighid)
##    Data: data
```

```
##
## REML criterion at convergence: 4840
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9277 -0.6378 -0.0329  0.5699  3.5300
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  neighid  (Intercept) 0.006758 0.08221
##  Residual             0.457945 0.67672
## Number of obs: 2310, groups:  neighid, 524
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  0.080752   0.023027   3.507
## deprive     -0.148122   0.025439  -5.823
## p7vrq        0.027761   0.002260  12.282
## p7read       0.026062   0.001750  14.896
## dadocc       0.008232   0.001366   6.026
## dadunemp    -0.114919   0.046921  -2.449
## daded        0.140874   0.040890   3.445
## momed        0.062401   0.037527   1.663
## male        -0.055392   0.028488  -1.944
##
## Correlation of Fixed Effects:
##          (Intr) depriv p7vrq  p7read dadocc dadnmp daded  momed
## deprive  -0.026
## p7vrq    -0.108  0.057
## p7read    0.097  0.083 -0.767
## dadocc    0.100  0.155 -0.056 -0.083
## dadunemp -0.228 -0.110  0.038  0.017  0.123
## daded    -0.214  0.038  0.002 -0.063 -0.212 -0.009
## momed    -0.240  0.023 -0.017 -0.023 -0.065  0.002 -0.419
## male     -0.596  0.006  0.085 -0.050  0.011  0.020 -0.005 -0.013
```

```r
display(m3, digits = 3)
```

```
## lmer(formula = attain ~ deprive + p7vrq + p7read + +dadocc +
##     dadunemp + daded + momed + male + (1 | neighid), data = data)
##             coef.est coef.se
## (Intercept)  0.081    0.023
## deprive     -0.148    0.025
## p7vrq        0.028    0.002
## p7read       0.026    0.002
## dadocc       0.008    0.001
## dadunemp    -0.115    0.047
## daded        0.141    0.041
## momed        0.062    0.038
## male        -0.055    0.028
##
## Error terms:
##  Groups   Name        Std.Dev.
##  neighid  (Intercept) 0.082
##  Residual             0.677
## ---
## number of obs: 2310, groups: neighid, 524
## AIC = 4862, DIC = 4710
## deviance = 4775.0
```

e) Change in SDs

Compared to model 2, standard deviation of the neighborhood random effect with all covariates added is now 0.082 (vs 0.259 in model 2). A lot of the between neighborhood variation can be explained by the added student-level covariates. The standard deviation of the residual (within neighborhood variation) is now 0.676 (vs 0.901 in model 2); there is also a decline in within neighborhood variation. The standard error on the fixed effect deprivation declined to 0.025 (vs 0.038 in model 2).

f) Coefficient of determinations R^2

For model 2, 10.68% of the variation in the education attainment is explained by the fixed effect of social deprivation score and the random effect of neighborhoods.

For model 3, 53.84% of the variation in the education attainment is explained by the fixed effect of social deprivation score, the fixed effects of all the student-level covariates, and the random effect of neighborhoods.

The R^2 greatly improved by adding the individual level covariates, indicating that these covarariates are important in explaining variation in student educational attainment.

```
total.null = (sigma.alpha^2 + sigma.y^2)

get.r2 <- function(model, total.null){
  sigma.alpha = sigma.hat(model)$sigma$neighid

  sigma.y = sigma(model)

  total = (sigma.alpha^2 + sigma.y^2)

  r2 = 1 - total/total.null

print(r2)
}

# total variance for c
get.r2(m2, total.null)
```

```
## (Intercept)
##   0.1068378
```

```
# total variance for d
get.r2(m3, total.null)
```

```
## (Intercept)
##   0.5384004
```

2. Model building and method selection

Goal: Describe difference in starting salaries of teachers over a decade from different schools nested in different districts. Depends on characteristics of hiring school–median household income of the neighborhood school is in. (level 1 predictors) Level 2 predictors– district level median household income varies.

Per district–4 to 7 schools with different neighborhood income (Teachers in each respective school.) Sample of schools nested within 50 different districts (Allowing us to do random effects)

a) Incentive pay in lower income schools

If we just broadly wanted to investigate whether the average teacher salary is predicted by schools income status, we could use a fixed effect model with cluster robust standard errors where data is completely unpooled no district or school is using information from another to predict the line. In this model we would use teacher starting salary as the outcome and regress this on the median household income of the neighborhood each school is in, fitting . Using fixed effects would allow each school to have their own parameter

We are assuming that income has the same relationship with teacher salary across all schools in all districts, but that some some schools may have higher or lower salary starting points, allowing us to fit a line to each school in each district.

The cluster robust standard errors help alllow for heteroskedasticity and take into the clustered data that we are using.

Outcome: Average Teacher Salary (Continuous) Treatment: (1: High, 0: Low *based on some objective cutoff predicted by neighborhood income)

b) Richer and poorer districts

It may be the case that different districts may have different starting points for salary based on income, and thus look like they pay more or less. To investigate this we could use a random intercept model with mixed effects so that we allow the intercept to vary for each district, while holding the slope constant (fixed). Each school i in a given district j:

We will use district-level median household income as a district-level covariate for district income.

The level 1 regression would look someting like this:

$$salary_{ij} = \alpha_j + neighboroodincome_{ij} + \epsilon_{ij} \quad with\ \epsilon_{ij}\ \sim \mathcal{N}(0, \sigma_y^2)$$

where

$$\alpha_j =\ \mu + medianhouseholdincome_j \quad with\ \mu_j \sim \mathcal{N}(0, \sigma_\alpha^2)$$

Here we are essentially pulling the school district averages towards a grand mean, reducing the variability for smaller clusters, and are essentially borrowing information from other schools and districts to do this. The con of this is that these estimates will be biased but more stable estimates of each district.

c) Alternate approach for b

Run fixed effects model with cluster robust standard errors for each teacher salary predicted by median district income (aggregate neighborhood incomes into one average per district) and compare teacher salary. Given teacher salary is a level 1 variable (per school) you may need to aggregate teacher salaries to get an average for the district to get around the collinearity issue that could arise.

The pro of this is that with fixed effects at least, you are allowing each district to be estimated, rather than aggregated as in a traditional OLS.

A con of this would obviously be loss of precision in our estimates large variance as we are collapsing data. This also means we are making alot of assumptions regarding the intervariation of a county, and only focusing on intra-variation which may minimize inequities that also exist within a county.