Centering in Cross Sectional Data

Andy Grogan-Kaylor

25 May 2021 14:54:52

# Introduction

These notes represent a brief discussion of centering with cross sectional data. Since so much of my current work focuses on cross national work on parenting and child development, I use these ideas as my substantive example.

Consider a cross-national data set where we are attempting to understand predictors of *behavior problems* as a function of *per capita income* and *parental use of physical punishment*.

# Simulate Some Data

Show / Hide Data Simulation Code

. clear all

. set obs 100  
Number of observations (\_N) was 0, now 100.

. generate income = runiform(10000, 70000)

. generate physical\_punishment = rbinomial(1,.3)

. generate country = int(\_n/10) + 1

. generate e = rnormal(0,1) // individual error

. generate u = country - 5 // random intercept

. generate behavior\_problems = 110 + -.0001 \* income + 10 \* physical\_punishment + u + e // plausible regression relationship

. list in 1/10, abb(20) // list out some data  
  
 ┌───────────────────────────────────────────────────────────────────────────────┐  
 │ income physical\_punishment country e u behavior\_problems │  
 ├───────────────────────────────────────────────────────────────────────────────┤  
 1. │ 19044.56 0 1 .4024389 -4 104.498 │  
 2. │ 11784.18 0 1 -.7492127 -4 104.0724 │  
 3. │ 54456.47 0 1 1.950699 -4 102.5051 │  
 4. │ 10323.12 0 1 -.2365216 -4 104.7312 │  
 5. │ 31238.75 0 1 .5317396 -4 103.4079 │  
 ├───────────────────────────────────────────────────────────────────────────────┤  
 6. │ 25779.69 1 1 .3833449 -4 113.8054 │  
 7. │ 46237.15 1 1 -1.336349 -4 110.0399 │  
 8. │ 22314.24 1 1 -.1576294 -4 113.6109 │  
 9. │ 24329.21 1 1 -1.299845 -4 112.2672 │  
 10. │ 35722.85 0 2 -.0299588 -3 103.3978 │  
 └───────────────────────────────────────────────────────────────────────────────┘

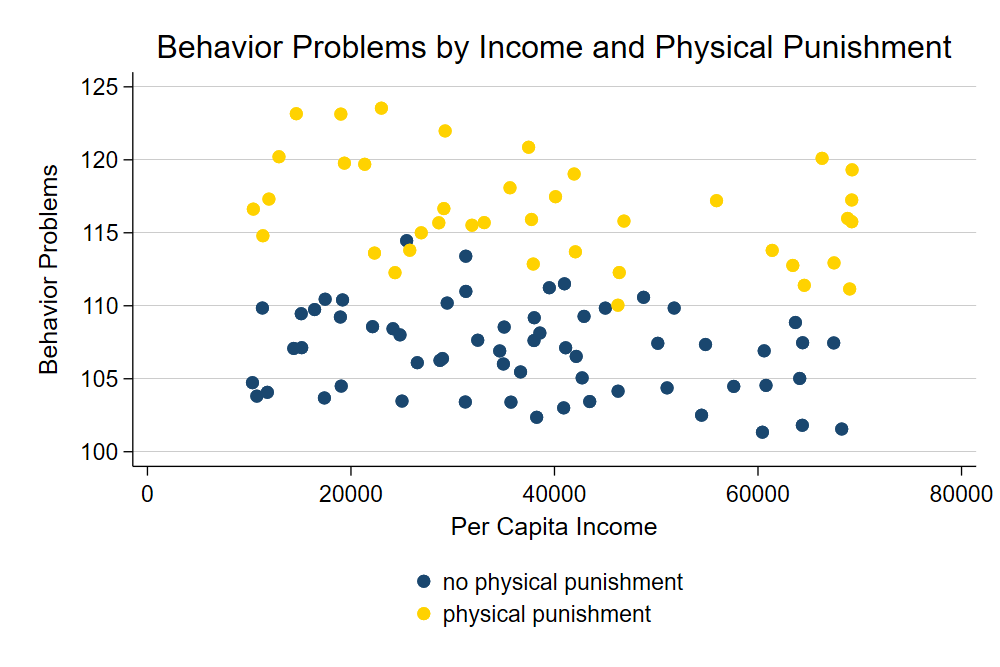
# Uncentered Data

## Equation

## Graph

. twoway (scatter behavior\_problems income if physical\_punishment ==0) ///  
> (scatter behavior\_problems income if physical\_punishment == 1), ///  
> legend(order(1 "no physical punishment" 2 "physical punishment") pos(6)) ///  
> title("Behavior Problems by Income and Physical Punishment") ///  
> xtitle("Per Capita Income") ///  
> ytitle("Behavior Problems") ///  
> scheme(michigan)

. graph export myscatter.png, width(1000) replace  
file myscatter.png saved as PNG format



Scatterplot

## Multilevel Model

. mixed behavior\_problems income physical\_punishment || country:  
  
Performing EM optimization:   
  
Performing gradient-based optimization:   
  
Iteration 0: log likelihood = -165.77618   
Iteration 1: log likelihood = -165.77618   
  
Computing standard errors:  
  
Mixed-effects ML regression Number of obs = 100  
Group variable: country Number of groups = 11  
 Obs per group:  
 min = 1  
 avg = 9.1  
 max = 10  
 Wald chi2(2) = 2395.48  
Log likelihood = -165.77618 Prob > chi2 = 0.0000  
  
────────────────────┬────────────────────────────────────────────────────────────────  
 behavior\_problems │ Coefficient Std. err. z P>|z| [95% conf. interval]  
────────────────────┼────────────────────────────────────────────────────────────────  
 income │ -.0000898 6.68e-06 -13.44 0.000 -.0001029 -.0000767  
physical\_punishment │ 10.08122 .211381 47.69 0.000 9.666924 10.49552  
 \_cons │ 110.447 .9834314 112.31 0.000 108.5195 112.3744  
────────────────────┴────────────────────────────────────────────────────────────────  
  
─────────────────────────────┬────────────────────────────────────────────────  
 Random-effects parameters │ Estimate Std. err. [95% conf. interval]  
─────────────────────────────┼────────────────────────────────────────────────  
country: Identity │  
 var(\_cons) │ 9.711427 4.253129 4.11622 22.91224  
─────────────────────────────┼────────────────────────────────────────────────  
 var(Residual) │ .9959301 .1494585 .7421465 1.336497  
─────────────────────────────┴────────────────────────────────────────────────  
LR test vs. linear model: chibar2(01) = 173.94 Prob >= chibar2 = 0.0000

We note that -0.898 is the effect of every additional $10,000 of per capita income. 10.081 is the effect of physical punishment. Notably, for this handout, 110.447 is the level of behavior problems for a child who did *not receive physical punishment* living in a family with **$0** *income*.

# Grand Mean Centering

Grand mean centering helps us to have more meaningful intercepts of our continuous variables.

Essentially, we are going to create

## Equation

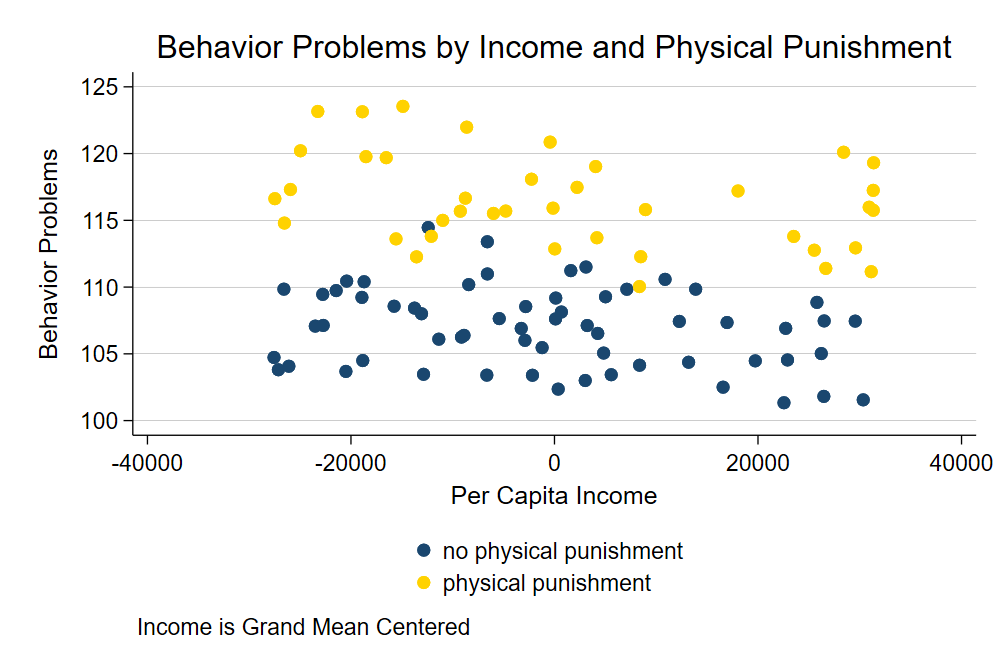
## Graph

. egen m\_income = mean(income) // grand mean of income

. generate c\_income = income - m\_income // grand mean centered income

. twoway (scatter behavior\_problems c\_income if physical\_punishment ==0) ///  
> (scatter behavior\_problems c\_income if physical\_punishment == 1), ///  
> legend(order(1 "no physical punishment" 2 "physical punishment") pos(6)) ///  
> title("Behavior Problems by Income and Physical Punishment") ///  
> caption("Income is Grand Mean Centered") ///  
> xtitle("Per Capita Income") ///  
> ytitle("Behavior Problems") ///  
> scheme(michigan)

. graph export myscatter2.png, width(1000) replace  
file myscatter2.png saved as PNG format



Scatterplot With Grand Mean Centering

In a graph, we see that *grand mean centering* has transformed the intercept so now the term is the level of behavior problems for the *average* child who did *not recieve physical punishment*.

## Multilevel Model

. mixed behavior\_problems c\_income physical\_punishment || country:  
  
Performing EM optimization:   
  
Performing gradient-based optimization:   
  
Iteration 0: log likelihood = -165.77618   
Iteration 1: log likelihood = -165.77618   
  
Computing standard errors:  
  
Mixed-effects ML regression Number of obs = 100  
Group variable: country Number of groups = 11  
 Obs per group:  
 min = 1  
 avg = 9.1  
 max = 10  
 Wald chi2(2) = 2395.48  
Log likelihood = -165.77618 Prob > chi2 = 0.0000  
  
────────────────────┬────────────────────────────────────────────────────────────────  
 behavior\_problems │ Coefficient Std. err. z P>|z| [95% conf. interval]  
────────────────────┼────────────────────────────────────────────────────────────────  
 c\_income │ -.0000898 6.68e-06 -13.44 0.000 -.0001029 -.0000767  
physical\_punishment │ 10.08122 .211381 47.69 0.000 9.666924 10.49552  
 \_cons │ 107.0459 .9511417 112.54 0.000 105.1817 108.9101  
────────────────────┴────────────────────────────────────────────────────────────────  
  
─────────────────────────────┬────────────────────────────────────────────────  
 Random-effects parameters │ Estimate Std. err. [95% conf. interval]  
─────────────────────────────┼────────────────────────────────────────────────  
country: Identity │  
 var(\_cons) │ 9.711427 4.253128 4.11622 22.91224  
─────────────────────────────┼────────────────────────────────────────────────  
 var(Residual) │ .9959301 .1494585 .7421465 1.336497  
─────────────────────────────┴────────────────────────────────────────────────  
LR test vs. linear model: chibar2(01) = 173.94 Prob >= chibar2 = 0.0000

We see that the and regression coefficients have not changed. However, the intercept, has changed, and is now more meaningful.

# Group Mean Centering

In group mean centering, we are doing something slightly different. We are creating a mean for each *group*, which in this data is *country*: e.g. , where is the index for *group* or *country*.

. bysort country: egen m\_g\_income = mean(income) // GROUP mean of income

. generate c\_g\_income = income - m\_g\_income // GROUP mean centered income

. bysort country: egen m\_g\_physical\_punishment = mean(physical\_punishment) // GROUP mean of physical punishment

. generate c\_g\_physical\_punishment = physical\_punishment - m\_g\_physical\_punishment // GROUP mean centered physical punishment

Interestingly, *group mean centering* has many implications. Here I focus on how employing different variables might provide *conceptually* or *theoretically* different results. For the sake of parismony, in the brief discussion below I focus on these *conceptual* or *theoretical* differences, and do not provide output. I use the quietly prefix to suppress output.

## Equation

Two versions of the equation are equally appropriate. Both address *conceptually* or *theoretically* different questions.

### Covariate and Group Mean

One parameterization of the multilevel model is to enter the *covariate* and its *group level mean* i.e. and .

### Group Mean Centered Covariate and Group Mean

A second, equally valid, but conceptually different parameterization of the multilevel model is to enter the *covariate deviated from its group level mean* and the *group level mean* i.e. and .

## Multilevel Model

Again, for the sake of parsimony, I use the quietly prefix to suppress output of the multilevel models.

### Covariate and Group Mean

This first parameterization focuses on *individual scores on covariates* and their *country level means*.

What is the effect of *income*, *country level mean income*, *physical punishment* and *country level mean of physical punishment* on *behavior problems*?

. quietly: mixed behavior\_problems income m\_g\_income physical\_punishment m\_g\_physical\_punishment || country:

### Group Mean Centered Covariate and Group Mean

This second parameterization focuses on how *individuals differ from their country level means*, and *country level means*.

What is the effect of *income deviated from its country level mean*, *country level mean income*, *physical punishment deviated from its country level punishment*, and *country level mean of physical punishment* on *behavior problems*?

. quietly: mixed behavior\_problems c\_g\_physical\_punishment m\_g\_income c\_g\_physical\_punishment m\_g\_physical\_punishment || country: