POLSCI 9592

Lecture 9: Dependent Data

Dave Armstrong

Goals for This Session

1. What are the consequences for our models of having *dependent* data? 2. When we have dependent data, we can make hypotheses about within-unit effects and between-unit effects, what are the consequences of this choice?
2. When we use unit effects to deal with dependent data, are we better off using fixed or random unit effects?
3. How do things get complicated when time is an important variable in our analysis?

Dependent Data

Multilevel data are different than the data we have considered thus far as they have observed data at different levels of aggregation. Some examples:

- Children in classrooms, in schools, in districts, in states.
- Voters in towns, in states
- Democracy measured at different years within each country.

The prominent feature here is that observations can be thought of as "nested" in groups.

- These groups can have attributes of their own that we might care about.

The Independence Assumption

- Recall that OLS assumes that each observation is independent of the others
- More specifically, the error term, or equivalently the Y-values, are independent of each other
- Although the assumption of independence is rarely perfect, in practice a random sample from a large population provides a close approximation
- Time-series data, panel data and clustered data often do not satisfy this condition
- In these cases, dependencies among the errors can be quite strong
- If independence is violated, OLS is no longer the optimal estimation method as standard errors are biased downwards

What happens if we ignore the multilevel nature of the data?

- Calculation of standard errors involves consideration of the sample size in the denominator of the formula:
- Answer to Q1: When the observations are not independent, the effective sample size is smaller than what we observe and thus an adjustment must be made to the standard errors or they will be biased downwards.

$$SE(\bar{x}) = \frac{S_x}{\sqrt{n}}$$

$$\begin{aligned} SE(B) &= \frac{S_E}{\sqrt{\sum(x_i - \bar{x})^2}} \\ &= \frac{S_y}{S_x}\sqrt{\frac{1 - r_{xy}^2}{n-2}} \end{aligned}$$

Within and Between Effects

The choice of estimator for grouped data comes down to a choice about what sorts of things you want to know. If you want to know something about between unit variance, then you need a between estimator, e.g.:

$$\bar{Y}_{\cdot j} = \delta_0 + \delta_1 \bar{X}_{\cdot j} + \nu_j$$

However, if you want to know something about the relationship of Y and X *within* groups (i.e., among individuals in each group), then you need a within estimator, e.g.:

$$(Y_{ij} - \bar{Y}_{\cdot j}) = \delta_0 + \delta_1 (X_{ij} - \bar{X}_{\cdot j}) + (u_{ij} - \bar{u}_{\cdot j})$$

Some models are compromises that allow us to learn something about both types of relationships.

Snijders and Bosker's Example (1)

- Assume the following two-level artificial data, with 5 groups containing 2 observations each:
- We now fit several models: (1) a total regression, (2) a regression between group means, (3) within group regressions, (4) multilevel model

j	i	X_{ij}	$\bar{X}_{\cdot j}$	Y_{ij}	$\bar{Y}_{\cdot j}$
1	1	1	2	5	6
1	2	3	2	7	6
2	1	2	3	4	5
2	2	4	3	6	5
3	1	3	4	3	4
3	2	5	4	5	4
4	1	4	5	2	3
4	2	6	5	4	3
5	1	5	6	1	2
5	2	7	6	3	2

S&B Example (2)

- Equation 3 stems from the individual regressions of Y_i on X_i within each group. It simplifies to this equation only when the slopes for each group are the same
- The multilevel regression (4) writes Y_{ij} as a function of the within group and between group relations between Y and X

1. Total regression $(Y_{ij} \sim X_{ij})$

$$\hat{Y}_{ij} = 5.33 - 0.33X_{ij}$$

2. Between Group Means $(Y_{\cdot j} \sim X_{\cdot j})$

$$\hat{Y}_{\cdot j} = 8 - 1\bar{X}_{\cdot j}$$

3. Within Groups $(Y_{ij} - \bar{Y}_{\cdot j} \sim X_{ij} - \bar{X}_{\cdot j})$

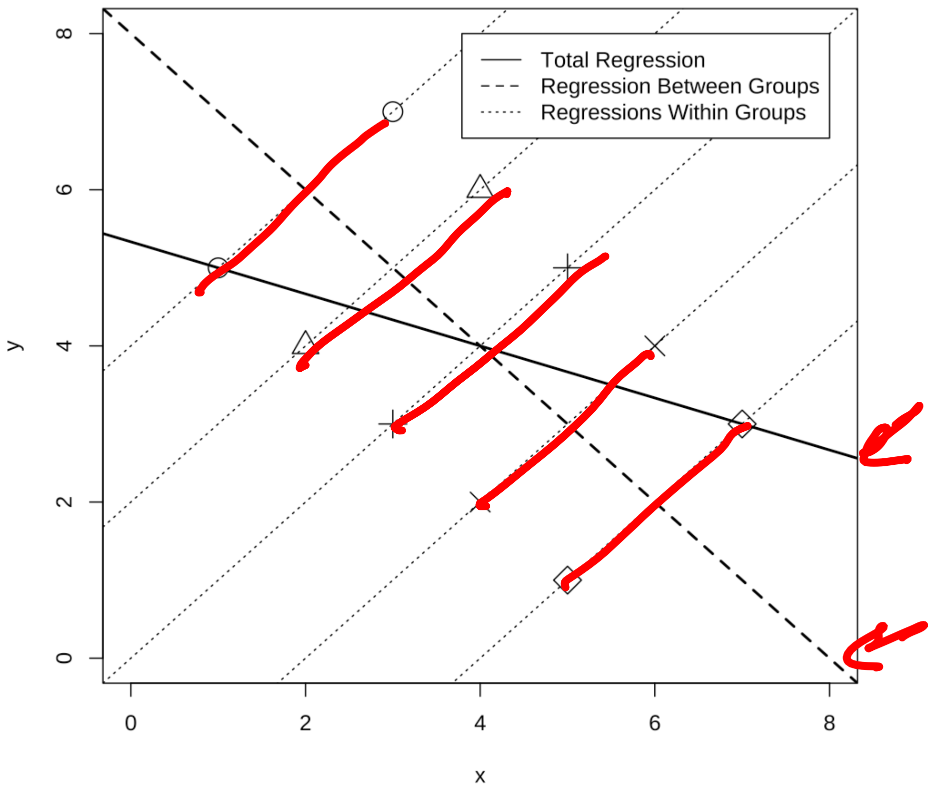
$$\hat{Y}_{ij} = \bar{Y}_{\cdot j} + 1(X_{ij} - \bar{X}_{\cdot j})$$

4. Multilevel Regression:

$$\begin{aligned}\hat{Y}_{ij} =& 8.00 - 1.00\bar{X}_{\cdot j} \\ &+ 1.00(X_{ij} - \bar{X}_{\cdot j})\end{aligned}$$

Snijders and Bosker's Example (3)

- Notice that both the total regression and the between group regression have done a poor job of capturing the trend in the data
- In fact the regression between groups is in the opposite direction to the individual within group regressions



Multilevel Regression Model: Varying Intercepts

Imagine we have the following model:

$$Y_{ij} = \beta_{0j} + \beta_1 x_{ij} + R_{ij}$$

where

$$\beta_{0j} = \gamma_{00} + U_{0j}$$

The new equation is:

$$Y_{ij} = \gamma_{00} + \beta_1 x_{ij} + U_{0j} + R_{ij}$$

The U_{0j} are deviations in the group intercepts from the overall mean γ_{00}.

U_{0j} - fixed or random?

If we think of U_{0j} as fixed, we estimate a parameter for each of U_{0j} (save 1 for identification).

- These group level parameters are by definition going to be perfectly collinear with any explanatory variable that only varies by group.

  If we think of U_{0j} as random, then we think of them as "group residuals", given $\mathbf{X}$.

- This is appropriate if groups are "exchangeable" - all drawn from a population of groups.
- There is *one* parameter associated with the groups in this model and that is the variance of the group effects.

Fixed or Random?

1. If the groups are of interest and researchers want to make inferences about the differences between groups, the fixed effects should be used.

2. If the groups are considered to be sampled from a population of groups and the research wants to make inferences pertaining to the population of groups, random effects should be used.

3. If the researcher wants to test propositions about the effects of group level variables - random coefficients should be used (we'll talk more about this later with Pluemper and Troger's piece).

4. If group sizes are small - random effects can help leverage strength across the groups if the assumption of exchangeability holds.

5. Random effects models work best when the assumption of approximate normality of U_{0j} and R_{ij} holds.

Bell and Jones

Bell and Jones make the following suggestion:

- Use both the within and between transformations to identify the different effects.

Mixed Models Example: British Context Data

The data are a subset from the 1997 British Election Study using a multistage sample where first parliamentary constituencies were randomly selected and then voters were randomly selected from within the selected constituencies (N=2141, 136 Constituencies)

- LRSCALE: left-right values scale
- PROFMAN: $\%$ professionals in constituency
- AGE
- SEX
- Degree (respondent obtained a univ degree)
- INCOME (household income)
- PANO: identifies the constituency

Hypotheses

1. Household income affects attitudes
2. More specifically, the richer one is, the more right-wing on average their attitudes will be
3. This relationship will differ depending on the area in which one lives - if they live in a rich constituency, they will be more likely to hold right-wing attitudes regardless of their own wealth
4. In other words, we expect their to be random effects for income

We will examine both whether the intercept varies across constituencies (indicating on average that overall attitudes differ) and whether the slope for income varies (indicating that income has different effects according to constituency)

Multilevel Regression Models: Within and Between Unit Effects

Hierarchical or Multilevel Models offer us a way to learn something about both within- and between-unit effects. We can formulate the model as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + R_{ij}$$

Here:

- γ_{00} represents the grand mean of Y_{ij}
- γ_{10} is the coefficient relating x_{ij} to Y_{ij}
- U_{0j} is the unit-specific deviation from the grand-mean (a unit-specific intercept shift)
- R_{ij} is the observation-specific residual from the regression line $\gamma_{00} + \gamma_{10} x_{ij} + U_{0j}$

Modeling both Between and Within Effects

We can allow these to be different by including the group specific variables - in this case, the group mean of income:

```
remotes::install_github("davidaarmstrong/uwo9592")
```

```
library(rio)
library(uwo9592)
dat <- import("data/context.dta")
dat <- dat %>% mutate(SEX = rio::factorize(SEX))
betweendat <- make_between_data(LRSCALE ~ AGE + SEX + INCOME + PROFMAN, dat, id="PANO")

library(lme4)
mod <- lmer(LRSCALE ~ INCOME_b1 + INCOME_w1 + AGE_b1 + AGE_w1 +
    SEXmen_b1 + SEXmen_w1 + PROFMAN_b1 + (1|PANO), data=betweendat)
s <- summary(mod, corr=FALSE)
round(s$coefficients, 3)
```

```
##               Estimate Std. Error t value
## (Intercept)      9.410      1.021   9.219
## INCOME_b1        0.341      0.058   5.848
## INCOME_w1        0.157      0.020   7.831
## AGE_b1           0.041      0.019   2.179
## AGE_w1           0.026      0.005   5.236
## SEXmen_b1        1.246      0.679   1.836
## SEXmen_w1        0.148      0.161   0.921
## PROFMAN_b1       0.032      0.015   2.117
```

Interpretation

- In the example above, the coefficient on `INCOME_w1` term suggests that within groups, as individuals have higher levels of income, they have more right-wing attitudes.
- The Between-unit coefficient (`INCOME_b1`) suggests that individuals who are in more affluent constituencies (as measured by the mean of income) have intercepts that are on average more right-wing. So, on the whole, richer constituencies are on average more right-wing in nature.

Obtaining Estimates of U_{0j}

I will give the formula for the empty model (without variables). It gets increasingly more complicated with other data in the model, but R will do that for us.

$$
\begin{aligned}
\hat{U}_0 &= \lambda_j \hat{\beta}_{0j} + (1 - \lambda_j)\hat{\gamma}_{00} \\
\hat{\gamma_{00}} &= \sum_{j=1}^{N} \frac{n_j}{M} \bar{Y}_{\cdot j} \\
\beta_{0j} &= \bar{Y}_{\cdot j} \\
\lambda_j &= \frac{\tau_0^2}{\tau_0^2 + \frac{\sigma^2}{n_j}}
\end{aligned}
$$

The consequence of this estimation procedure is to pull all observations toward the estimated population mean of γ_{00}.

Estimates of U_{0j} from R

```
head(ranef(mod)$PANO)
```

```
##   (Intercept)
## 1  0.03742999
## 2 -0.22809604
## 3 -0.47200633
## 4 -0.09535114
## 5 -0.01363689
## 6 -0.47573646
```

p-values

Note that none of the model summaries come with p-values (this is considered a feature rather than a flaw, see here for a discussion). There are a few ways you can accomplish this.

```
library(LMERConvenienceFunctions)
pamer.fnc(mod)
```

```
##            npar   Sum Sq  Mean Sq F value upper.den.df upper.p.val lower.den.df
## INCOME_b1     1 872.7246 872.7246 69.6000         2133      0.0000         1997
## INCOME_w1     1 540.3118 540.3118 43.0900         2133      0.0000         1997
## AGE_b1        1 101.4241 101.4241  8.0886         2133      0.0045         1997
## AGE_w1        1 349.6833 349.6833 27.8873         2133      0.0000         1997
## SEXmen_b1     1  40.8009  40.8009  3.2539         2133      0.0714         1997
## SEXmen_w1     1  10.6267  10.6267  0.8475         2133      0.3574         1997
## PROFMAN_b1    1  56.2162  56.2162  4.4833         2133      0.0343         1997
##            lower.p.val expl.dev.(%)
## INCOME_b1       0.0000       2.8885
## INCOME_w1       0.0000       1.7883
## AGE_b1          0.0045       0.3357
## AGE_w1          0.0000       1.1574
## SEXmen_b1       0.0714       0.1350
## SEXmen_w1       0.3574       0.0352
## PROFMAN_b1      0.0344       0.1861
```

P-values II

There is a nice discussion online of a few different methods here

```
library(lmerTest)
mod <- update(mod)
car::Anova(mod)
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: LRSCALE
##               Chisq Df Pr(>Chisq)    
## INCOME_b1  34.1982  1  4.978e-09 ***
## INCOME_w1  61.3188  1  4.854e-15 ***
## AGE_b1      4.7468  1    0.02935 *  
## AGE_w1     27.4188  1  1.638e-07 ***
## SEXmen_b1   3.3714  1    0.06634 .  
## SEXmen_w1   0.8475  1    0.35727    
## PROFMAN_b1  4.4833  1    0.03423 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing Equality of Between and Within Effects

```r
library(car)
linearHypothesis(mod, "INCOME_b1 = INCOME_w1")
```

```
##
## Linear hypothesis test:
## INCOME_b1 - INCOME_w1 = 0
##
## Model 1: restricted model
## Model 2: LRSCALE ~ INCOME_b1 + INCOME_w1 + AGE_b1 + AGE_w1 + SEXmen_b1 +
##     SEXmen_w1 + PROFMAN_b1 + (1 | PANO)
##
##   Df  Chisq Pr(>Chisq)
## 1
## 2  1 8.9427   0.002786 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Robust Standard Errors

Robust standard errors are a post-hoc fix to the standard errors that takes into account the non-independence of observations within cluster.

- This is often used as an alternative to a multilevel model.
    - This doesn't inherently estimate a between-and-within effects model, though you can use between- and within-transformed variables.
    - It doesn't account for the entirety of the unit effect, only that part of the unit effect that could be explained by the between-variables.
- A similar result can be obtained with the multilevel model, but it provides more information and a better statistical foundation.

Below is an example:

RSE Example

Multilevel Model

```
library(lmtest)
printCoefmat(summary(mod)$coefficients[,-3], digits = 3)
```

```
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  9.40996    1.02069    9.22  6.5e-16 ***
## INCOME_b1    0.34111    0.05833    5.85  3.2e-08 ***
## INCOME_w1    0.15670    0.02001    7.83  7.8e-15 ***
## AGE_b1       0.04089    0.01877    2.18    0.031 *  
## AGE_w1       0.02644    0.00505    5.24  1.8e-07 ***
## SEXmen_b1    1.24623    0.67872    1.84    0.068 .  
## SEXmen_w1    0.14776    0.16050    0.92    0.357    
## PROFMAN_b1   0.03173    0.01498    2.12    0.036 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Robust Standard Errors

```
mod1 <- lm(LRSCALE ~ INCOME_b1 + INCOME_w1 + AGE_b1 + AGE_w1 +
    SEXmen_b1 + SEXmen_w1 + PROFMAN_b1, data=betweendat)
coeftest(mod1, vcov. = sandwich::vcovCL, df=135, cluster=~PANO)
```

```
## 
## t test of coefficients:
## 
##              Estimate Std. Error t value  Pr(>|t|)    
## (Intercept) 9.3736256  1.0581934  8.8581 4.128e-15 ***
## INCOME_b1   0.3382678  0.0548124  6.1714 7.396e-09 ***
## INCOME_w1   0.1566964  0.0167873  9.3342 2.731e-16 ***
## AGE_b1      0.0406847  0.0191866  2.1205   0.03579 *  
## AGE_w1      0.0264416  0.0053893  4.9063 2.629e-06 ***
## SEXmen_b1   1.2701148  0.7436063  1.7080   0.08993 .  
## SEXmen_w1   0.1477559  0.1555917  0.9496   0.34399    
## PROFMAN_b1  0.0336335  0.0146932  2.2891   0.02363 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Review

1. What are the consequences for our models of having *dependent* data? 2. When we have dependent data, we can make hypotheses about within-unit effects and between-unit effects, what are the consequences of this choice?
2. When we use unit effects to deal with dependent data, are we better off using fixed or random unit effects?
3. How do things get complicated when time is an important variable in our analysis?

Exercise

We're using data from the World Values Survey 2005-2009 Wave. The variables in the dataset are:

- age - Respondent age
- country - Respondent country/region of residence
- democ_duties - Scale that increases in the duties required by democracy
- educ - Highest educational level attained
- income - Household income scale
- lrself - Position on left-right ideological spectrum
- moral - Traditional moral values scale
- religimp - Importance of religion in respondent's life
- religperson - Whther R considers him/herself a religious person
- sex - Sex
- survyear - Year of the survey

1: WVS

There are two variables we're interested in modeling here `moral` or `democ_duties`. Use the variables `age`, `sex`, `educ`, `income` and `religimp` as the independent variables.

```r
library(rio)
dat <- import("data/wvs2005_2009_mlm.dta")
```

1. Fit a fixed effects model for the country effects. Fixed effects means "within".
2. Fit a random effects model to the data.
3. Fit a random effects model using the method described by Bell and Jones.
4. Compare the effects, what do you learn from the models.