POIR 613: Measurement Models and Statistical Computing

Pablo Barberá

School of International Relations
University of Southern California
`pablobarbera.com`

Course website:

pablobarbera.com/POIR613/

Today

1. Solutions for last week's challenge
2. Discussion: survey research in the digital age
3. Multilevel regression
4. Multilevel regression and post-stratification
5. MRP with R

Multilevel regression and post-stratification

Credit where it's due

References for the materials in these slides:

- Course materials for "Applied Multilevel Regression" by Zoltan Fazekas
- Gelman and Hill, 2007, "Data Analysis Using Regression and Multilevel/Hierarchical Models", Cambridge University Press
- Kastellec et al, 2016, "Estimating State Public Opinion With Multi-Level Regression and Poststratification using R"

How many survey respondents would you need to identify with 0.05 significance and 80% power a difference of 4 percentage points in vote share?

```
> power.prop.test(p1=0.48, p2=0.52, power=0.80,
sig.level=0.05)

     Two-sample comparison of proportions power
      calculation

              n = 2451.596
             p1 = 0.48
             p2 = 0.52
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

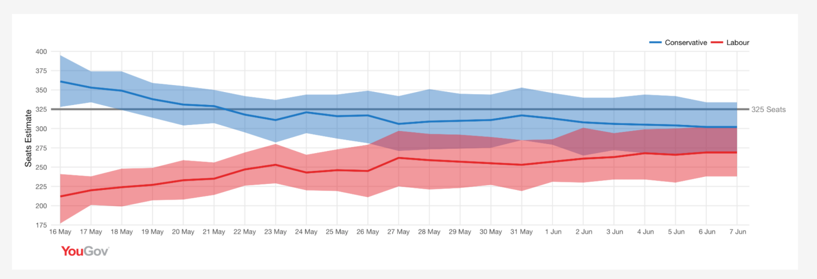
Small-area estimation

Common issue in survey research: interest in estimates of public opinion for subnational units, but small sample size means high margins of error.

Solutions?

1. Aggregate multiple small-sample polls and compute weighted average to reduce noise
   - Fivethirtyeight model
   - Add prior information and model correlations across states (assuming common shifts in vote shares)
2. Multilevel-regression with post-stratification (MRP):
   - Model relationship between demographic/political variables and outcomes of interest
   - Compute cell-level estimates of outcome variable
   - Use population weights to aggregate cell-level estimates

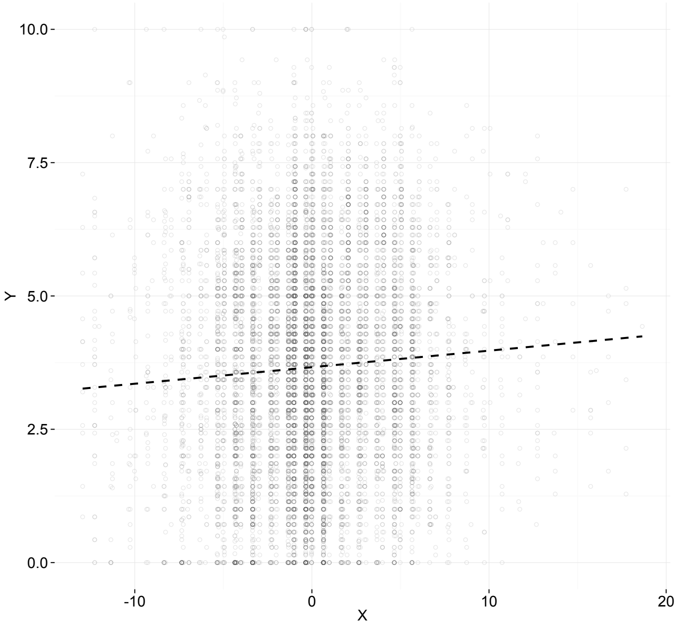
MRP works



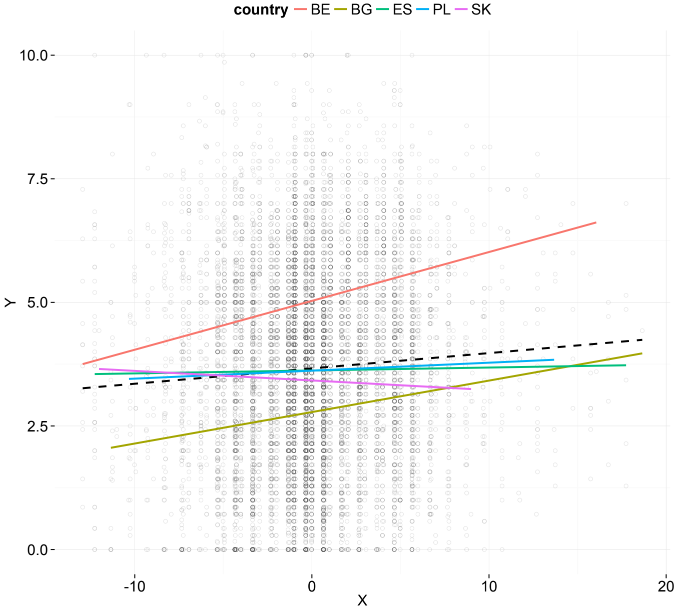
YouGov's 2017 UK election predictions, by Ben Lauderdale, Doug Rivers, and Jack Blumenau

Multilevel regression models

Motivation



Motivation



Quick summary

- Multilevel (hierarchical) modeling is a generalization of linear and generalized linear modeling in which data is structured in groups and regression coefficients themselves can vary by group, as a function of parameters also estimated from data (Gelman and Hill 2007).
- ... researchers should be aware that multilevel models are data intensive & [...] we should be equally aware that multilevel models are theory intensive (Steenbergen and Jones 2002, p.234).
- This does not only mean estimation complexity, it also means complexity in interpretation: how to get quantities of interest, how these *should be* interpreted, which element of the model tells the real story, etc.

Basics of multilevel modeling

- Applicable when our data has a *hierarchical* structure, where level-1 observations are *nested* within level-2
- For example:
  - L1 individuals - L2 countries, districts, regions, etc. - very common in comparative political science
  - L1 measurement - L2 individuals - panel setting
  - Or just TSCS models, for example
- Customarily, n is the sample size/number of observations for L1, and J is the the sample size/number of observations for L2
- Do we have n independent observations in reality, assuming that the data is indeed clustered (or observations are nested within L2)?

Notation

Standard OLS regression:

$$y_i = X_i\beta + \varepsilon_i \qquad \text{for } i = 1, \ldots, n$$

OR

$$y_i \sim \mathrm{N}(X_i\beta, \sigma^2) \qquad \text{for } i = 1, \ldots, n$$

- y is a vector of length n; we use $_i$ to denote the i^{th} row (observation, i.e., individual)
- k predictors, including constant (we will also denote this α, or in other notational convention β_0)
- X is an $n \times k$ matrix of predictors, where $X_i^1 = 1$
- ε, errors assumed to stem from $\mathrm{N}(0, \sigma^2)$ – normal distribution with a mean of 0 and standard deviation σ

Generalizing OLS

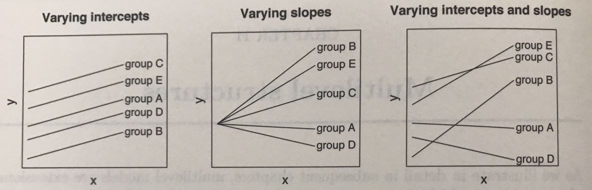


Figure 11.1 *Linear regression models with (a) varying intercepts ($y = \alpha_j + \beta x$), (b) varying slopes ($y = \alpha + \beta_j x$), and (c) both ($y = \alpha_j + \beta_j x$). The varying intercepts correspond to group indicators as regression predictors, and the varying slopes represent interactions between x and the group indicators.*

varying-intercept model:

$$y_i = \alpha_{j[i]} + X_i\beta + \varepsilon_i \tag{1}$$

varying-slope model:

$$y_i = \alpha + X\beta_{j[i]} + \varepsilon_i \tag{2}$$

varying-intercept, varying-slope model:

$$y_i = \alpha_{j[i]} + X\beta_{j[i]} + \varepsilon_i \tag{3}$$

Costs and benefits of multilevel modeling

Motivations:

- Accounting for individual- and group-level variation in estimating group-level regression coefficients
- Modeling variation among individual-level regression coefficients
- Estimating regression coefficients for particular groups

Drawbacks:

- Complexity
- Additional modeling assumptions: each level of the model must meet regression assumptions

Pooling

Why do multilevel models tend to work well? Equilibrium between two extremes:

- No pooling: run regression for each group independently (or fixed effect models)
- Complete pooling: ignore hierarchical structure of the data (OLS without fixed effects)

In contrast, partial pooling implies groups with fewer observations *borrow strength* for groups with more observations, e.g.:

$$
\begin{aligned}
y_i &\sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2) && \text{for } i = 1, \ldots, n \\
\alpha_j &\sim N(\mu_\alpha, \sigma_\alpha^2) && \text{for } j = 1, \ldots, J
\end{aligned}
$$

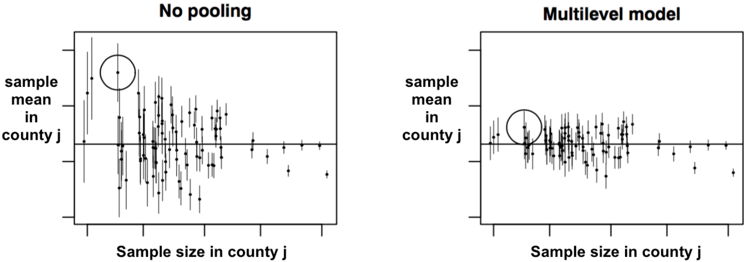


Figure 12.1 *Estimates ± standard errors for the average log radon levels in Minnesota counties plotted versus the (jittered) number of observations in the county: (a) no-pooling analysis, (b) multilevel (partial pooling) analysis, in both cases with no house-level or county-level predictors. The counties with fewer measurements have more variable estimates and larger higher standard errors. The horizontal line in each plot represents an estimate of the average radon level across all counties. The left plot illustrates a problem with the no-pooling analysis: it systematically causes us to think that certain counties are more extreme, just because they have smaller sample sizes.*

Fitting multilevel models in R (lme4 package)

Varying-intercept model with no predictors:

```
m1 <- lmer (y ~ 1 + (1 | county) )
```

Varying-intercept model with individual-level predictor:

```
m1 <- lmer (y ~ x + (1 | county) )
```

Varying-slope model with individual-level predictor:

```
m1 <- lmer (y ~ x + (1 + x | county) )
```

Varying-intercept, varying-slope model with individual-level predictor:

```
m1 <- lmer (y ~ x + (1 + x|county) + (1|county))
```

MRP

MRP

Intuitions:

- Model individual survey responses as a function of *demographic and geographic predictors*, partially pooling respondents across units
- Post-stratify: estimates are weighted by the percentage of each type in the unit

Why is it recommended?

- Outperforms disaggregation when working with small and medium-sized samples
- Produces reasonably accurate estimates of state public opinion using as little as $N = 1,400$
- Poststratification corrects for clustering and other statistical issues that may bias estimates obtained via disaggregation
- More informative about determinants of public opinion
- Estimate opinion in units rarely surveyed

Mechanics of MRP

1. Gather national opinion poll(s)
2. Partition the population into cells based on sociodemographic characteristics
3. Create a separate dataset of state-level predictors.
4. Collect census data to enable poststratification.
5. Fit a regression model for an individual survey response given demographics and geography

$$Pr(y_i = 1) = \mathrm{logit}^{-1}(\alpha + \beta^{\text{race, gender}}_{j[i]} + \beta^{\text{age}}_{k[i]} + \beta^{\text{educ}}_{l[i]} + \beta^{\text{state}}_{s[i]} + \beta^{\text{year}}_{p[i]})$$

$$\beta^{\text{state}}_s \sim N(\alpha^{\text{region}}_{m[s]} + \beta^{\text{relig}} \cdot \mathrm{relig}_s, \sigma^2_{\text{state}}),\ \text{for } s = 1, \ldots, 51$$

6. Poststratify the demographic-geographic types: compute cell-level estimates and weight by proportion in population