Longitudinal Data I

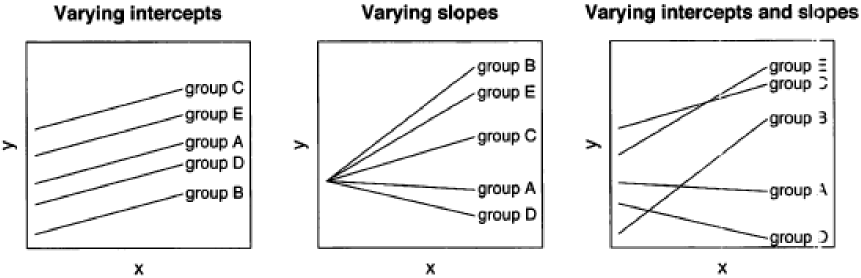
Roberta De Vito

BROWN
Public Health

Multi-level Structure

- grouped data
- repeated measurements
- time-series

Varying-intercept and varying-slope models (Q1)



Data on child support (Q2)

ID	dad age	mom race	informal support	city ID	city name	enforce intensity	benefit level	city indicators 1	2	⋯	20
1	19	hisp	1	1	Oakland	0.52	1.01	1	0	⋯	0
2	27	black	0	1	Oakland	0.52	1.01	1	0	⋯	0
3	26	black	1	1	Oakland	0.52	1.01	1	0	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
248	19	white	1	3	Baltimore	0.05	1.10	0	0	⋯	0
249	26	black	1	3	Baltimore	0.05	1.10	0	0	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
1366	21	black	1	20	Norfolk	−0.11	1.08	0	0	⋯	1
1367	28	hisp	0	20	Norfolk	−0.11	1.08	0	0	⋯	1

Data on child support

ID	dad age	mom race	informal support	city ID
1	19	hisp	1	1
2	27	black	0	1
3	26	black	1	1
⋮	⋮	⋮	⋮	⋮
248	19	white	1	3
249	26	black	1	3
⋮	⋮	⋮	⋮	⋮
1366	21	black	1	20
1367	28	hisp	0	20

city ID	city name	enforce-ment	benefit level
1	Oakland	0.52	1.01
2	Austin	0.00	0.75
3	Baltimore	−0.05	1.10
⋮	⋮	⋮	⋮
20	Norfolk	−0.11	1.08

Ways of analyzing these data

Individual-Level Regression

- Informal support: binary outcome
- Several individual- and city-level predictors
- Enforcement is the treatment
- The model (Q3):

Ways of analyzing these data

Group-Level Regression

city ID	city name	enforce-ment	benefit level	# in sample	avg. age	prop. black	proportion with informal support
1	Oakland	0.52	1.01	78	25.9	0.67	0.55
2	Austin	0.00	0.75	91	25.8	0.42	0.54
3	Baltimore	−0.05	1.10	101	27.0	0.86	0.67
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	Norfolk	−0.11	1.08	31	27.4	0.84	0.65

Ways of analyzing these data

Individual-level regression with city indicators, followed by group-level regression

- logistic regression: 22 predictors
- linear regression

Multilevel models

- the model

- the city coefficient

$$\alpha_j \sim N(U_j \gamma, \sigma^2_{alpha}), \;\; j = 1, \ldots, 20$$

Repeated Measurements

- 2000 Australian adolescents

person		parents smoke?		wave 1		wave 2		⋯
ID	sex	mom	dad	age	smokes?	age	smokes?	
1	f	Y	Y	15:0	N	15:6	N	⋯
2	f	N	N	14:7	N	15:1	N	⋯
3	m	Y	N	15:1	N	15:7	Y	⋯
4	f	N	N	15:3	N	15:9	N	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Repeated Measurements

```
y <- data[,seq(6,16,2)]
female <- ifelse (data[,2]=="f", 1, 0)
mom.smoke <- ifelse (data[,3]=="Y", 1, 0)
dad.smoke <- ifelse (data[,4]=="Y", 1, 0)
psmoke <- mom.smoke + dad.smoke
```

Repeated Measurements

Q4

```
y <- data[,seq(6,16,2)]
female <- ifelse (data[,2]=="f", 1, 0)
mom.smoke <- ifelse (data[,3]=="Y", 1, 0)
dad.smoke <- ifelse (data[,4]=="Y", 1, 0)
psmoke <- mom.smoke + dad.smoke
```

$$\begin{aligned}\Pr(y_{jt}=1) &= \operatorname{logit}^{-1}(\beta_0+\beta_1\text{psmoke}_j+\beta_2\text{female}_j+ \\ &\qquad +\beta_3(1-\text{female}_j)\cdot t+\beta_4\text{female}_j\cdot t+\alpha_j),\end{aligned}$$

Repeated Measurements

Different Table

age	smokes?	person ID	wave
15:0	N	1	1
14.7	N	2	1
15:1	N	3	1
15:3	N	4	1
⋮	⋮	⋮	⋮
15:6	N	1	2
15:1	N	2	2
15:7	Y	3	2
15:9	N	4	2
⋮	⋮	⋮	⋮

person		parents smoke?	
ID	sex	mom	dad
1	f	Y	Y
2	f	N	N
3	m	Y	N
4	f	N	N
⋮	⋮	⋮	⋮

Repeated Measurements

Different Table

```
y <- obs.data[,2]
person <- obs.data[,3]
wave <- obs.data[,4]
female <- ifelse (person.data[,2]=="f", 1, 0)
mom.smoke <- ifelse (person.data[,3]=="Y", 1, 0)
dad.smoke <- ifelse (person.data[,4]=="Y", 1, 0)
psmoke <- mom.smoke + dad.smoke
```

$$\begin{aligned}\Pr(y_i = 1) \;=\; & \operatorname{logit}^{-1}(\beta_0 + \beta_1 \mathrm{psmoke}_{j[i]} + \beta_2 \mathrm{female}_{j[i]} + \\ & + \beta_3 (1 - \mathrm{female}_{j[i]}) \cdot t[i] + \beta_4 \mathrm{female}_{j[i]} \cdot t[i] + \alpha_{j[i]}).\end{aligned}$$

Time-series cross-sectional data

- Repeated measurements could easily have irregular patterns
- Time-series cross-sectional data commonly have overall time patterns
- One must consider the state-year data as clustered within states and also within years

Motivation

- Accounting for individual- and group-level variation
- Modeling variation among individual-level regression
- Estimating regression coefficients for particular groups
- Complexity
- When does multilevel modeling make a difference?

Notation

- The units
- Outcome
- Regression Predictor

Notation

- The units
- Outcome
- Regression Predictor

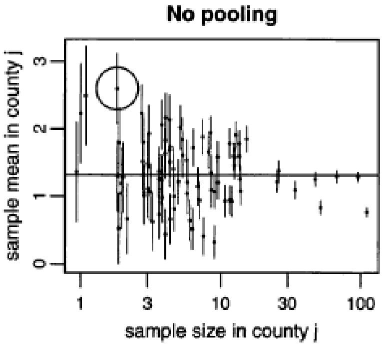
- Groups
- Index variables $j[i]$ (Q5)
- Varying-intercept and Varying-slope

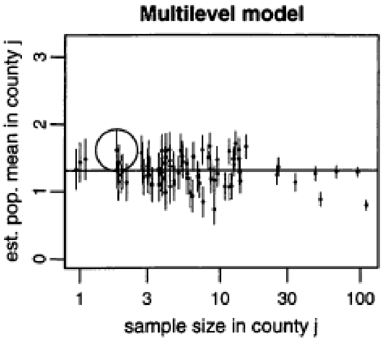
Radon example

- Estimate the distribution of radon
- 85 counties in Minnesota

Radon example

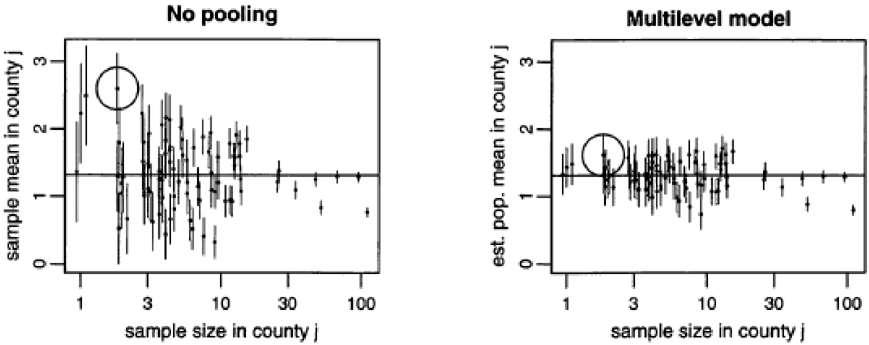
- Estimate the distribution of radon
- 85 counties in Minnesota





Radon example

$$\widehat{\alpha}_j^{multilevel} \approx \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\bar{y}_{all}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \quad \text{Q6}$$



Complete-pooling and no-pooling analyses

- Log home radon
- Floor of measurement
- Measurements were taken in the lowest living area of each house (with basement as 0, and first floor as 1)

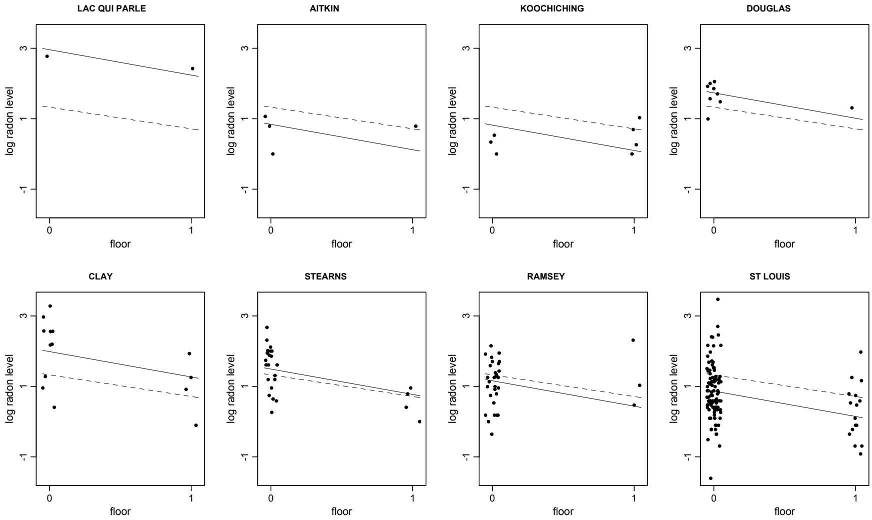
Complete-pooling and no-pooling analyses

```
lm(formula = y ~ x)
            coef.est coef.se
(Intercept)  1.33     0.03
x           -0.61     0.07
  n = 919, k = 2
  residual sd = 0.82
```

Complete-pooling and no-pooling analyses

```
lm(formula = y ~ x + factor(county) - 1)
                    coef.est  coef.sd
x                   -0.72      0.07
factor(county)1      0.84      0.38
factor(county)2      0.87      0.10
. . .
factor(county)85     1.19      0.53
  n = 919, k = 86
  residual sd = 0.76
```

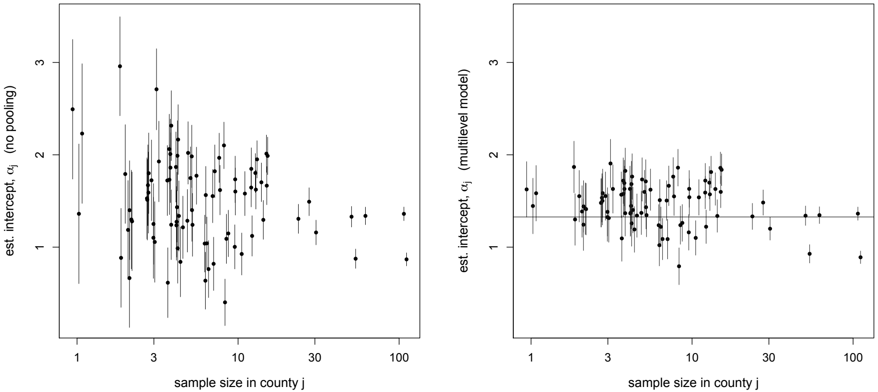
Complete-pooling (dashed line) and no-pooling (solid line)



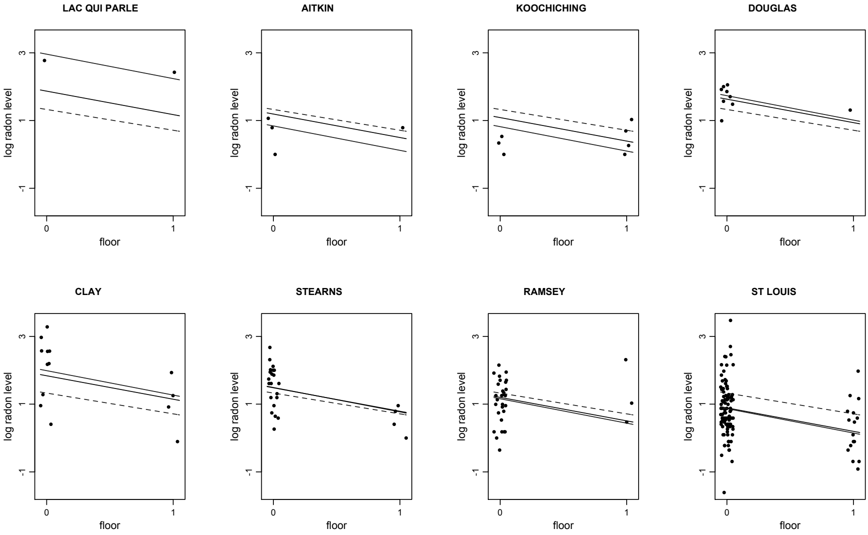
The simplest multilevel model

- $y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$
- $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$

The simplest multilevel model



The simplest multilevel model



Interpret the coefficients

- $\widehat{\mu}_\alpha = 1.46$, $\widehat{\beta} = -0.69$, $\widehat{\sigma}_y = 0.76$, $\widehat{\sigma}_\alpha = 0.33$
- average regression for each counties

- The variance ratio $\widehat{\sigma}_\alpha^2 / \widehat{\sigma}_y^2$

- The intraclass correlation

$$\frac{\widehat{\sigma}_\alpha^2}{\widehat{\sigma}_y^2 + \widehat{\sigma}_\alpha^2}$$

Allowing regression coefficients to vary across groups

- Allow β to vary across groups

$$y_i = \beta_{0j[i]} + \beta_{1j[i]} X_{i1} + \beta_{2j[i]} X_{i2} + \cdots + \epsilon_i$$

- Varying-intercept

$$y_i = \alpha_{j[i]} + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

- $\alpha_j \sim N(\mu_{alpha}, \sigma^2_\alpha)$ or also

$$\alpha_j = \mu_\alpha + \eta_j \quad \eta_j \sim N(0, \sigma^2_\alpha)$$

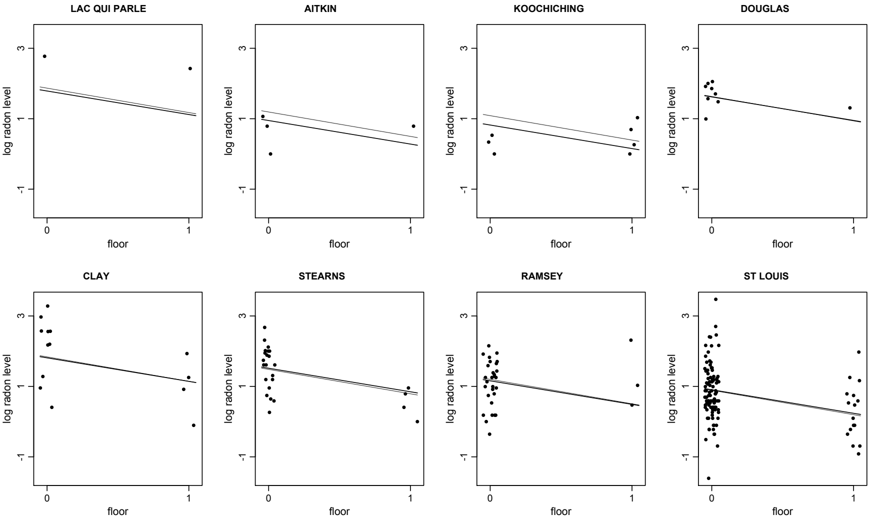
Adding a group-level predictor

$$y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$$

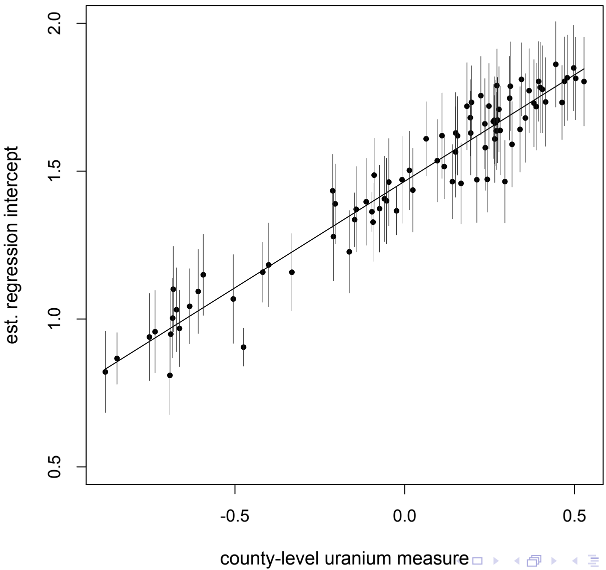
and

$$\alpha_j \sim N(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2)$$

Adding a group-level predictor



Adding a group-level predictor



Adding a group-level predictor

```
lmer(formula = y ~ x + u.full + (1 | county))
            coef.est coef.se
(Intercept)  1.47     0.04
x           -0.67     0.07
u.full       0.72     0.09
Error terms:
 Groups   Name        Std.Dev.
 county   (Intercept) 0.16
 Residual             0.76
# of obs: 919, groups: county, 85
deviance = 2122.9
```

```
$county
    (Intercept)     x u.full
1          1.45 -0.67   0.72
2          1.48 -0.67   0.72
. . .
85         1.42 -0.67   0.72
```