Multilevel/Hierarchical Models

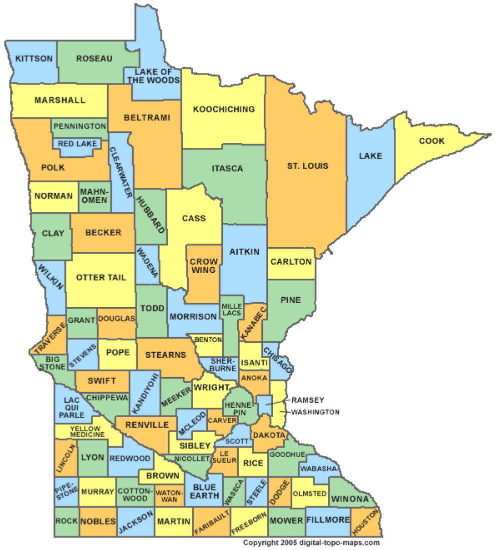
STA303/STA1002: Methods of Data Analysis II, Summer 2016

Michael Guerzhoy

Case Study: Radon Levels in Minnesota

- Radon is a radioactive gas that is known to cause lung cancer, and is responsible for several thousand of lung cancer deaths per year in the US
- Radon levels vary in different homes, and also vary in different counties

Minnesota



Minnesota counties

Goal

- Based on a limited set of measurements, want to know the log(radon levels) in the different counties

Complete Pooling

- Combine all the information from all the counties into a single “pool” of data
- Problem with complete pooling: the levels might differ for the different counties

No-Pooling Estimate

- Compute the average radon level for measurements in each county
- Compare pairs of counties using t-tests
- Equivalent to

```
lm(log_radon~county, data=mn)
```

and looking at the coefficients for each county

No-Pooling Estimate: Problem

- Let’s look at Lac Qui Parle County
    - (in R)
- We have just two data points for Lac Qui Parle, so we shouldn’t necessarily trust the data from there as much
- If we want to get at an estimate of the average log-radon level in Lac Qui Parle County, we probably want some kind of weighted average between what we observe in Lac Qui Parle and the overall average

Multilevel Model

- Consider how the data is generated
- $y_i \sim N\left(\alpha_{j[i]}, \sigma_y^2\right)$
- y_i is the i-th measurement
- j[i] is the county in which the i-th measurement was taken
- $\alpha_{j[i]}$ is the true log-radon level in county j[i]
- NEW:

$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$

- Estimate the best $\mu_\alpha, \sigma_\alpha^2$ from the data

Multilevel Model

$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$
$$y_i \sim N\left(\alpha_{j[i]}, \sigma_y^2\right)$$

- Fake-data generation in R

Partial Pooling

$$y_i \sim N\left(\alpha_{j[i]}, \sigma_y^2\right)$$
$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$

- Let $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$
- (Approximate) Likelihood used by lme in R:
$$P\left(y_1, y_2, \dots, y_n \middle| \mu_\alpha, \sigma_y^2, \sigma_\alpha^2\right)$$
$$= (\Pi_j f\left(\alpha_j \middle| \mu_\alpha, \sigma_\alpha^2\right)) \left(\Pi_i f\left(y_i \middle| \alpha_{j[i]}, \sigma_y^2\right)\right)$$
- lme finds the α_j, σ_y^2, $\mu_\alpha, \sigma_\alpha^2$ which maximize the likelihood
- Can now look at the different α_j

Partial Pooling with No Predictors

- (in R)

Complete/Partial/No-Pooling

$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$
$$y_i \sim N\left(\alpha_{j[i]}, \sigma_y^2\right)$$

- No-Pooling: $\sigma_\alpha^2 = \infty$. That is, we assume that there is no connection at all between the log-radon levels in the different counties
  - `lm(log.radon~county, data=mn)`
- Complete pooling: $\sigma_\alpha^2 = 0$. Assume the true mean log-radon levels in all counties are the same
  - `lm(log.radon~1, data=mn)`
- Partial pooling: assume the mean log-radon levels are different in different counties, but their SD is σ_α (so they don't differ by that much

R output

Random effects: coefficients that are *modelled* (i.e., generated by a distribution)
Fixed effects: coefficients that are note modelled

Note: the terminology is inconsistent in different places

```
summary(lmer(log.radon~(1|county), data=mn))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log.radon ~ (1 | county)
##    Data: mn
##
## REML criterion at convergence: 2259.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.4661 -0.5734  0.0441  0.6432  3.3516
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  county   (Intercept) 0.09581  0.3095
##  Residual             0.63662  0.7979
## Number of obs: 919, groups:  county, 85
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.31258    0.04891   26.84
```

$\hat{\sigma}_\alpha^2$

$\hat{\sigma}_\alpha$

$\hat{\sigma}_y^2$

$\hat{\mu}_\alpha$

R output

The $\alpha_j's$ for the different counties that are most likely

```
coef(lmer(log.radon~(1|county), data=mn))
```

```
## $county
##                       (Intercept)
## AITKIN                  1.0674994
## ANOKA                   0.8875568
## BECKER                  1.2303812
## BELTRAMI                1.2245444
## BENTON                  1.2899760
## BIG STONE               1.3749235
## BLUE EARTH              1.7171954
## BROWN                   1.4315991
## CARLTON                 1.0833131
## CARVER                  1.2608819
## CASS                    1.3506019
## CHIPPEWA                1.4695309
```

Complete/Partial/No-Pooling

- No-Pooling
  - Doesn’t share information between data points
  - Estimates for different counties will be completely different from each other
- Complete pooling
  - Fully shares information between data points
  - Estimates for the different counties are all the same
- Partial pooling
  - Tries to share information between data points in an optimal way
  - Estimates for different counties are generally closer together than for the no-pooling estimate

Partial pooling with Predictors

- Let's use the floor predictor (x) as well
  - The floor on which the measurement was taken
- Simplest variant:

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

- Advantage: better estimates for the levels for the various counties would lead to better estimates for the β
- Interpretation of β: keeping everything else constant, the increase in radon levels going up one floor
- Better estimate of β is obtained by partially pooling information when estimating $\alpha_{j[i]}$

Rewriting the model so that it makes sense in terms of lmer

- Instead of:

$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$
$$y_i \sim N\left(\alpha_{j[i]}, \sigma_y^2\right)$$

- Write

$$\alpha_{j[i]} \sim N(0, \sigma_\alpha^2)$$
$$y_i \sim N\left(\mu_a + \alpha_{j[i]}, \sigma_y^2\right)$$

$$\alpha_{j[i]} \sim N(0, \sigma_\alpha^2)$$
$$y_i \sim N\left(\mu_a + \alpha_{j[i]}, \sigma_y^2\right)$$

```
summary(lmer(log.radon~(1|county), data=mn))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log.radon ~ (1 | county)
##    Data: mn
##
## REML criterion at convergence: 2259.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.4661 -0.5734  0.0441  0.6432  3.3516
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  county   (Intercept) 0.09581  0.3095
##  Residual             0.63662  0.7979
## Number of obs: 919, groups:  county, 85
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.31258    0.04891   26.84
```

$\hat{\sigma}_\alpha^2$

$\hat{\sigma}_y$

$\hat{\mu}_\alpha$

Random Slopes

$$y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2\right)$$
$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right)$$

- Multivariate Normal Distribution (not going into details): keeping β constant, α is normally distributed and vice versa. α and β are correlated. E.g., if $\rho > 0$, larger α means β will probably be large too
  - Not going into details here
- Interpretation: in each county, the effect of moving one floor up on the radon levels is different
  - Perhaps different in one county, the ceilings are 2.5m high, and in another county, the ceilings are 2.2m high
    - What is the effect of that on the βs?
- Rewrite:

$$y_i \sim N\left((\mu_\alpha + \alpha_{j[i]}) + (\mu_\beta + \beta_{j[i]}) x_i, \sigma_y^2\right)$$
$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right)$$

R Output

```
lmer(log.radon~floor+(floor|county) , data=mn)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log.radon ~ floor + (floor | county)
##    Data: mn
## REML criterion at convergence: 2168.325
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  county   (Intercept) 0.3487
##           floor       0.3436   -0.34
##  Residual             0.7462
## Number of obs: 919, groups:  county, 85
## Fixed Effects:
## (Intercept)        floor
##      1.4628      -0.6811
```

$\hat{\rho}$ $\hat{\sigma}_\alpha$

$\hat{\sigma}_\beta$

$\hat{\sigma}_y$

$\hat{\mu}_\alpha$ $\hat{\mu}_\beta$

Prediction for a new observation in an existing group

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

- Know α, β, and x, want to predict new y
- Simulate multiple y's from the distribution
- (in R)

Prediction for a new observation in a new group

- For each simulation,
- First, generate

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right)$$

- Next, generate the new data

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