# ENV 710: Lecture 15

multilevel models

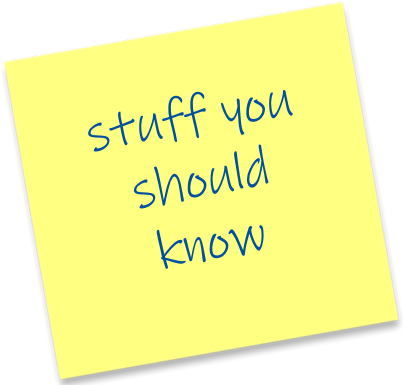
Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

linear models

mixed models

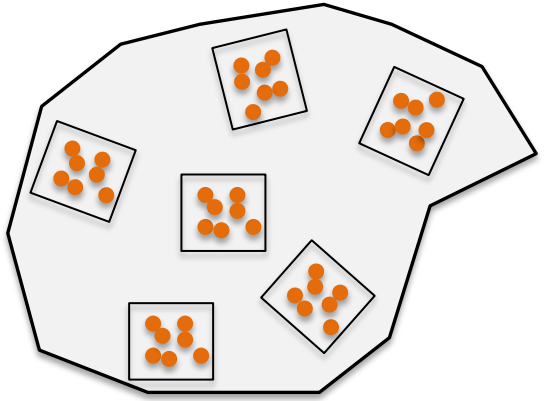
# learning goals

- what are fixed and random effects?
- when and how are multilevel models conducted?
  - pooled, unpooled, and partially pooled models
  - models with varying slopes and intercepts
- practical application of multilevel models
  - interpreting coefficients
  - model diagnostics and validation
  - comparing models



# a problem

- environmental experiments/studies often collect more than one data point at a site/location or of a subject under different conditions
- most models (lm, glm) assume that the data points are independent and identically distributed (*iid*)
- repeated measures or block designs violate this assumption because *observations coming from the same location or site are usually correlated*



- monthly measures of water quality at testing stations along a river
- repeated measures of generator efficiency under different conditions across power plants
- response of patients to a drug under the treatment of different doctors or in different health facilities

# linear mixed effects models (lmm)

- compared to violations of other assumptions (e.g., normality and homogeneity of variances) linear models are not robust to violations of *iid*
  - lead to increased Type I error (false positives)
  - produce overconfident results (narrow standard errors, CI's)

- the solution is to use lmm's or mixed models (also called multi-level models) to explicitly capture the dependency among data points using random effects

# fixed and random effects

- fixed effects: the effects of the variables are interesting in themselves

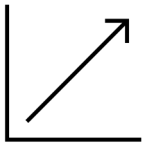
- random effects: interest is in the underlying population
  - "group" effect is random if we can think of the levels in that group to be samples from a larger population
  - in random effects model, the observations are not independent

# fixed and random effects

am I interested in the variable (vs. just want to account for the correlation)?

- **yes** = fixed effect
- **no** = random effect

e.g. block effect



do I want an effect size for every level of the variable (vs. just want a single effect size)?

- **yes** = fixed effect
- **no** = random effect

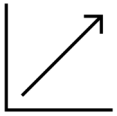
e.g. 195 countries

do I want to make inference to an entire population (vs. interested in specific subjects)?

- **no** = fixed effect
- **yes** = random effect

e.g. random selection of tree species to study phenology

# other advantages of random effects



*accounting for hidden structure in the data*: (e.g. block effect in a randomized block ANOVA), random effect accommodates the correlations that exist and prevents pseudoreplication

*increased scope of inference and assessment of variability*: treating the effects of population as random allows us to make an inference not only about the sampled populations, but about an entire "population of populations"

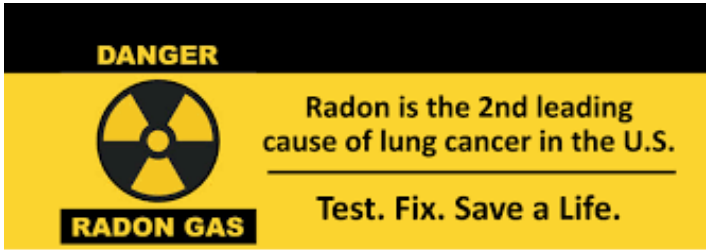
*partitioning of variability*: can estimate the variability among populations and explain the differences among populations by measured covariates

*model spatial, temporal and spatiotemporal correlation*: spatial or temporal autocorrelation can be accommodated

*partial pooling, borrowing strength, and "shrinkage"*: treating population effects as random can be seen as a compromise between assuming all populations are equal (complete pooling) and assuming that they are totally unrelated (no pooling); the random effects estimates will not be the same as fixed-effects estimates; rather, they will be pulled in towards their overall mean - "shrinkage"

# example

## radon testing




data on radon levels in houses in the state of Minnesota: does floor of house affect radon reading?

- `log.radon`: radon measurement from the house (log scale)
- `floors`: indicator for radon measurement made on the first floor of the house (0 = basement, 1 = first floor)
- `county`: county name (85 counties)
- `log.uranium`: uranium level in the county (log scale)

| | log.radon | floor | county |
|---|---|---|---|
| 1 | 0.7884574 | 1 | 1 |
| 2 | 0.7884574 | 0 | 1 |
| 3 | 1.0647107 | 0 | 1 |
| 4 | 0.0000000 | 0 | 1 |
| 5 | 1.1314021 | 0 | 2 |
| 6 | 0.9162907 | 0 | 2 |

Gelman and Hill 2005

# example

## radon testing

data on radon levels in houses in the state of Minnesota: does floor of house affect radon reading?

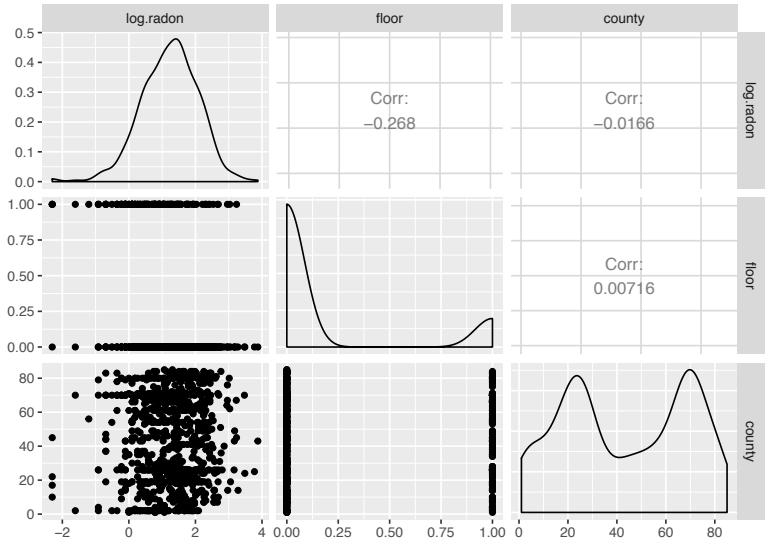
- `log.radon`: radon measurement from the house (log scale)
- `floors`: indicator for radon measurement made on the first floor of the house (0 = basement, 1 = first floor)
- `county`: county name (85 counties)
- `log.uranium`: uranium level in the county (log scale)

### pooled model

```
lm(log.radon ~ floor)
```

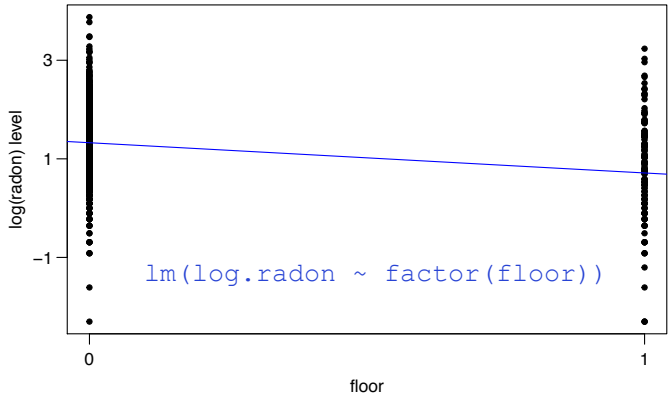
### unpooled model

```
lm(log.radon ~ floor + county)
```

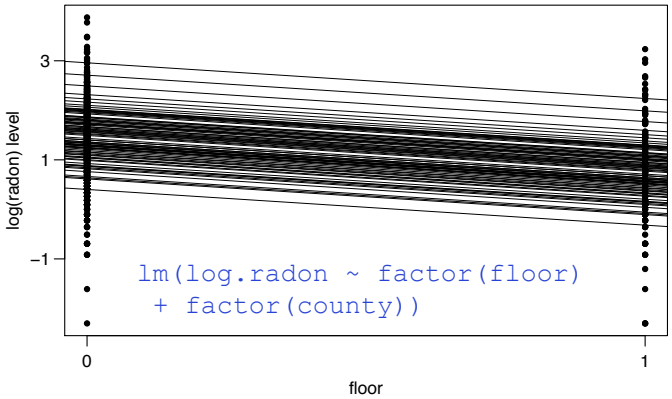



Gelman and Hill 2005

# pooled and unpooled models

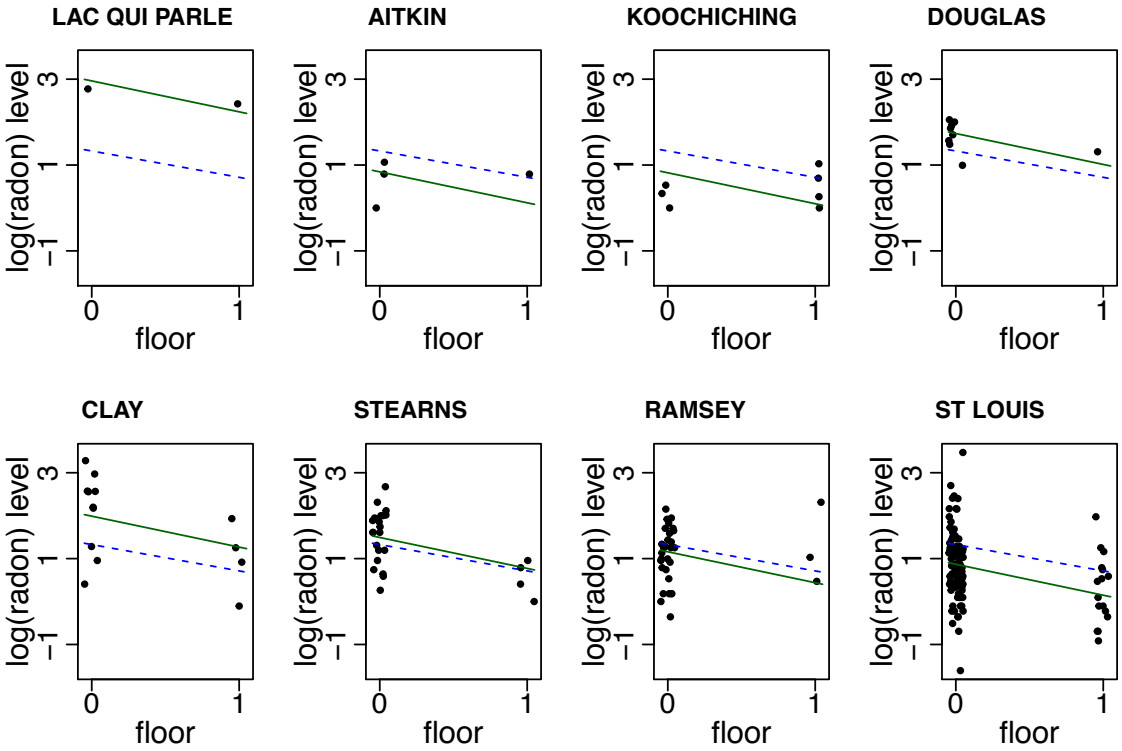



```
Coefficients:
                Estimate Std. Error t value  Pr(>|t|)
(Intercept)      1.32674    0.02972  44.640  <2e-16 ***
factor(floor)1 -0.61339    0.07284  -8.421  <2e-16 ***
---

Residual standard error: 0.8226 on 917 degrees of freedom
Multiple R-squared:  0.07178, Adjusted R-squared:  0.07077
F-statistic: 70.91 on 1 and 917 DF,  p-value: < 2.2e-16
```




```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.84054    0.37866   2.220  0.02670 *
factor(floor)1     -0.72054    0.07352  -9.800  < 2e-16 ***
factor(county)2     0.03428    0.39274   0.087  0.93047
...
factor(county)81    1.86899    0.57854   3.231  0.00128 **
factor(county)82    1.38947    0.84590   1.643  0.10084
factor(county)83    0.78238    0.43250   1.809  0.07082 .
factor(county)84    0.80481    0.43269   1.860  0.06323 .
factor(county)85    0.34598    0.65534   0.528  0.59768 ---

Residual standard error: 0.7564 on 833 degrees of freedom
Multiple R-squared:  0.287, Adjusted R-squared:  0.2142
F-statistic: 3.945 on 85 and 833 DF,  p-value: < 2.2e-16
```

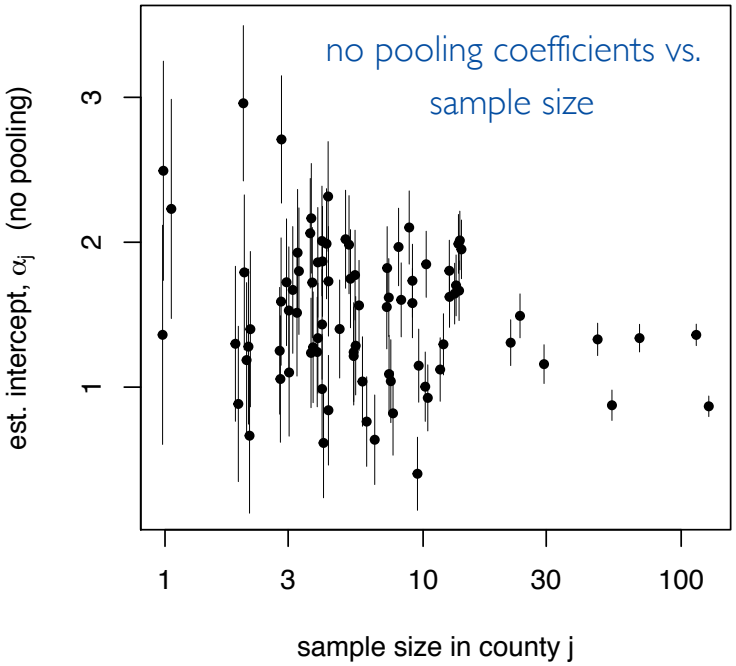
# pooled and unpooled models

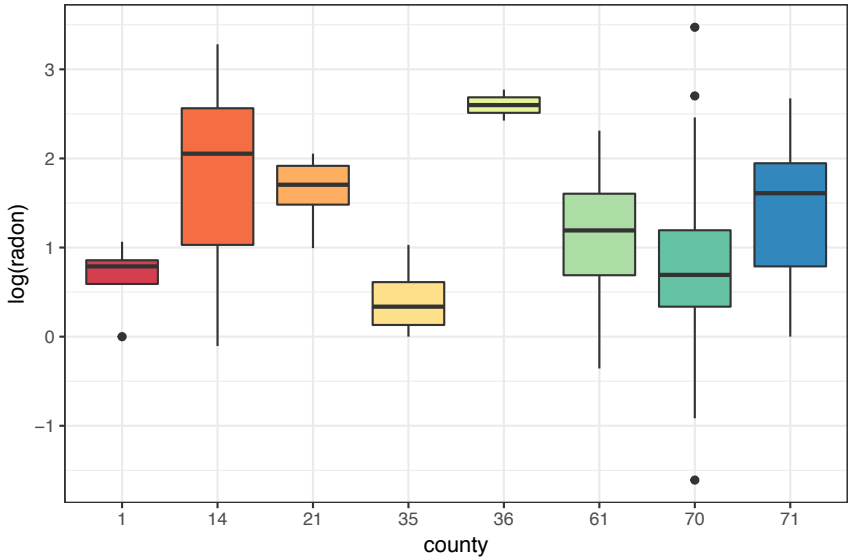



pooled model

unpooled model

# pooled and unpooled models




- there is group-level variation (e.g. counties vary in their radon levels)
- needs to be taken into account
- what's the best way?

# multi-level modeling

- varying intercept model
- estimates the effect of all the counties but just reports the variance of all those effects
- tells you the amount of variation in radon among counties

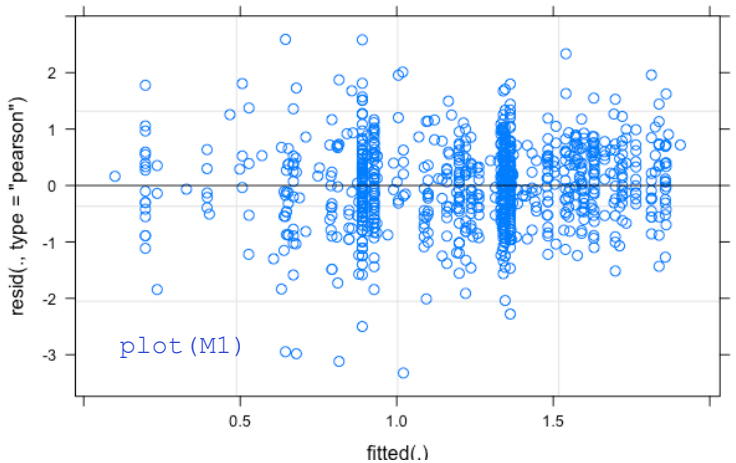
```
lmer(log.radon ~ floor + (1|county))
```

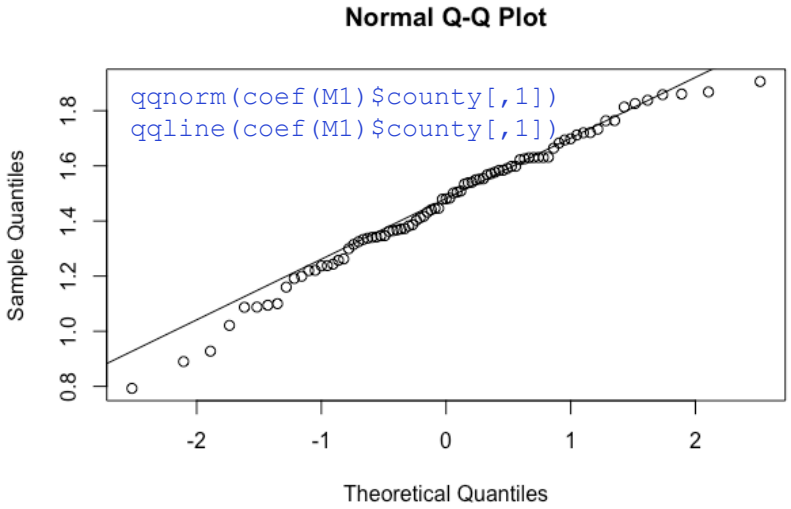
- (1|county) notation: 1 stands for intercept, model expects multiple responses per subject (county) and these responses will depend on each subject's baseline level (intercept)

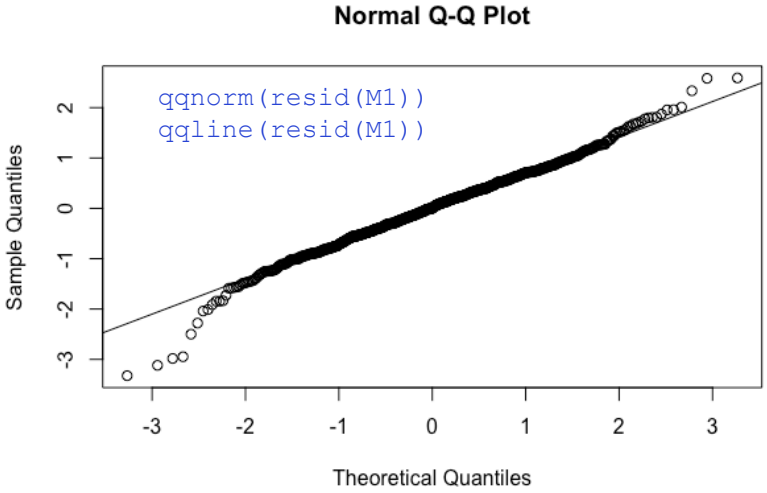
# multi-level modeling

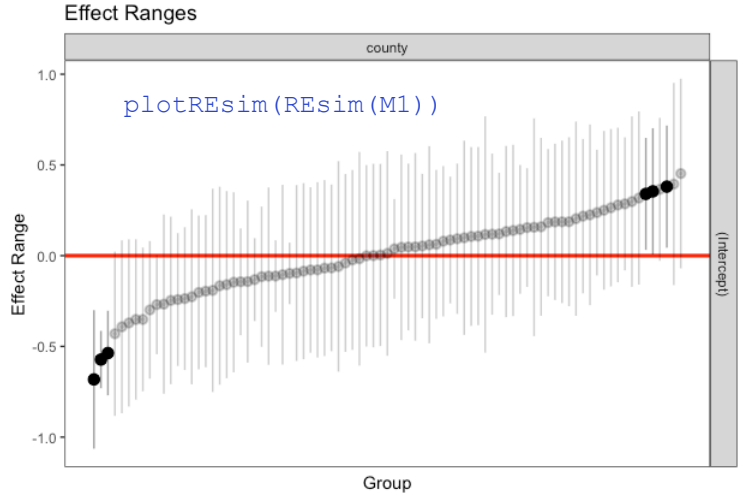
```
summary(lmer(log.radon ~ factor(floor) + (1|county)))
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 county   (Intercept) 0.1077   0.3282
 Residual             0.5709   0.7556
Number of obs: 919, groups:  county, 85

Fixed effects:
                Estimate Std. Error        df t value Pr(>|t|)
(Intercept)      1.46160    0.05158  87.66221  28.339   <2e-16 ***
factor(floor)1  -0.69299    0.07043 914.86987  -9.839   <2e-16 ***
```

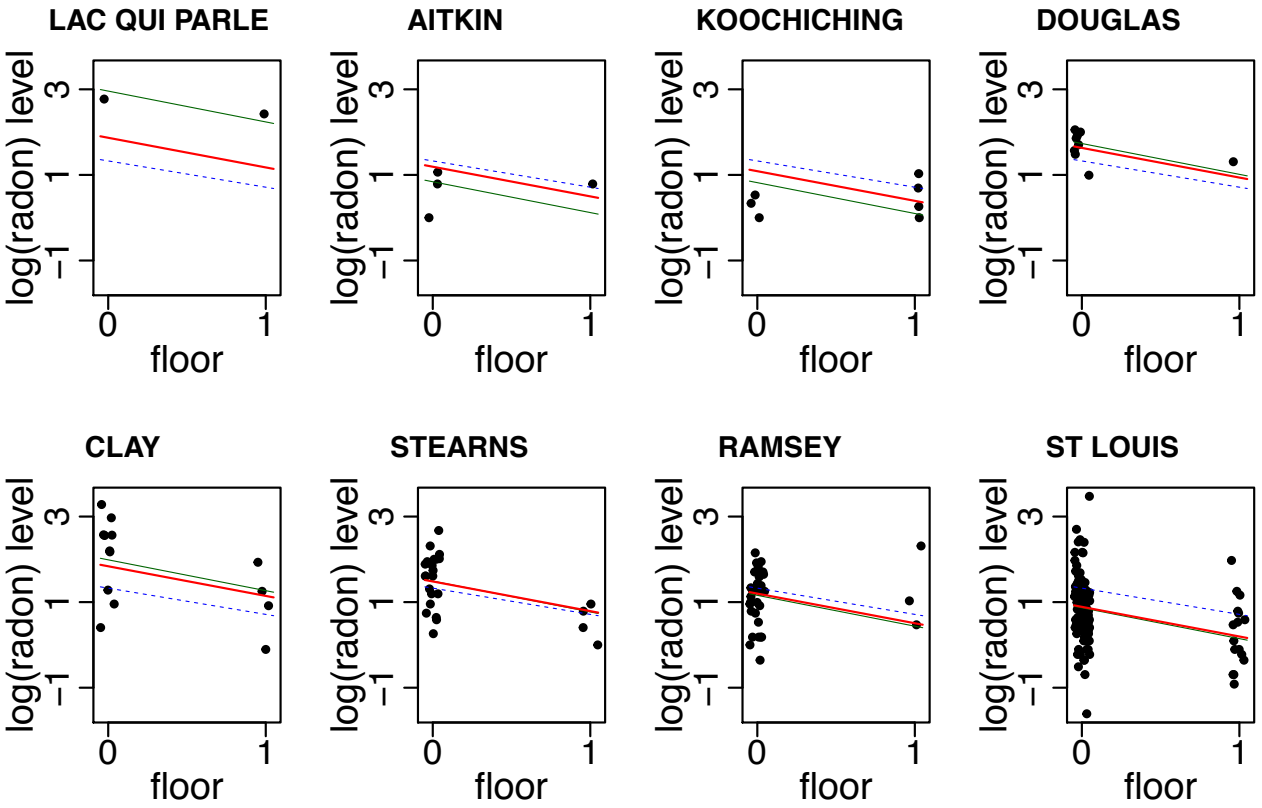

plot(M1)
resid(., type = "pearson")
fitted(.)


Normal Q-Q Plot
qqnorm(coef(M1)$county[,1])
qqline(coef(M1)$county[,1])
Sample Quantiles
Theoretical Quantiles


Normal Q-Q Plot
qqnorm(resid(M1))
qqline(resid(M1))
Sample Quantiles
Theoretical Quantiles


Effect Ranges
county
plotREsim(REsim(M1))
Effect Range
Group
(Intercept)

# multi-level modeling




pooled model
unpooled model
partial pooling:
varying intercept model

# multi-level modeling

- varying intercept, varying slope model

```
lmer(log.radon ~ floor + (1 + floor|county))
```

- (1+floor|county) tells the model to expect differing baseline-levels (the intercept, represented by 1) as well as differing responses (slopes) to the main factor in question - floor in this problem

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 county   (Intercept) 0.1216   0.3487
          floor       0.1180   0.3436   -0.34
 Residual             0.5567   0.7462
Number of obs: 919, groups:  county, 85

Fixed effects:
               Estimate Std. Error       df t value Pr(>|t|)
(Intercept)     1.46277    0.05387 71.66148  27.155  < 2e-16 ***
factor(floor)1 -0.68110    0.08758 39.44284  -7.777 1.75e-09 ***
```

# multi-level modeling

- varying intercept, varying slope model

```
lmer(log.radon ~ floor + (1 + floor|county))
```

- variance among houses is 0.5567
- random effects tell us if there is variation in a fixed effect for the different levels of the random effects term
- compare magnitude of random effects and fixed effects by comparing the standard deviation of the random effects to the estimates of the fixed effects

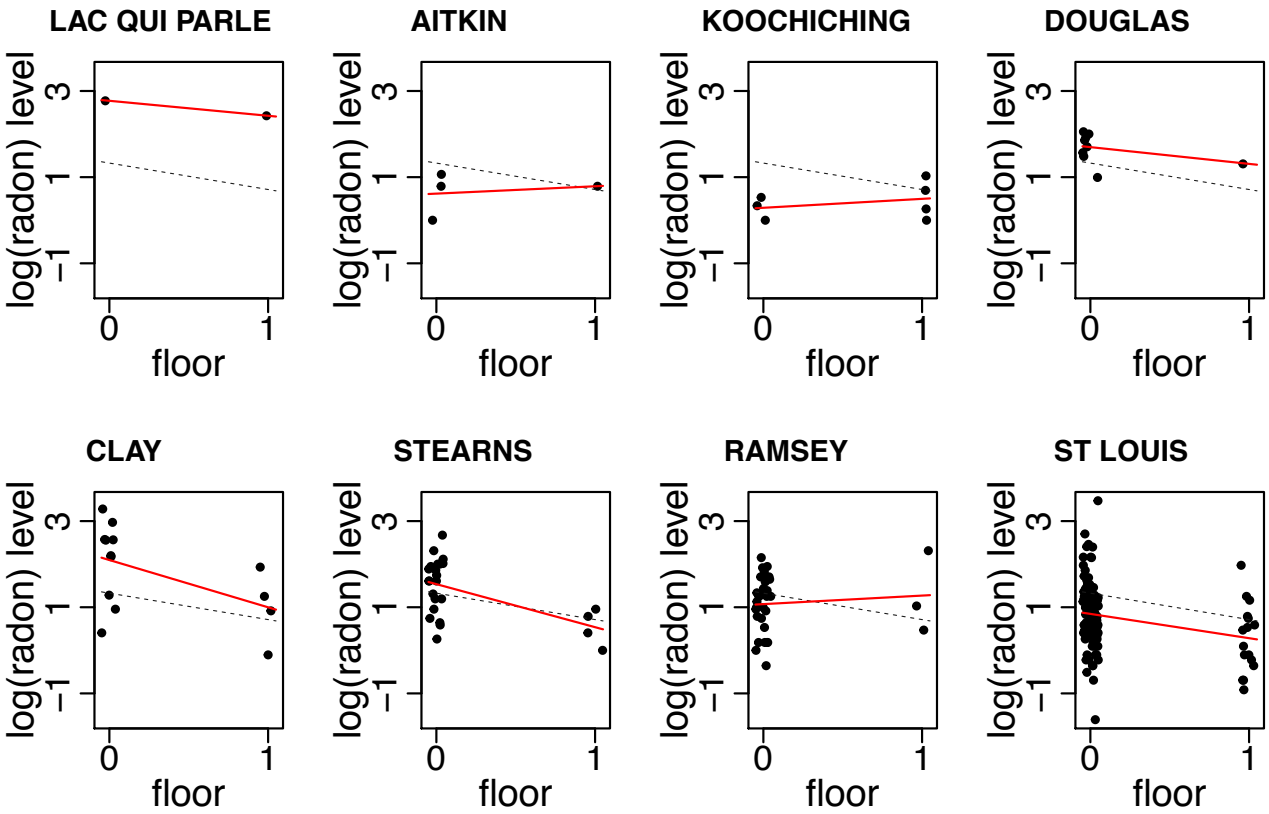
```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 county   (Intercept) 0.1216   0.3487
          floor       0.1180   0.3436   -0.34
 Residual             0.5567   0.7462
Number of obs: 919, groups:  county, 85

Fixed effects:
               Estimate Std. Error       df t value Pr(>|t|)
(Intercept)     1.46277    0.05387 71.66148  27.155  < 2e-16 ***
factor(floor)1 -0.68110    0.08758 39.44284  -7.777 1.75e-09 ***
```

Intraclass correlation

```
(0.1216+.1181)/
(0.1216+.1181+.5567) = 30.1%.
```

# multi-level modeling

pooled model
varying intercept & slope model



# 'significant' effect of random effect?

- use the Likelihood Ratio Test (more coming soon)

```
M1 <- lmer (log.radon ~ floor + (1 | county))
M2 <- lmer(log.radon ~ floor + (1 + floor|county))

require(lmtest)
lrtest(M1, M2)
```

```
Likelihood ratio test

Model 1: log.radon ~ floor + (1 | county)
Model 2: log.radon ~ floor + (1 + floor | county)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   4 -1085.7
2   6 -1084.2  2 2.9807     0.2253
```

```
AIC(lm.pooled, M1, M2)

           df      AIC
lm.pooled  3 2253.025
M1         4 2179.305
M2         6 2180.325
```

# Questions?