

# Statistical Methods for Insurance: Multilevel Models

Di Cook & Souhaib Ben Taieb, Econometrics and Business Statistics, Monash University  
W6.C2

# Overview of this class

- Fixed effects vs random effects
- Mixed effects models
- Diagnostics

# What is a multilevel model?

- Observations are not independent, but belong to a hierarchy
- Example: individual level demographics (age, gender), and school level information (location, course offerings, classroom resources)
- Multilevel model enables fitting different types of dependencies

# Fixed vs random

- **Fixed effects** can be used when you know all the categories, e.g. age, gender, smoking status
- **Random effects** are used when not all groups are captured, and we have a random selection of the groups, e.g. individuals (if you have multiple measurements), schools, hospitals

# Mixed effects models - a type of multilevel model

For data organized in  $g$  groups, consider a continuous response linear mixed-effects model (LME model) for each group  $i, i = 1, \dots, g$ :

$$\underset{(n_i \times 1)}{\mathbf{y}_i} = \underset{(n_i \times p)(p \times 1)}{\mathbf{X}_i \boldsymbol{\beta}} + \underset{(n_i \times q)(q \times 1)}{\mathbf{Z}_i \mathbf{b}_i} + \underset{(n_i \times 1)}{\boldsymbol{\varepsilon}_i}$$

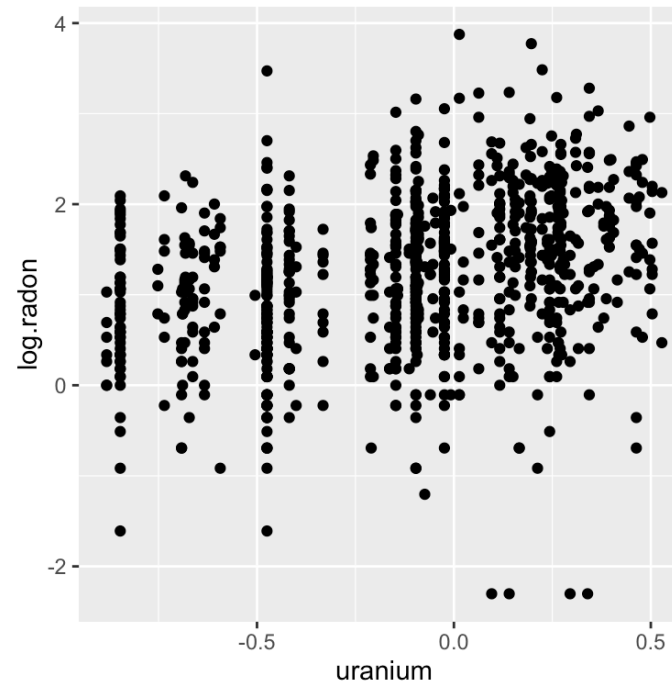
- $\mathbf{y}_i$  is the vector of outcomes for the  $n_i$  level-1 units in group  $i$
- $\mathbf{X}_i$  and  $\mathbf{Z}_i$  are design matrices for the fixed and random effects
- $\boldsymbol{\beta}$  is a vector of  $p$  fixed effects governing the global mean structure
- $\mathbf{b}_i$  is a vector of  $q$  random effects for between-group covariance
- $\boldsymbol{\varepsilon}_i$  is a vector of level-1 error terms for within-group covariance

# Example

- Data: *radon*, 919 owner-occupied homes in 85 counties of Minnesota. Available in the **HLMdiag** package
- Response: *log.radon*
- Fixed: *storey* (categorical), *uranium* (quantitative)
- Random: *county* (house is a member of county)

```
## Observations: 919
## Variables: 5
## $ log.radon    <dbl> 0.7885, 0.7885, 1.0647, 0.0000, 1.1314, 0.9163, 0....
## $ storey      <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ uranium     <dbl> -0.689, -0.689, -0.689, -0.689, -0.847, -0.847, -0...
## $ county      <int> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,...
## $ county.name <fctr> AITKIN, AITKIN, AITKIN, AITKIN, ANOKA, ANOKA, ANO...
```

# Take a look



Plot of response vs covariate. What do you see?

# Here's what we see

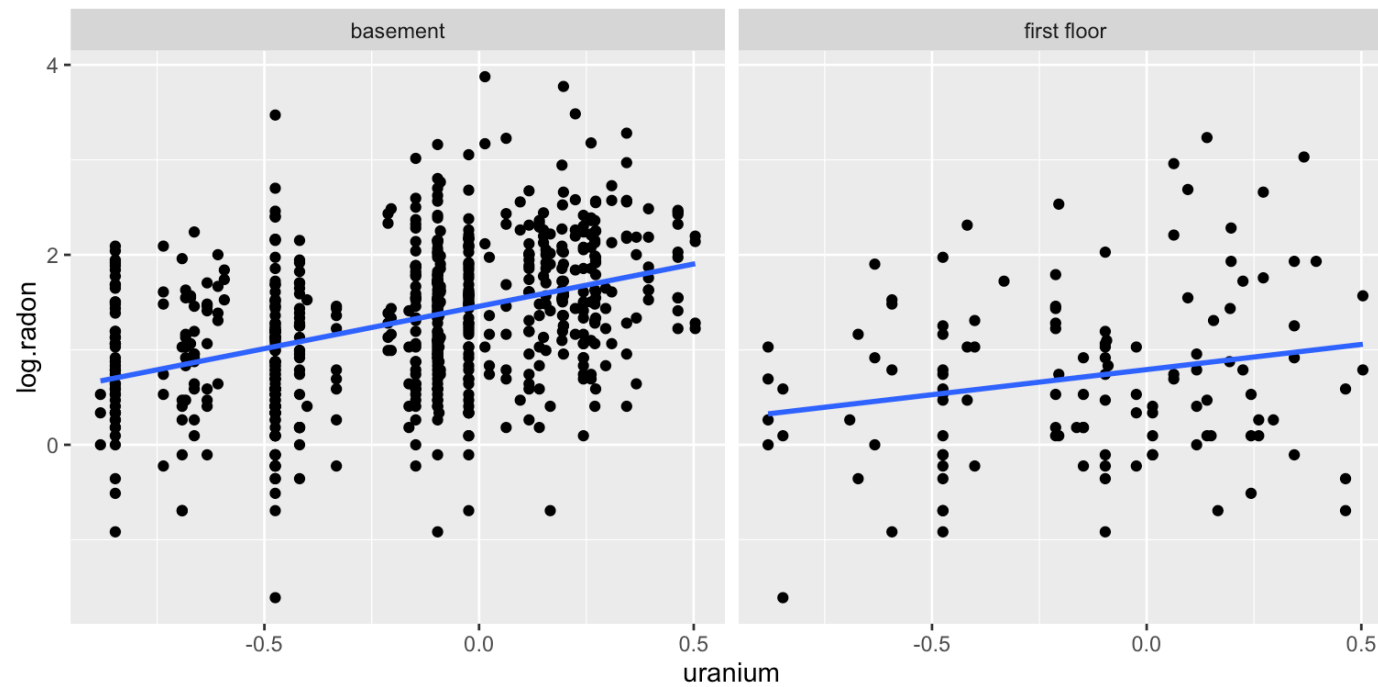
- Vertical stripes: each county is represented by an average uranium value
- Weak linear association, lots of variation for houses within county
- Four points inline horizontally at the base (be suspicious)
- Some counties only have 2, 3 points
- Scales?



# Pre-processing

- Counties with less than 4 observations removed
- Four flat-line observations should be removed, really suspect these were erroneously coded missing values

# Look again



# Fit a simple model

$$\log.\text{radon} = \beta_0 + \beta_1 \text{storey} + \beta_2 \text{uranium} + \varepsilon$$

```
##
## Call:
## glm(formula = log.radon ~ storey + uranium, data = radon_sub)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.6610  -0.4928   0.0191   0.4745   2.4205
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.4483     0.0313   46.25 < 2e-16 ***
## storeyfirst floor -0.6112     0.0733  -8.34 3.3e-16 ***
## uranium           0.8359     0.0742  11.26 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.547)
##
```

11/39

# Your turn

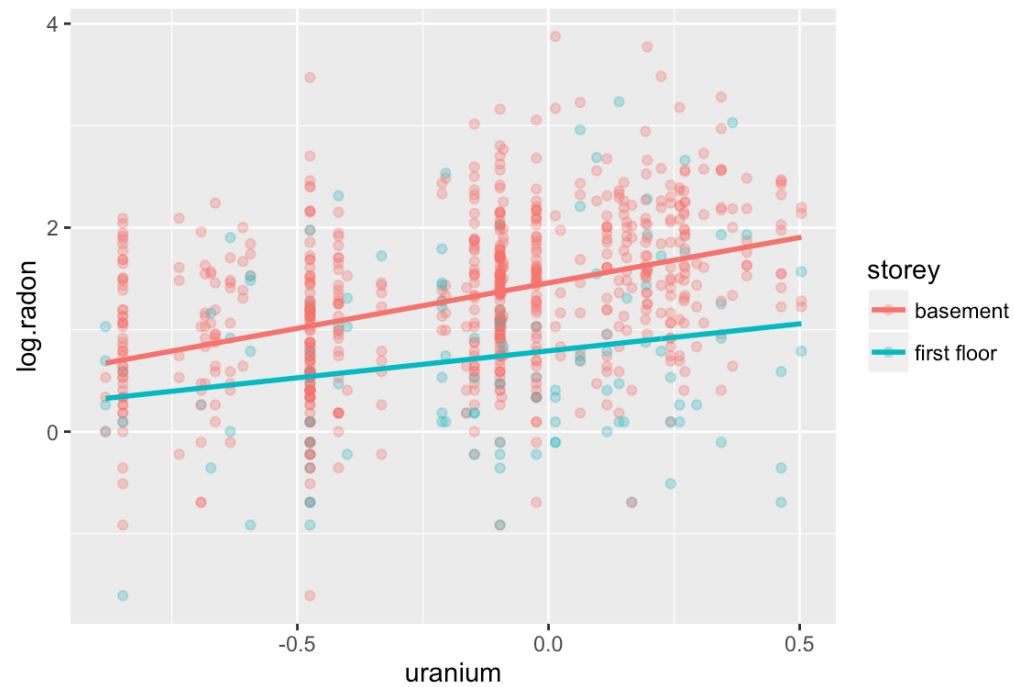
Make a sketch of what this model looks like.

# Fit an interaction term

```
##
## Call:
## glm(formula = log.radon ~ storey * uranium, data = radon_sub)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.6445  -0.4898   0.0131   0.4653   2.4369
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.4580     0.0318   45.91 < 2e-16 ***
## storeyfirst floor    -0.6659     0.0796   -8.37 2.7e-16 ***
## uranium              0.8909     0.0805   11.07 < 2e-16 ***
## storeyfirst floor:uranium -0.3620     0.2066   -1.75  0.08  .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.546)
##
##      Null deviance: 538.51  on 795  degrees of freedom
```

13/39

# What does this model look like?



# Your turn

Write down the equation of the fitted model

# Mixed effects model

$$\log.\text{radon}_{ij} = \beta_0 + \beta_1 \text{storey}_{ij} + \beta_2 \text{uranium}_i + b_{0i} + b_{1i} \text{storey}_{ij} + \varepsilon_{ij}$$

$$i = 1, \dots, \#counties; j = 1, \dots, n_i$$

```
library(lme4)
radon_lmer <- lmer(log.radon ~ storey + uranium +
  (storey | county.name), data = radon_sub)
summary(radon_lmer)
radon_lmer_fit <- augment(radon_lmer)
```



# Your turn

For the radon data:

- What is  $p, q, g$ ?
- And hence  $n_i, i = 1, \dots, g$ ?

$$\log. radon_{ij} = \beta_0 + \beta_1 storey_{ij} + \beta_2 uranium_i + b_{0i} + b_{1i} storey_{ij} + \varepsilon_{ij}$$

$$i = 1, \dots, \#counties; j = 1, \dots, n_i$$

# Examining the model output: fixed effects

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	1.48066	0.03856	38.40
storeyfirst floor	-0.59011	0.11246	-5.25
uranium	0.84600	0.09532	8.88

How do these compare with the simple linear model estimates?

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.44830	0.03131	46.254	< 2e-16 ***
storeyfirst floor	-0.61125	0.07332	-8.337	3.35e-16 ***
uranium	0.83591	0.07422	11.262	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Examining the model output: random effects

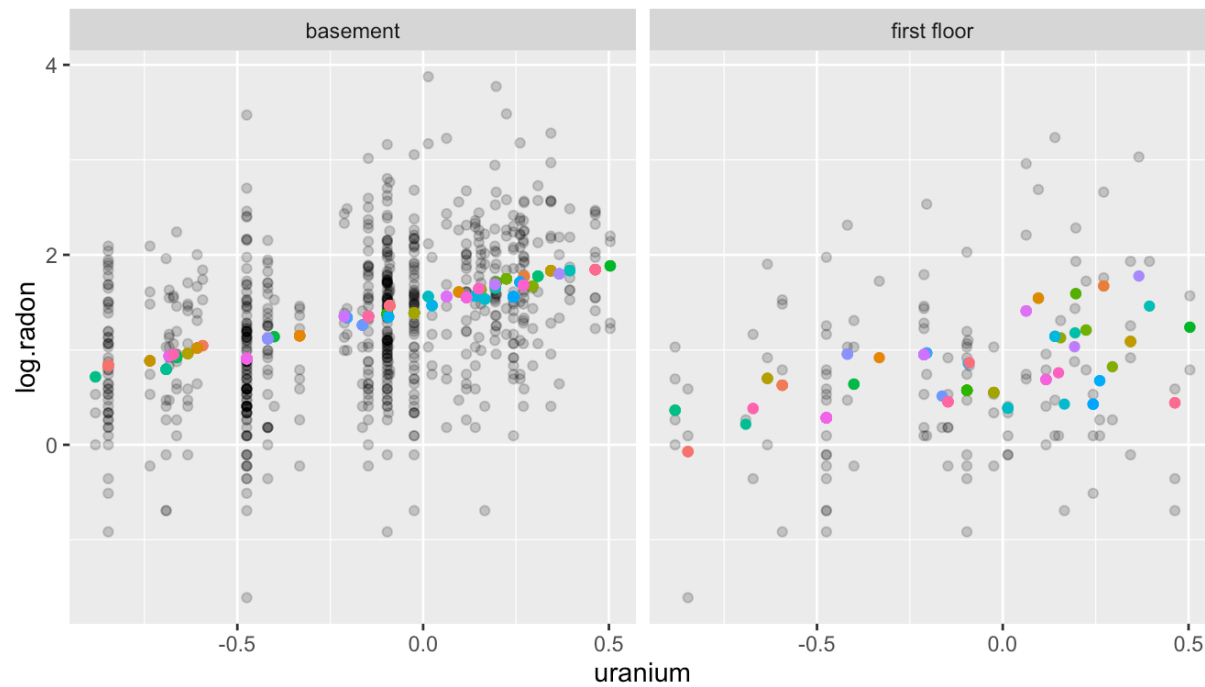
Random effects:

Groups	Name	Variance	Std.Dev.	Corr
county.name	(Intercept)	0.01388	0.1178	
	storeyfirst floor	0.22941	0.4790	0.02
Residual		0.50694	0.7120	

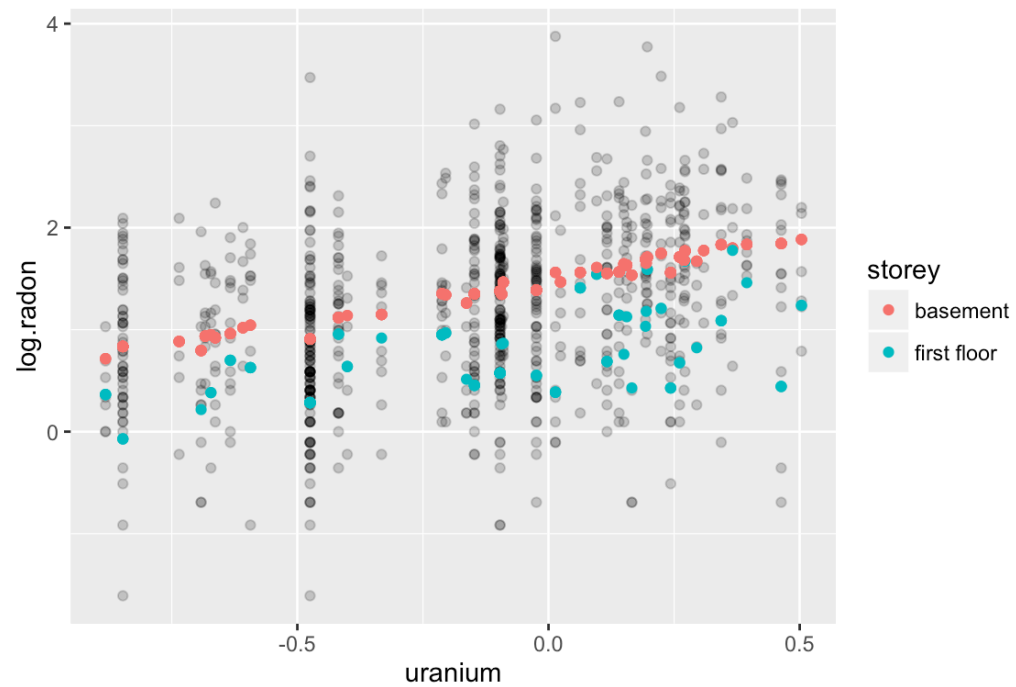
Number of obs: 796, groups: county.name, 46

This is saying that the variance of the estimates for first floor observations is larger than the storey.

# What it looks like

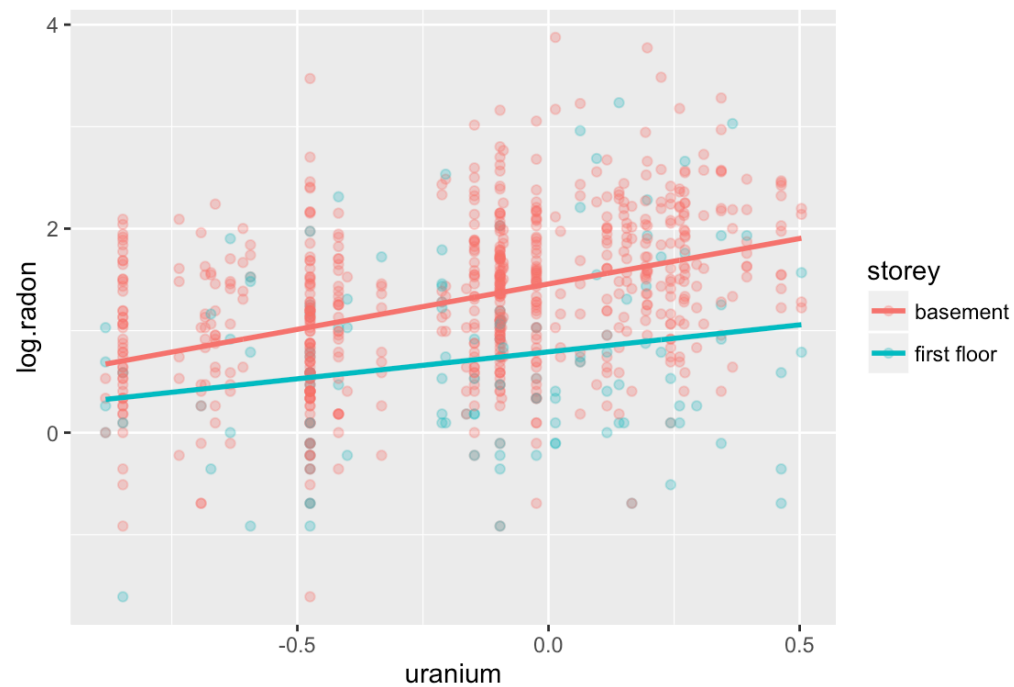


# Or like this



# Your turn

How does the mixed effects model differ from the simple linear model? (Hint: Think about the variance.)



# Assumptions

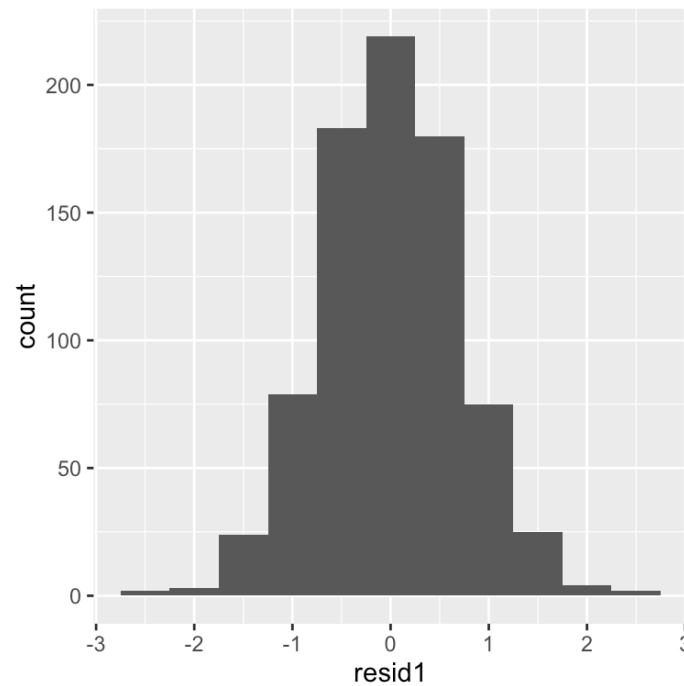
Recall:

$$\underset{(n_i \times 1)}{\mathbf{y}_i} = \underset{(n_i \times p)(p \times 1)}{\mathbf{X}_i \boldsymbol{\beta}} + \underset{(n_i \times q)(q \times 1)}{\mathbf{Z}_i \mathbf{b}_i} + \underset{(n_i \times 1)}{\boldsymbol{\varepsilon}_i}$$

- $\mathbf{b}_i$  is a random sample from  $\mathcal{N}(\mathbf{0}, \mathbf{D})$  and independent from the level-1 error terms,
- $\boldsymbol{\varepsilon}_i$  follow a  $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{R}_i)$  distribution
- $\mathbf{D}$  is a positive-definite  $q \times q$  covariance matrix and  $\mathbf{R}_i$  is a positive-definite  $n_i \times n_i$  covariance matrix

# Extract and examine level-1 residuals

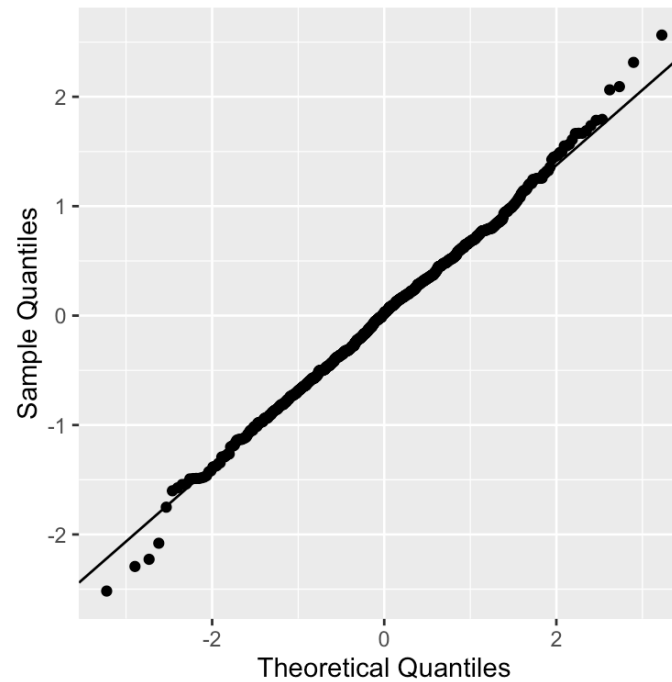
$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{R}_i)$$



Level-1 (observation level) look normal.



# qqplot

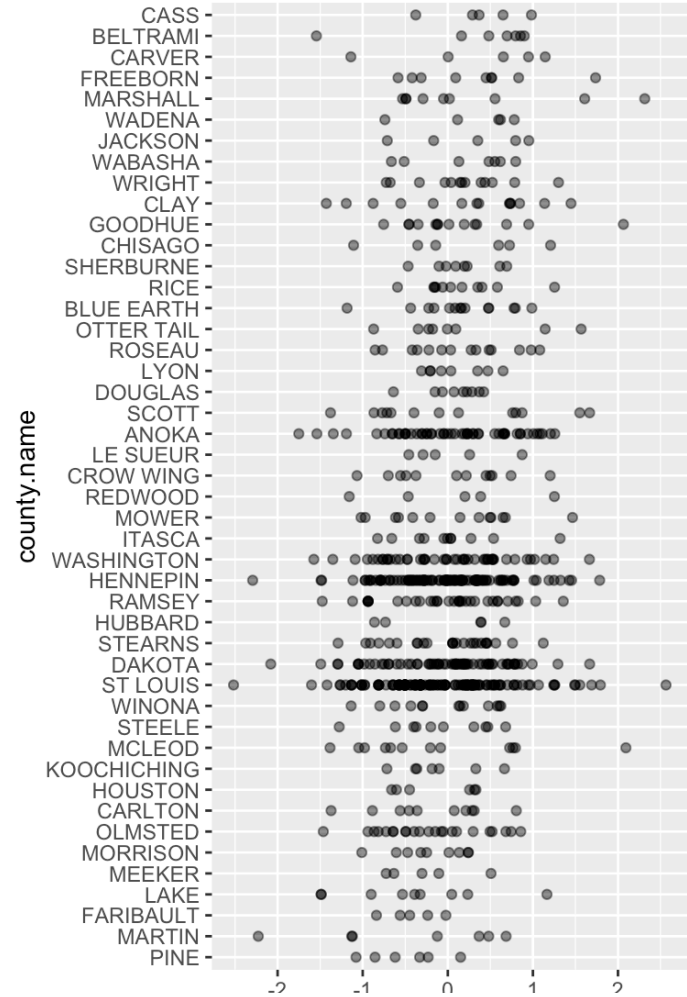


Level-1 (observation level) do look nearly normal.

# Examine within group

## Summary statistics

```
## # A tibble: 6 x 4
##   county.name      m      s      n
##   <fctr>    <dbl> <dbl> <int>
## 1      ANOKA  0.051 0.719   52
## 2    BELTRAMI  0.335 0.867    7
## 3 BLUE EARTH  0.152 0.562   14
## 4    CARLTON -0.194 0.651   10
## 5     CARVER  0.322 0.924    5
## 6      CASS   0.383 0.504    5
```





resid1

27/39

# Learn

There is some difference on average between counties, which means that residuals still have some structure related to the county location.

# Normality tests

Anderson-Darling, Cramer-von Mises, Lilliefors (Kolmogorov-Smirnov)

```
##  
## Anderson-Darling normality test  
##  
## data: radon_lmer_fit$resid1  
## A = 0.4, p-value = 0.4
```

all believe that the residuals are consistent with normality.

# Conclusion about level-1 residuals

The assumption:

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{R}_i)$$

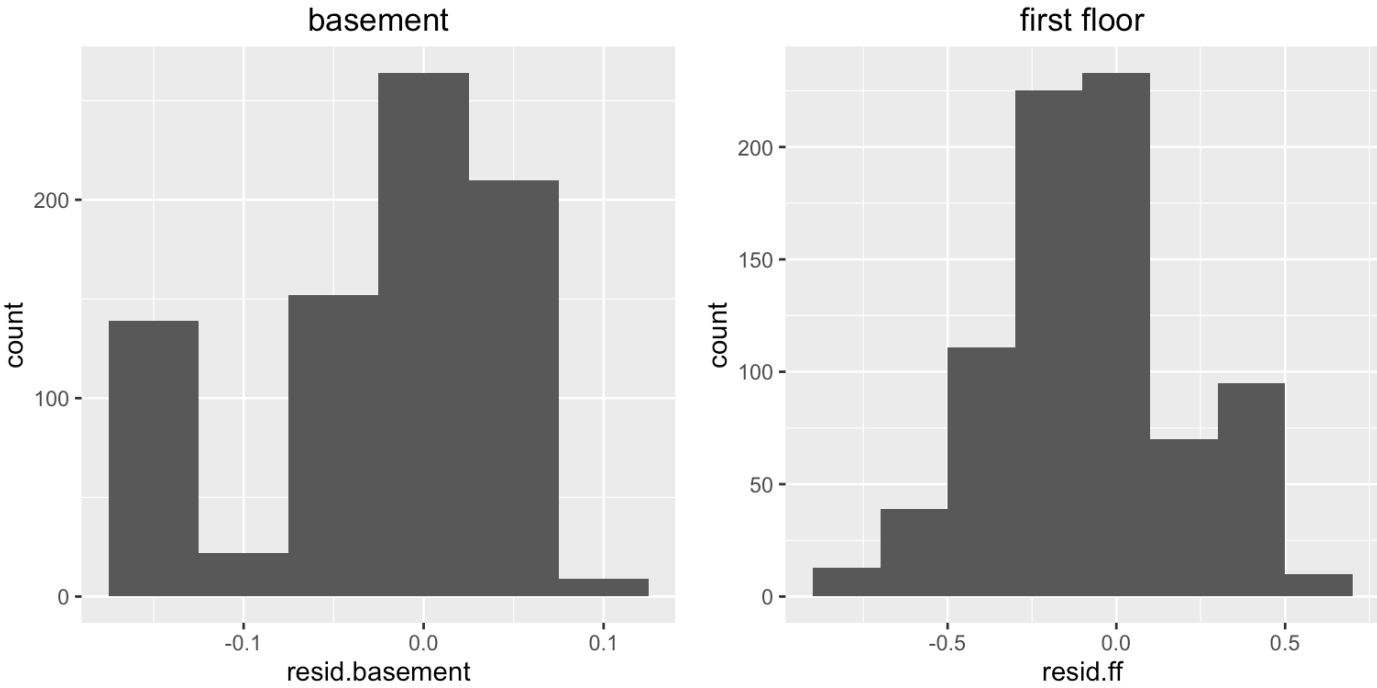
is probably ok, at the worst it is not badly violated.

# Random effects

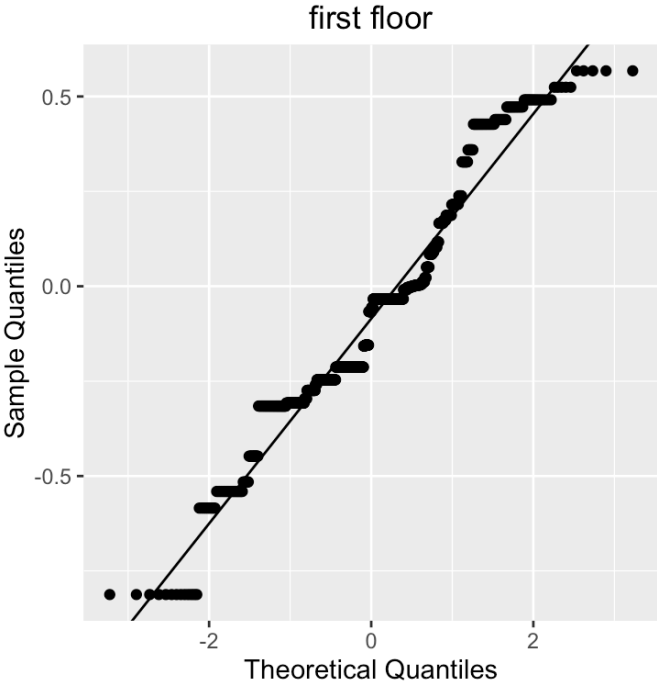
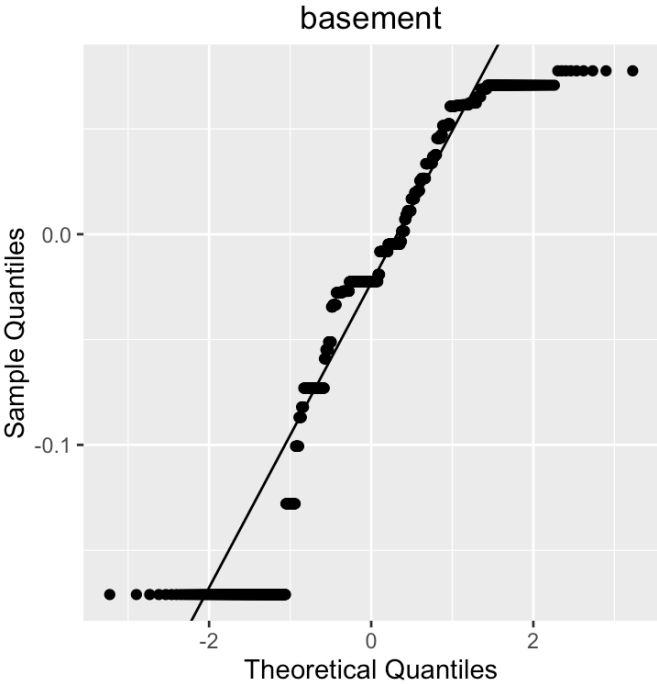
$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \quad i = 1, \dots, g$$

where  $\mathbf{D}$  allows for correlation between random effects within group, and these should be independent from the level-1 error

We have both intercepts (basement) and slopes (first floor)

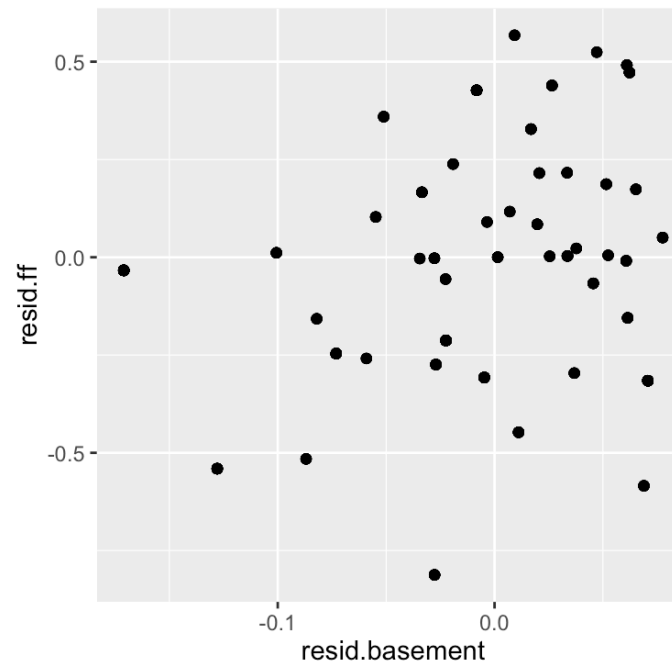






33/39

# Should be no correlation



# Fitted vs Observed

Plotting the observed vs fitted values, gives a sense for how much of the response is explained by the model. Here we can see that there is still a lot of unexplained variation.



35/39

## Goodness of fit

From the linear model

```
##      null.deviance df.null logLik  AIC  BIC deviance df.residual
## 1              539      795  -887 1783 1806      432          792
```

From the random effects model

```
##      sigma logLik  AIC  BIC deviance df.residual
## 1 0.712   -885 1784 1817      1760          789
```

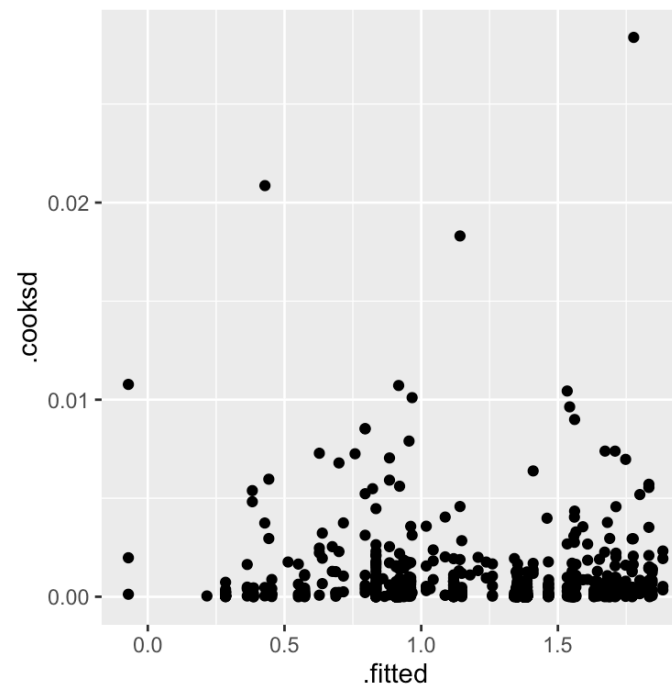
Hmmm... deviance looks strange! Compute sum of squares of residuals instead:

```
## [1] 387
```

Which model is best?

36/39

# Influence



No overly influential observations

# Resources

- [HLMDiag package explanation](#)
- [HLM package](#)

# Share and share alike

This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 United States License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/3.0/us/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

