Multilevel Models

Professor Diane Lambert

June 2010

Supported by MOE-Microsoft Key Laboratory of Statistics and Information Technology and the Beijing International Center for Mathematical Research, Peking University.

With many thanks to Professor Bin Yu of University of California Berkeley, and Professor Yan Yao and Professor Ming Jiang of Peking University.

A Simple Multilevel Model

Data are observed in groups

Many groups!

e.g., take 2 mice from each of L litters.

The groups may not be interesting, but they increase randomness in the outcome,

which affects the other estimated coefficients

$$Y_i = b_0 + b_1X_1 + b_2X_2 + e_{G[i]} + e_i$$

$X_1$and X_2 are fixed effects

$e_{G[i]}$ is a random effect -- one for each group

Other Multilevel Models

There can be more than one random effect

pairs of litters (litter effect)

each mouse observed 5 times (time effect)

$$Y_i = b_0 + b_1X_1 + b_2X_2 + e_{G[i]} + e_{t[i]} + e_i$$

$X_1$and X_2 are fixed effects

$e_{G[i]}$ is a random effect -- one for each group

$e_{t[i]}$ is the same for all mice measurements at t

Other Multilevel Models

Random effects can be nested

observe houses in counties

counties are part of states

$$Y_i = b_0 + b_1X_1 + b_2X_2 + e_{S[i]} + e_{[C|S][i]} + e_i$$

effect of county is random, and nested in state

first draw the random effect for state S

then draw the random effect for county given state.

(Could also have random effects in a glm.)

Example

Radon is a gas that seeps naturally from the ground everywhere

radon measurements (need to log) on

12,687 houses

in 386 counties

in 9 states

Other variables

region (properties of the earth)

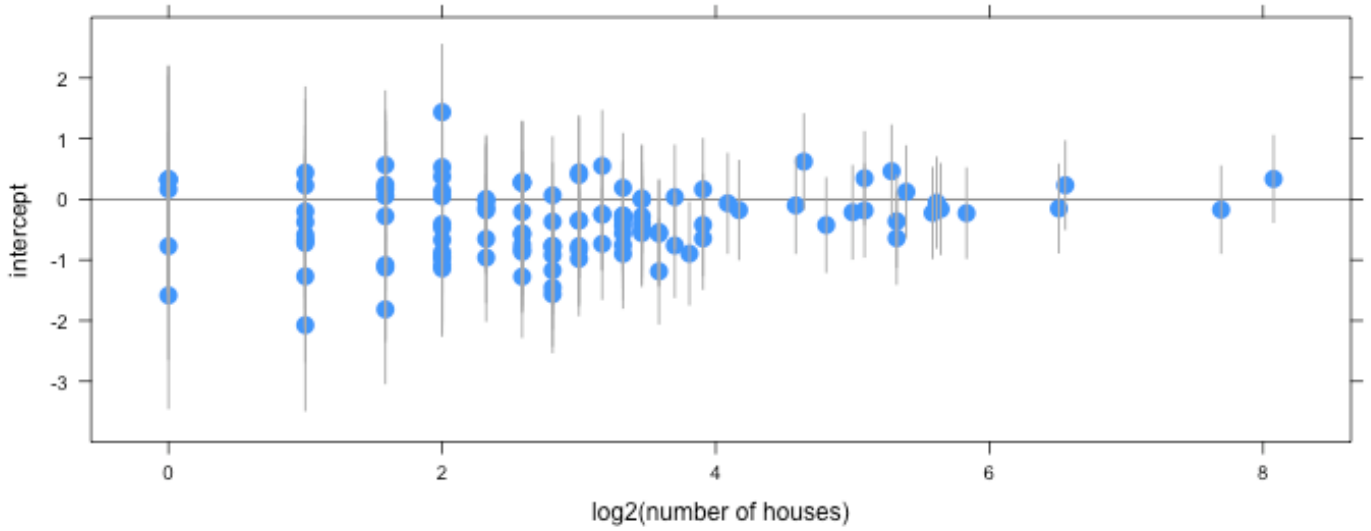
floor of house monitored, room, wave

Consider One State: MO

1850 houses in 115 counties

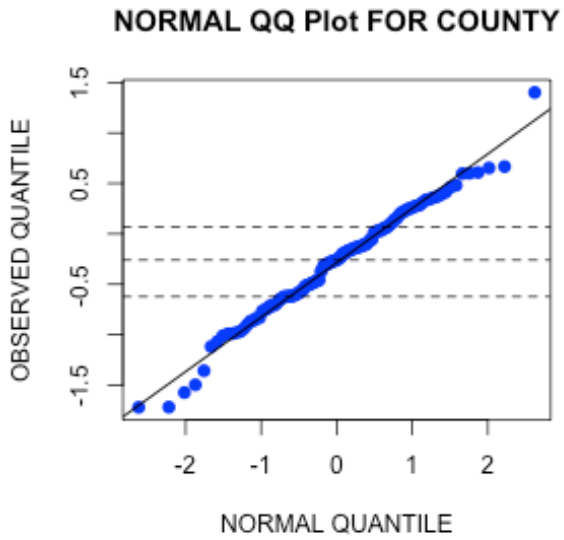
- 31% have less than 5 observations (houses)
- 80% have less than 15 observations

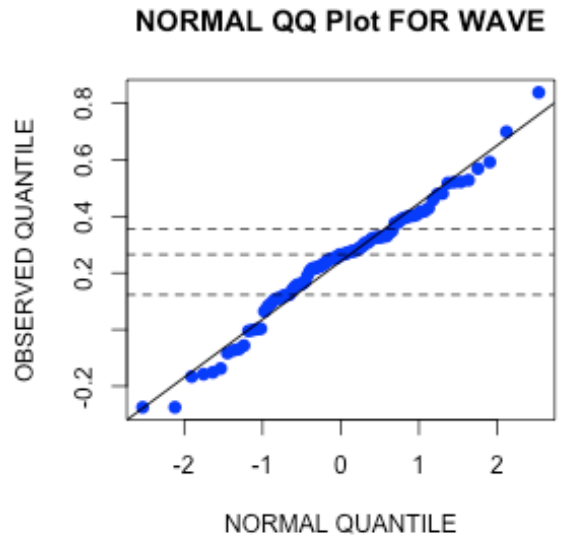
```
z <- lm(radon ~ county + floor +
                typebldg , data = MO)
```



County and Wave Effects

```
z <- lm(radon ~ county + floor +
                  typebldg , data = MO)
```





The intercept of the gray line is the mean effect. The slope is the standard deviation of the effects.

MO continued

most of county and wave coefs are not stat sig

with so many, we'd expect some to be stat sig

with a 5% test

- wave: 5% are stat sig on left; 5% on right
- county: 4% on the left; 0.8% on the right

The level effects act like a random sample.

- complete pooling:
  - ignore the effect, use a distribution wp1 on 0
- no pooling:
  - exaggerates the differences in county -- by chance some will be stat sig

Simple Multilevel Models in R

Use the lme4 package

may take some work to get this installed

runs under R2.11.1, not under the version in the lab

`lmer` fits a multilevel linear model

`glmer` fits a multilevel glm

Formulas are a little different

```
lmer(y ~ x1 + x2 + (1|a1) + (1|a2), data)
```

fixed effects

random intercepts

LM Example

```
z <- lm(radon ~ floor + typebldg, data = MO)
```

	Estimate	Std. Error	t value
(Intercept)	1.07	0.10	10.61
floor	-0.27	0.03	-9.79
typebldg1	-0.45	0.10	-4.48
typebldg2	-0.80	0.18	-4.53
typebldg3	-2.00	0.96	-2.08
typebldg5	-1.23	0.29	-4.22

```
Residual standard error: 0.9533 on 1853
degrees of freedom
Multiple R-squared: 0.1069
```

Random County Effects

```
zc <- lmer(radon ~ floor + typebldg +
                  (1|county), data = MO)

Fixed effects:
            Estimate Std. Error t value  oldEst
(Intercept)  0.73     0.107     6.810
floor       -0.21     0.027    -7.775     -.27
typebldg1   -0.30     0.095    -3.162     -.45
typebldg2   -0.71     0.166    -4.249     -.80
typebldg3   -1.53     0.902    -1.696    -2.00
typebldg5   -0.94     0.278    -3.391    -1.23

Random effects:
 Groups   Name          Variance Std.Dev.
 county   (Intercept)   0.15457  0.39316
 Residual               0.78677  0.88700
```

Random County and Wave Effects

```
zcw <- lmer(radon ~ floor + typebldg +
       (1|county) + (1|wave), data = MO)
```

```
Fixed effects hardly change

Random effects:
 Groups   Name        Variance  Std.Dev.
 county   (Intercept) 0.1531118 0.391295
 wave     (Intercept) 0.0093978 0.096942
 Residual             0.7777066 0.881877
```

.78/.15 ~ 5, so the standard devn of radon between counties is like the standard devn of the mean of 5 radon measurements in a county.

Wave is responsible for only a little variation in radon

Random Slope Models

```
zcwSlope <- lmer(radon ~ floor + typebldg +
        (1|county) + (floor|wave), data = MO)
```

One interpretation:

- counties
  - randomly sampled (or adds random noise)
  - affects the level of the response
- wave
  - randomly sampled (or adds random noise)
  - affects the level and the slope of floor

```
radon ~ floor + typebldg + (1|county) + (floor|wave)
```

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 county   (Intercept) 0.137659 0.37102
 wave     (Intercept) 0.024146 0.15539
          floor       0.093155 0.30521  -0.732
 Residual             0.729538 0.85413

Fixed effects:
             Estimate Std. Error t value
(Intercept)  0.78527    0.10572   7.427
floor       -0.46577    0.05125  -9.088
typebldg1   -0.22637    0.09342  -2.423
typebldg2   -0.56726    0.16304  -3.479
typebldg3   -1.40099    0.88756  -1.578
typebldg5   -0.73675    0.27344  -2.694
```

Is (floor|wave) better than (1|wave)?

Use anova to compare the fit after (floor|wave) to the fit from the simpler model (1|wave)

```
zcf.w:
    radon ~ floor + typebldg + (1|county) + (floor|wave)
zcw:
    radon ~ floor + typebldg + (1|county) + (1|wave)


anova(zcf.w, zcw)
```

	Df	logLik	Chisq	Df	Pr(>Chisq)
zcw	9	-2470.7			
zcf.w	11	-2443.5	54.282	2	1.633e-12

Nested Models

```
zcw <- lmer(radon ~ floor + typebldg +
        (1|county) + (wave|county), data = MO)
```

Warning: this uses a lot of memory

52,110 random intercepts for our data!

One interpretation:

counties

affect the level of the response

wave

each county has its own distribution of random wave effects

Probably not a sensible model for our data.

Estimated Effects

Model is `zcw`

`coef(zcw)` gives the fixed and random models in a list

`fixef(zcw)` gives fixed effects (vector)

`ranef(zcw)` gives random effects (list)

vector for wave, vector for county

To get mean estimates within counties

wave is averaged out

random intercept = fixed intercept + county effect

fixed slopes for floor, typebldg

Mean estimates within waves

county is averaged out, so the coefs are

random intercept = fixed intercept + wave effect

fixed slopes for floor typebldg

Uncertainty in Estimated Means

For lm's and glm's we used simulation

- simulate (b_0, …, b_k) from its approx distribution
- compute $b_0 + b_1X_1 + \ldots + b_kX_k$ for each simulated **b**

For multilevel models

- need to include the random effects
- easier to run a full simulation

```
simRadon <- simulate(zcw, nsim)
```

  - `zcw` is an object created by a call to lmer
  - creates nsim simulated trials of the **data** used to fit zcw

Simulating a Multilevel Model

```
zcw <- lmer(radon ~ floor + typebldg +
                    (1|county) + (1|wave))

# Produce new responses for the fitted model
  zcwYSim <- simulate(zcw, nsim = 100)

# Fit the same model to the new responses.
# Put each new model in a list.
  zcwSim <- vector('list', length = 100)
  for (i in 1:nsim) {
    zcwSim[[i]] <- refit(zcw, newresp =
                         zcwYSim[,i])
  }
```

Estimated Means

The fixed effects (and **X**) give the means averaged over all random terms (counties, waves, and noise)

To show simulated errors in those means, show simulated values of $b_0 + b_1X_1 + \ldots + b_kX_k$

(simulation sd's are similar to standard errors)