Multilevel Models

David M. Rocke

June 6, 2017

Multilevel Models

- A good reference on this topic is *Data Analysis using Regression and Multilevel/Hierarchical Models* by Andrew Gelman and Jennifer Hill, 2007, Cambridge University Press.
- The software orientation is both with using `lmer` in R or using `bugs` called from R.
- Bugs is a set of programs for Bayesian analysis of statistical problems. It can sometimes solve problems that are not easily handled in frequentist statistics, but it also can be very slow, and does not always give an answer.
- We will concentrate on analysis using `lmer`.

Multilevel Models

- Multilevel models are those in which individuals observations exist in groups.
- The individuals have potential predictors, but the relationship of the predictor to the prediction can be different in different groups.
- The intercepts may be different, so that all individuals in one group may have on the average higher levels of the response.
- The slopes (coefficients) may be different between groups as well, as in a group-by-predictor interaction.

Radon Data Set

This is a processed subset of the srrs2.dat data set of individual home radon levels in the US. These values are for Minnesota only, and we are interested in household and county level analysis.

Variable	Definition
radon	Radon level in individual home
log.radon	Log-radon or log(0.1) if radon=0
floor	0 = basement, 1 = first floor
county.name	Name of each of 85 counties
county	county number, 1–85

Types of Analysis

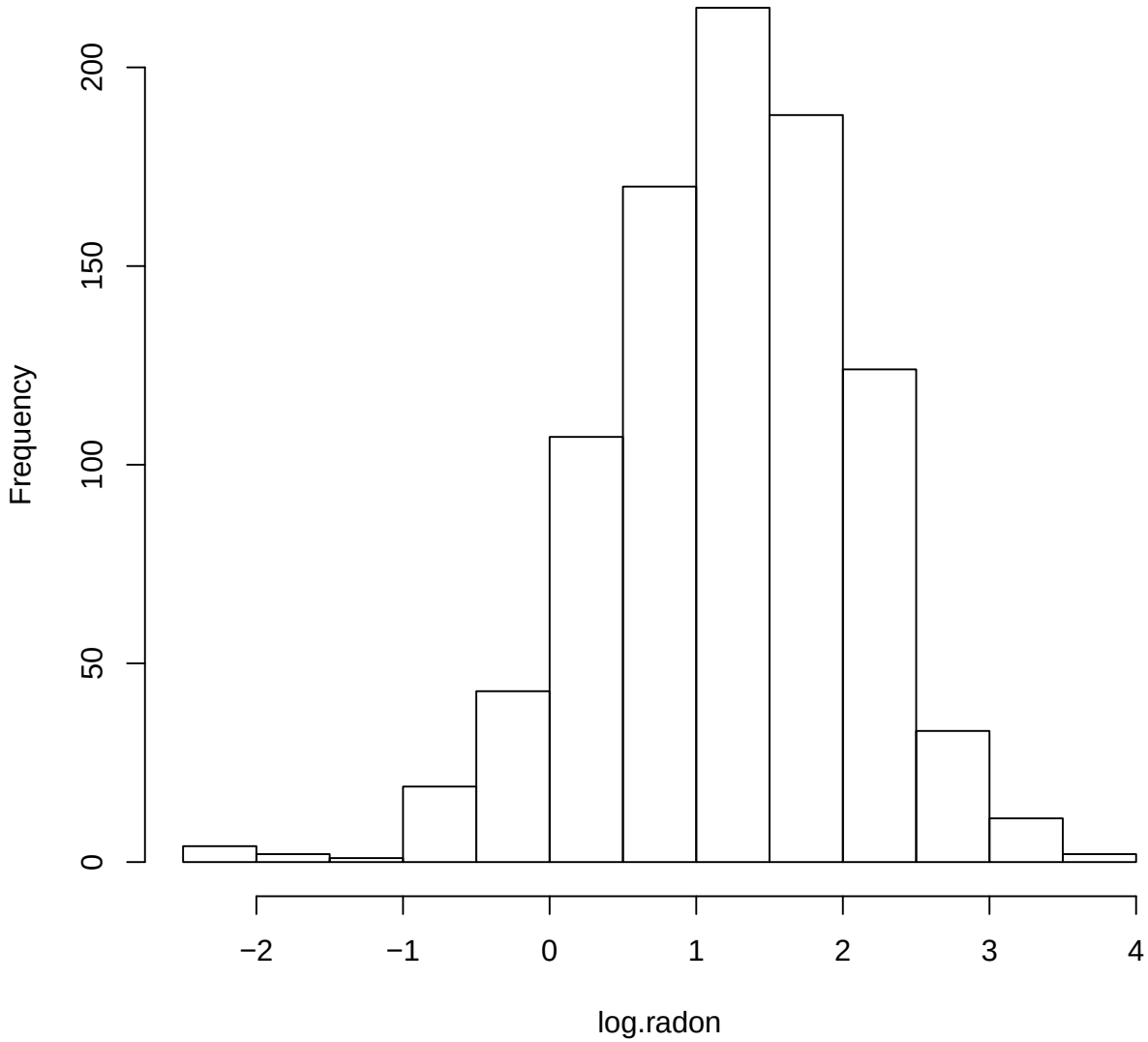
- If we want to know the distribution of radon levels, we can pool the data from all 85 counties.
- Or we can analyze each county separately.
- We can also have a varying intercept for county, but use a pooled error variance.
- Or we can use a two-level model for houses and counties, which is in effect partially pooled.
- In each case, we can add one or more covariates.

Pooled Analysis

```
pool1 <- function(){
# pooled analysis
  print(mean(log.radon))
  print(sd(log.radon))
  pdf("pooled.hist.pdf")
  hist(log.radon)
  dev.off()
}

> pool1()
[1] 1.224623        #mean log radon level across all 919 households
[1] 0.8533272       #standard deviation of log radon level
```

This does not allow any analysis of which counties have the highest radon levels.

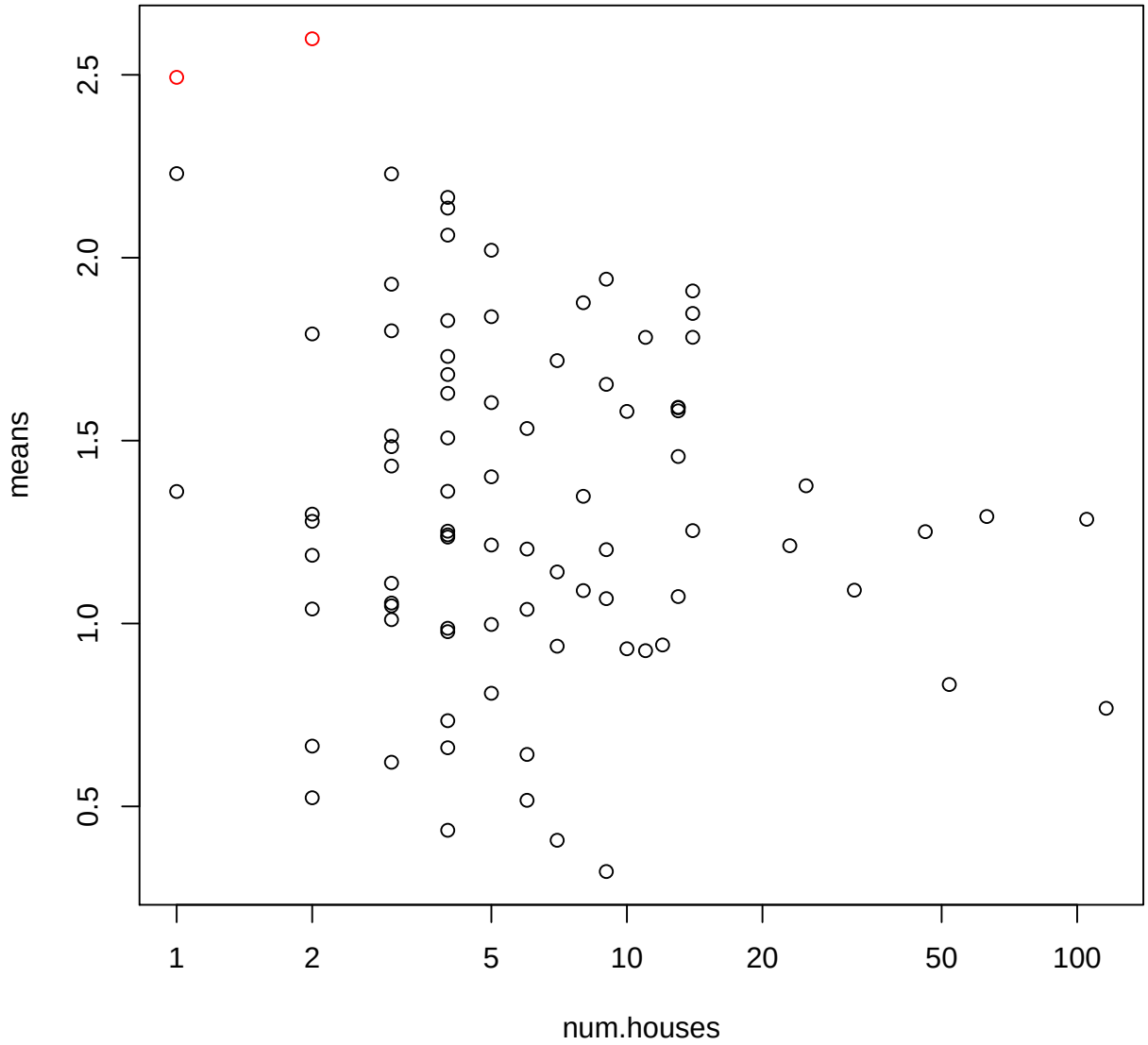

Histogram of log.radon
Frequency
0
50
100
150
200
−2
−1
0
1
2
3
4
log.radon

Unpooled Analysis

```
nopool1 <- function(){
  num.houses <- as.vector(table(county))
  means <- tapply(log.radon,county,mean)
  sds <- tapply(log.radon,county,sd)
  pdf("meanVN.pdf")
  plot(num.houses,means,log="x")
  dev.off()
  print(which(means > 2.3))
  print(county.name[county==50])
  print(county.name[county==36])
}
```

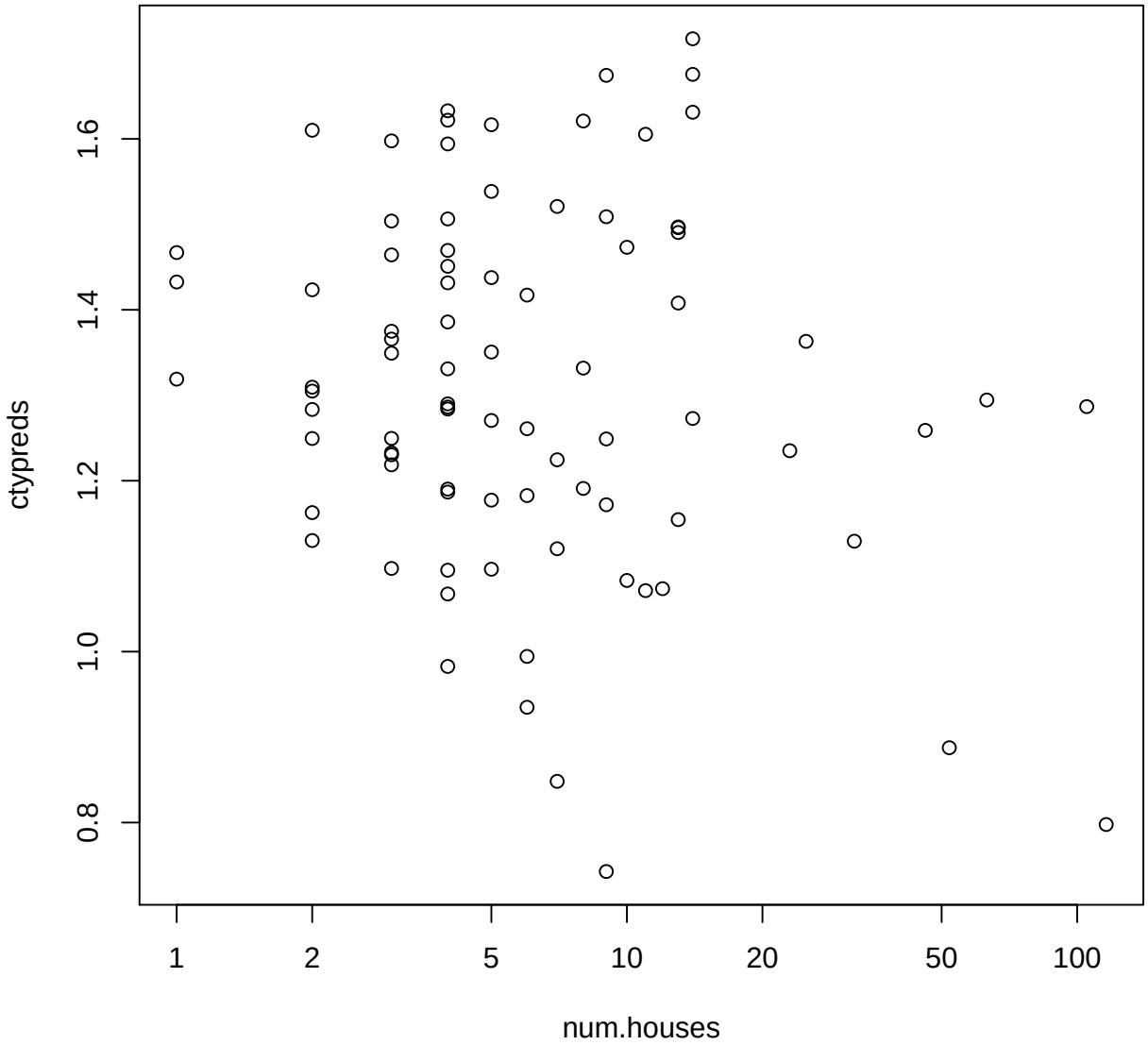
```
> nopool1()
36 50
36 50
[1] "MURRAY              "
[1] "LAC QUI PARLE       " "LAC QUI PARLE       "
```

Two highest radon means have one or two houses per county.
This is probably chance variation.


means
0.5
1.0
1.5
2.0
2.5
1
2
5
10
20
50
100
num.houses

Partial Pooling

```
partialpool1 <- function(){
  require(lme4)
  radon.lmer <- lmer(log.radon ~ 1 + (1|county))
  preds <- predict(radon.lmer)
  num.houses <- as.vector(table(county))
  ctypreds <- tapply(preds,county,mean)
  pdf("ctypredsVN.pdf")
  plot(num.houses,ctypreds,log="x")
  dev.off()
}
```

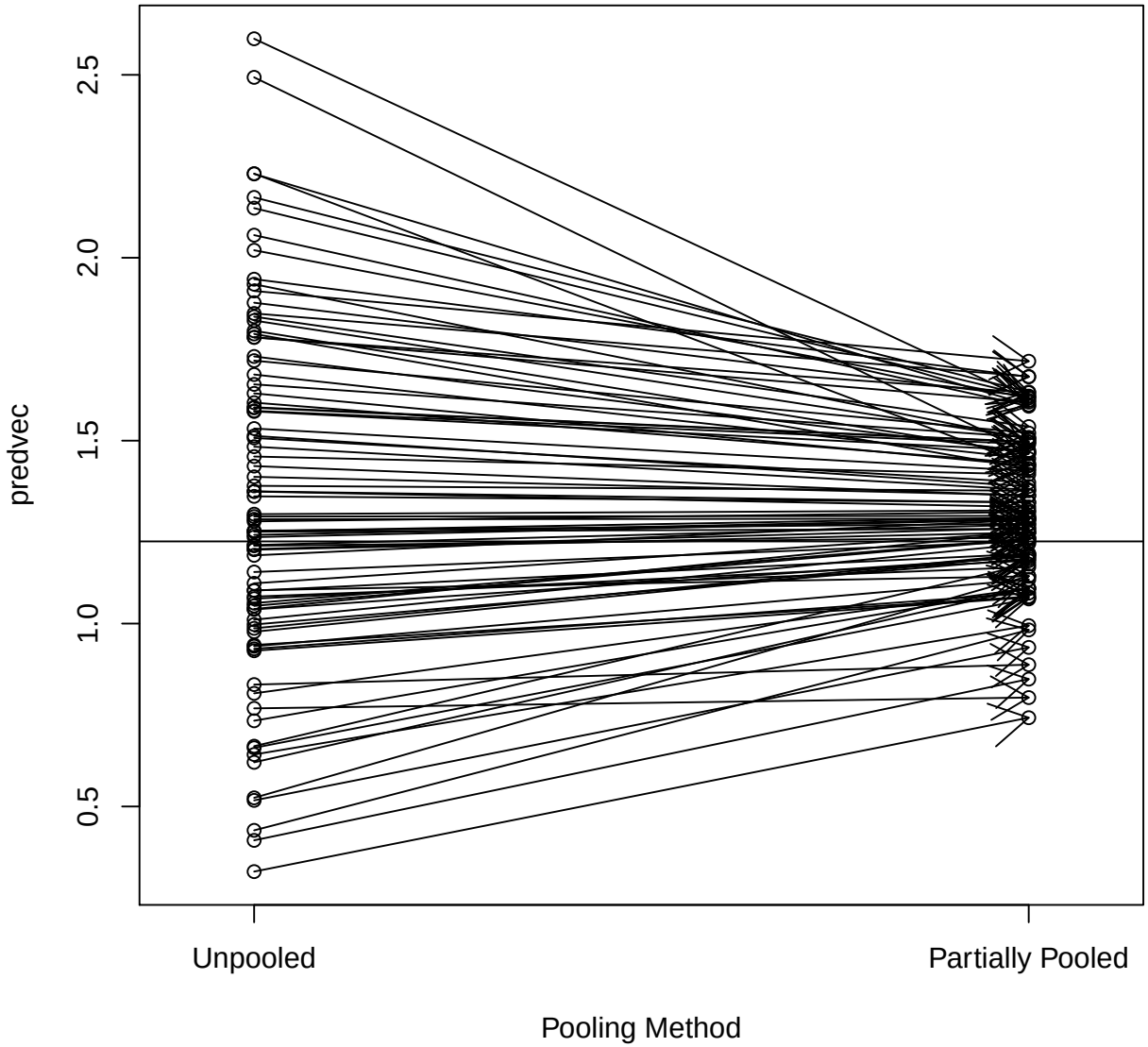

ctypreds
0.8
1.0
1.2
1.4
1.6
1
2
5
10
20
50
100
num.houses

Comparison of Pooling, No Pooling, Partial Pooling

The mean log radon level across all counties is 1.225.

County	Pooled	Unpooled	Partially Pooled
Lac Qui Parle	1.225	2.599	1.610
Murray	1.225	2.493	1.467
Waseca	1.225	0.435	0.983
Koochiching	1.225	0.407	0.848
Lake	1.225	0.322	0.743

```
poolcomp <- function(){
  require(lme4)
  radon.lmer <- lmer(log.radon ~ 1 + (1|county))
  preds <- predict(radon.lmer)
  num.houses <- as.vector(table(county))
  ctypreds <- tapply(preds,county,mean)
  poolpred <- mean(log.radon)
  unpoolpred <- tapply(log.radon,county,mean)
  predvec <- c(unpoolpred,ctypreds)
  n <- length(ctypreds)
  poolmeth <- rep(0:1,each=n)
  pdf("poolcomp.pdf")
  plot(poolmeth,predvec,xlab="Pooling Method",type="p",xaxt="n",xlim=c(-.1,1.1))
  axis(1,at=c(0,1),labels=c("Unpooled","Partially Pooled"))
  abline(h=poolpred)
  arrows(0,unpoolpred,1,ctypreds)
  dev.off()
}
```


predvec
0.5
1.0
1.5
2.0
2.5
Unpooled
Partially Pooled
Pooling Method

Partially Pooled via `lmer`

- The predicted value for each observation in a county is a linear combination of the individual county mean (unpooled) and the pooled grand mean.
- Each county mean is "shrunk" towards the center.
- The county individual mean has a weight of the samples size in the county, which is inversely proportional to the variance of the county mean.

Using Individual-Level Covariates

- The variable `floor` indicates whether the radon reading was taken in the basement, where it likely would be higher, or on the first floor.
- We could add this as a covariate and also if we chose we could make the coefficient of this covariate depend on the county.
- Individual county analysis might not be able to estimate the coefficient of `floor` because 25 of the 85 counties have no houses with data from the first floor.

```
> summary(lmer(log.radon~floor+(1|county)))
Linear mixed model fit by REML ['lmerMod']
Formula: log.radon ~ floor + (1 | county)

REML criterion at convergence: 2171.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.3989 -0.6155  0.0029  0.6405  3.4281

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 t value
(Intercept)  1.46160    0.05158  28.339
floor       -0.69299    0.07043  -9.839

Correlation of Fixed Effects:
      (Intr)
floor -0.288
```

```
> summary(lmer(log.radon~floor+(1+floor|county)))
Linear mixed model fit by REML ['lmerMod']
Formula: log.radon ~ floor + (1 + floor | county)

REML criterion at convergence: 2168.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.4044 -0.6224  0.0138  0.6123  3.5682

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 county   (Intercept) 0.1216   0.3487
          floor       0.1181   0.3436   -0.34
 Residual             0.5567   0.7462
Number of obs: 919, groups:  county, 85

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.46277    0.05387  27.155
floor       -0.68110    0.08758  -7.777

Correlation of Fixed Effects:
      (Intr)
floor -0.381
```

```
> radon.lmer1 <- lmer(log.radon~floor+(1|county))
> radon.lmer2 <- lmer(log.radon~floor+(1+floor|county))
> anova(radon.lmer1,radon.lmer2)
refitting model(s) with ML (instead of REML)
Data: NULL
Models:
radon.lmer1: log.radon ~ floor + (1 | county)
radon.lmer2: log.radon ~ floor + (1 + floor | county)
            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
radon.lmer1  4 2171.7 2190.9 -1081.8   2163.7
radon.lmer2  6 2173.1 2202.1 -1080.5   2161.1 2.5418      2     0.2806
```

Although this test is not reliable because the null hypothesis is on the boundary, the p-value is not near significant and the simpler model has a lower AIC and BIC. The df = 2 because the larger model computes one extra variance and one correlation.

REML (restricted maximum likelihood) vs. ML is like using n - 1 as the denominator for the variance instead of n.