IDS 702: Module 4.3

Multilevel/hierarchical linear models (illustration I)

Dr. Olanrewaju Michael Akande

IDS 702

The radon analysis

There are 919 total observations in the data. The data is in the file `Radon.txt` on Sakai.

Variable	Description
radon	radon levels for each house
log_radon	log(radon)
state	state
floor	lowest living area of each house: 0 for basement, 1 for first floor
countyname	county names
countyID	ID for the county names (1-85)
fips	state + county fips code
uranium	county-level soil uranium
log_uranium	log(uranium)

IDS 702

The radon analysis

```r
Radon <- read.csv("data/Radon.txt", header = T,sep="")
Radon$floor <- factor(Radon$floor,levels=c(0,1),labels=c("Basement","First Floor"))
str(Radon)
```

```
## 'data.frame':	919 obs. of  9 variables:
##  $ radon      : num  2.2 2.2 2.9 1 3.1 2.5 1.5 1 0.7 1.2 ...
##  $ state      : chr  "MN" "MN" "MN" "MN" ...
##  $ log_radon  : num  0.788 0.788 1.065 0 1.131 ...
##  $ floor      : Factor w/ 2 levels "Basement","First Floor": 2 1 1 1 1 1 1 1 1 1 ...
##  $ countyname : chr  "AITKIN" "AITKIN" "AITKIN" "AITKIN" ...
##  $ countyID   : int  1 1 1 1 2 2 2 2 2 2 ...
##  $ fips       : int  27001 27001 27001 27001 27003 27003 27003 27003 27003 27003 ...
##  $ uranium    : num  0.502 0.502 0.502 0.502 0.429 ...
##  $ log_uranium: num  -0.689 -0.689 -0.689 -0.689 -0.847 ...
```

```r
head(Radon)
```

```
##   radon state log_radon       floor countyname countyID  fips  uranium
## 1   2.2    MN 0.7884574 First Floor     AITKIN        1 27001 0.502054
## 2   2.2    MN 0.7884574    Basement     AITKIN        1 27001 0.502054
## 3   2.9    MN 1.0647107    Basement     AITKIN        1 27001 0.502054
## 4   1.0    MN 0.0000000    Basement     AITKIN        1 27001 0.502054
## 5   3.1    MN 1.1314021    Basement      ANOKA        2 27003 0.428565
## 6   2.5    MN 0.9162907    Basement      ANOKA        2 27003 0.428565
##   log_uranium
## 1  -0.6890476
## 2  -0.6890476
## 3  -0.6890476
## 4  -0.6890476
## 5  -0.8473129
## 6  -0.8473129
```

IDS 702

THE RADON ANALYSIS

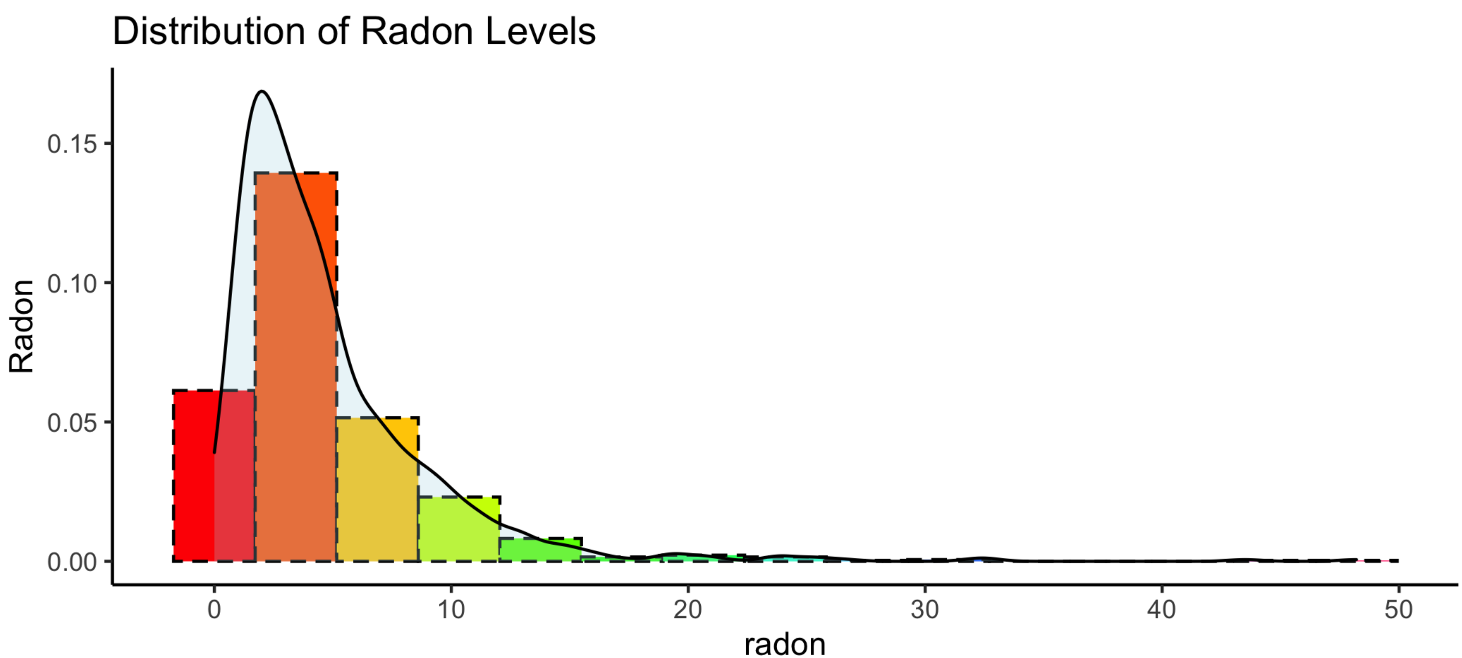
```
table(Radon$countyname) #we don't have enough data in some counties, so we should look to borrow information across counties.
```

```
## 
##           AITKIN            ANOKA           BECKER         BELTRAMI 
##                4               52                3                7 
##           BENTON        BIG STONE      BLUE EARTH            BROWN 
##                4                3               14                4 
##          CARLTON           CARVER             CASS         CHIPPEWA 
##               10                6                5                4 
##          CHISAGO             CLAY       CLEARWATER             COOK 
##                6               14                4                2 
##       COTTONWOOD        CROW WING           DAKOTA            DODGE 
##                4               12               63                3 
##          DOUGLAS        FARIBAULT         FILLMORE         FREEBORN 
##                9                6                2                9 
##          GOODHUE         HENNEPIN          HOUSTON          HUBBARD 
##               14              105                6                5 
##           ISANTI           ITASCA          JACKSON          KANABEC 
##                3               11                5                4 
##        KANDIYOHI          KITTSON      KOOCHICHING   LAC QUI PARLE 
##                4                3                7                2 
##             LAKE LAKE OF THE WOODS         LE SUEUR          LINCOLN 
##                9                4                5                4 
##             LYON         MAHNOMEN         MARSHALL           MARTIN 
##                8                1                9                7 
##           MCLEOD           MEEKER      MILLE LACS         MORRISON 
##               13                5                2                9 
##            MOWER           MURRAY         NICOLLET           NOBLES 
##               13                1                4                3 
##           NORMAN         OLMSTED       OTTER TAIL       PENNINGTON 
##                3               23                8                3 
##             PINE        PIPESTONE             POLK             POPE 
##                6                4                4                2 
##           RAMSEY          REDWOOD        RENVILLE             RICE 
##               32                5                3               11 
##             ROCK           ROSEAU            SCOTT        SHERBURNE 
##                2               14               13                8 
##           SIBLEY         ST LOUIS          STEARNS           STEELE 
##                4              116               25               10 
##          STEVENS            SWIFT             TODD         TRAVERSE 
##                2                4                3                4 
##          WABASHA           WADENA           WASECA       WASHINGTON 
##                7                5                4               46 
##         WATONWAN           WILKIN           WINONA           WRIGHT 
##                3                1               13               13 
##   YELLOW MEDICINE 
```

The radon analysis

The raw radon levels can only take on positive values.

```
ggplot(Radon,aes(radon)) +
  geom_histogram(aes(y=..density..),color="black",linetype="dashed",
                 fill=rainbow(15),bins=15) + theme(legend.position="none") +
  geom_density(alpha=.25, fill="lightblue") + scale_fill_brewer(palette="Blues") +
  labs(title="Distribution of Radon Levels",y="Radon") + theme_classic()
```



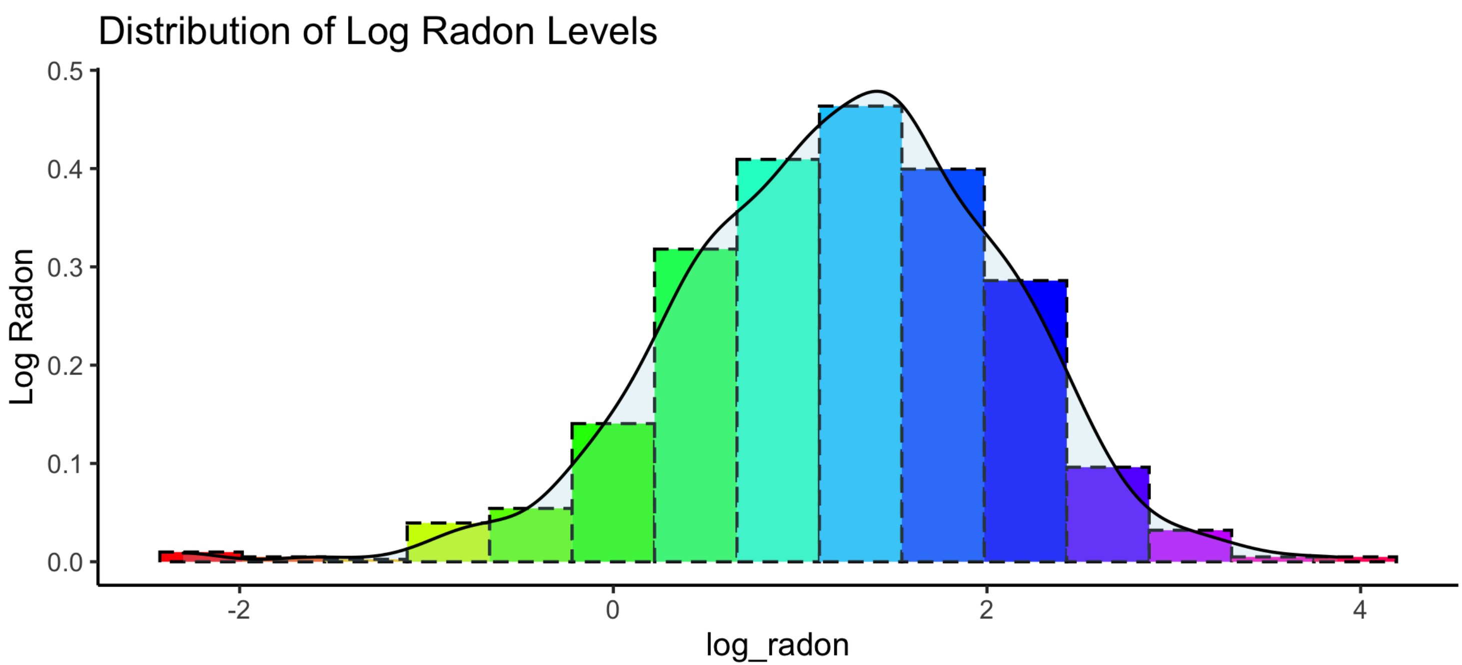
Obviously very skewed.

IDS 702

The radon analysis

Let's look at `log_radon` instead.

```
ggplot(Radon,aes(log_radon)) +
  geom_histogram(aes(y=..density..),color="black",linetype="dashed",
                 fill=rainbow(15),bins=15) + theme(legend.position="none") +
  geom_density(alpha=.25, fill="lightblue") + scale_fill_brewer(palette="Blues") +
  labs(title="Distribution of Log Radon Levels",y="Log Radon") + theme_classic()
```



Much better! Let's go with log radon for now.

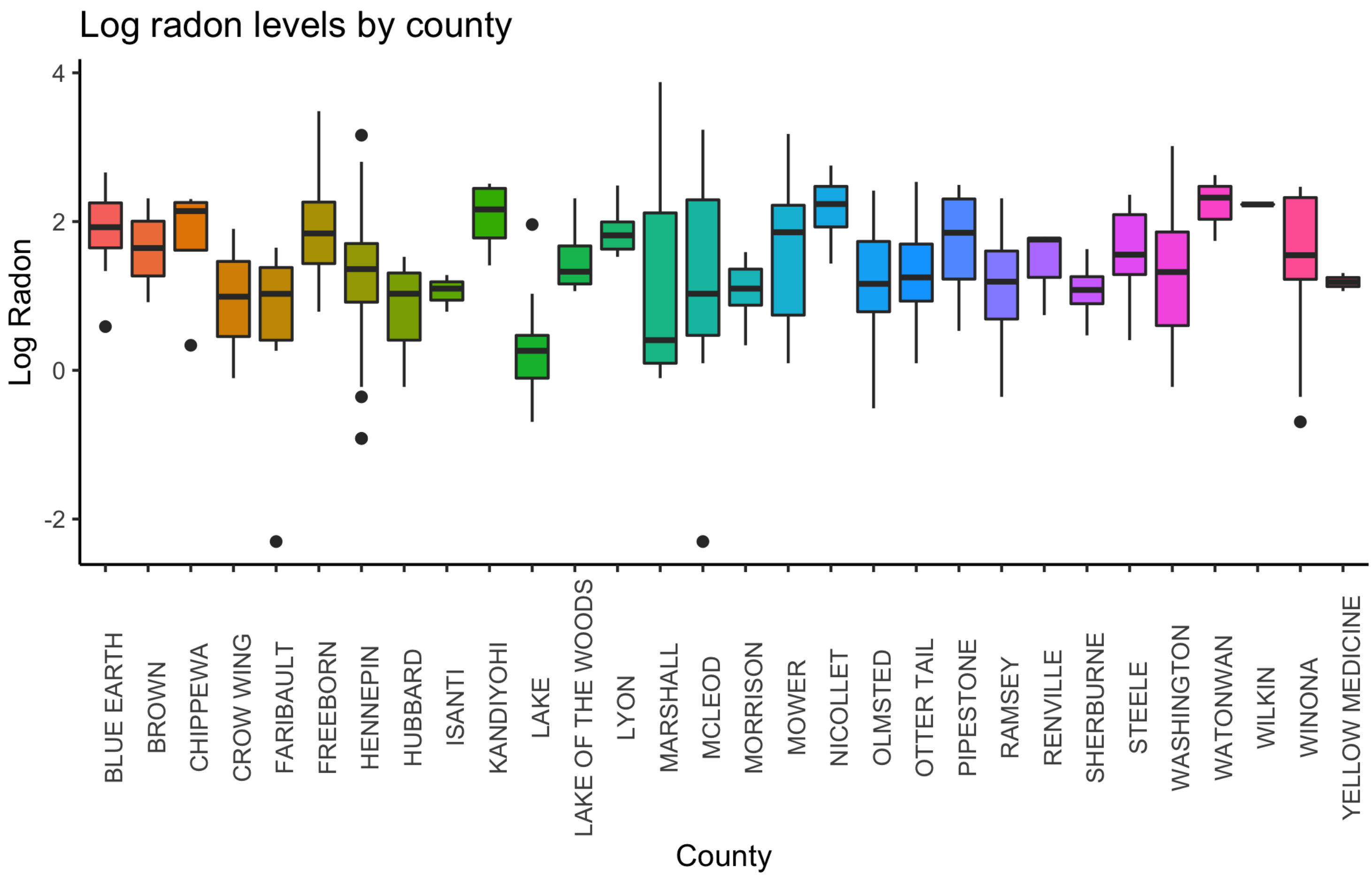
The radon analysis

Are there any variations of radon levels by county? There are too many counties, so, let's do it for a random sample of counties.

```
set.seed(1000)
sample_county <- sample(unique(Radon$countyname),25,replace=F)
ggplot(Radon[is.element(Radon$countyname,sample_county),],
       aes(x=countyname, y=log_radon, fill=countyname)) +
  geom_boxplot() +
  labs(title="Log radon levels by county",
       x="County",y="Log Radon") + theme_classic() +
  theme(legend.position="none",axis.text.x = element_text(angle = 90))
```

IDS 702

The radon analysis



Looks like the levels vary by county. However, there are many counties with very little data.

IDS 702

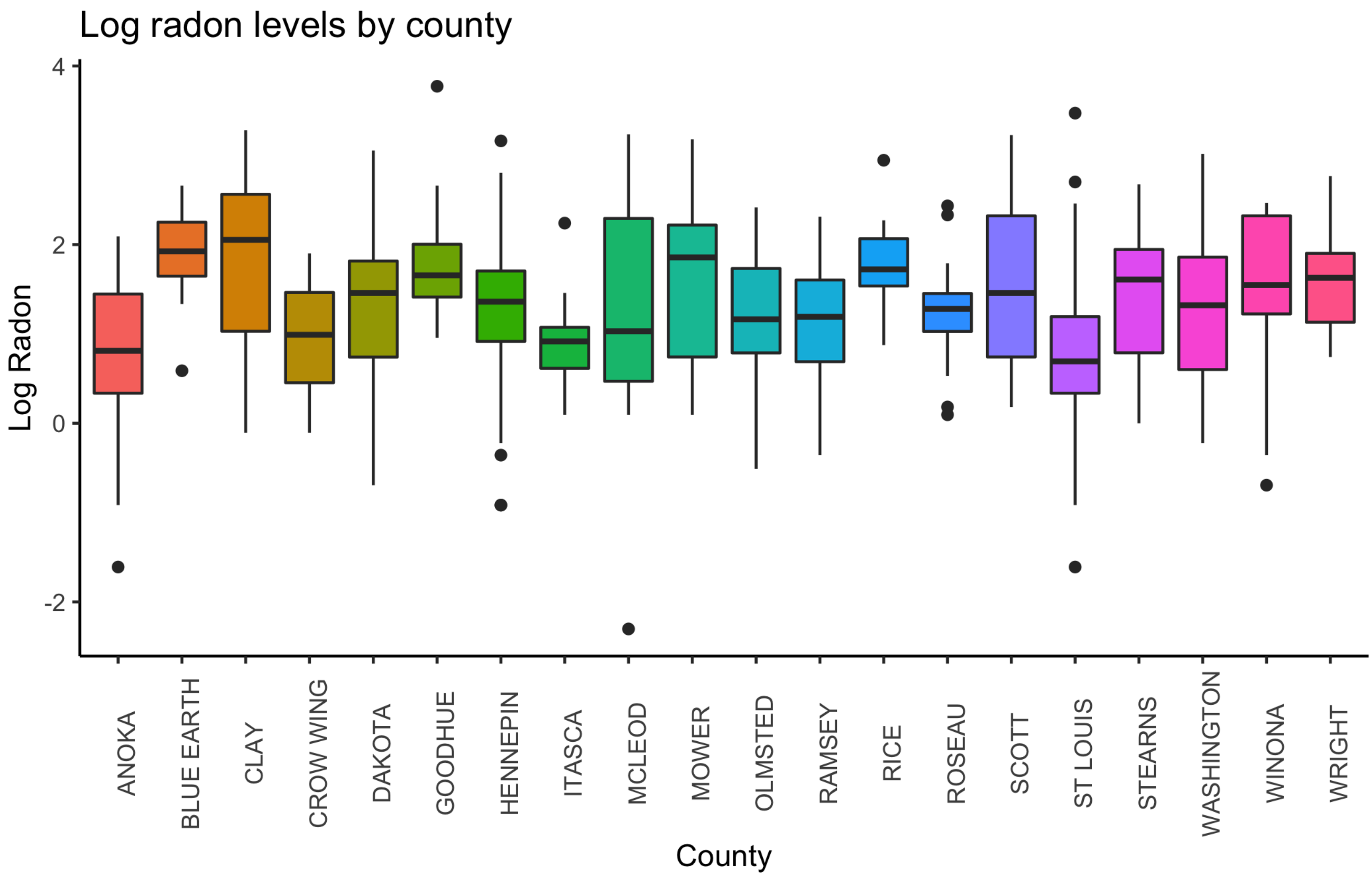
The radon analysis

Let's focus on counties with at least 11 houses.

```
sample_county <- which(table(Radon$countyID) > 10)
ggplot(Radon[is.element(Radon$countyID,sample_county),],
       aes(x=countyname, y=log_radon, fill=countyname)) +
  geom_boxplot() +
  labs(title="Log radon levels by county",
       x="County",y="Log Radon") + theme_classic() +
  theme(legend.position="none",axis.text.x = element_text(angle = 90))
```

IDS 702

The radon analysis



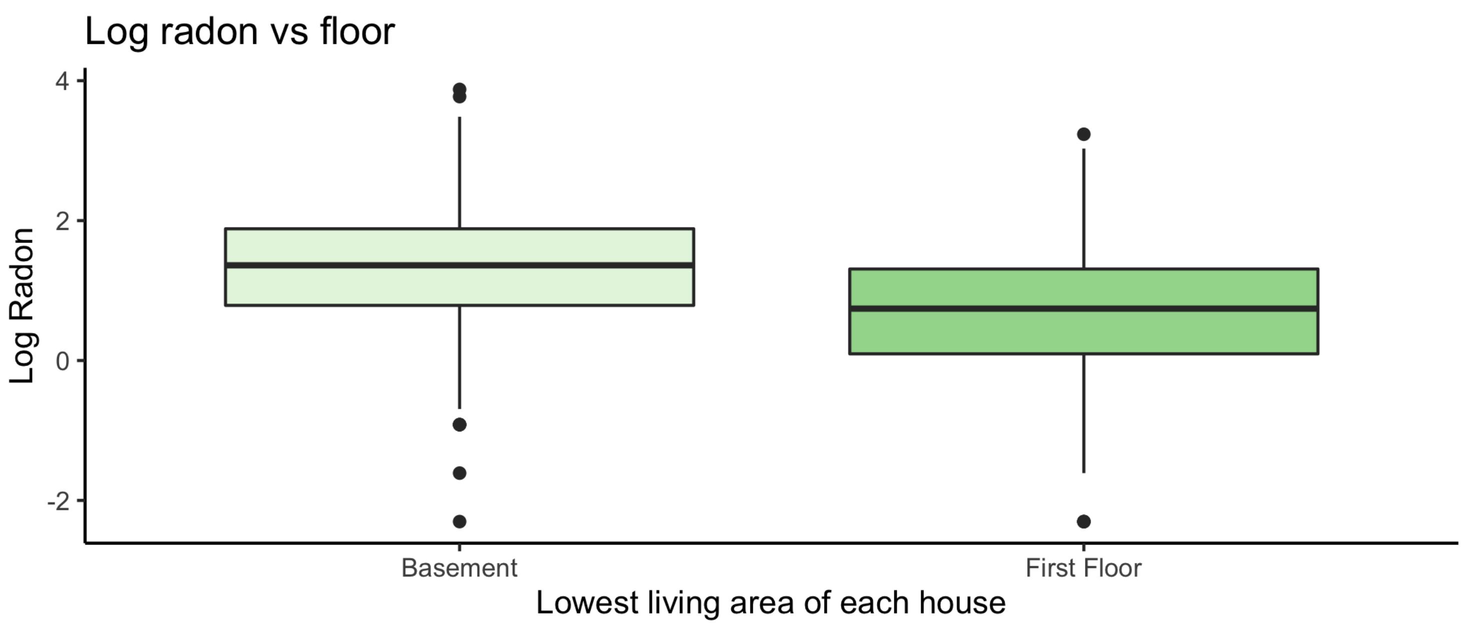
What can you conclude from this plot?

IDS 702

The radon analysis

Next, the relationship with `floor`, the only individual-level (different observation for each house) variable we have.

```
ggplot(Radon,aes(x=floor, y=log_radon, fill=floor)) +
  geom_boxplot() + scale_fill_brewer(palette="Greens") +
  labs(title="Log radon vs floor", x="Lowest living area of each house",y="Log Radon") +
  theme_classic() + theme(legend.position="none")
```



Looks like radon levels are higher for houses with the basement as the lowest living area.

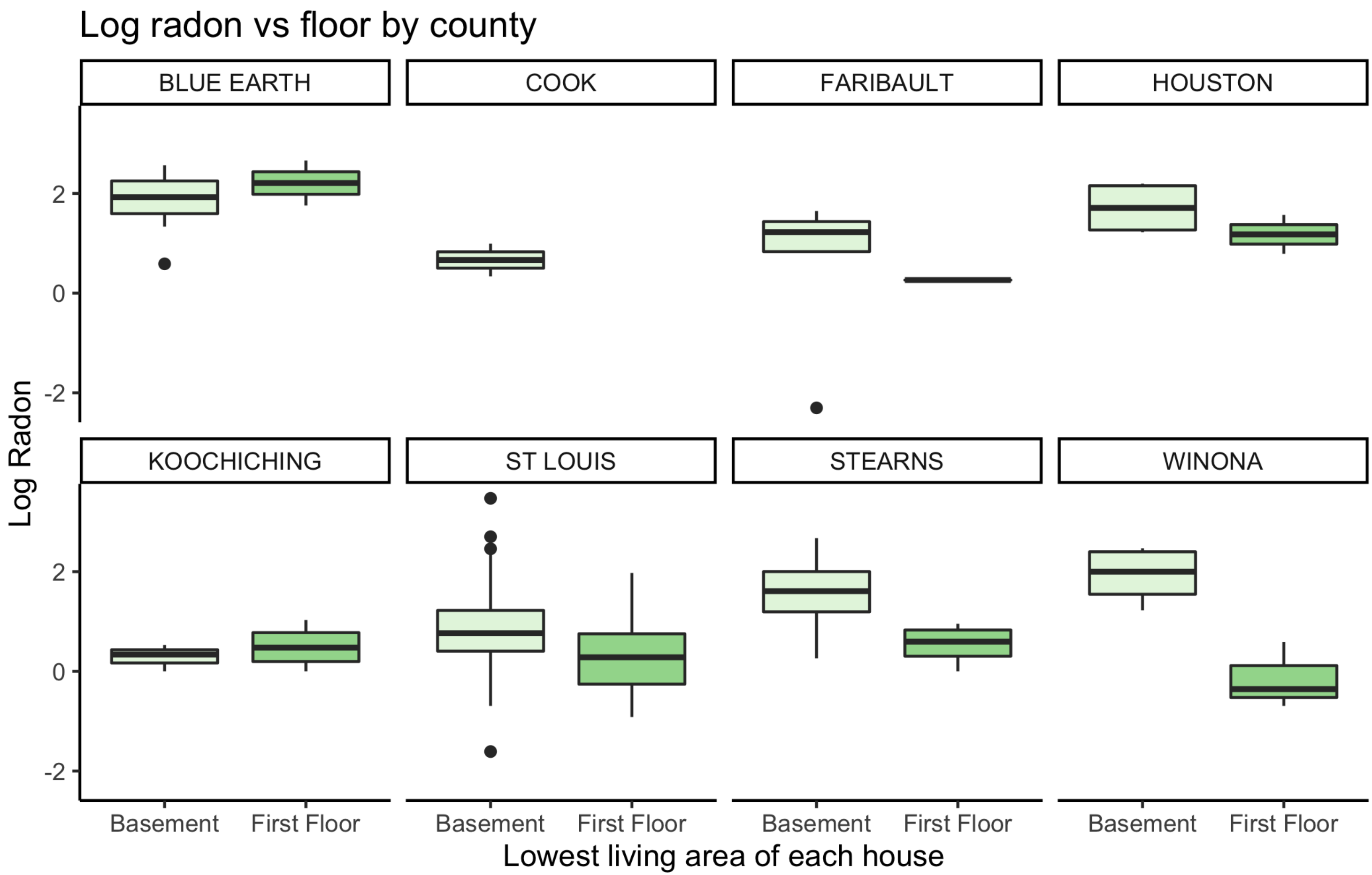
The radon analysis

Let's look at the same relationship for a random sample of counties.

```
sample_county <- sample(unique(Radon$countyname),8,replace=F)
ggplot(Radon[is.element(Radon$countyname,sample_county),],
       aes(x=floor, y=log_radon, fill=floor)) +
  geom_boxplot() +
  scale_fill_brewer(palette="Greens") +
  labs(title="Log radon vs floor by county",
       x="Lowest living area of each house",y="Log Radon") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ countyname,ncol=4)
```

IDS 702

The radon analysis



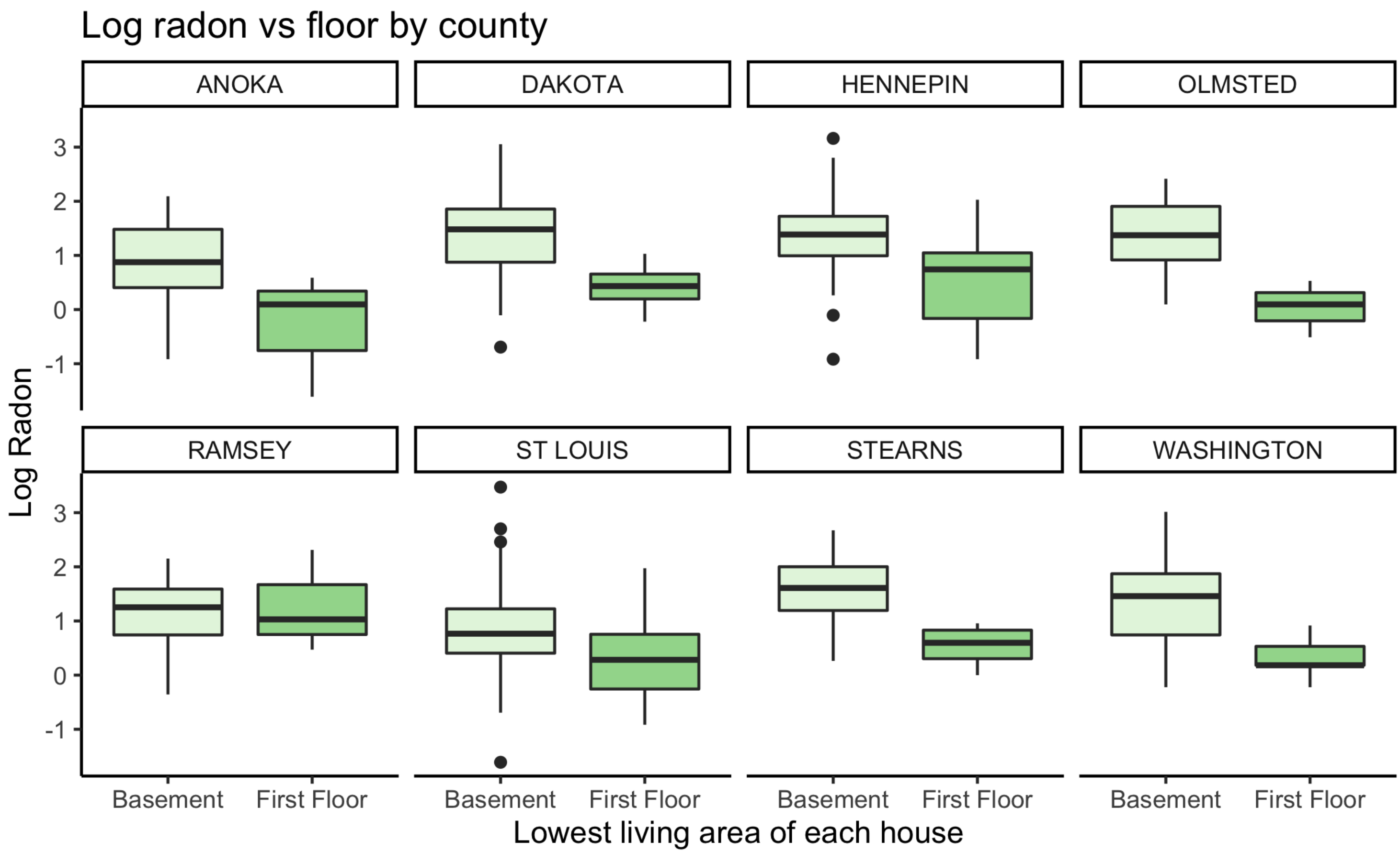
Again, not enough data for some counties.

The radon analysis

Let's focus on counties with at least 16 houses.

```
sample_county <- which(table(Radon$countyID) > 15)
ggplot(Radon[is.element(Radon$countyID,sample_county),],
       aes(x=floor, y=log_radon, fill=floor)) +
  geom_boxplot() +
  scale_fill_brewer(palette="Greens") +
  labs(title="Log radon vs floor by county",
       x="Lowest living area of each house",y="Log Radon") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ countyname,ncol=4)
```

The radon analysis



Even though the overall direction is the same, it looks like the actual differences between floor = 0 and floor = 1 differs for some counties.

IDS 702

The radon analysis

- Let's start by only focusing on `floor`.
- We will try a varying-slope, varying-intercept linear model.
- Let y_{ij} and x_{1ij} be the log radon level and indicator variable `floor` respectively for house i in county j.
- Mathematically, we have

$$\begin{aligned} y_{ij} &= (\beta_0 + \gamma_{0j}) + (\beta_1 + \gamma_{1j}) x_{1ij} + \epsilon_{ij}; \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, 85 \\ \epsilon_{ij} &\sim N(0, \sigma^2) \\ (\gamma_{0j}, \gamma_{1j}) &\sim N_2(\mathbf{0}, \Sigma). \end{aligned}$$

- Alternative representation:

$$\begin{aligned} \log(\text{radon}_{ij}) &= (\beta_0 + \gamma_{0j}) + (\beta_1 + \gamma_{1j}) \text{ floor}_{ij} + \epsilon_{ij}; \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, 85 \\ \epsilon_{ij} &\sim N(0, \sigma^2) \\ (\gamma_{0j}, \gamma_{1j}) &\sim N_2(\mathbf{0}, \Sigma). \end{aligned}$$

The radon analysis

- We skipped this before but Σ actually takes the form

$$\Sigma = \begin{bmatrix} \tau_0^2 & \rho\tau_0\tau_1 \\ \rho\tau_0\tau_1 & \tau_1^2 \end{bmatrix}$$

where

- τ_0^2 describes the across county variation attributed to the random/varying intercept,
- τ_1^2 describes the across county variation attributed to the random/varying slope (that is, floor), and
- ρ describes the correlation between γ_{0j} and γ_{1j}.

IDS 702

The radon analysis

In R, we have

```
Model1 <- lmer(log_radon ~ floor + (floor | countyname), data = Radon)
summary(Model1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log_radon ~ floor + (floor | countyname)
##    Data: Radon
##
## 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
##  countyname (Intercept)      0.1216   0.3487
##             floorFirst Floor 0.1181   0.3436   -0.34
##  Residual                    0.5567   0.7462
## Number of obs: 919, groups:  countyname, 85
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)       1.46277    0.05387  27.155
## floorFirst Floor -0.68110    0.08758  -7.777
##
## Correlation of Fixed Effects:
##             (Intr)
## florFrstFlr -0.381
```

IDS 702

Interpretation of fixed effects

- Intuitively, we have an overall "average" regression line for all houses across all counties in Minnesota which has slope -0.68 and intercept 1.46.
- That is, the general estimated line for any of the houses in Minnesota is:

$$\widehat{\log(\text{radon}_i)} = 1.46 - 0.68 \times \text{floor}_i$$

- For any house in Minnesota with a basement as the lowest living area, the baseline radon level is $e^{1.46} = 4.31$.
- Then, for any house in Minnesota, having a first floor as the lowest living area, instead of a basement, reduces the radon level by a multiplicative effect of $e^{-0.68} = 0.51$, that is, about a 49% reduction.
- However, if the house is in Dakota county for example, we also need to add on the random intercepts and slopes for that county.

Interpretation of fixed effects

- For Dakota county, we have

```
(ranef(Model1)$countyname)["DAKOTA",]
```

```
##          (Intercept) floorFirst Floor
## DAKOTA  -0.1099052      -0.08786805
```

so that the estimated regression line for Dakota county is actually

$$\widehat{\log(\text{radon}_i)} = (1.46 - 0.11) + (-0.68 - 0.09) \times \text{floor}_i = 1.35 - 0.77 \times \text{floor}_i$$

- Thus, for any house in Dakota county in Minnesota with a basement as the lowest living area, the baseline radon level is actually $e^{1.35} = 3.86$, which is lower than the overall state wide average.

- And for any house in Dakota county in Minnesota, having the first floor be the lowest living area then reduces the radon level by a multiplicative effect of $e^{-0.77} = 0.46$, that is about a 54% reduction, more than the overall state wide effect.

IDS 702

The radon analysis

Again,

```
summary(Model1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log_radon ~ floor + (floor | countyname)
##    Data: Radon
##
## 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
##  countyname (Intercept)      0.1216   0.3487
##             floorFirst Floor 0.1181   0.3436   -0.34
##  Residual                    0.5567   0.7462
## Number of obs: 919, groups:  countyname, 85
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)       1.46277    0.05387  27.155
## floorFirst Floor -0.68110    0.08758  -7.777
##
## Correlation of Fixed Effects:
##             (Intr)
## florFrstFlr -0.381
```

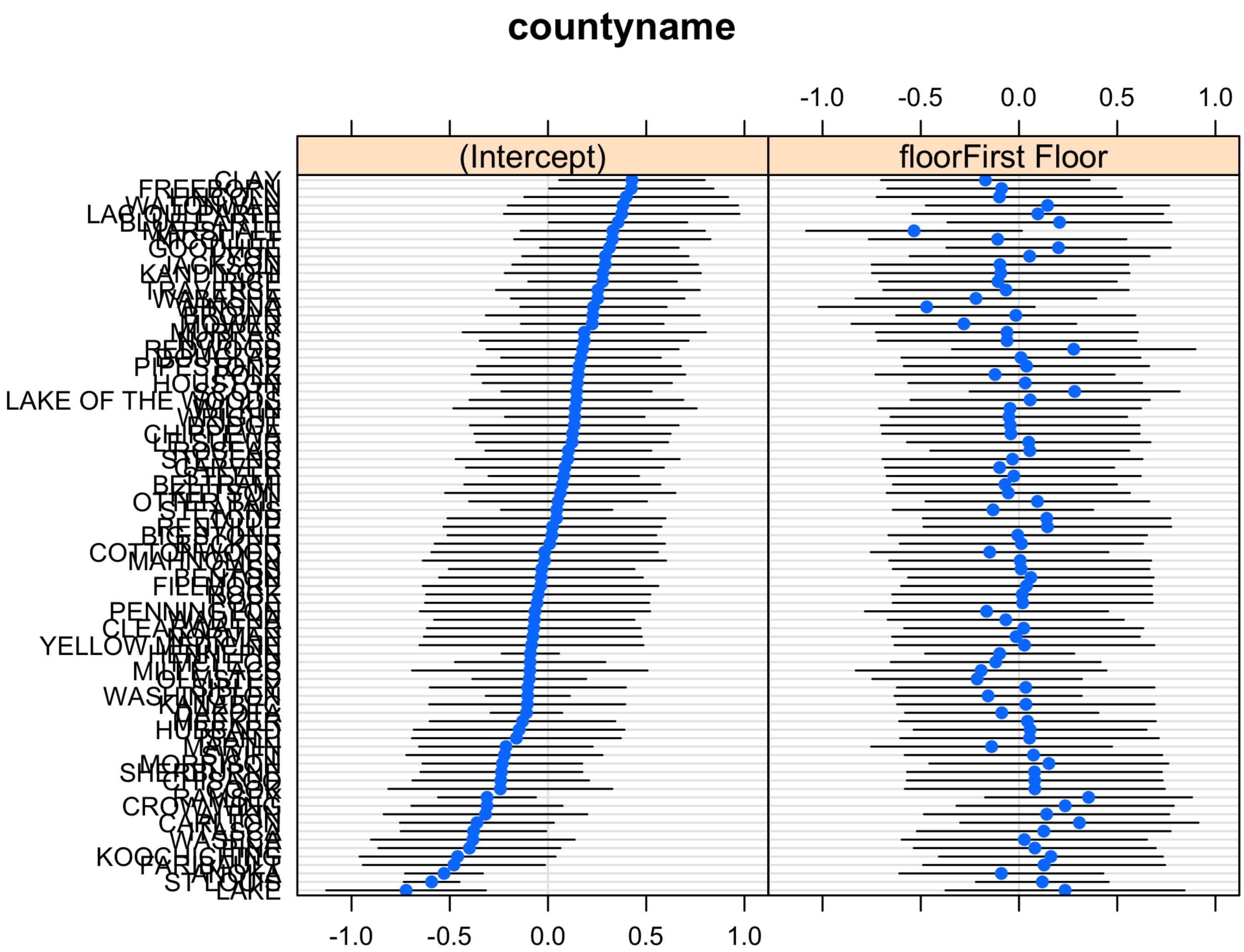
IDS 702

Interpretation of random effects

- The estimated standard error $\hat{\sigma} = 0.75$ describes the within-county or remaining unexplained variation.
- The estimated $\hat{\tau}_0 = 0.35$ describes the across-county variation attributed to the random intercept.
- The estimated $\hat{\tau}_1 = 0.34$ describes the across-county variation attributed to the random slope (the predictor, floor).
- Those two sources of county variation are actually quite similar.
- The estimated correlation between γ_{0j} and γ_{1j} is $\hat{\rho} = -0.34$.
- You can visualize the random effects by typing `dotplot(ranef(Model1, condVar=TRUE))$countyname` in R.
- So many counties! So, you will need to zoom out on your computer.

IDS 702

Interpretation of random effects



IDS 702

What's next?

Move on to the readings for the next module!