36-617: Applied Linear Regression

Random slopes, correlation & centering, sample size

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Announcements

- Quiz 07 available ca. 6pm, due tomorrow 7pm
  - Covers lectures 17, 18, 19 (last week & today)
- HW07 – due Weds 1159pm
- HW08 – out tomorrow I think/hope
- This week
  - Today G&H Ch 13: multiple random effects, sample size
  - Weds G&H Ch 14: multilevel logistic regression models
- Project: I will share a rough schedule later this week
- Tues: Midterm elections in the US
  - If you are eligible to vote, be sure your vote counts!
  - If you are not eligible to vote, remind others to vote!

Outline

- More than one random effect
  - Random slope, random intercept with group-level predictor
  - Multiple random effects and $Cor(\eta_{0j}, \eta_{1j})$
- Correlation and Centering
  - "Fake Data" Example
  - Heights Data Example
- Sample Sizes
- A Note on Notation
- Sections I am skipping:
  - G&H 13.3, G&H 13.6, G&H 13.7, …

More Than One Random Effect

- Last time we looked at

Level 1 $\left\{ \; y_i = \alpha_{0j[i]} + \alpha_1 x_i + \epsilon_i \right.$

Level 2 $\left\{ \; \alpha_{0j} = \beta_0 + \beta_1 u_j + \eta_j \right.$

- But we could also consider, e.g.,

Level 1 $\left\{ \; y_i = \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i \right.$

Level 2 $\left\{ \begin{aligned} \alpha_{0j} &= \beta_{00} + \beta_{01} u_j + \eta_{0j} \\ \alpha_{1j} &= \beta_{10} + \eta_{1j} \end{aligned} \right.$

- Any coefficient at Level 1 that varies across groups will have a model equation at Level 2

From MLM to Variance Components to lmer()

$$
\begin{aligned}
y_i &= \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i \\
\alpha_{0j} &= \beta_{00} + \beta_{01} u_j + \eta_{0j} \\
\alpha_{1j} &= \beta_{10} + \eta_{1j}
\end{aligned}
$$

$$
\begin{aligned}
y_i &= (\beta_{00} + \beta_{01} u_{j[i]} + \eta_{0j[i]}) + (\beta_{10} + \eta_{1j[i]}) x_i + \epsilon_i \\
y_i &= (\beta_{00} + \beta_{01} u_{j[i]} + \beta_{10} x_i) + (\eta_{0j[i]} + \eta_{1j[i]} x_i) + \epsilon_i
\end{aligned}
$$

$$
\begin{aligned}
\texttt{y} &\sim \texttt{1} + \texttt{u} + \texttt{x} + (\texttt{1} + \texttt{x} | \texttt{county}) \\
\texttt{y} &\sim \texttt{u} + \texttt{x} + (\texttt{x} | \texttt{county})
\end{aligned}
$$

The "random slope, random intercept with group predictor model" for the Radon data

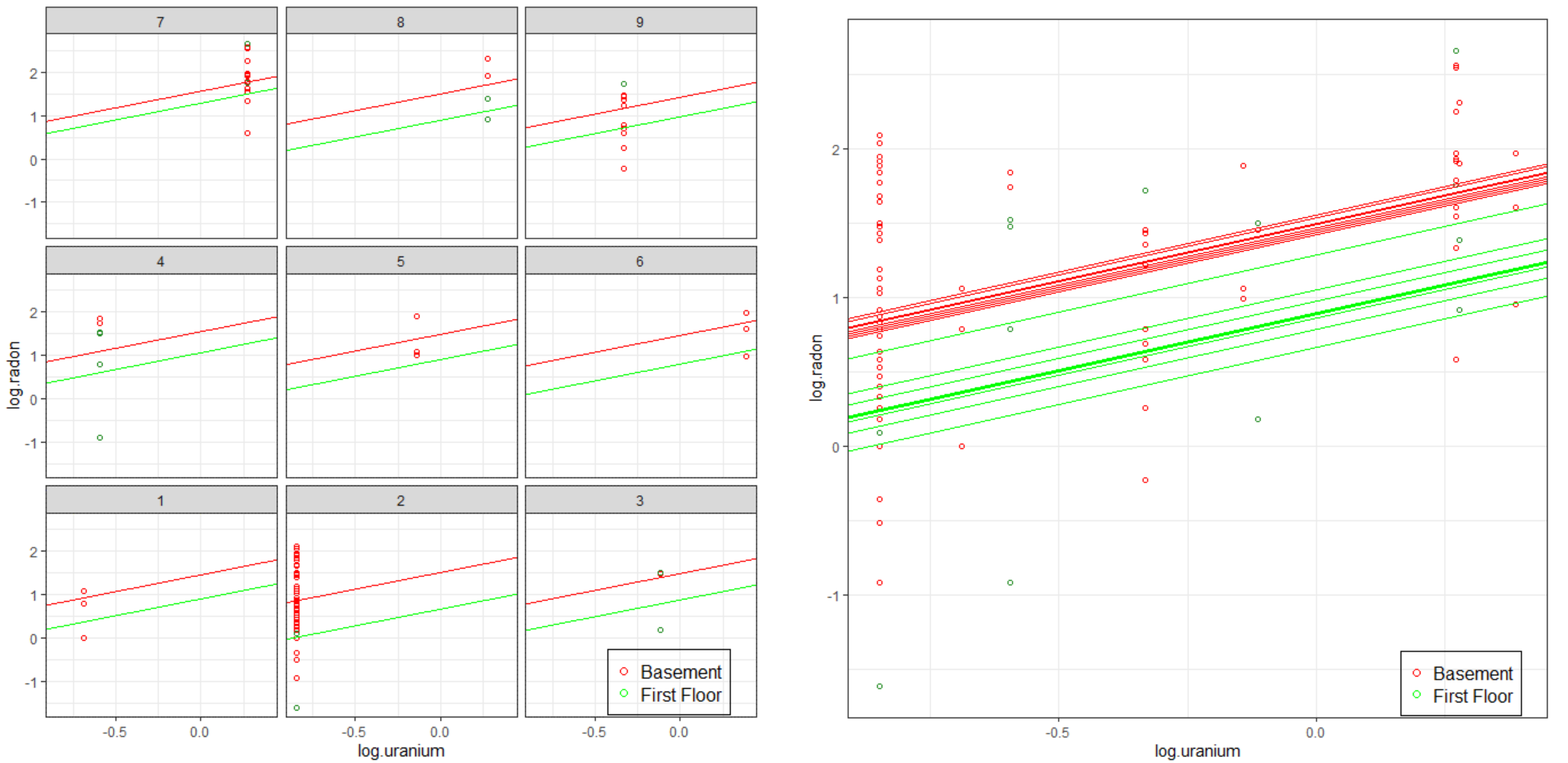
$$
\begin{aligned}
y_i &= \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i \\
\alpha_{0j} &= \beta_{00} + \beta_{01} u_j + \eta_{0j} \\
\alpha_{1j} &= \beta_{10} + \eta_{1j} \\
& \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \\
& \eta_{0j} \overset{iid}{\sim} N(0, \tau_0^2) \\
& \eta_{1j} \overset{iid}{\sim} N(0, \tau_1^2)
\end{aligned}
$$

```
> y <- log.radon
> x <- floor
> u.full <- log.uranium

> M3 <- lmer(y ~
  x + u.full + (1 + x | county) )
```

```
> display(M3)
            coef.est  coef.se
(Intercept)  1.46      0.04
x           -0.64      0.09
u.full       0.77      0.09

Error terms:
 Groups   Name        SD   Corr
 county   (Intcpt) 0.13
          x           0.36  0.21
 Residual             0.75

number of obs: 919, groups:
   county, 85
AIC = 2142.6, DIC = 2106.7
deviance = 2117.7
```


Details in mn-radon.r

Displaying the “random slope, random intercept with group predictor model”



More on Multiple Random Effects

$$
\begin{aligned}
y_i &= \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i & \epsilon_i &\overset{iid}{\sim} N(0, \sigma^2) \\
\alpha_{0j} &= \beta_{00} + \beta_{01} u_j + \eta_{0j} & \eta_{0j} &\overset{iid}{\sim} N(0, \tau_0^2) \\
\alpha_{1j} &= \beta_{10} + \eta_{1j} & \eta_{1j} &\overset{iid}{\sim} N(0, \tau_1^2)
\end{aligned}
$$

- We always model ϵ_i as "independent of everything" because it is the "unexplainable variation"
- η_{0j} and η_{1j} might be dependent on each other!
  - Perhaps underground radon levels are relatively high in some counties (α_{0j} large) but dissipate to a relatively constant level above ground (α_{1j} large too).
  - Suggests η_{0j} and η_{1j} might be correlated.

Multiple Random Effects

- Thus we often do (and lmer() does) model the correlation between random effects, e.g.:

$$\begin{aligned} y_i &= \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \\ \alpha_{0j} &= \beta_{00} + \beta_{01} u_j + \eta_{0j} \\ \alpha_{1j} &= \beta_{10} + \eta_{1j} \end{aligned}$$

$$\begin{pmatrix} \eta_{0j} \\ \eta_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \rho_{01}\tau_0\tau_1 \\ \rho_{01}\tau_1\tau_0 & \tau_1^2 \end{pmatrix} \right)$$

ρ: “rho”, rhymes with “row”

Multiple Random Effects

$$
\begin{aligned}
y_i &= \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i \\
\alpha_{0j} &= \beta_{00} + \beta_{01} u_j + \eta_{0j} \\
\alpha_{1j} &= \beta_{10} + \eta_{1j} \\
& \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \\
& \eta_{0j} \overset{iid}{\sim} N(0, \tau_0^2) \\
& \eta_{1j} \overset{iid}{\sim} N(0, \tau_1^2) \\
& \mathrm{Cor}(\eta_{0j}, \eta_{1j}) = \rho
\end{aligned}
$$

```
> y <- log.radon
> x <- floor
> u.full <- log.uranium

> M3 <- lmer(y ~
  x + u.full + (1 + x | county) )
```

```
> display(M3)
            coef.est  coef.se
(Intercept)  1.46      0.04
x           -0.64      0.09
u.full       0.77      0.09

Error terms:
 Groups   Name        SD   Corr
 county   (Intcpt)    0.13
          x           0.36 0.21
 Residual             0.75

number of obs: 919, groups:
   county, 85
AIC = 2142.6, DIC = 2106.7
deviance = 2117.7
```

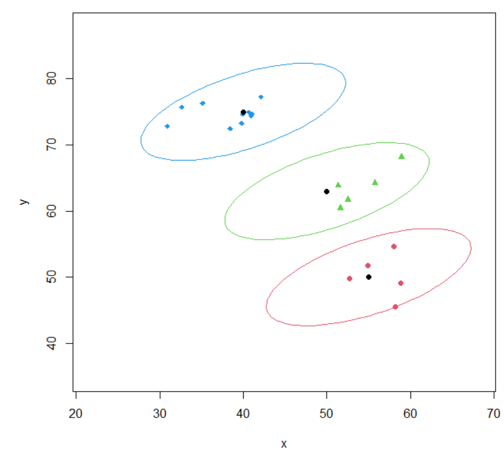
Correlation and Centering

- Forcing $Cor(\eta_{0j}, \eta_{1j}) = 0$ can radically change τ_0^2 and τ_1^2 in the model.
  - It is generally better to allow $Cor(\eta_{0j}, \eta_{1j})$ to be estimated
- Centering x can sometimes reduce the correlation,
  - Then the effect of forcing $Cor(\eta_{0j}, \eta_{1j}) = 0$ is not as great
- There are two ways to center x, with different interpretations
  - Center at Grand Mean (CGM): $x - \bar{x}$
    - Slopes unchanged, intercepts adjusted by overall $\bar{x}$
  - Center Within Clusters (CWC): $x - \bar{x}_{group}$
    - Slopes unchanged, intercepts adjusted by $\bar{x}_{group}$ within groups

More detail in Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological methods*, *12*(2), 121.

A "Fake Data" Example

- 3 groups, indexed by gp in the data set (*j* in math)
  - gp=1: 5 observations
  - gp=2: 5 observations
  - gp=3: 10 observations
- $x_i \sim N\left(\mu^x_{j[i]}, 25\right), y_i \sim N\left(\mu^y_{j[i]}, 9\right), \rho_{xy} = 0.6$
  - $(\mu^x_1, \mu^y_1) = (55, 50)$
  - $(\mu^x_2, \mu^y_2) = (50, 63)$
  - $(\mu^x_3, \mu^y_3) = (40, 75)$



See "centering and correlation.r" for details…

The MLM for the "Fake Data" Example

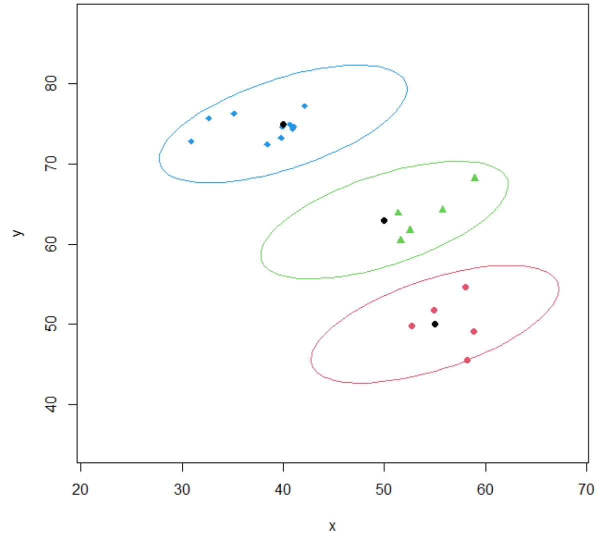
$$
\begin{aligned}
y_i &= \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i \ , \quad \epsilon_i \sim N(0, \sigma^2) \\
\alpha_{0j} &= \beta_{00} + \eta_{0j} \ , \quad \eta_{0j} \sim N(0, \tau_0^2) \\
\alpha_{1j} &= \beta_{10} + \eta_{1j} \ , \quad \eta_{1j} \sim N(0, \tau_1^2) \\
\mathrm{Cor}\left(\eta_{0j}, \eta_{1j}\right) &= \rho
\end{aligned}
$$

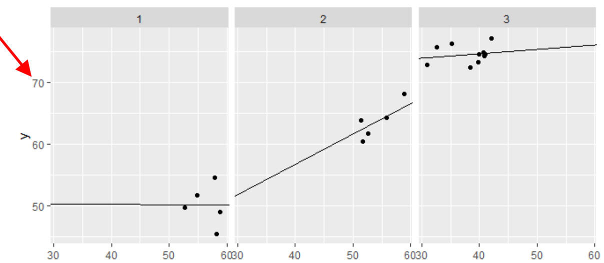
$$
\begin{aligned}
y_i &= (\beta_{00} + \eta_{0j[i]}) + (\beta_{10} + \eta_{1j[i]}) x_i + \epsilon_i \\
y_i &= (\beta_{00} + \beta_{01}) x_i + (\eta_{0j[i]} + \eta_{1j[i]} x_i) + \epsilon_i
\end{aligned}
$$

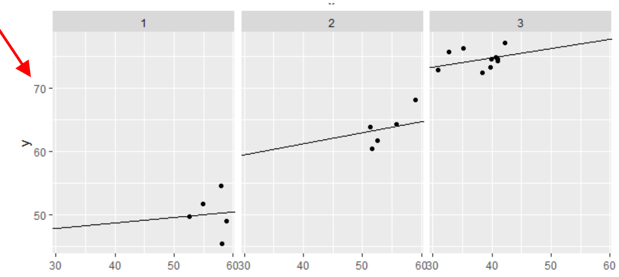
$$
\begin{aligned}
\mathtt{y} &\sim \mathtt{1} + \mathtt{x} + (\mathtt{1} + \mathtt{x}|\mathtt{gp}) \\
\mathtt{y} &\sim \mathtt{x} + (\mathtt{x}|\mathtt{gp}) \quad (\text{estimate } \rho) \\
\mathtt{y} &\sim \mathtt{x} + (\mathtt{x}||\mathtt{gp}) \quad (\text{force } \rho \equiv 0)
\end{aligned}
$$

Uncentered x, ρ estimated or $\rho \equiv 0$

```
> samp <- data.frame(x,y,gp)
> m.raw <-  lmer(y ~ x + (x | gp),
+   data=samp)
Model failed to converge with max|grad| =
0.00353063 (tol = 0.002, component 1)

> m0.raw <- lmer(y ~ x + (x || gp),
+   data=samp)

> round(rbind(
+   cbind("m.raw "= fixef(m.raw),
+         "m0.raw"= fixef(m0.raw)),
+   cbind(extract.sigs(m.raw),
+         extract.sigs(m0.raw))),2)
```







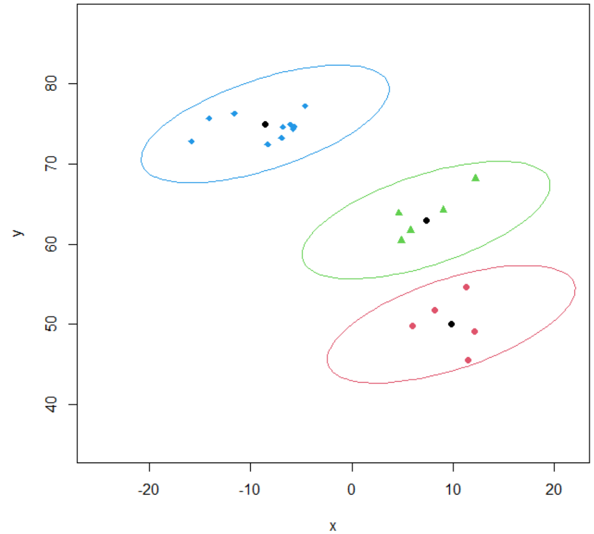
		m.raw	m0.raw
$\hat{\beta}_{00}$	(Intercept)	53.08	56.14
$\hat{\beta}_{10}$	x	0.19	0.14
$\hat{\sigma}$	sigma	2.29	2.41
$\hat{\tau}_0$	tau0	21.80	12.97
$\hat{\tau}_1$	tau1	0.36	0.10
$\hat{\rho}$	cor	-0.81	0.00

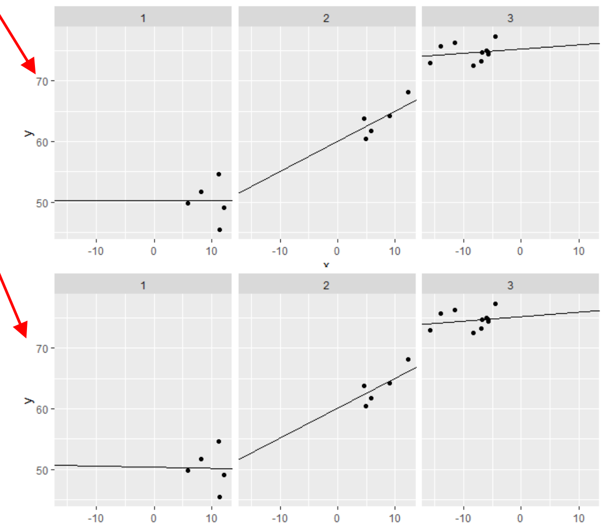
See “centering and correlation.r” for details…

CGM centered x, ρ estimated or $\rho \equiv 0$

```
> x.cgm <- with(samp, x - mean(x))
> m.cgm <-  lmer(y ~ x.cgm + (x.cgm | gp),
+   data=samp)

> m0.cgm <- lmer(y ~ x.cgm + (x.cgm || gp),
+   data=samp)

> round(rbind(
+   cbind("m.cgm" = fixef(m.cgm),
+         "m0.cgm"= fixef(m0.cgm)),
+   cbind(extract.sigs(m.cgm),
+         extract.sigs(m0.cgm))),2)
```

		m.cgm	m0.cgm
$\hat{\beta}_{00}$	(Intercept)	61.81	61.89
$\hat{\beta}_{10}$	x.cgm	0.19	0.18
$\hat{\sigma}$	sigma	2.28	2.28
$\hat{\tau}_0$	tau0	12.82	12.77
$\hat{\tau}_1$	tau1	0.36	0.36
$\hat{\rho}$	cor	-0.07	0.00

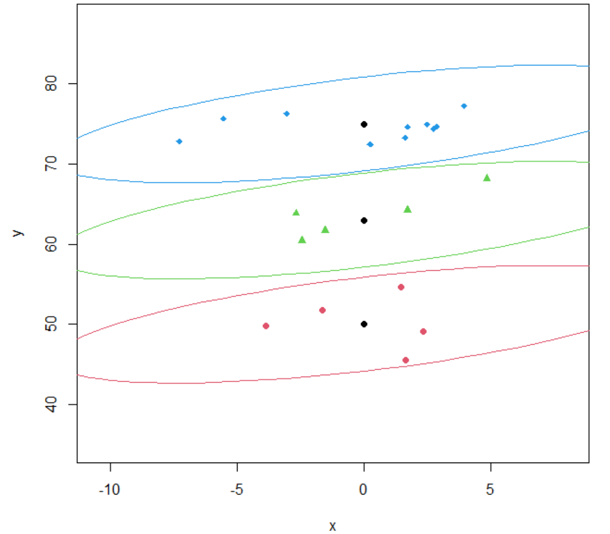


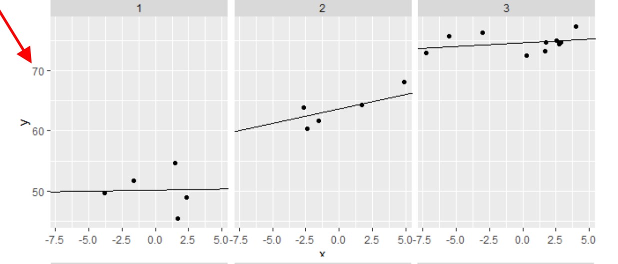


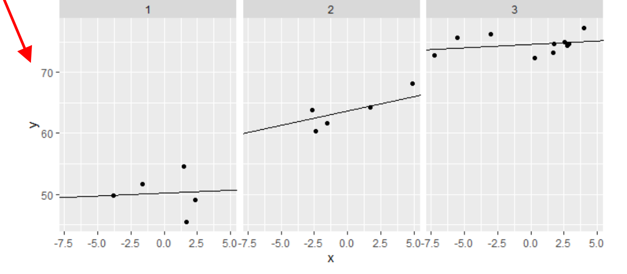
CWC centered x, ρ estimated or $\rho \equiv 0$

```
> x.cwc <- with(samp,
+   unlist(sapply(split(x,gp),
+   function(x) {x - mean(x)})))
> m.cwc <-  lmer(y ~ x.cwc + (x.cwc | gp),
+   data=samp)

> m0.cwc <- lmer(y ~ x.cwc + (x.cwc || gp),
+   data=samp)

> round(rbind(
+   cbind("m.cwc" = fixef(m.cwc),
+         "m0.cwc"= fixef(m0.cwc)),
+   cbind(extract.sigs(m.cwc),
+         extract.sigs(m0.cwc))),2)
```

		m.cwc	m0.cwc
$\hat{\beta}_{00}$	(Intercept)	62.82	62.82
$\hat{\beta}_{10}$	x.cwc	0.21	0.23
$\hat{\sigma}$	sigma	2.30	2.31
$\hat{\tau}_0$	tau0	12.25	12.25
$\hat{\tau}_1$	tau1	0.33	0.31
$\hat{\rho}$	cor	0.17	0.00







Try this with the heights data…

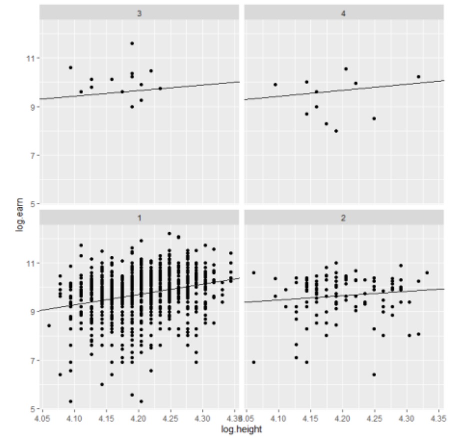
- We’ll fit the random-intercepts, random-slopes model to predict log(earn) from log(height), using the J=4 race categories as the “groups”:

$$
\begin{aligned}
log(earn_i) &= \alpha_{0j[i]} + \alpha_{1j[i]} log(height_i) + \epsilon_i \\
\alpha_{0j} &= \beta_{00} + \eta_{0j} \\
\alpha_{1j} &= \beta_{10} + \eta_{1j}
\end{aligned}
$$

- We’ll again look at
  - Uncentered log(height)
  - CGM centered log(height)
  - CWC centered log(height)

Uncentered log(heights)

```
> lmer.1 <- lmer(log.earn ~ log.height + (log.height|race),data=heights)
boundary (singular) fit: see ?isSingular
> fixef(lmer.1); VarCorr(lmer.1)
(Intercept)  log.height
  -1.688080    2.705316
 Groups   Name        Std.Dev. Corr
 race     (Intercept) 7.1933
          log.height  1.7220   -1.000
 Residual             0.8912
```




```
> params <- data.frame(sort(unique(heights$race)),coef(lmer.1)$race[,1:2])
> names(params) <- c("race","int1","slo1") ## ; round(params,2)
> round(data.frame(race=params$race,exp.earn1=exp(params$int1),
+                  exp.height1=exp(0), slo1=params$slo1),2)
  race exp.earn1 exp.height1 slo1
1    1      0.00           1 4.25
2    2      9.70           1 1.76
3    3      0.97           1 2.31
4    4      0.43           1 2.50
```

% increase in earnings per 1% increase in height, in group j

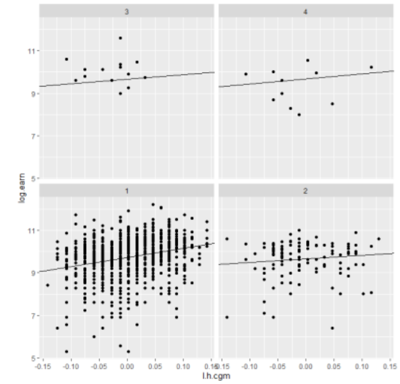
(see log transform handout on Canvas)

Expected earnings ($)… …for a person in group j of this height (inches)

Details in heights-mlm.r

CGM centered log(heights)

```
> lmer.2 <- lmer(log.earn ~ l.h.cgm + (l.h.cgm|race),data=heights)
> fixef(lmer.2); VarCorr(lmer.2)
(Intercept)     l.h.cgm
   9.678727    2.618731
 Groups   Name        Std.Dev. Corr
 race     (Intercept) 0.044119
          l.h.cgm     1.878514 1.000
 Residual             0.891133
```



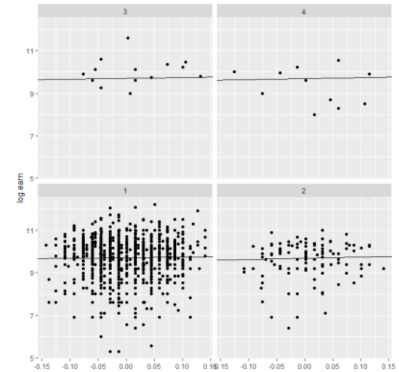
```
> params <- data.frame(sort(unique(heights$race)),coef(lmer.2)$race[,1:2])
> names(params) <- c("race","int2","slo2") ## ; round(params,2)
> round(data.frame(race=params$race,exp.earn2=exp(params$int2),
+    exp.height2=exp(mean(heights$log.height)),
+    slo2=params$slo2),2)
  race exp.earn2 exp.height2 slo2
1    1  16603.69       66.82 4.26
2    2  15620.08       66.82 1.66
3    3  15794.41       66.82 2.14
4    4  15895.68       66.82 2.41
```

% increase in earnings per 1% increase in height, in group j

Expected earnings ($)… …for a person in group j of this height (inches)

CWC centered log(heights)

```
> lmer.3 <- lmer(log.earn ~ l.h.cwc + (l.h.cwc|race),data=heights)
> fixef(lmer.3); VarCorr(lmer.3)
(Intercept)      l.h.cwc
  9.6980503   0.4265615
 Groups    Name         Std.Dev. Corr
 race      (Intercept) 0.045748
           l.h.cwc      0.314637 -1.000
 Residual               0.921047
> params <- data.frame(sort(unique(heights$race)),coef(lmer.3)$race[,1:2])
> names(params) <- c("race","int3","slo3") ## ; round(params,2)
> round(data.frame(race=params$race,exp.earn3=exp(params$int3),
+     exp.height3=exp(sapply(split(heights$log.height,heights$race),mean)),
+     slo3=params$slo3),2)
  race exp.earn3 exp.height3 slo3
1    1  16594.63       66.89 0.30
2    2  15931.58       66.51 0.58
3    3  16430.73       64.87 0.37
4    4  16194.05       65.80 0.47
```



% increase in earnings per 1% increase in height, in group j

Expected earnings ($)…

…for a person in group j of this height (inches)

Sample Sizes (for “nested” models)

- Bare minima:
  - To estimate τ^2’s, have to have $J \geq 2$ groups
  - To estimate σ^2, some groups have to have $n_j \geq 2$ observations
  - When n_j and J are both small, hard to estimate τ^2’s, $\sigma^2, \eta's$ & $\alpha's$
  - When either n_j’s or J are bigger than about 10, estimation is easier
- If J is small, not much advantage over totally unpooled model
- The larger J is
  - The easier it is to interpret the MLM
  - The MLM is more efficient to estimate than the totally unpooled model
- As n_j’s and J grow, it make sense to build more complex models (more Level 1 and Level 2 covariates)
  - My R.O.T: roughly 10 “observations” “available” for each $\beta, \eta, \sigma \; and \; \tau$

Sample Size Recommendations

Table 1 Minimum number of clusters recommended for accurate estimates by effect (assuming likelihood estimation methods and no small sample size adjustments)

Effect of interest	Continuous outcomes	Binary outcomes[a]
Level-1 fixed-effect point estimates	5	10
Level-2 fixed-effect point estimates	15	30
Fixed-effect standard errors	30	50
Level-1 variance estimate	10	30
Level-2 variance estimate	10/30 (REML/FML)*	10/50 (REML/FML)*
Level-2 variance standard error	50	100

[a] In addition to a minimum number of clusters, a cluster size greater than 5 is also recommended for binary outcomes

Source: McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295-314.

* REML and FML are estimation methods that we will talk about next week.

A Note on Notation

- We have been following, and will follow, Gelman & Hill's i & $j[i]$ notation for MLMs and Variance Components Models

$$
\begin{aligned}
y_i &= \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i \\
i &= 1, \ldots, n \quad \text{observations} \\
j &= 1, \ldots, J \quad \text{groups with } n_j \text{ observations each} \\
j[i] &= \text{the group } j \text{ that } i \text{ is in}
\end{aligned}
$$

- Most books and journal articles use an ij notation[1] instead

$$
\begin{aligned}
y_{ij} &= \alpha_{0j} + \alpha_{1j} x_{ij} + \epsilon_{ij} \\
i &= 1, \ldots, n_j \quad \text{observations in group } j \\
j &= 1, \ldots, J \quad \text{groups}
\end{aligned}
$$

- The ij notation is less cluttered, but can become ambiguous
- The i & $j[i]$ notation never is ambiguous

[1]Even more confusingly, some authors interchange the roles of i and j, so that i refers to groups and j refers to individuals!

Summary

- More than one random effect
  - Random slope, random intercept with group-level predictor
  - Multiple random effects and $Cor(\eta_{0j}, \eta_{1j})$
- Correlation and Centering
  - "Fake Data" Example
  - Heights Data Example
- Sample Sizes
- A Note on Notation
- Sections I am skipping:
  - G&H 13.3, G&H 13.6, G&H 13.7, …