36-463/663: Multilevel & Hierarchical Models

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

Outline

- Announcements
- Using R: London Schools Example
- ggplot2 Graphs
  - Multilevel Models
  - The London Data
- Start looking at Gelman & Hill, Ch's 1-2
  - Not everything can or will be covered in class!

Announcements

- HW01: Due next Tue Sep 6, on Blackboard.
- Office Hours:
  - Nick (Loc posted on www.stat.cmu.edu/~brian/463-663):
    - Mon 5-6pm
  - BJ (BH132E):
    - Tue & Thu, 3-4pm
    - Occasional Tuesdays after class
- We are both also able to meet by appointment; so if the above hours don't work, please email for an appointment as needed.

Using R: London Schools Example

- Goldstein et al (1993) try to rank schools in London to distinguish the "best" from the "worst" of them.
- london-schools.r

ggplot2 Graphs

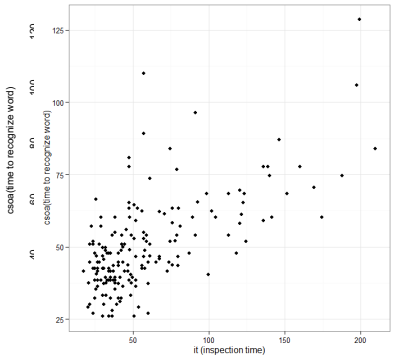
- Do only once:
  - Packages -> Install Packages
  - Select a CRAN mirror
  - Select “ggplot2” from the list that appears, installs quickly
- Do once to start each session using lattice graphs:
  - library(ggplot2)
- Help:
  - library(help=ggplot2)
  - help(qplot)
  - help.search(“facet”)
  - …
- To get rid of a library:
  - detach(package:ggplot2)

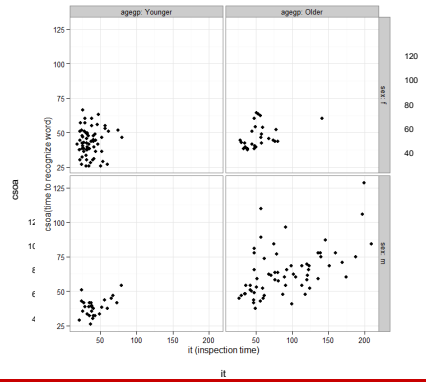
ggplot2 Graphs - Resources

- 02a-intro-ggplot-graphs.r
- The Web! [e.g. Google ‘using ggplot’]
  - http://www.cookbook-r.com/Graphs/
  - http://sape.inf.usi.ch/quick-reference/ggplot2
- Books:
  - Chang, W. (2013). *The R Graphics Cookbook.* Sebastopol, CA: O'Reilly Media.
  - Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* NY: Springer.

ggplot2 Graphs

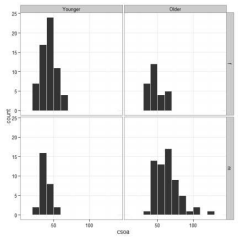
```
load("usingR.RData")

names(tinting)
[1] "case"   "id"     "age"    "sex" "tint"   "target" "it"
       "csoa" "agegp"

str(tinting)
'data.frame':   182 obs. of  9 variables:
 $ case  : num  1 1 1 1 1 1 1 2 2 2 ...
 $ id    : num  1 1 1 1 1 1 1 2 2 2 ...
 $ age   : num  22.4 22.4 22.4 22.4 22.4 22.4 22.4 26.8 26.8
       26.8 ...
 $ sex   : Factor w/ 2 levels "f","m": 1 1 1 1 1 1 1 1 1 1 ...
 $ tint  : Ord.factor w/ 3 levels "no"<"lo"<"hi": 1 2 3 1 2 3 1
       1 2 3 ...
 $ target: Factor w/ 2 levels "locon","hicon": 2 2 2 1 1 1 2 2
       2 2 ...
 $ it    : num  26 32.2 27 17.7 20.8 ...
 $ csoa  : num  46.8 37.4 42.6 41.6 37.4 ...
 $ agegp : Factor w/ 2 levels "Younger","Older": 1 1 1 1 1 1
       1 1 1 1 ...

qplot(it,csoa,data=tinting) +
 xlab("it (inspection time)") + ylab("csoa(time to recognize
       word)")

.Last.value  + facet_grid(sex ~ agegp, labeller=label_both)
```



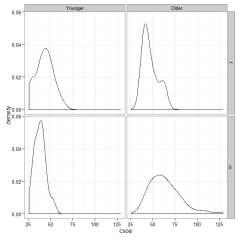


ggplot2 Graphs

gh <- ggplot(tinting,aes(x=csoa))

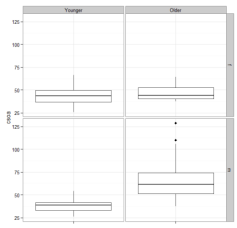


gh + geom_histogram(color="white", binwidth=10) + facet_grid(sex ~ agegp)

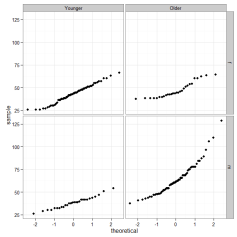


gh + geom_density() + facet_grid(sex ~ agegp)

gv <- ggplot(tinting,aes(x="",y=csoa))



gv + geom_boxplot() + facet_grid(sex ~ agegp) + xlab(" ")



ggplot(tinting,aes(sample=csoa)) + stat_qq() + facet_grid(sex ~ agegp)

More on Multilevel Models…

- Useful when information comes to us in clumps of individuals that are more like each other within a clump than across clumps
  - Classrooms within schools or schools within a city
  - States or geographic areas within a nation
  - Election precincts within a larger election
- Useful when a different linear regression should be fitted within each clump, but there is not enough information to separately estimate all clumps
  - Deducing state opinions from a national opinion survey
  - Fitting separate regressions to rank schools in London – some schools are represented by only 1 or 2 students!

More on Multilevel Models…

- Traditional linear regression can either
  - Ignore the clumps completely and fit a single model to all the data
  - Treat each clump completely separately but fail to share information across clumps when some clumps "need help"
  - Both of these are examples of "Fixed Effects"
- Multilevel models allow
  - treating clumps separately, ***and***
  - sharing information across clumps to make better estimates
  - These are examples of "Random Effects"
- Most MLM's have both fixed and random effects – "Mixed Effects" models
- Gelman & Hill do not like the "Fixed/Random" nomenclature, so they try to avoid it.

Back to London Schools…

- Earlier we looked briefly at three linear models

```
mean.lm <- lm(Y ~ school - 1, data=school.frame)
adj.1.lm <- lm(Y ~ school + LRT - 1, data=school.frame)
adj.2.lm <- lm(Y ~ school*LRT - 1, data=school.frame)
```

- And we compared them, seeing that 'adj.2.lm' fit better than 'adj.1.lm' and 'adj.1.lm' fit better than 'mean.lm'

```
anova(mean.lm,adj.1.lm,adj.2.lm)
Analysis of Variance Table

Model 1: Y ~ school - 1
Model 2: Y ~ school + LRT - 1
Model 3: Y ~ school * LRT - 1
  Res.Df    RSS Df Sum of Sq        F    Pr(>F)
1   1940 1801.8
2   1939 1218.9  1    582.91 940.8659 < 2.2e-16 ***
3   1906 1180.9 33     38.03   1.8602  0.002189 **
```

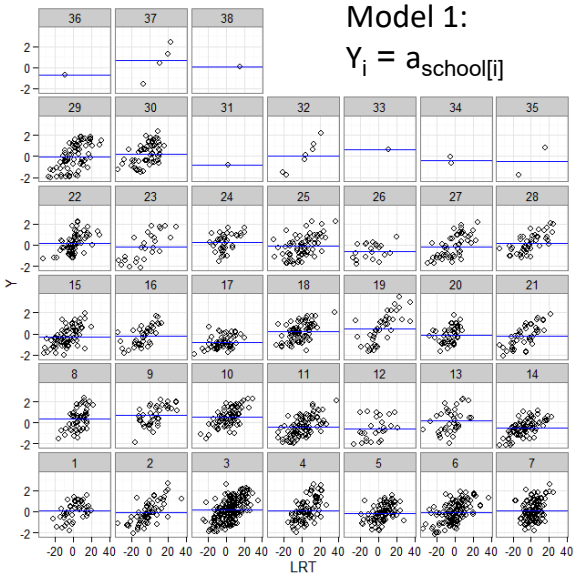

Back to London Schools…

- We won't take time to understand these models deeply now (later in the course!)
- But I wanted to show you, graphically, how these (and similar models) are representing the relationship between
  - Y (end of year score)
  - LRT (beginning of year score)

  in the data
- Lattice graphs are very useful for this!
  - Code in handout, posted on-line at ~brian/463/week01
  - Takes some fussing and trial-and-error, but results interesting!

London Schools: Ignore LRT and only look at mean(y) in each school

```
facet_data <-
split(school.frame,
  school.frame$school)

g <- ggplot(school.frame,
  aes(x=LRT,y=Y)) +
  facet_wrap( ~ school,
  as.table=F) +
  geom_point(pch=1)

for (i in 1:38)
 g <- g +
  geom_hline(data=
  facet_data[[i]],
  aes(yintercept=mean(Y)),
  color="blue")

plot(g)
```



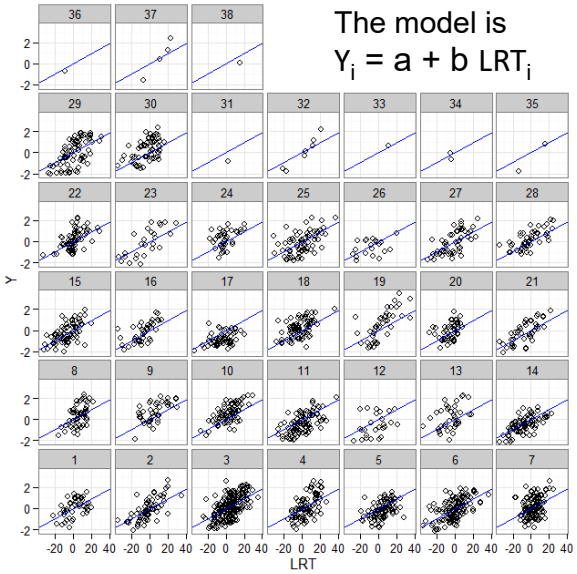
Model 1:

$Y_i = a_{school[i]}$

We would rank schools by mean(y) in this case. This ignores the status of students at the beginning of the school year.

London Schools: Ignore Schools and fit a single linear regression Y ~ LRT

```
b <- coef(lm(Y~LRT,
  data=school.frame))

ggplot(school.frame,
  aes(x=LRT,y=Y)) +
  facet_wrap( ~ school,
  as.table=F) +
  geom_point(pch=1) +
  geom_abline(intercept=
  b[1],slope=b[2],
  color="blue")
```



The model is

$Y_i = a + b\ LRT_i$

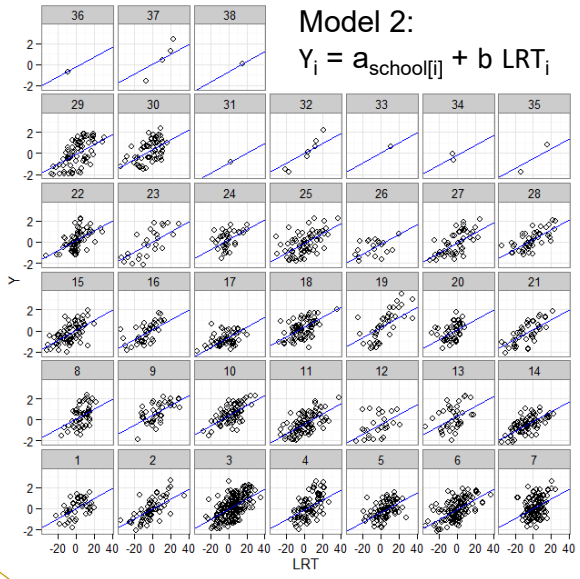
This is not one of our three earlier models, but it is sort of a pre-cursor to Model 2.

We really don't have anything to rank schools with here…

Everyone has the same slope and the same intercept.

London Schools: Use same slope on LRT for all schools, different intercepts

```
facet_data <-
  split(school.frame,
  school.frame$school)

coef <- lm(Y ~ school +
  LRT - 1, data =
  school.frame)$coef

g <- ggplot(school.frame,
    aes(x=LRT,y=Y)) +
    facet_wrap( ~ school,
    as.table=F) +
  geom_point(pch=1)

for (i in 1:38) {
  g <- g + geom_abline(
  data=facet_data[[i]],
  intercept=coef[i],
  slope=coef[39],
  color="blue")
}

plot(g)
```



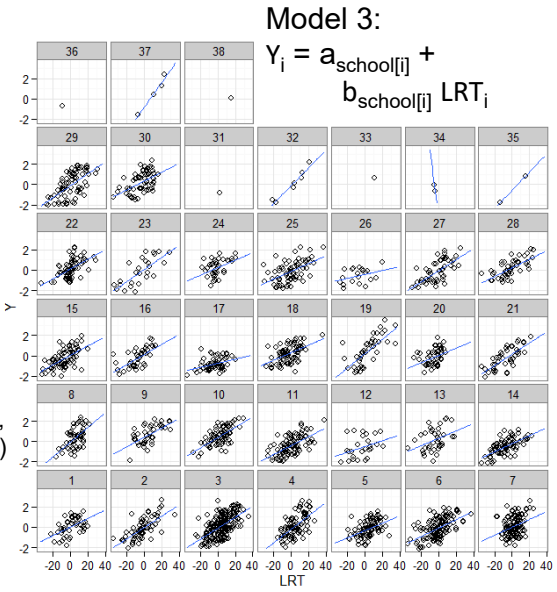
With this model, we could rank schools based on their intercepts.

However, the model clearly fits some schools better than others!

Note: this code does not work in the current version of ggplot2, but it worked in previous versions….

London Schools: Different slope and intercept for each school

```
yrange <-
range(school.frame$Y)

ggplot(school.frame,
  aes(x=LRT,y=Y)) +
  facet_wrap( ~ school,
  as.table=F) +
  geom_point(pch=1) +
  geom_smooth(method=
  "lm", formula = y ~ x, se=F,
  fullrange=T) + ylim(yrange)
```



Now, we let slopes and intercepts vary from school to school, to get the best fit. We would still like to rank based on interecepts.

However some schools have crazy regressions or cannot be fitted (too small a sample in that school!)

This is a problem with fixed effects models, and it is something MLM's are good at fixing!

Summary

- Announcements
  - HW01: Due next Tue Sep 6, on Blackboard.
  - **Office Hours** (Nick M 5-6; BJ TTh 3-4; or by appt)
- Using R: London Schools Example
- ggplot2 Graphs
- For next week: Gelman & Hill, Ch's 1-2
  - Not everything can or will be covered in class!