W271 Live Session 13: Mixed models

Devesh Tiwari

8/7/2017

Main topics covered in Week 14 (Async Unit 13)

```
- Linear mixed-effect model
- The notion of fixed and random effects in the context of linear mixed effect model
- The independence assumption
-  Modeling random intercepts, slopes, and both random intercepts and slopes Mathematical formulation
```

Readings:

BMBW Douglas Bates, Martin Machler, Benjamin Bolker, and Steve Walker. *Fitting Linear Mixed Effect Models Using lme4*

Agenda:

1. Review of terminonlogy and concepts
2. Group R - demo

Review of terminology and concepts

1. Panel data: Fixed and random effects review.

Panel data has multiple observations (J) for cross-sectional units (I). Within these data, we are interested in the relationship between a dependent, or response, variable, and an independent variable of interest.

$$y_{i,j} = \alpha_0 + \beta_1 * x_{i,j} + \epsilon_{i,j}$$

QUESTIONS: (1) What challenges do we face if we try to implement the above model? What is a fixed effects estimator in this context? What is a random effects model in this context?

2. Multi-level data structures

Suppose that you were interested in understanding the 2016 Presidential election better. In particular, you want to know if poorer counties in the US tended to vote for the Democratic Party or not. You compile a dataset that has each counties' average income and the share of the county vote that were for the Democratic Party. You want to estimate the following:

$$voteshare_{i,s} = \alpha_0 + \beta_1 income_{i,s} + \epsilon_{i,s}$$

As a student of American politics, you suspect that state level charactersistics have an effect on county level vote-share, which is why we included subscript s. Therefore, you are dealing with a multi-level data set. In the OLS framework, the best we could would be do include a dummy variable for each state.

QUESTION: Social scientists often call this type of regression a "fixed effects" regression. Even though this is not a panel dataset, why do you think this is the case? What does the inclusion of state-level dummy variables do to the model above?

It is useful, though, to think about the ways in which a county's state effects it's vote-share:

- Some states might have a history of supporting one party over another. So we can think of each state as having a separate mean for vote-share.
- The relationship between income and Democratic vote-share might differ across states. So we can think of each state as having a separate slope coefficient for the income variable.
- Because we are dealing with states now, it is likely the case that there is some state-level errors that are unaccounted for in the model.
- Because each county belongs to a given state, it is more than likely the case that error terms within each state are correlated.

OLS is not well suited to deal with these issues, so instead we turn to linear mixed models as follows:

$$voteshare_{i,s} = \alpha_s + \beta_1 income_{i,s} + \epsilon_{i,s}$$

where

$$\alpha_s \sim N(\mu_\alpha, \sigma_\alpha)$$

Now, we are saying that each state in the data-set gets it's own intercept AND that those values are drawn from a random variable itself! In this setup, we call income a fixed effect (because it's effect is constant across states) and we would call α a random effect because it varies across states (or groups).

We can further estimate random intercept models, where each state gets it's own intercept term, and we can estimate random slope models, where each state gets it's own beta coefficient denoting the relationship between income and vote-share. We can also include group level (in this case state level) parameters into the model if we wanted to, and we could include multiple group level variables.

Group Discussion: Sleep study data 1

1. Briefly explore the data. What do you notice about both plots?
2. Given the heterogeneity across subjects, what is a better measure of the average reaction time, the global mean or subject specific mean?

```r
rm(list = c(ls()))
library(lme4)
```

```
## Loading required package: Matrix
```

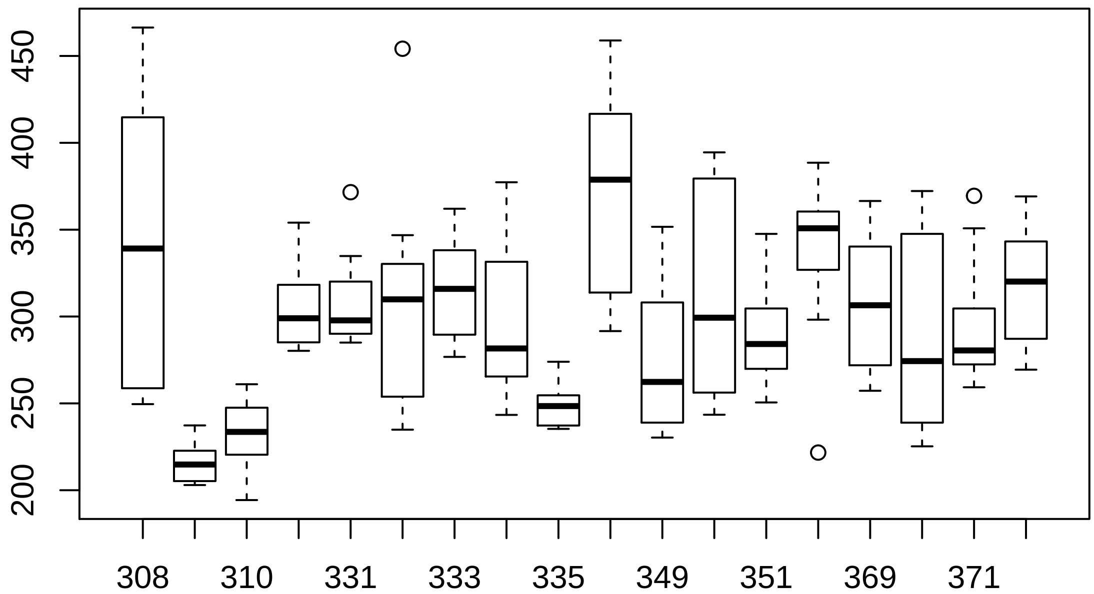
```r
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

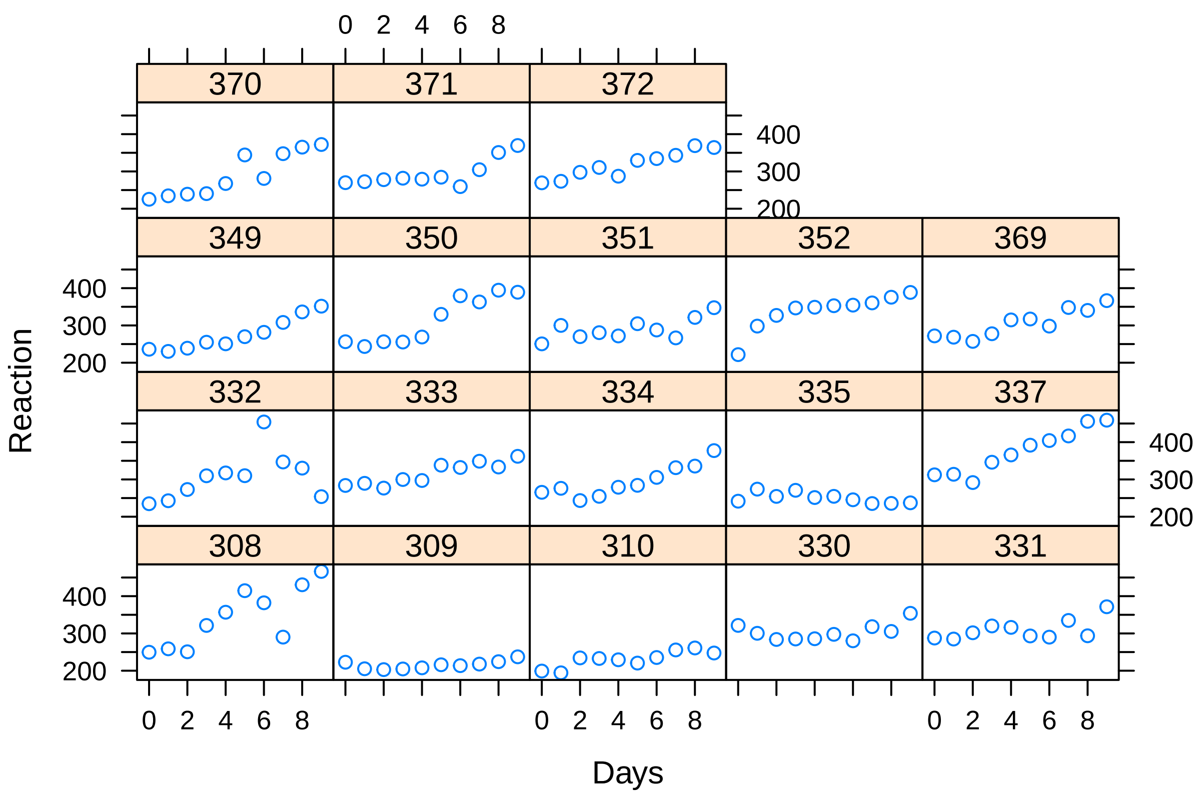
```r
library(lattice)
library(arm)
```

```
## Loading required package: MASS

##
## arm (Version 1.9-3, built: 2016-11-21)

## Working directory is /Users/DKT/Documents/Projects/MIDS/Summer 2017/live_sessions/week13
```

```r
data("sleepstudy")
boxplot(sleepstudy$Reaction ~ sleepstudy$Subject)
```




```r
xyplot(Reaction ~ Days | Subject, data = sleepstudy)
```




```r
# Pause for question

mean(sleepstudy$Reaction) #Global mean
```

```
## [1] 298.5079
```

```r
subjectMeans <- aggregate(sleepstudy$Reaction, by = list(sleepstudy$Subject), mean)
subjectMeans$deviation_from_mean <- subjectMeans$x - mean(sleepstudy$Reaction)
subjectMeans
```

```
##    Group.1        x deviation_from_mean
## 1      308 342.1338           43.625938
## 2      309 215.2330          -83.274912
## 3      310 231.0013          -67.506622
## 4      330 303.2214            4.713528
## 5      331 309.4361           10.928158
## 6      332 307.3021            8.794178
## 7      333 316.1583           17.650418
## 8      334 295.3021           -3.205842
## 9      335 250.0700          -48.437852
## 10     337 375.7210           77.213118
## 11     349 275.8345          -22.673422
## 12     350 313.6027           15.094788
## 13     351 290.0978           -8.410142
## 14     352 337.4215           38.913648
## 15     369 306.0346            7.526748
## 16     370 291.7018           -6.806122
## 17     371 294.9840           -3.523852
## 18     372 317.8861           19.378238
```

```r
s.mean <- lmer(Reaction ~ 1 + (1 | Subject), data = sleepstudy)
summary(s.mean)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ 1 + (1 | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1904.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4983 -0.5501 -0.1476  0.5123  3.3446
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 1278     35.75
##  Residual             1959     44.26
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   298.51       9.05   32.98
```

```r
fixef(s.mean) ## Corresponds to the global mean above
```

```
## (Intercept)
##    298.5079
```

```r
ranef(s.mean) ## Corresponds to the subject level impact on reaction
```

```
## $Subject
##     (Intercept)
```

```
## 308   37.829172
## 309  -72.209815
## 310  -58.536725
## 330    4.087221
## 331    9.476087
## 332    7.625658
## 333   15.305131
## 334   -2.779868
## 335  -42.001705
## 337   66.953478
## 349  -19.660706
## 350   13.089079
## 351   -7.292650
## 352   33.743024
## 369    6.526637
## 370   -5.901763
## 371   -3.055622
## 372   16.803368
```

```
coef(s.mean)$Subject  ## Subject level means. Note that they are slightly different!
```

```
##     (Intercept)
## 308    336.3371
## 309    226.2981
## 310    239.9712
## 330    302.5951
## 331    307.9840
## 332    306.1335
## 333    313.8130
## 334    295.7280
## 335    256.5062
## 337    365.4614
## 349    278.8472
## 350    311.5970
## 351    291.2152
## 352    332.2509
## 369    305.0345
## 370    292.6061
## 371    295.4523
## 372    315.3113
```

Group Discussion 2: Mixed modeling with the sleep study data

1. Does sleep deprivation correspond to higher reaction times?
2. What is the difference between lm.2 and model.random_intercept?

```
lm.1 <- lm(Reaction ~ Days, data = sleepstudy)
lm.2 <- lm(Reaction ~  Days + as.factor(Subject), data = sleepstudy)
stargazer(lm.1, lm.2, type = "text", summary = FALSE)
```

```
##
## ====================================================================
##                                      Dependent variable:
```

```
##                         ----------------------------------------------
##                                            Reaction
##                                   (1)                     (2)
## ----------------------------------------------------------------------
## Days                           10.467***               10.467***
##                                 (1.238)                 (0.804)
##
## as.factor(Subject)309                                -126.901***
##                                                        (13.860)
##
## as.factor(Subject)310                                -111.133***
##                                                        (13.860)
##
## as.factor(Subject)330                                 -38.912***
##                                                        (13.860)
##
## as.factor(Subject)331                                 -32.698**
##                                                        (13.860)
##
## as.factor(Subject)332                                 -34.832**
##                                                        (13.860)
##
## as.factor(Subject)333                                  -25.976*
##                                                        (13.860)
##
## as.factor(Subject)334                                 -46.832***
##                                                        (13.860)
##
## as.factor(Subject)335                                 -92.064***
##                                                        (13.860)
##
## as.factor(Subject)337                                  33.587**
##                                                        (13.860)
##
## as.factor(Subject)349                                 -66.299***
##                                                        (13.860)
##
## as.factor(Subject)350                                 -28.531**
##                                                        (13.860)
##
## as.factor(Subject)351                                 -52.036***
##                                                        (13.860)
##
## as.factor(Subject)352                                   -4.712
##                                                        (13.860)
##
## as.factor(Subject)369                                 -36.099**
##                                                        (13.860)
##
## as.factor(Subject)370                                 -50.432***
##                                                        (13.860)
##
## as.factor(Subject)371                                 -47.150***
##                                                        (13.860)
```

```
##
## as.factor(Subject)372                                      -24.248*
##                                                            (13.860)
##
## Constant                     251.405***                  295.031***
##                               (6.610)                     (10.447)
##
## ----------------------------------------------------------------------
## Observations                    180                         180
## R2                             0.286                       0.728
## Adjusted R2                    0.282                       0.697
## Residual Std. Error      47.715 (df = 178)        30.991 (df = 161)
## F Statistic           71.464*** (df = 1; 178) 23.908*** (df = 18; 161)
## ======================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

```
model.random_intercept <- lmer(Reaction ~ Days + (1 | Subject), data = sleepstudy)
summary(model.random_intercept)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (1 | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1786.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2257 -0.5529  0.0109  0.5188  4.2506
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 1378.2   37.12
##  Residual              960.5   30.99
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 251.4051     9.7467   25.79
## Days         10.4673     0.8042   13.02
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.371
```

```
fixef(model.random_intercept) # Impact that is consistent across groups
```

```
## (Intercept)        Days
##   251.40510    10.46729
```

```
ranef(model.random_intercept) # varies across gruops
```

```
## $Subject
##     (Intercept)
## 308   40.783710
## 309  -77.849554
## 310  -63.108567
```

```
## 330    4.406442
## 331   10.216189
## 332    8.221238
## 333   16.500494
## 334   -2.996981
## 335  -45.282127
## 337   72.182686
## 349  -21.196249
## 350   14.111363
## 351   -7.862221
## 352   36.378425
## 369    7.036381
## 370   -6.362703
## 371   -3.294273
## 372   18.115747
```

```
coef(model.random_intercept)  # These are the coefficients for each subject. Note that the only thing t
```

```
## $Subject
##     (Intercept)     Days
## 308    292.1888 10.46729
## 309    173.5556 10.46729
## 310    188.2965 10.46729
## 330    255.8115 10.46729
## 331    261.6213 10.46729
## 332    259.6263 10.46729
## 333    267.9056 10.46729
## 334    248.4081 10.46729
## 335    206.1230 10.46729
## 337    323.5878 10.46729
## 349    230.2089 10.46729
## 350    265.5165 10.46729
## 351    243.5429 10.46729
## 352    287.7835 10.46729
## 369    258.4415 10.46729
## 370    245.0424 10.46729
## 371    248.1108 10.46729
## 372    269.5209 10.46729
##
## attr(,"class")
## [1] "coef.mer"
```

```
                              # is the intercept, which is what we wanted!

# Question: Once we have incorporated subject level effects, is Days still "statistically significant?
s.mean <- lmer(Reaction ~ 1 + (1 | Subject), data = sleepstudy, REML = FALSE)
model.random_intercept <- lmer(Reaction ~ Days + (1 | Subject), data = sleepstudy, REML = FALSE)
anova(s.mean, model.random_intercept)
```

```
## Data: sleepstudy
## Models:
## s.mean: Reaction ~ 1 + (1 | Subject)
## model.random_intercept: Reaction ~ Days + (1 | Subject)
##                        Df    AIC    BIC  logLik deviance  Chisq Chi Df
## s.mean                  3 1916.5 1926.1 -955.27   1910.5
```

```
## model.random_intercept  4 1802.1 1814.8 -897.04   1794.1 116.46      1
##                          Pr(>Chisq)
## s.mean
## model.random_intercept  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Group Discussion 3: Random - slope model

1. What does the random slope model tell you?
2. How can you tell if you actually "need" the random slopes?

```r
model.random_slope <- lmer(Reaction ~ Days + (1 + Days|Subject), data = sleepstudy)
fixef(model.random_slope)
```

```
## (Intercept)        Days
##   251.40510    10.46729
```

```r
ranef(model.random_slope) # Note here that both the intercept and Days vary. Which is by design
```

```
## $Subject
##     (Intercept)        Days
## 308   2.2585654   9.1989719
## 309 -40.3985770  -8.6197032
## 310 -38.9602459  -5.4488799
## 330  23.6904985  -4.8143313
## 331  22.2602027  -3.0698946
## 332   9.0395259  -0.2721707
## 333  16.8404312  -0.2236244
## 334  -7.2325792   1.0745761
## 335  -0.3336959 -10.7521591
## 337  34.8903509   8.6282839
## 349 -25.2101104   1.1734143
## 350 -13.0699567   6.6142050
## 351   4.5778352  -3.0152572
## 352  20.8635925   3.5360133
## 369   3.2754530   0.8722166
## 370 -25.6128694   4.8224646
## 371   0.8070397  -0.9881551
## 372  12.3145394   1.2840297
```

```r
coef(model.random_slope)
```

```
## $Subject
##     (Intercept)       Days
## 308    253.6637 19.6662579
## 309    211.0065  1.8475828
## 310    212.4449  5.0184061
## 330    275.0956  5.6529547
## 331    273.6653  7.3973914
## 332    260.4446 10.1951153
## 333    268.2455 10.2436615
## 334    244.1725 11.5418620
## 335    251.0714 -0.2848731
```

```
## 337    286.2955 19.0955699
## 349    226.1950 11.6407002
## 350    238.3351 17.0814910
## 351    255.9829  7.4520288
## 352    272.2687 14.0032993
## 369    254.6806 11.3395026
## 370    225.7922 15.2897506
## 371    252.2121  9.4791309
## 372    263.7196 11.7513157
##
## attr(,"class")
## [1] "coef.mer"
```