36-401 Class Notes and Exercise for 11/21/00

Mixed Models

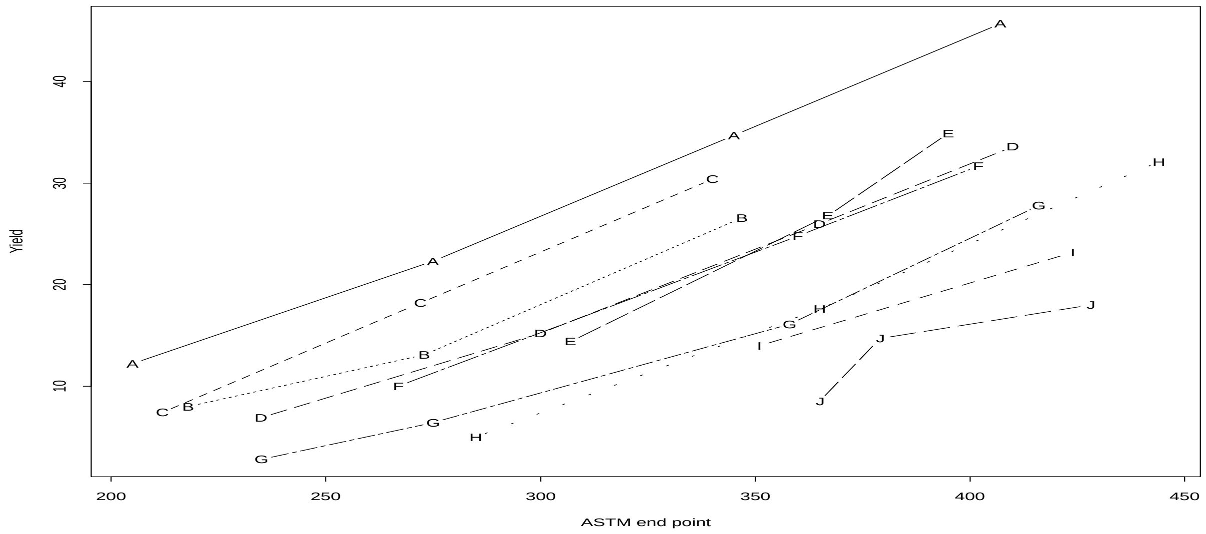
Howard Seltman

Main Idea: Mixed models are regression models with both *fixed* effects and *random* effects. So far we have only dealt with fixed effects, i.e. categorical covariates for which the levels seen are the specific levels of interest. Random effects are covariates for which the levels seen in the data are a random selection from a larger set of potential levels. We will deal with mixed models in the context of multiple measurements drawn over time on each of several randomly selected subjects, i.e. the subjects are the random effects. This topic includes repeated measures, hierarchical linear models, and longitudinal analysis.

References: *Modern Applied Statistics with S-Plus*, Venables and Ripley, 1997, chapter 10; you may want to skim section 10.4. *Linear mixed models in practice : an SAS-oriented approach*, Verbeke and Molenberghs, 1997. *SAS system for Mixed Models*, Littell, et al., 1996.

Example: Prater (1956) wanted to estimate gasoline yield for oil refinement. The "petrol" data set (in library MASS) has ten crude oil samples (No). Yield (Y) for each sample was measured for 2 to 4 different refinement conditions called ASTM end point (EP). For each of the 10 samples we know the specific gravity (SG), vapor pressure (VP) and V10 measurements.
Note the hierarchical nature of the data, a feature that should suggest possible use of mixed models.

```
# Plot data by oil sample (No)
x_split(petrol$EP, petrol$No)
y_split(petrol$Y, petrol$No)
plot(petrol$EP, petrol$Y, type="n", xlab="ASTM end point", ylab="Yield")
for (i in 1:length(x)) lines(x[[i]], y[[i]], lty=i, pch=names(x)[i], type="b")
```



```
# Do simple regression for each subject
pet1.lm _ lm(Y ~ No/EP -1, petrol)
matrix(round(coef(pet1.lm),2), 2, 10, byrow=T, dimnames=list(c("b0","b1"),levels(petrol$No)))
#         A      B      C      D      E      F      G      H      I      J
# b0 -22.67 -24.97 -30.69 -29.86 -56.16 -32.88 -30.29 -43.91 -30.24 -36.66
# b1   0.17   0.15   0.18   0.15   0.23   0.16   0.14   0.17   0.13   0.13

# Test for common slope (parallel lines)
pet2.lm_lm(Y ~ No -1 + EP, petrol)
anova(pet2.lm, pet1.lm)
# Analysis of Variance Table    Response: Y
#         Terms Resid. Df      RSS    Test Df Sum of Sq  F Value     Pr(F)
# 1 No - 1 + EP        21 74.13199
# 2   No/EP - 1        12 30.32915 1 vs. 2  9  43.80284 1.925665 0.1439048

# Test for common intercept
anova(lm(Y~EP,petrol), pet2.lm)
# Analysis of Variance Table     Response: Y
#         Terms Resid. Df      RSS Test Df Sum of Sq  F Value        Pr(F)
# 1          EP        30 1759.694
# 2 No - 1 + EP        21   74.132  +No  9  1685.562 53.05371 1.984302e-12

# Try linear mixed model  (Why?)
pet3.lme _ lme(Y ~ SG +VP +V10 + EP, random=~1, cluster=~No, data=petrol)
# Estimation Method: RML
# Convergence at iteration: 6
# Restricted Loglikelihood: -76.19098
# Restricted AIC: 166.382
# Restricted BIC: 176.6421
#
#   Standard Deviation(s) of Random Effect(s)
#  (Intercept)
#      1.44474
#
#  Cluster Residual Variance: 3.505164
#
# Fixed Effects Estimate(s):
#                   Value Approx. Std.Error  z ratio(C)
# (Intercept) -6.1347915       14.55416371  -0.4215145
#          SG  0.2193982        0.14693771   1.4931372
#          VP  0.5458631        0.52052725   1.0486735
#         V10 -0.1542427        0.03996237  -3.8596986
#          EP  0.1571768        0.00558800  28.1275578
#
# Random Effects (Conditional Modes):
#   (Intercept)
# A -0.05943842   B -0.21857083   C  1.92029373   D -1.92760485
# E -0.21650121   F  0.56931066   G  0.06701571   H  0.19194435
# I -0.40277963   J  0.07633048
#
# Number of Observations: 32   Number of Clusters: 10
```

Theory: The mixed effects model for the i^{th} subject is

$$\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\boldsymbol{\eta_i} + \boldsymbol{\epsilon_i}$$

$$\boldsymbol{\eta_i} \sim N(\mathbf{0}, \mathbf{D}) \qquad \boldsymbol{\epsilon_i} \sim N(\mathbf{0}, \boldsymbol{\Sigma_i})$$

$$\mathbf{Y_1} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \end{bmatrix} \quad \mathbf{X_1} = \begin{bmatrix} 1 & x_{111} & x_{112} & x_{113} \\ 1 & x_{121} & x_{122} & x_{123} \\ 1 & x_{131} & x_{132} & x_{133} \\ 1 & x_{141} & x_{142} & x_{143} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

In the simplest case $\epsilon_1 \sim N(\mathbf{0}, \sigma^2\mathbf{I_4})$.
For a random intercept only, $\mathbf{Z_1'} = [1111]$, $\eta_i \sim N(0, \delta^2)$.
Other cases include autocorrelation,

$$\boldsymbol{\Sigma_1} = \sigma^2\mathbf{I_4} + \tau^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

and compound symmetry (uniform correlation),

$$\boldsymbol{\Sigma_1} = \sigma^2\mathbf{I_4} + \begin{bmatrix} \tau^2 & \tau^2 & \tau^2 & \tau^2 \\ \tau^2 & \tau^2 & \tau^2 & \tau^2 \\ \tau^2 & \tau^2 & \tau^2 & \tau^2 \\ \tau^2 & \tau^2 & \tau^2 & \tau^2 \end{bmatrix}$$

A compound symmetry $\boldsymbol{\Sigma_i}$ is used *instead* of including $\mathbf{Z_i}\boldsymbol{\eta_i}$ to achieve random intercepts; the two are equivalent.

For random intercept and random slope, we can use 2 (or more) columns in $\mathbf{Z_i}$, $\boldsymbol{\eta_i}$ is a vector of random effects, and $\mathbf{D}$ can have zero or non-zero off diagonal elements.

Estimation of random effects is best done by restricted maximum likelihood (REML), which is the S-Plus default. I suggest BIC for testing random effects. Nested F tests for fixed effects require ML, not REML.

The reasons for using mixed models to fit random slopes and intercepts are

1. to fit real variation in level or rate of change of response between subjects (and to assess the size of that variation)

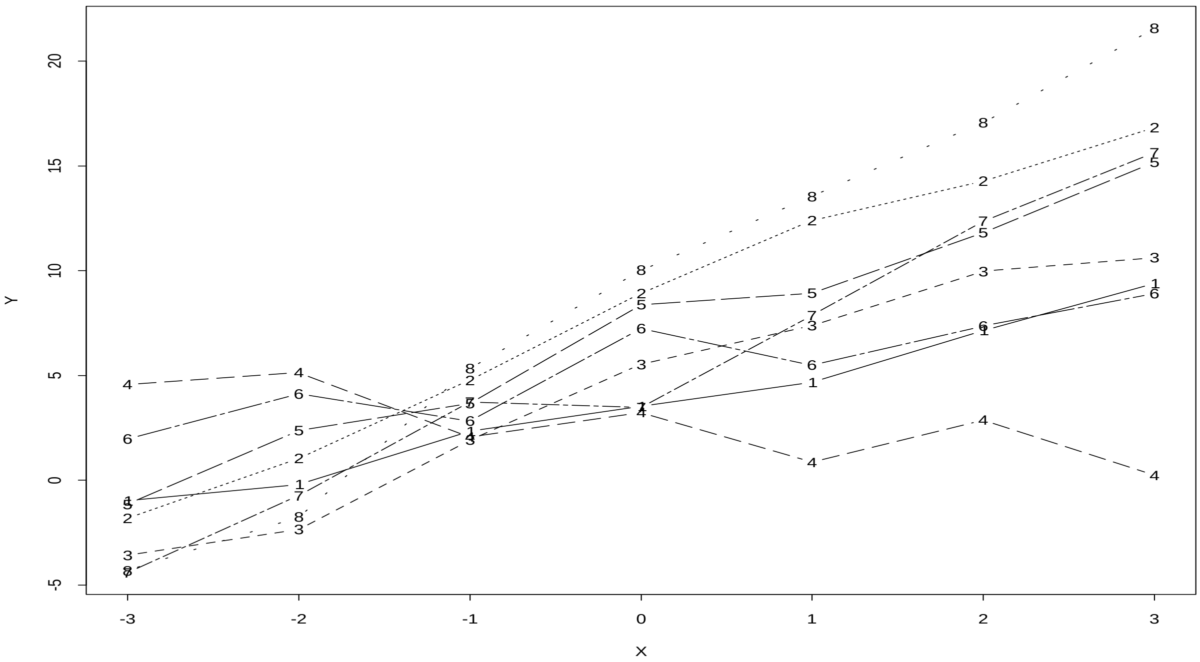
2. to improve prediction for a given subject

3. to assure that standard errors, and therefore tests and confidence intervals for fixed effects are correct; if the random effects are incorrectly modeled, the standard errors are incorrect.

Breakout Session for 11/21/00

1. What is the correlation between samples in the uniform correlation model?

2. What model might fit these data?



3. Interpret the results (same data as #2 above):

```
rslt1_lme(y~x, data=dtf, random=~1, cluster=~case)
summary(rslt1)
# Restricted BIC: 315.9685
#
#   Standard Deviation(s) of Random Effect(s)
#  (Intercept)
#     1.735946
#
#  Cluster Residual Variance: 11.04525
#
# Fixed Effects Estimate(s):
#                Value Approx. Std.Error z ratio(C)
# (Intercept) 5.667399         0.7575784    7.48094
#           x 2.274274         0.2220567   10.24186

rslt2_lme(y~x, data=dtf, random=~1+x, cluster=~case)
summary(rslt2)
# Restricted BIC: 236.0035
#
#   Standard Deviation(s) of Random Effect(s)
#  (Intercept)        x
#     2.106907 1.546896
#
#  Correlation of Random Effects
#   (Intercept)
# x 0.8228476
#
#  Cluster Residual Variance: 1.066404
#
# Fixed Effects Estimate(s):
#                Value Approx. Std.Error z ratio(C)
# (Intercept) 5.667399         0.7575784   7.480940
#           x 2.274274         0.5512456   4.125701

anova(rslt2, rslt1)
# Response: y
# rslt2 fixed: (Intercept), x    random: (Intercept), x   serial corr struct: identity
# rslt1
#  fixed: (Intercept), x   random: (Intercept)    serial corr struct: identity
#
#       Model Df    AIC    BIC  Loglik    Test Lik.Ratio P value
# rslt2     1  6 223.85 236.00 -105.93
# rslt1     2  4 307.87 315.97 -149.93 1 vs. 2    88.016       0
```