Random and Mixed Effects Models (Ch. 10)

Random effects models are very useful when the observations are sampled in a highly structured way. The basic idea is that the error associated with any linear,

$$E(y_i) = x_i^T \beta$$

or nonlinear

$$E(y_i) = \eta(x_i, \beta)$$

model has more structure than simply $N(0, \sigma^2)$. Sometimes, for example, the y's are obtained by *two stage* sampling: a batch of chemical is sampled from a production run, and then several smaller samples are taken from each batch. Or a group of schools is chosen at random, and then classes are sampled within each school. Or a sample of patients is followed over time, so that successive measurements on an individual patient might be expected to be correlated. This last case is sometimes called 'repeated measures', modelling; in social science examples such as the schools example this is sometimes called multi-level modelling.

I will follow the discussion in the text fairly closely, filling in where I think it is helpful. This handout only considers linear models (§10.1).

The *gasoline* data in library `MASS` (`data(petrol)`), is an example of the first type. There were 10 batches of crude oil (called samples in the book), and several measurements were made on each batch. The measurements are:

response `Y`	yield of the refined product as a percentage of crude
covariate `SG`	specific gravity
covariate `VP`	vapour pressure
covariate `V10`	ASTM 10% point
covariate `EP`	ASTM end point in degrees F

There is another variable in the data frame, `No`, which records the batch. It is a factor variable with 10 levels `"A"` through `"J"`. The first three covariates were measured on the batch, and then within each batch there were several (between 2 and 4) measurements taken of `EP` and `Y`. My first steps were to try to get a sense of the data from various plots and summaries.

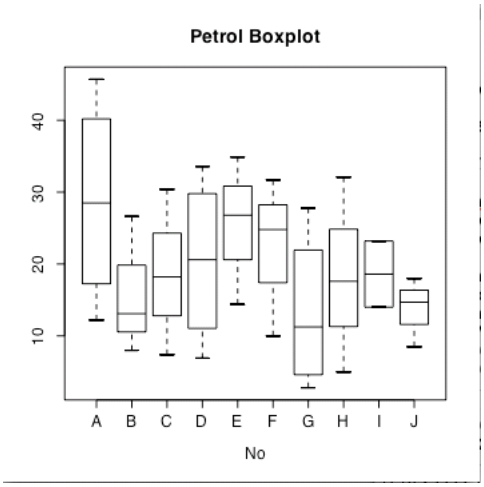
```
> library(MASS)
> data(petrol)
> dim(petrol)
[1] 32  6
> petrol
   No   SG  VP V10  EP    Y
1   A 50.8 8.6 190 205 12.2
2   A 50.8 8.6 190 275 22.3
3   A 50.8 8.6 190 345 34.7
```

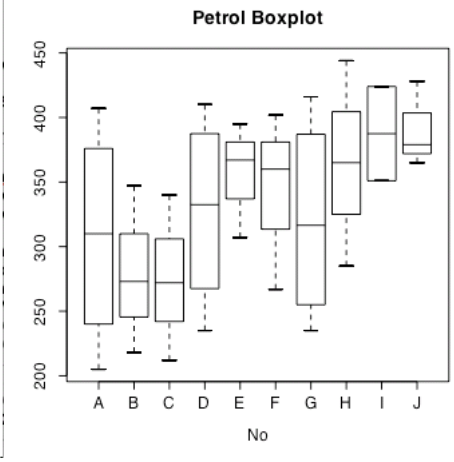
```
4   A 50.8 8.6 190 407 45.7
5   B 40.8 3.5 210 218  8.0
6   B 40.8 3.5 210 273 13.1
7   B 40.8 3.5 210 347 26.6
8   C 40.0 6.1 217 212  7.4
9   C 40.0 6.1 217 272 18.2
10  C 40.0 6.1 217 340 30.4
11  D 38.4 6.1 220 235  6.9
12  D 38.4 6.1 220 300 15.2
13  D 38.4 6.1 220 365 26.0
14  D 38.4 6.1 220 410 33.6

...

> tapply(petrol$Y,petrol$No,mean)
       A        B        C        D        E        F        G        H
28.72500 15.90000 18.66667 20.42500 25.36667 22.16667 13.27500 18.23333
       I        J
18.60000 13.73333

> petrol.mean <- cbind(tapply(petrol$Y,petrol$No,mean),tapply(petrol$SG,petrol$No,mean),
+ tapply(petrol$VP,petrol$No,mean),tapply(petrol$V10,petrol$No,mean),
+ tapply(petrol$EP,petrol$No,mean) )

> petrol.mean <- data.frame(petrol.mean)
> names(petrol.mean)<-c("Y","SG","VP","V10","EP")
> petrol.mean
         Y   SG  VP V10       EP
A 28.72500 50.8 8.6 190 308.0000
B 15.90000 40.8 3.5 210 279.3333
C 18.66667 40.0 6.1 217 274.6667
D 20.42500 38.4 6.1 220 327.5000
E 25.36667 40.3 4.8 231 356.3333
F 22.16667 32.2 5.2 236 343.0000
G 13.27500 41.3 1.8 267 321.0000
H 18.23333 38.1 1.2 274 364.6667
I 18.60000 32.2 2.4 284 387.5000
J 13.73333 31.8 0.2 316 390.6667
```

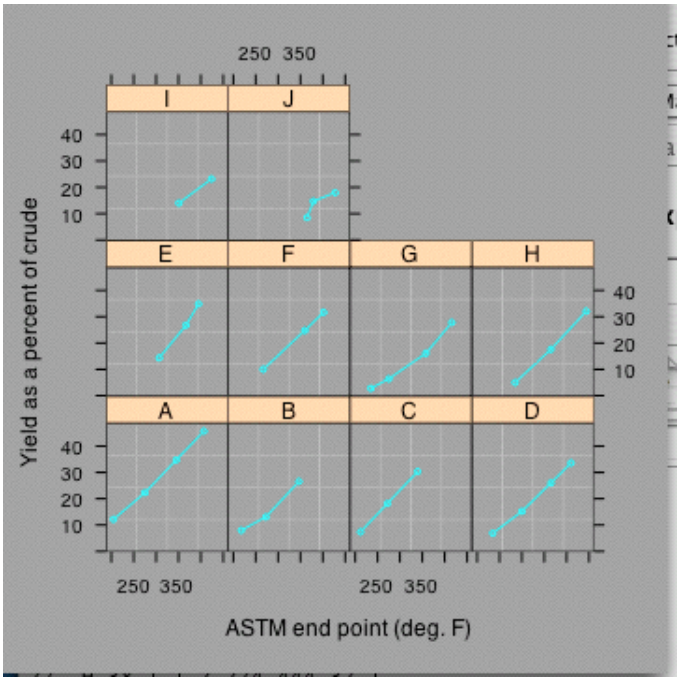
Then I tried a few elementary plots, and the code given in the book (p.272) for Figure 10.1.

```
> plot(petrol$No,petrol$Y,main="Petrol Boxplot", xlab="No")
> plot(petrol$No,petrol$EP,main="Petrol Boxplot", xlab="No")
```

```
> library(lattice)
> xyplot(Y~EP | No, data=petrol,
+ xlab="ASTM end point (deg. F)",
+ ylab="Yield as a percent of crude",
+ panel=function(x,y){
+ panel.grid()
+ m<-sort.list(x)
+ panel.xyplot(x[m],y[m],type="b",cex=0.5)})
```







The left boxplot is yield and the right boxplot is EP. The grey plot shows the regression of `Y` on `EP` in each group. (See Figure 10.1 for a clearer picture.)

The first regression model fit in the text is separate linear regressions for each of the groups. A new data frame was created that replaced each covariate by $(x_i - \bar{x})$. This just changes the estimates of the intercepts; the claim is that these are then easier to interpret.

```
> ##
> ## Fit separate regressions for each of the 10 groups
> ##
> ## (replace each covariate by 'covariate - mean(covariate)'
> ##
> Petrol <- petrol
> Petrol[,2:5] <- scale(Petrol[,2:5],scale=F)
> pet1.lm <- lm(Y ~ No/EP -1, data=Petrol)
> coef(pet1.lm)
       NoA        NoB        NoC        NoD        NoE        NoF
32.7452257 23.6201614 28.9852457 21.1303143 19.8208227 20.4164797
       NoG        NoH        NoI        NoJ     NoA:EP     NoB:EP
14.7805644 12.6824864 11.6172945  6.1775520  0.1668576  0.1463249
    NoC:EP     NoD:EP     NoE:EP     NoF:EP     NoG:EP     NoH:EP
 0.1796814  0.1535378  0.2287929  0.1604756  0.1357129  0.1704130
    NoI:EP     NoJ:EP
 0.1260274  0.1289979
> 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 32.75 23.62 28.99 21.13 19.82 20.42 14.78 12.68 11.62 6.18
b1  0.17  0.15  0.18  0.15  0.23  0.16  0.14  0.17  0.13 0.13
```

The model formula `No/EP`, or in general `a/b` is explained on p.150 near the bottom. It is a shorthand for `a + a:b`, which means a separate model `1 + b` for each level of `a`. (Although it can be used if `a` is not a factor, this doesn't usually make sense.) The mathematical model is

$$y_{ij} = \beta_{0i} + \beta_{1i} EP_{ij} + \epsilon_{ij} \tag{1}$$

where $j = 1, \ldots, n_i; i = 1, \ldots, 10$ and we assume the ϵ_{ij} are independent, mean 0 and variance σ^2. (We haven't used any random effects yet.)

Since the slopes are all fairly similar, but the intercepts are very different, the second model tried is

$$y_{ij} = \beta_{0i} + \beta EP_{ij} + \epsilon_{ij}. \tag{2}$$

```
> pet2.lm <- lm(Y~No -1 +EP, data=Petrol)
> summary(pet2.lm)

Call:
```

```
lm(formula = Y ~ No - 1 + EP, data = Petrol)

Residuals:
     Min       1Q   Median       3Q      Max
-3.13601 -0.93477 -0.08414  1.16652  3.39579

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
NoA 32.549392   0.949476  34.281  < 2e-16 ***
NoB 24.274641   1.125928  21.560 8.31e-16 ***
NoC 27.782046   1.133370  24.513  < 2e-16 ***
NoD 21.154164   0.939794  22.509 3.48e-16 ***
NoE 21.519127   1.093576  19.678 5.19e-15 ***
NoF 20.435522   1.086548  18.808 1.28e-14 ***
NoG 15.035907   0.941567  15.969 3.20e-13 ***
NoH 13.063047   1.100631  11.869 8.92e-11 ***
NoI  9.805387   1.365804   7.179 4.46e-07 ***
NoJ  4.436077   1.135286   3.907  0.00081 ***
EP   0.158730   0.005718  27.759  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.879 on 21 degrees of freedom
Multiple R-Squared: 0.9953,Adjusted R-squared: 0.9929
F-statistic: 408.4 on 11 and 21 DF,  p-value: < 2.2e-16

> anova(pet2.lm,pet1.lm)
Analysis of Variance Table

Model 1: Y ~ No - 1 + EP
Model 2: Y ~ No/EP - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     21 74.132
2     12 30.329  9    43.803 1.9257 0.1439
```

The last command compares the full model (1) with the reduced model (2). In (1) there are 10 intercepts and 10 slopes estimated, leaving 32-20=12 residual degrees of freedom. In (2) there are 10 intercepts and 1 slope estimated, leaving 21 residual degrees of freedom. The improvement in residual sum of squares by fitting 10 different slopes is not enough to justify all these extra parameters.

The next step is to see if there is any structure in the intercepts. We haven't yet used the covariates `SG`, `VP` and `V10`. So we now try the model

$$y_{ij} = \mu + \beta_1 SG_i + \beta_2 VP_i + \beta_3 V10_i + \beta_4 EP_{ij} + \epsilon_{ij} \tag{3}$$

where we have replaced β_{0i} in (2) with a linear structure. (Still no random effects!) This model has 5 parameters, leaving 27 residual degrees of freedom. The details are at the bottom of p.273 and the top of p.274. This model doesn't seem to be as good as model (2), so we still haven't really explained the variation in the intercepts (β_{0i}).

So now we try a different explanation, we model the intercepts as random variables with a common mean and variance σ_1^2, say:

$$y_{ij} = \mu + \zeta_i + \beta_1 SG_i + \beta_2 VP_i + \beta_3 V10_i + \beta_4 EP_{ij} + \epsilon_{ij} \tag{4}$$

where we assume $\zeta_i \sim N(0, \sigma_1^2)$ independently of ϵ_{ij}. The model means that each group has a random intercept. The variance component σ_1^2 measures the variation due to the choice of batch of crude oil, whereas σ^2 measures the variation in the process to measure the yield. Linear mixed models are fit using `lme`, in the library `nlme`. Model (3) is called a *mixed effects* model, as it has some random effects (intercept) and some fixed effects (everything else).

```
> ?lme
> library(nlme)
> ?lme
> pet3.lme <- lme(fixed = Y ~ SG + VP + V10 + EP,
+ random = ~ 1 | No, data=Petrol)
> summary(pet3.lme)
Linear mixed-effects model fit by REML
 Data: Petrol
       AIC      BIC    logLik
  166.3820 175.4528 -76.19098

Random effects:
 Formula: ~1 | No
        (Intercept) Residual
StdDev:    1.445028 1.872146

Fixed effects: Y ~ SG + VP + V10 + EP
                Value Std.Error DF  t-value p-value
(Intercept) 19.706795 0.5683413 21 34.67423  0.0000
SG           0.219397 0.1469559  6  1.49295  0.1861
VP           0.545861 0.5205881  6  1.04855  0.3348
V10         -0.154244 0.0399668  6 -3.85929  0.0084
EP           0.157177 0.0055878 21 28.12841  0.0000
 Correlation:
    (Intr) SG     VP     V10
SG   0.059
VP   0.013  0.067
V10  0.015  0.433  0.836
EP  -0.004  0.023 -0.116 -0.197
```

```
Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-1.7807117 -0.6063671 -0.1069013  0.4571818  1.7811918

Number of Observations: 32
Number of Groups: 10
```

Note that the degrees of freedom for some of the covariates (`SG, VP, V10`) are 6, whereas for `EP` it is 21. This is because the first set only takes 10 different values, and there are four fixed effects parameters estimated. But it's a bit tricky to figure out why the intercept and `EP` have 21.

The output gives us estimates of σ_1^2 and σ^2:

$$\hat{\sigma}_1^2 = (1.444)^2 = 2.09 \qquad \hat{\sigma}^2 = (1.872)^2 = 3.51.$$

Note also that the estimates of $\beta_1, \ldots \beta_4$ are not very different between the fixed and mixed effects models, but that the standard errors of the estimates for `SG, VP, V10` are much smaller. This is because we have separated out batch-to-batch variation from within batch variation.

On p.275 the book compares the fixed and mixed effects models using `anova`, but before doing this they had to refit the mixed effects model using a different method for estimating the variance components (σ^2 and σ_1^2). This is related to a theoretical point about likelihood ratio tests; but the bottom line is if you are interested in *estimating* the components of variance σ^2 and σ_1^2 then it is better to use method `REML`, but if you are interested in comparing models, it is better to use method `ML`. Their main conclusion is that the mixed effects model doesn't really fit any better, but we're pressing on anyway.

Note from the summary of `pet3.lme` that the variables `SG` and `VP` do not seem to have much effect on `Y`, so the next model they tried omits these variables.

```
> pet4.lme <- update(pet3.lme, fixed=Y~V10 +EP)
> anova(pet4.lme,pet3.lme)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
pet4.lme     1  5 163.9555 170.7920 -76.97775
pet3.lme     2  7 166.3819 175.4528 -76.19098 1 vs 2 1.573551  0.4553
Warning message:
Fitted objects with different fixed effects. REML comparisons are not meaningful. in: an
```

I just tried the above `anova` statement for fun, but note that I got a warning telling me it was not valid, because of the fitting method used. This is a reminder to refit both models using `method="ML"` if we want to do likelihood ratio tests to compare nested models.

```
> pet3.lme <- update(pet3.lme, method="ML")
> pet4.lme <- update(pet3.lme, fixed=Y~V10+EP)
> anova(pet4.lme,pet3.lme)
         Model df      AIC      BIC    logLik   Test L.Ratio p-value
pet4.lme     1  5 149.6119 156.9406 -69.80594
pet3.lme     2  7 149.3833 159.6435 -67.69166 1 vs 2 4.22855  0.1207

> summary(pet4.lme)
Linear mixed-effects model fit by maximum likelihood
 Data: Petrol
       AIC      BIC    logLik
  149.6119 156.9406 -69.80594

Random effects:
 Formula: ~1 | No
        (Intercept) Residual
StdDev:    1.381100 1.823660

Fixed effects: Y ~ V10 + EP
                Value Std.Error DF   t-value p-value
(Intercept) 19.651589 0.5733608 21  34.27439       0
V10         -0.210805 0.0160972  8 -13.09575       0
EP           0.157586 0.0056728 21  27.77945       0
 Correlation:
    (Intr) V10
V10 -0.046
EP  -0.004 -0.285

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-1.8146954 -0.4421839 -0.1487166  0.4532735  1.8221910

Number of Observations: 32
Number of Groups: 10
```

There are many components to an `lme` object; see `?lme.Object` for details. For example,

```
> fixed.effects(pet4.lme)
(Intercept)         V10          EP
 19.6515891  -0.2108053   0.1575859
> coef(pet4.lme)
  (Intercept)        V10        EP
A    21.05404 -0.2108053 0.1575859
```

```
B    18.33760 -0.2108053 0.1575859
C    21.48568 -0.2108053 0.1575859
D    17.53792 -0.2108053 0.1575859
E    19.45035 -0.2108053 0.1575859
F    19.42200 -0.2108053 0.1575859
G    20.17194 -0.2108053 0.1575859
H    19.84125 -0.2108053 0.1575859
I    19.21158 -0.2108053 0.1575859
J    20.00352 -0.2108053 0.1575859
```

The intercepts are actual 'estimates' of the individual random effects ζ_i, although for random variables we usually think of *predicting* them, rather than estimating them. They are not explicitly used for summarizing the data; for that we use the estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}^2$.

The final model is to let the slope on `EP` have a random component:

$$y_{ij} = \mu + \zeta_i + \beta_3 V10_i + (\beta_4 + \eta_i) EP_{ij} + \epsilon_{ij} \tag{5}$$

where we have as before $\epsilon_{ij} \sim N(0, \sigma^2)$, $\zeta_i \sim N(0, \sigma_1^2)$ and now $\eta_i \sim N(0, \sigma_2^2)$ and $\mathrm{cov}(\zeta_i, \eta_i) = \sigma_{12}$, i.e. we don't assume the random effects for the intercept and slope are necessarily independent.

```
> pet5.lme <- update (pet4.lme, random=~1 + EP | No)
> pet5.lme
Linear mixed-effects model fit by maximum likelihood
  Data: Petrol
  Log-likelihood: -69.80776
  Fixed: Y ~ V10 + EP
(Intercept)         V10          EP
 19.6514774  -0.2108117   0.1575926

Random effects:
 Formula: ~1 + EP | No
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev       Corr
(Intercept) 1.3820901440 (Intr)
EP          0.0008064208 0.002
Residual    1.8225962064

Number of Observations: 32
Number of Groups: 10
> anova(pet4.lme,pet5.lme)
         Model df      AIC      BIC    logLik   Test     L.Ratio p-value
pet4.lme     1  5 149.6119 156.9406 -69.80594
pet5.lme     2  7 153.6155 163.8757 -69.80776 1 vs 2 0.003646107  0.9982
```

This shows that model (5) fits the data no better than model (4).

The `anova` command gives different output for `lme` than for `lm`, although it uses the same general theory. In this last comparison, `pet4.lme` has fitted 5 parameters $(\mu, \beta_3, \beta_4, \sigma^2, \sigma_1^2)$ and `pet5.lme` has fitted 7 parameters: these 5 plus σ_2^2 and σ_{12}. So the difference in loglikelihoods has 2 degrees of freedom.

The hardest part in specifying the model for `lme` is getting the random effects part correctly specified. I didn't find the help file very helpful. The general format for `lme` is

```
lme(fixed = ...  [formula], data= ...  [data.frame],
             random = ... [formula] , ... [other stuff] ).
```

Once you see a formula for a given example it's kind of obvious, but I find it hard to construct it from scratch. The definitive reference is the book by Pinheiro and Bates (exact reference in the help file).

The next two linear model examples are a multi-level sampling example involving students (in classes, in schools), and a growth curve model, where measurements are repeated on the same experimental unit (in this case trees) at several time points.

A very general formulation of the mixed effects linear model is given on p.279:

$$y_{ij} = x_{ij}\beta + z_{ij}\zeta_i + \epsilon_{ij} \tag{6}$$

where we assume we have the responses in groups indexed by i, and then the responses within each group are indexed by j. In fact the schools example has more structure, I think it should be indexed by ijk, where $i = 1, 2$ indexes levels of `COMB`, j indexes schools and k indexes pupils; but I'm not completely sure. In (6) x_{ij} and z_{ij} are row vectors of explanatory variables. The most general assumption about ϵ_{ij} is that they are independent among levels of i ($\mathrm{cov}(\epsilon_{ij}, \epsilon_{i'j}) = 0$), but possibly dependent within groups:

$$\mathrm{var}(\epsilon_{ij}) = \sigma^2 g(\mu_{ij}, z_{ij}, \theta), \quad \mathrm{corr}(\epsilon_{ij}) = \Gamma(\alpha).$$

Note this is a very general formulation of the variance; in our example above we had a much simpler structure. It is usually assumed as well that ζ_i are independent of the ϵ_{ij} and have variance-covariance matrix (note that ζ_i are vectors)

$$\mathrm{var}(\zeta_i) = D(\alpha_\zeta).$$

Usually $g \equiv 1$, $\Gamma = I$, and only D is unrestricted (as we had above for `pet5.lme`).