APTS Statistical Modelling: Practical 2

Helen Ogden

The data in the file `hip.txt` (available from the APTS web site) are taken from Crowder and Hand (*Analysis of Repeated Measures*, 1990, Chapman and Hall) and can be read into R by using

```
hip <- read.table("hip.txt",
                  col.names = c("y", "age", "sex", "subj", "time"))
```

Variable `y` represents measurements of response variable *haematocrit* on 30 patients (`subj`) on up to three occasions (`time`), one before a hip-replacement operation, and two afterwards. The `age` and `sex` (0=male, 1=female) of the patients is also recorded.

We will investigate these data using linear mixed models of the form $y_{ij} \sim N(\mu_{ij}, \sigma^2)$ where y_{ij} is the response for subject i, time j and

$$\mu_{ij} = x_{ij}^T \beta + z_{ij}^T b_i, \quad b_i \sim N(0, \Sigma_b).$$

You should consider including `age`, `sex` and `time` (and possibly interactions) within x_{ij} and `time` within z_{ij}. We will treat `time` as a categorical variable.

LMMs for clustered data can be fitted in R using the `lmer` function from the `lme4` library:

```
library(lme4)
```

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

For example

```
hip_lmm1 <- lmer(y ~ age + sex + factor(time) + (1 | subj), data = hip)
```

fits the model with `1`, `age`, `sex`, `I(time=2)` and `I(time=3)` in x_{ij}, and just the intercept `1` in z_{ij}.

The default estimation method is REML. If you want to obtain maximum likelihood estimates (for example, for use in model comparison), they can be obtained using the additional argument `REML = FALSE`.

You might find some of the following functions useful – they all take an `lmer` fit as their first argument: `summary`, `fitted`, `residuals`, `fixef` (fixed effects estimates), `ranef` (random effects estimates), `VarCorr` (variance estimates) `coef` (coefficient estimates at cluster level, incorporating fixed and random effects), `AIC`, `BIC` and `predict`.

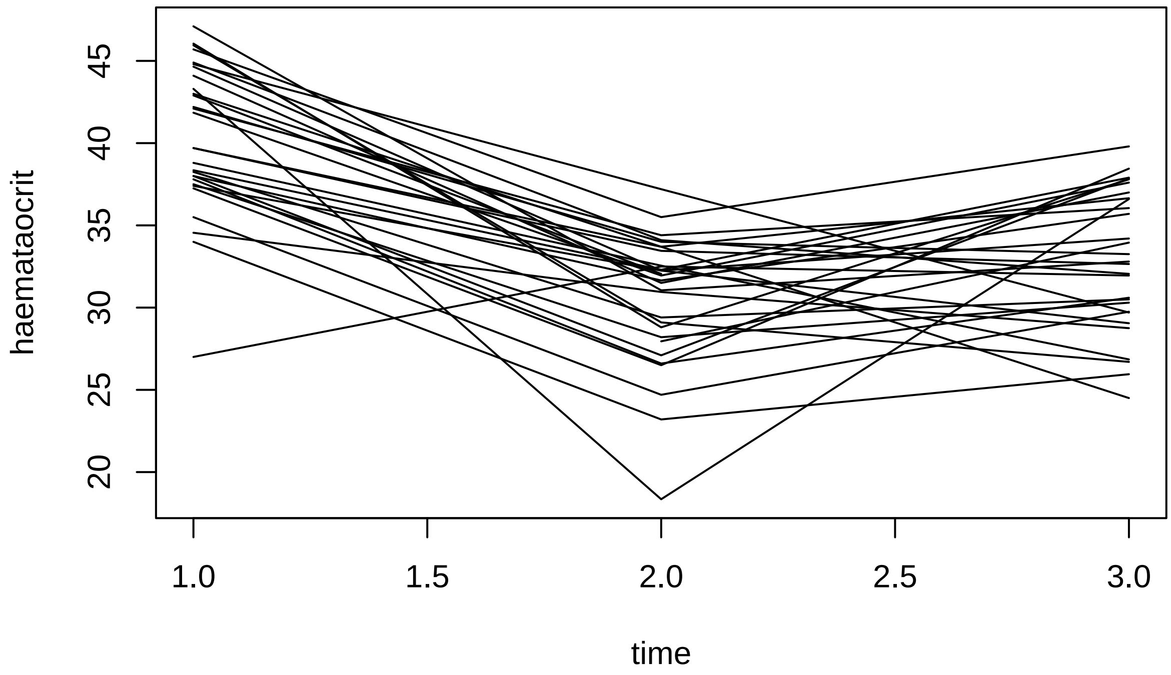
Tasks

1. Plot the time profiles of the response variable for each subject on a single plot (as we did for the rat growth data in Example 2.4 in the lecture notes). Do you think you think it will be necessary to include a random intercept for the subject? What about a random slope for time?
2. Find your preferred LMM for this data.
3. For your preferred LMM, plot the predicted haematocrit levels for each subject against time.

Solutions

We can plot out the variation in the response over time for each subject:

```
plot(hip$time, hip$y, type="n", xlab="time", ylab="haemataocrit")
for (i in 1:30)
    lines(hip$time[hip$subj==i], hip$y[hip$subj==i])
```



We can then fit various possible LMMs. We use maximum likelihood rather than REML so that we can compare models with AIC.

```
hip_lmm1_ML <- lmer(y ~ age + sex + factor(time) + (1|subj),
                    data = hip, REML = FALSE)
summary(hip_lmm1_ML)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: y ~ age + sex + factor(time) + (1 | subj)
##    Data: hip
##
##      AIC      BIC   logLik deviance df.resid
##    508.8    526.1   -247.4    494.8       81
##
## Scaled residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.3004 -0.5844  0.0343  0.6366  1.7127
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subj     (Intercept)  2.471   1.572
##  Residual             14.052   3.749
## Number of obs: 88, groups:  subj, 30
##
## Fixed effects:
##                 Estimate Std. Error t value
## (Intercept)     39.07979    3.80458  10.272
## age              0.03518    0.05630   0.625
## sex             -1.91758    0.99871  -1.920
## factor(time)2 -9.75246    0.97756  -9.976
## factor(time)3 -7.34666    0.98714  -7.442
##
## Correlation of Fixed Effects:
##             (Intr) age    sex    fct()2
## age         -0.969
## sex         -0.054 -0.093
## factor(tm)2 -0.150  0.022 -0.013
## factor(tm)3 -0.144  0.018 -0.027  0.505
```

The summary suggests to try dropping age, since it has the smallest t value.

```r
hip_lmm2_ML <- lmer(y ~ sex + factor(time) + (1|subj),
                    data = hip, REML = FALSE)
summary(hip_lmm2_ML)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: y ~ sex + factor(time) + (1 | subj)
##    Data: hip
##
##      AIC      BIC   logLik deviance df.resid
##    507.1    522.0   -247.6    495.1       82
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2634 -0.5907  0.0413  0.6412  1.7047
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subj     (Intercept)  2.567   1.602
##  Residual             14.050   3.748
## Number of obs: 88, groups:  subj, 30
```

```
##
## Fixed effects:
##                Estimate Std. Error t value
## (Intercept)     41.3844     0.9389  44.077
## sex             -1.8595     1.0009  -1.858
## factor(time)2   -9.7657     0.9773  -9.992
## factor(time)3   -7.3578     0.9870  -7.455
##
## Correlation of Fixed Effects:
##               (Intr) sex    fct()2
## sex           -0.592
## factor(tm)2 -0.524 -0.011
## factor(tm)3 -0.510 -0.026  0.505
```

```
hip_lmm3_ML <- lmer(y ~ factor(time) + (1|subj),
                    data = hip, REML = FALSE)
c(AIC(hip_lmm1_ML), AIC(hip_lmm2_ML), AIC(hip_lmm3_ML))
```

```
## [1] 508.7600 507.1478 508.4233
```

On the basis of AIC, we prefer `hip_lmm2_ML` of these models. We could also consider an interaction between `sex` and `time`

```
hip_lmm4_ML <- lmer(y ~ sex*factor(time) + (1|subj),
                    data = hip, REML = FALSE)
c(AIC(hip_lmm2_ML), AIC(hip_lmm4_ML))
```

```
## [1] 507.1478 507.9042
```

AIC slightly prefers the model `hip_lmm2_ML`, without an interaction.

We could try including a random slope for `time` in the model:

```
hip_lmm5_ML <- lmer(y ~ factor(time) + (factor(time)|subj),
                    data = hip, REML = FALSE)
```

```
## Error: number of observations (=88) <= number of random effects (=90) for term (facto
```

We get an error message, because there are now too many different random effect terms in the model to be able to estimate them all from the data available.

We can then refit our chosen model with REML:

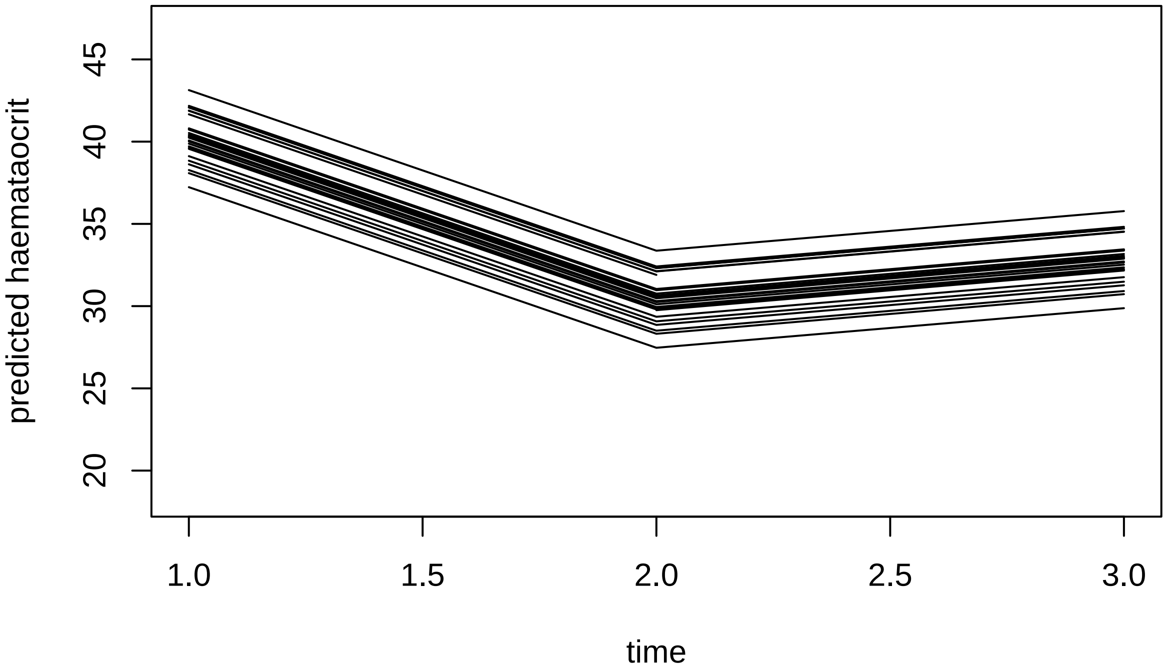
```
hip_lmm2 <- lmer(y ~ sex + factor(time) + (1|subj), data = hip)
summary(hip_lmm2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ sex + factor(time) + (1 | subj)
##    Data: hip
##
```

```
## REML criterion at convergence: 489.5
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -3.1899 -0.5637  0.0305  0.6154  1.6744 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subj     (Intercept)  2.92    1.709   
##  Residual             14.55    3.815   
## Number of obs: 88, groups:  subj, 30
## 
## Fixed effects:
##                 Estimate Std. Error t value
## (Intercept)      41.3847     0.9661  42.838
## sex              -1.8600     1.0360  -1.795
## factor(time)2    -9.7657     0.9947  -9.817
## factor(time)3    -7.3572     1.0047  -7.323
## 
## Correlation of Fixed Effects:
##             (Intr) sex    fct()2
## sex         -0.596              
## factor(tm)2 -0.518 -0.011       
## factor(tm)3 -0.504 -0.026  0.505
```

We can plot the predicted haematocrit levels for each of our subjects, and compare with our earlier plot of actual haematocrit levels against `time`.

```r
hip$pred_haematocrit <- predict(hip_lmm2)
plot(hip$time, hip$y, type="n", xlab="time", ylab="predicted haemataocrit")
for (i in 1:30)
    lines(hip$time[hip$subj==i], hip$pred_haematocrit[hip$subj==i])
```



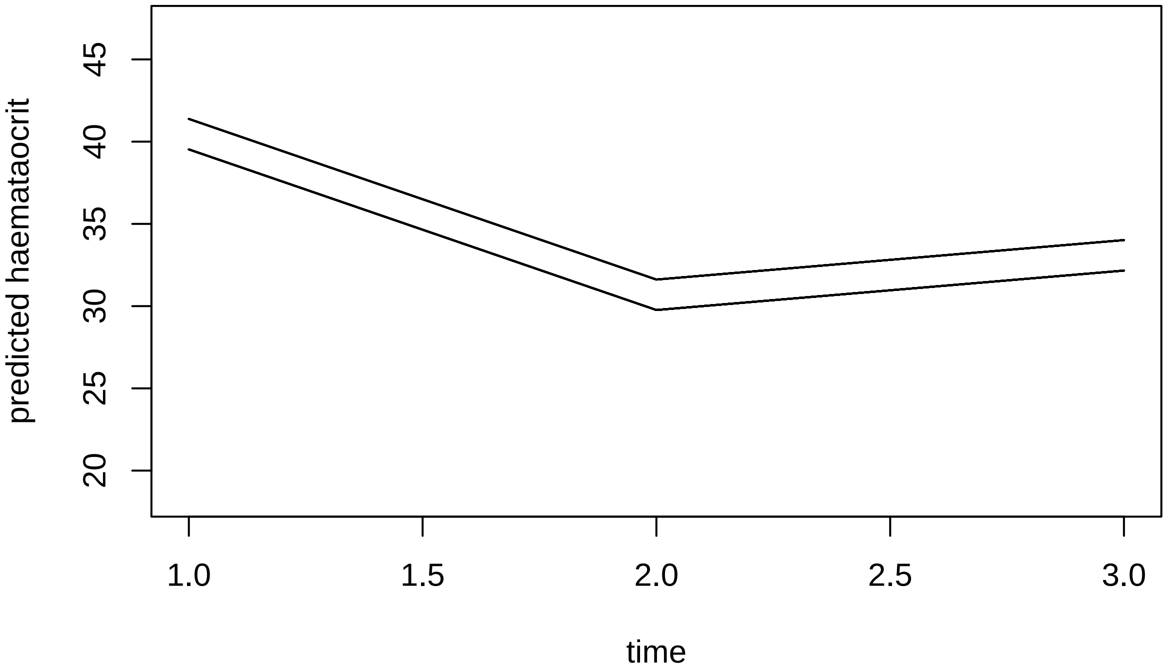
We can also consider a linear model, without the random intercept term.

```r
hip_lm2 <- lm(y ~ sex + factor(time), data = hip)
c(AIC(hip_lm2), AIC(hip_lmm2_ML))
```

```
## [1] 507.1334 507.1478
```

The AIC for the two models are very close, but AIC slightly prefers the simpler linear model, without a random intercept. We can plot the predicted haematocrit levels for each of our subjects according to this model, giving two lines, for male and female subjects:

```r
hip$pred_haematocrit_lm2 <- predict(hip_lm2)
plot(hip$time, hip$y, type="n", xlab="time", ylab="predicted haemataocrit")
for (i in 1:30)
    lines(hip$time[hip$subj==i], hip$pred_haematocrit_lm2[hip$subj==i])
```


predicted haemataocrit
20
25
30
35
40
45
1.0
1.5
2.0
2.5
3.0
time