Week 5 Practical Session

David Barron

Trinity Term 2018

```r
ess <- read_csv("C:\\Users\\dbarron\\Dropbox\\Teaching\\MSc teaching\\Advanced Quant\\data\\ESS/ess.csv
```

```
Parsed with column specification:
cols(
  ESS5_id = col_character(),
  cntry = col_character(),
  ESS5_reg = col_character(),
  NUTS1 = col_character(),
  NUTS2 = col_character(),
  NUTS3 = col_character(),
  happy = col_character(),
  gndr = col_character(),
  agea = col_character(),
  marsts = col_character(),
  eduyrs = col_character(),
  dweight = col_double(),
  pweight = col_double(),
  c_gdppc_2010 = col_double(),
  c_giaftot_2010 = col_double()
)
```

```r
Happy <- recode(ess$happy, "'Extremely happy' = 10; 'Extremely unhappy' = 0;
                '.a'=NA; '.b'=NA; '.c'=NA")
Happy <- as.numeric(as.character(Happy))
ess$Happy <- Happy
rm(Happy)
```

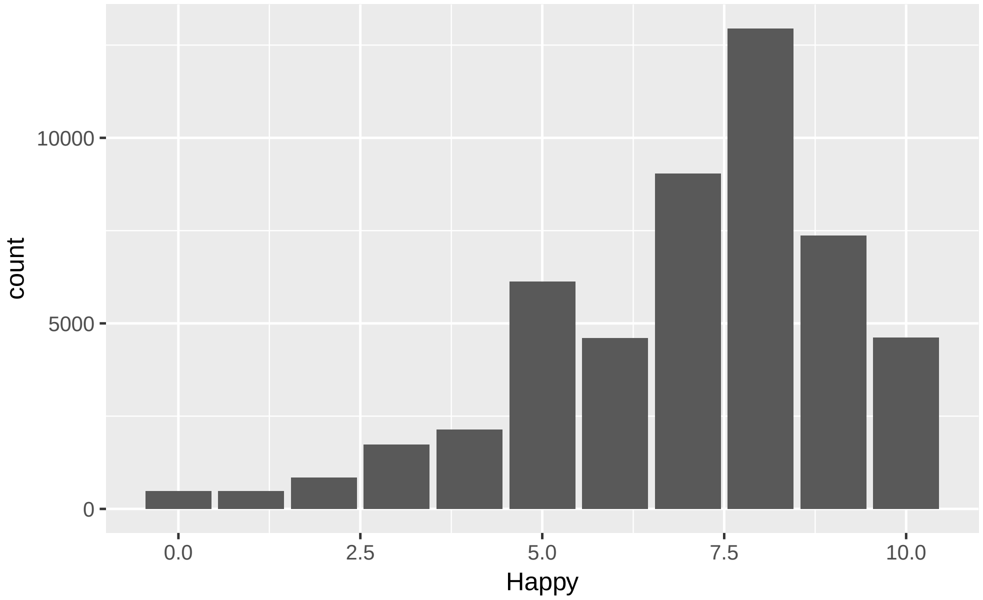
Multilevel regression

These data come from the European Social Survey. There are 50,781 respondents in 26 countries. Data were collected in 2010/11. The outcome variable of interest consistes of responses to the question “Taking all things together, how happy would you say you are?” Possible responses range from 0 (extremely unhappy) to 10 (extremely happy). Intermediate levels aren’t labelled on the questionnaire.

```r
library(ggplot2)
xtabs(~Happy, ess)
```

```
Happy
    0     1     2     3     4     5     6     7     8     9    10 
  473   472   845  1733  2133  6128  4597  9031 12949  7365  4615 
```

```r
ggplot(ess, aes(x = Happy)) + geom_bar()
```



We are going to use linear regression. Strictly speaking, this is an ordinal variable, but with that many levels I think most people would opt to use linear regression. We are going to investigate a range of individual level (age, sex) and country level (inequality, gdp per capita) variables for association with happiness. The data need a little bit of cleaning first. This is mainly to remove missing data that use STATA coding.

```r
toNA <- function(var, val = ".a") {
    var[var %in% val] <- NA
    var
}

ess$Age <- as.numeric(toNA(ess$agea, ".a"))
ess$Sex <- factor(toNA(ess$gndr))

ess$Marital <- recode(ess$marsts, "'.a' = NA; '.b' = NA; '.c' = NA; '.d' = NA;
                      'In a legally registered civil union' = 'Married';
                      'Legally divorced/civil union dissolved' = 'Divorced';
                      'Legally married' = 'Married'; 'Legally separated' = 'Divorced';
                      'None of these (NEVER married or in legally registered civil' = 'Single';
                      'Widowed/civil partner died' = 'Widowed'; '' = NA")

ess$Marital <- factor(ess$Marital)
ess$Marital2 <- recode(ess$Marital, "'Married' = 'Married';
                       'Divorced' = 'Single'; 'Single' = 'Single'; 'Widowed' = 'Single'")
names(ess)[14:15] <- c("GDP", "Gini")

ess$EDUyears <- as.numeric(toNA(ess$eduyrs, c(".a", ".b", ".c")))
```

You would normally want to do some descriptive statistics on these variables to check that they are OK. For example, there are some people who say they had 40 or more years of education, which seems excessive!

Simple model, random intercept

We are treating country as the level 2 variable. Let's try a model with sex, age, marrital status and gini coefficient.

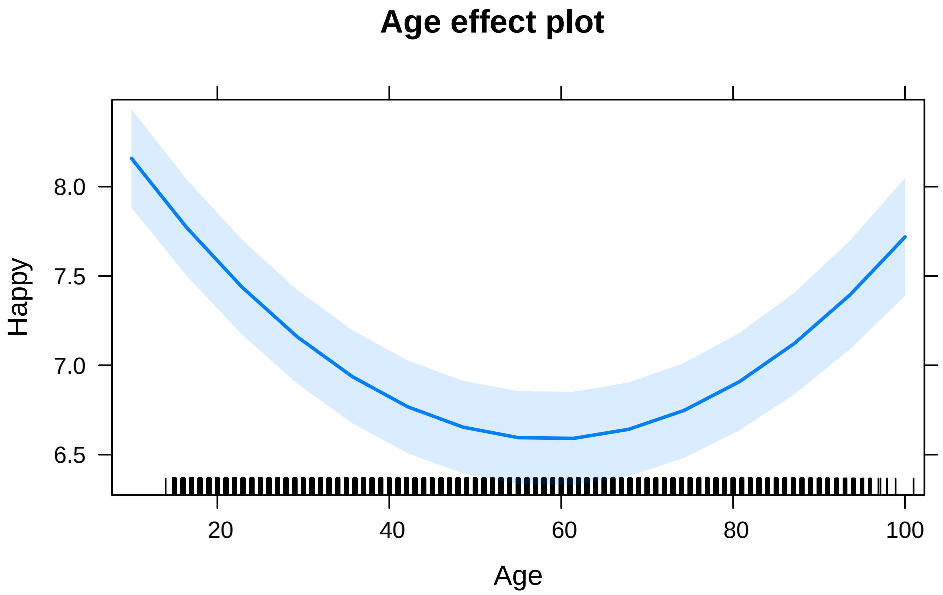
```r
require(arm)
m1 <- lmer(Happy ~ Age + I(Age^2) + Marital + Sex + Gini + (1 | cntry), data = ess)
```

```
Warning: Some predictor variables are on very different scales: consider
rescaling
```

```r
display(m1)
```

```
lmer(formula = Happy ~ Age + I(Age^2) + Marital + Sex + Gini +
    (1 | cntry), data = ess)
               coef.est coef.se
(Intercept)     9.78     1.07
Age            -0.08     0.00
I(Age^2)        0.00     0.00
MaritalMarried  0.22     0.08
MaritalSingle   0.20     0.04
MaritalWidowed -0.10     0.05
SexMale        -0.10     0.03
Gini           -3.32     3.63

Error terms:
 Groups   Name        Std.Dev.
 cntry    (Intercept) 0.58
 Residual             1.91
---
number of obs: 19416, groups: cntry, 20
AIC = 80475.7, DIC = 80362.5
deviance = 80409.1
```

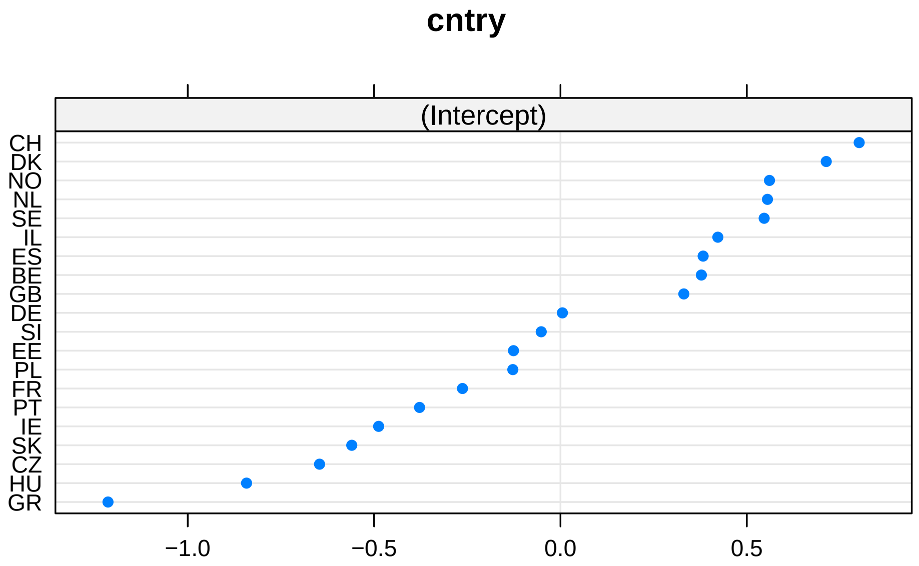
```r
plot(Effect("Age", m1))
```



The warning message is because A^2 is much larger than the other variables. ML estimators can have a problem with this, so it is a wise precaution to rescale (by dividing the variable by a constant). We might want to look at how the random intercepts are distributed.

```
lattice::dotplot(ranef(m1, condVar = FALSE))
```

```
$cntry
```



Now try adding some more variables.

```
m2 <- lmer(Happy ~ Age + I(Age^2/1000) + Marital2 + Sex + Gini + I(GDP/1000) +
    EDUyears + (1 | cntry), data = ess)
display(m2)
```

```
lmer(formula = Happy ~ Age + I(Age^2/1000) + Marital2 + Sex +
    Gini + I(GDP/1000) + EDUyears + (1 | cntry), data = ess)
               coef.est coef.se
(Intercept)     7.99     0.89
Age            -0.10     0.00
I(Age^2/1000)   0.85     0.04
Marital2Single -0.15     0.07
SexMale        -0.07     0.03
Gini           -0.29     2.79
I(GDP/1000)     0.02     0.01
EDUyears        0.07     0.00

Error terms:
 Groups   Name        Std.Dev.
 cntry    (Intercept) 0.43
 Residual             1.90
---
number of obs: 19192, groups: cntry, 20
AIC = 79237.7, DIC = 79130.8
deviance = 79174.2
```

Apart from Gini, these are all statistically significant. Let's see if the effect of, say, education, varies by country.

```
ess$GDPk <- ess$GDP/1000
m3 <- lmer(Happy ~ Age + I(Age^2/1000) + Marital2 + Sex + Gini + GDPk + EDUyears +
```

```r
    (1 + EDUyears | cntry), data = ess)
display(m3)
```

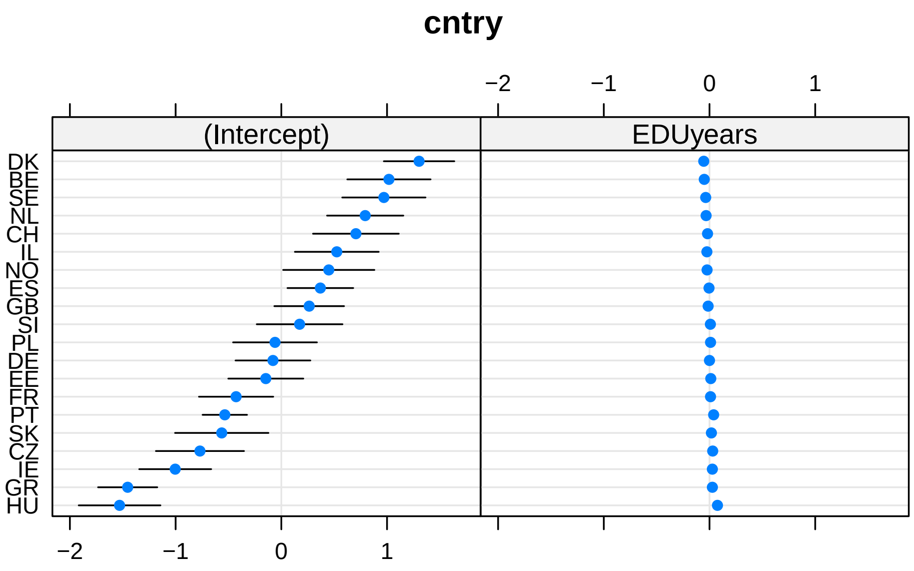
```
lmer(formula = Happy ~ Age + I(Age^2/1000) + Marital2 + Sex + 
    Gini + GDPk + EDUyears + (1 + EDUyears | cntry), data = ess)
               coef.est coef.se
(Intercept)     8.30     0.78  
Age            -0.10     0.00  
I(Age^2/1000)   0.85     0.04  
Marital2Single -0.14     0.07  
SexMale        -0.07     0.03  
Gini           -0.21     2.37  
GDPk            0.01     0.00  
EDUyears        0.07     0.01  

Error terms:
 Groups   Name        Std.Dev. Corr  
 cntry    (Intercept) 0.84           
          EDUyears    0.04     -0.92 
 Residual             1.90           
---
number of obs: 19192, groups: cntry, 20
AIC = 79176.9, DIC = 79067.5
deviance = 79110.2 
```

```r
anova(m2, m3)
```

```
refitting model(s) with ML (instead of REML)

Data: ess
Models:
m2: Happy ~ Age + I(Age^2/1000) + Marital2 + Sex + Gini + I(GDP/1000) + 
m2:     EDUyears + (1 | cntry)
m3: Happy ~ Age + I(Age^2/1000) + Marital2 + Sex + Gini + GDPk + 
m3:     EDUyears + (1 + EDUyears | cntry)
   Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)    
m2 10 79194 79273 -39587    79174                             
m3 12 79134 79229 -39555    79110 64.042      2   1.24e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lattice::dotplot(ranef(m3, condVar = TRUE))
```

```
$cntry
```



It doesn't look like very much variation, but it is statistically significant. Might the variation in EDUyears' effect be associated with levels of GDP?

```
m4 <- update(m3, . ~ . + EDUyears:GDPk)
```

```
Warning: Some predictor variables are on very different scales: consider
rescaling
```

```
display(m4)
```

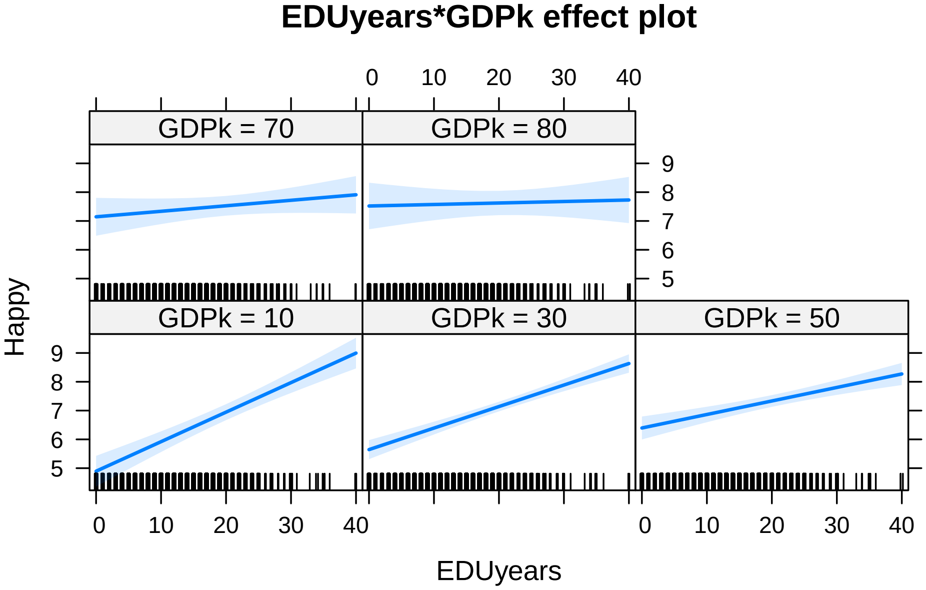
```
lmer(formula = Happy ~ Age + I(Age^2/1000) + Marital2 + Sex +
    Gini + GDPk + EDUyears + (1 + EDUyears | cntry) + GDPk:EDUyears,
    data = ess)
               coef.est  coef.se
(Intercept)     7.37     0.80
Age            -0.10     0.00
I(Age^2/1000)   0.84     0.04
Marital2Single -0.14     0.07
SexMale        -0.07     0.03
Gini           -0.36     2.34
GDPk            0.04     0.01
EDUyears        0.12     0.01
GDPk:EDUyears   0.00     0.00

Error terms:
 Groups   Name        Std.Dev. Corr
 cntry    (Intercept) 0.68
          EDUyears    0.02     -0.88
 Residual             1.90
---
number of obs: 19192, groups: cntry, 20
AIC = 79181.8, DIC = 79041
deviance = 79098.4
```

```
anova(m3, m4)
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: ess
Models:
m3: Happy ~ Age + I(Age^2/1000) + Marital2 + Sex + Gini + GDPk +
m3:     EDUyears + (1 + EDUyears | cntry)
m4: Happy ~ Age + I(Age^2/1000) + Marital2 + Sex + Gini + GDPk +
m4:     EDUyears + (1 + EDUyears | cntry) + GDPk:EDUyears
   Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
m3 12 79134 79229 -39555    79110
m4 13 79124 79227 -39549    79098 11.788      1  0.0005961 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(Effect(c("EDUyears", "GDPk"), m4), x.var = "EDUyears")
```



You can see that the effect of education on happiness does vary pretty substantially with GDP. However, remember that there aren't very many countries, so this has to be treated with a little bit of caution. However, it looks as though the effect of education on happiness has a much bigger effect in poorer countries than it does in richer ones.

Homework

1. Use the data `5.1.txt`, which is from the Scottish Youth Cohort Trends dataset. It is a comma-delimited file, so you can read it using `read.csv('5.1.txt')`. You should get 33,988 rows and 9 variables, as follows.

- caseid: student id
- schoolid: School id
- score: Point score calculated from awards in Standard grades taken at age 16. Scores range from 0 to 75, with a higher score indicating a higher attainment
- cohort90: The sample includes the following cohorts: 1984, 1986, 1988, 1990, 1996 and 1998. The cohort90 variable is calculated by subtracting 1990 from each value. Thus values range from -6 (corresponding to 1984) to 8 (1998), with 1990 coded as zero
- female: Sex of student (1 = female, 0 = male)

- sclass: Social class, defined as the higher class of mother or father (1 = managerial and professional, 2 = intermediate, 3 = working, 4 = unclassified)
- schtype: School type, distinguishing independent schools from state-funded schools
- (1 = independent, 0 = state-funded)
- schurban: Urban-rural classification of school (1 = urban, 0 = town or rural)
- schdenom: School denomination (1 = Roman Catholic, 0 = non-denominational)

2. The outcome variable is `score`. Try to develop a suitable multilevel model using these data.