mlm practical 2

January 20, 2020

1 Guided Practical 2 - Random Intercept Models

This session we will show you:

```
1) the syntax of multilevel models in R.
2) how to establish if, and to what extent, variation occurs at different levels of clusteri
3) how to compare the fit of different models
4) how to extract, and plot, random effects
5) how to add  explanatory variables to the model.
```

1.0.1 Introduction

We work with the hedonism data from the ESS.

The data have a two-level hierarchical structure with individual respondents at level 1 and countries at level 2.

We will treat country as a random classification (or grouping variable). The target of inference could be a wider population of countries from which those in the study can be considered a random sample.

It is not clear which countries such a population would contain. In this case, it is more natural to think of the sample data as if they were a set of realisations from some underlying process that could extend through time and possibly space. This process has driven the observations, but the statistics we compute from the observed data refer to a particular point in time and are subject to random fluctuations. We are interested in the underlying process that has generated the data we observe, and use the 'sample' data to make inferences about this process.

Whatever our view about the data generation process, it seems reasonable to argue that the responses provided by indivduals who reside in a particular country will not be independnet observations; breaking the assumption of independence of case which under traditional regression methods. These data are therefore a classic example of the type of data analysed using multilevel models.

1.0.2 Load Required Packages

Two main packages are needed for the analysis in this tutorial.

"foreign" is used to import datasets, created in different statistical software, into R. This package was also used in Tutorial 1.

"lme4" is the main package used to run multilevel models in R. The basic syntax for this package is introuced below. Later tutorials will build on the syntax introduced in this session to run models for different types of dependent variable.

"lattice" is one of the key packages for drawing graphs in R. As demonstrated below, the command "dotplot" provides a simple way through which to plot random effects

[ ]:
```
library (foreign)
library (lme4)
library (lattice)
```

1.0.3 The Dataset

We will analyse data from all 20 countries in the study. The following countries were included in the study: Austria, Belgium, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Netherlands, Norway, Poland, Portugal, Slovenia, Spain, Sweden, Switzerland, and the United Kingdom. The combined sample size for these countries is 36,537.

The dataset contains the following variables:

```
1) age -    Respondent's age in years
2) female - 0 if respondent male, 1 if respondent female
3) eduyrs - Number of years of education of respondent
4) income - Respondent's monthly household income in bands (less than €150, €150-300, €300-500
€1500-2000, €2000-2500, €2500-3000, €3000-5000, €5000-7500, €7500-10000, more than €10000).
```

The data are stored in the file "hedon_intercept.dta" (which is a Stata datafile). They can be imported into R using the "foreign" library. The sytax below imports the dataset and stores it in a dataframe called "hedon"

[ ]:
```
hedon <- read.dta("hedon_intercept.dta")
```

As in Tutorial 1, we can use the "ls()" command to check which objects in the environment. In this case, checking that "hedon" now exists.

[ ]:
```
ls()
```

As always, before undertaking any statstical modelling, it is important to get a sense of your data and the type of variables you have to play with.

The "str()" command provides any overview of an object. In the case of a dataframe, this includes the number of cases, the names of variables and the type of data they store.

The command "summary (hedon)" provides descriptive statistics for each variable in the dataframe, while the command "head (hedon)" displays the first few rows of data stored in the dataframe.

[ ]:
```
str(hedon)
```

[ ]:
```
summary(hedon)
```

[ ]:
```
head (hedon)
```

Looking at the above output, answer the following questions.

1) How many cases are in the dataset?
2) How many variables are in the dataset?
3) What type of variable is "country"?

1.0.4 Running a Variance Component Model

In R, a multilevel model for a continuous outcome can be estimated through the "lmer" command in the "lme4" package.

The syntax for "lmer" is broadly similar to that of the "lm" command used to run OLS regession models in R. That is to say, the output is written to an object, involves a formula listing the dependent variable and any independent variables, as well as a reference to the dataframe that should be read in order to estimate the model. In addition, the "lmer" command includes an argument to inform R of which variable represents the clustering of cases within the data.

When undertaking multilevel modelling, you should start by estimating a null model, or a Variance Component Model. This is a model with no independent variables, it provides an estimate of the extent to which the unexplained variation in an outcome can be attributed to the different levels of clustering within the data (in this case how much of the variation in hedonism score can be attributed to the respondent, and how much to the country in which they live.

The "lmer" command below works as follows :-

1) The estimated model is written to an object called "nullmodel"
2) The dependent variable is "hed" this is entered to the left of the ~ symbol.
3) Any independent variables are listed to the right of the ~, seperated by +. No independent
4) (1|country) defines the random part of the model. In this case, a random intercept is requ
5) The model is to be estimated using the dataframe "hedon"
6) Setting "REML=FALSE" means the model will be estimated through maximum, rather than restric

Models estimated with maximum likelihood will typically yield more robust results that those which use restricted likelihood; however the latter may be useful when estimating complex models since ML estimation can be time consuming or fail to converge.

[ ]:
```
nullmodel <-lmer (hed~(1|country), data = hedon, REML=FALSE)
```

Using the R "summary" command to view the "nullmodel" object provides an overview of the model that was just estimated.

[ ]:
```
summary (nullmodel)
```

The above output should be considered in four main sections.

The output begins by summarising the requested model, it shows model formula, the dataset used and notes that the model was fitted through maximum likelihood.

This is followed by a range of overall model fit statistics, including AIC, BIC and the log-likelihood value.

The section headed "Random effects" details any coefficients/variables that are allowed to vary between clusters (in this case simply the intercept which varies between countries). The figures in this section are used to calculate the Variance Partition Coefficient discussed below.

This section also notes the total number of cases used to estimate the model, and how many groups (clusters) were included in the model. In this case 36527 cases, spread across 20 countries. This information is useful for ensuring your model includes the expected data.

The final section of the output, title "Fixed Effects", provides details of the regression coefficients estimated by the model. These summarise the average impact of each explanatory variable included in the model and are interpreted in the same way as coefficients in a single level OLS regression model. In this case, the value of the Intercept suggests that the mean score on the hedonism variable is -0.20.

1.0.5 Calculating Variance Partition Coefficients (VPC)

The VPC is an estimate of the proportion of variance in the outcome that is accounted for by each level of data within the model.

The variance explained at level 2 (in this case between countries) is calculated as,

variance at level 2/ (variance at level 1+ variance at level 2).

In this case,

0.08977 / (0.88506 + 0.08977) = 0.09208 or 9.2%

That is to say 9.2% of the variation in hedonism within this dataset is attributable to differences between countries, while 90.8% is due to differences at the indivdual level.

1.0.6 Establishing the Significance of a Multilevel Model

The Variance Component Model can provide a formal statsitical test of whether clustering is a dataset is sufficient to require the use of a multilevel model, as oppossed to a single level regression model.

Specifically, the overall model fit of the Variance Component Model is compared to an equivilent single level regression model.

The "lm" command below estimates a single level OLS regression of hedonism with an intercept as the only indepedent variable. This model is stored in an object called "singlemodel".

[ ]:
```
singlemodel <- lm (hed~1, data = hedon)
```

Comparing the two models through a chi-squared comparrison of their log-likelihood values provides a formal statistical test as to which model best fits the data. Only if the multilevel model does not show a significant improvment in model fit could it be argued that a single level model might be appropriate for modelling the data.

The command "logLik()" reports the log-likelihood value for a previously estimated model, as shown in the command box below.

[ ]:
```
logLik(nullmodel)
logLik(singlemodel)
```

The likelihood ratio test statistic is calculated as the difference in the log likelihood values for the two models:

LR = 2(51294 – 49651)= 3288 on 1 d.f. (because there is only parameter difference between the models).

Bearing in mind that the 5% point of a chi-squared distribution on 1 d.f. is 3.84, there is overwhelming evidence of country effects on hedonism. This indicates that the multilevel model with country effects is most appropriate for modelling the data in this dataset.

An alternative method to the manual calculation shown above is to use the "anova" command as shown below. The key figure in this output is Pr(>Chisq) which indicates the statsitical significance of the difference in model fit between the two models. In this case the p-value indicates a highly significant improvmet in model fit. This reflects the calculation above, with any minor differences attributable to rounding error.

[ ]:
```
anova (nullmodel, singlemodel)
```

1.0.7 Exploring Counrty Level Effects

Having established that clustering does appear to exist within the data, an interesting question is which counries, on average, have the highest levels of hedonism? And how does each country vary from the "average" county?

These questions are addressed by reflecting on the random effects, and specifically the random intercepts, estimated within these models. Within a model object, such as "nullmodel", which has been estimated through the "lme4" package, random effects are accessed through the "ranef" command.

For example, the commands below create an object called "nullrand" which contains details of the random intercept for each country, as estimated within "nullmodel"

[ ]:
```
nullrand <- ranef(nullmodel)
nullrand
```

Random effects are often presented as caterpillar plots. These show not only the level 2 residual for each country with regards to the intercept, but also a 95% confidence interval and sorts level 2 clusters in terms of the nature/extent of their variation from the average.

The syntax below produces a caterpillar plot of the random intercepts associated with each country in "nullmodel". In the event that a model contains more than one random parameter (a random coefficient model, introduced in the next tutorial), seperate plots will be produced for each ransom effect.

[ ]:
```
dotplot(ranef(nullmodel))
```

Recall that random effects respresent how a particular level 2 unit (in this case a country) varies from the average, measured in standard deviations. Hence the figure 0.0 in the graph above represents the "average country". Countries for which the confidence interval of the random effect does not cross 0.0 can be said to be different from the average.

Countries with negative random effects have, on average, lower levels of hedonism. Positive random effects are assoicated with higher average levels of hedonism, i.e. on average Poland appears the least hedonistic country, while Denmark appears the most.

1.0.8 Running a Random Intercept Model

Adding explanatory varaibles to the Variance Component Model turns the model into a Random Intercept Model. Comparing Random Intercept models to the previously estimated Variance Component Model will indicate the extent to which differences between countries remain after controlling for the independent variabes thata re then included. Put another way, to what extent do any independent variables included in the Random Intercept Model help explain differences between countries?

The next model considers only one independent variable (the respondent's age) but individual income, education and gender will be considered below.

In essence, the research question has become 'Do differences in hedonism between countries remain after controlling for individual age?'

As highlighted in Tutorial 1, it is standard practice to centre any continuous independent before estimating a multilevel model. The following syntax calculates a new variable, "agecen" containing each respondent's age centered around the grand mean age of the sample. The final command confirms that variable is now part of the "hedon" dataframe.

Recall from Tutorial 1 that centering the age variable makes interpretation of the model easier since the coefficient associated with the intercept can be interpreted as the average level of hedonism for a respondent of mean age.

[ ]:
```
meanage <- mean (hedon$age, na.rm=TRUE)
hedon$agecen = hedon$age-meanage
summary (hedon)
```

The command box below shows the syntax needed to run a Random Intercept Model with centered age as the only independent variable. The syntax is esentially the same as for the Variance Component Model with the addition of "agecen +" to the right-hand side of the ~.

This is essentially the same as adding an explanatory variable to an OLS regression estimated using the "lm" command.

The Random Intercept model is stored in an object called "ri1".

[ ]:
```
ri1 <-lmer (hed~ agecen + (1|country), data = hedon, REML=FALSE)
summary (ri1)
```

The coefficient for age is -0.017 and can be interpreted as follows. Older people are less hedonistic; for every extra year, hedionism drops, on average, by 0.017. The t-value for this coefficient is highly statsitically significant.

Compared to the variance components model, the intercept has hardly changed, nor has the level 2 variance. This suggests that the answer to the question, "Do differences in hedonism between countries remain after controlling for individual age?" is 'Yes'.

However, as above, a formal answer to the question of whether significance variance exists at level 2 (the country level) of the model is best achieved by comparing the model fit of this model (ri1) with the equivalent single level model.

Using the command box below construct a single level OLS regression model of the effect of the centered age variable on hedonism; store that model in an object, and use the "anova" command to establish whether model "ri1" better fits the data compared to a single level model. What do you conclude?

[ ]:
```
sl1 <- lm (hed~1+agecen, data = hedon)
summary (sl1)
anova (ri1, sl1)
```

1.0.9 Checking if the inclusion of an explanatory variable improves model fit

Just as the Anova/chi-squared test can be used to compare the fit of a multilevel model to the eqivilent single level model, so it can be used to compare a Variance Component Model to a Random Intercept Model. This formally addresses the question, "Has adding an explanatory variable to the model improved the fit of the model?"

Recall that the Variance Component Model was stored in an object called "nullmodel". Hence, the "anova" command below should provide a test of which model best fits the data.

[ ]:
```
anova (ri1, nullmodel)
```

Log-likelihood ratio tests (the formal test behind the Anova command) require that both models have been run on identical samples. Yet due to missing data on the age variable, the sample size for the random intercept model is 36364 compared to 36527 for the Variance Component Model. This is reflected in the message "models were not all fitted to the same size of dataset"

The command below creates a new object "nullmodelx". This repeats the Variance Component Model estimaed above but restricts the data used to those cases which were also used in the "ri1" model. This is achieved through the argument,

data=na.omit(hedon[, all.vars(formula(ri1))])

This command reduces the dataset by removing any cases which have missing data "na.omit" on any of the variables used in the ri1 model "all.vars(formula(ri1)"

[ ]:
```
nullmodelx <-lmer (hed~(1|country), data=na.omit(hedon[ , all.
↪vars(formula(ri1))]), REML=FALSE)
summary (nullmodelx)
```

```
anova (ri1, nullmodelx)
```

Despite model "nullmodelx" reporting a sample size of 36364 (which is identical to the sample size in model "ri1") the "anova" command still fails, reporting "all models must be fit to the same data object".

As an added level of robustness, when conducting a Log-likelihood Ratio Test, R requires not only that the two models been compared have identical sample sizes, but also that the underlying dataset used in both models were identical. In order to conduct the comparrison of the two models it is therefore neccssary to rerun model "ri1" using the same dataset constraint as was employed to estimate model "nullmodelx".

This test formally confirms that the Random Intercept Model, including a respondent's age, is a better fit for the data than the null model which included no independent variables.

[ ]:
```
ri1x <-lmer (hed~ agecen + (1|country), data=na.omit(hedon[ , all.
 ↪vars(formula(ri1))]), REML=FALSE)
summary (ri1x)
anova (nullmodelx, ri1x)
```

While the Level 2 variance was little changed between the null model and the random intercept model including age, suggesting that adding respondent age to the model did little to explain differences in hedonism between countries, there is a reduction in the Level 1 variance (0.7869 in model ri1x compared to 0.8854 in nullmodelx). This suggests that including respondent age in the model does explain some of the variation in hedonism between individuals.

1.0.10 Controlling for a Respondent's Gender and Years of Education

Consider the variables "female" and "eduyrs". Expand the random intercept model to consider the impact of an individual's gender and experience of education on hedonism. For each model you chose to run you should :-

```
1) Centre independent variables as needed.
2) Establish if adding additional variable(s) improves your model fit.
3) Interpret the average impact of each independent variable in substantive terms
4) Reflect on the extent to which introducing further indepndent variables changes the amount
5) Calculate a VPC for each model you run, and interpret any changes in this between models.
6) Extract the random intercepts for your final model.  Compare the order of countries, and the
```

[ ]:
```

```

1.0.11 Advanced Issue

Reflecting on the diagnostic commands introduced in Tutorial 1, can you produce residual plots for one of the multilevel models run in this tutorial? You might consider a hedroskedasity check, and/or a q-q plot.

[ ]:
```

```