Workshop 1 - Getting Ready

JPS

Introduction

Today's practical will refresh some of the regression modelling techniques that you learnt with Albert. In doing so, we will provide a bit of a taster of the modelling challenges that we will encounter in this module. In the second part of this session you also have an introduction to the applied part of the syllabus, where we will review the kind of techniques, questions, R libraries, and datasets that we will explore through the module.

By the end of the day you should make sure to install the libraries that we will be using throughout the year, listed in the Syllabus below. This is to avoid problems of incompatibility during the practical sessions coming up. If you encounter any problems downloading or installing any of the recommended libraries today, we should make sure we resolve them before our next workshop. If any technical issues persist come see me or Jade during our support hours. Otherwise, you can also contact the IT Service Desk (you can get an in-person appointment here) so any issues are resolved in advance of the practical where that library is to be used.

Also, if you are using your own laptop, to avoid potential incompatibilities with the libraries used through the module, it would be important to make sure that you have installed the latest version of R, which in January 2025 is version 4.4.2 'Pile of leaves'. This is the R version that has been used to prepare this module's practicals. You can simply check the R version you are using as that information will be provided in the first line of the automatic output generated after opening R. If you need to update R you can do so by closing R and installing the latest version, available here.

Recap & Taster

Let's review some simple modelling techniques you have already seen and introduce a few new ones. To do so we are going to employ the European Social Survey (ESS) to explore some predictors of self-reported happiness. You can get the 2018/19 ESS from Minerva. The ESS is a truly unique dataset, in its geographical coverage but also in the wide range of topics covered, which might be quite relevant to some of your substantive areas of interest, e.g. social trust, attitudes towards the Criminal Justice system, etc. Furthermore, the data is really easy to download, no need to submit an application, just provide some information through a short registration process and it can be downloaded directly from their website (http://www.europeansocialsurvey.org/).

The version of the ESS available on Minerva has got a STATA format (.dta), so we will first need to install the package *foreign*. If you have not saved the dataset in the same folder where your R script is, then you will also need to modify the code below to provide the direction to the folder where you have saved the dataset.

I am going to import the data now, but before doing so I am also going to set scientific notation in R off, as I do not find it useful.

```
options(scipen=999, digits=5) #This is to remove scientific notation.
#setwd("C:/Users/...")
library(foreign)
ess = read.dta("ESS9e01_1.dta") #Importing the data.
```

There are 491 variables in the dataset, let's just keep the handful of variables that we need for our analysis so we can free a lot of memory. These are: gender ('gndr'), age ('agea'), self-reported happiness ('happy'), how

often meet with other people ('sclmeet'), number of people living in the same household ('hhmmb').

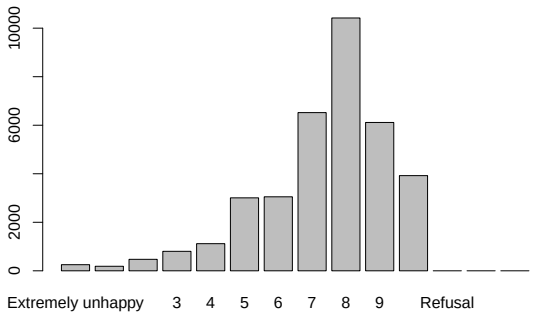
```
vars = c("gndr","agea","happy","sclmeet","hhmmb") #Selecting variables into an object.
ess = ess[vars]  #Keeping the variables selected above in the object 'vars'.
```

Next, we can take a quick look at these variables to see how they are measured, how they are distributed, whether they are affected by missing cases, etc. Here, we are only going to run some frequency tables for the categorical variables and histograms for the continuous data to move on to the modelling exercises quickly. However, you should dedicate far more time than this to explore your data before you start any modelling. This is key to familiarise yourself with the data, anticipate potential issues, and inform modelling decisions.

```
head(ess) #This command helps you gain a first impression of the data in a way
#that is not to overwhelming.
prop.table(table(ess$gndr, useNA = "ifany"))
prop.table(table(ess$sclmeet, useNA = "ifany"))
table(ess$happy, useNA="ifany")
hist(ess$agea)
hist(ess$hhmmb)
```

We can provide a first approximation to the research question by exploring the distribution of self-reported levels of happiness.

```
table(ess$happy, useNA="ifany")
plot(ess$happy)
```



More people report to be happy than not, with 8 (out of 10) as the most common value (the mode) of happiness reported. If we want to calculate the average of that distribution we first need to transform 'happy' into a numeric variable. This is a bit tricky since 'happy' is a factor with both characters and numbers as levels. One approach we can take to transform this variable is by combining two *ifelse* commands, one for each of the character levels ('extremely happy' and 'extremely unhappy'), with the rest of the values being transformed directly into numeric using *as.numeric(as.character)*.

```
ess$happyrec = ifelse(ess$happy=="Extremely unhappy",0,
               ifelse(ess$happy=="Extremely happy",10,as.numeric(as.character(ess$happy))))
class(ess$happyrec)   #This is to check that happyrec is now a numeric variable.
table(ess$happyrec, useNA="ifany")  #This is to check that the transformation went ok.
```

Now that we have 'happyrec' as a numeric variable, we can calculate the average using *mean*, but to do so we need to specify the option *na.rm=TRUE* to drop the 148 missing cases in 'happyrec'.

```
mean(ess$happyrec)
mean(ess$happyrec, na.rm=TRUE)
```

At this point we can start doing some modelling. Let's start with a standard linear model. As we saw above,

the variable is approximately normally distributed, so even though it is not truly a continuous variable a linear model should still be ok.

```
linear = lm(happyrec ~ gndr + agea + sclmeet + hhmmb, data=ess)
summary(linear)
```

```
##
## Call:
## lm(formula = happyrec ~ gndr + agea + sclmeet + hhmmb, data = ess)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.347 -0.877  0.336  1.273  4.852
##
## Coefficients:
##                                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                    5.182469   0.085654   60.50 <0.0000000000000002 ***
## gndrFemale                     0.039291   0.019667    2.00               0.0457 *
## agea                          -0.002138   0.000592   -3.61               0.0003 ***
## sclmeetLess than once a month  0.877017   0.077470   11.32 <0.0000000000000002 ***
## sclmeetOnce a month            1.562555   0.077316   20.21 <0.0000000000000002 ***
## sclmeetSeveral times a month   1.860119   0.073842   25.19 <0.0000000000000002 ***
## sclmeetOnce a week             2.026604   0.074383   27.25 <0.0000000000000002 ***
## sclmeetSeveral times a week    2.238998   0.073390   30.51 <0.0000000000000002 ***
## sclmeetEvery day               2.269775   0.076251   29.77 <0.0000000000000002 ***
## hhmmb                          0.155984   0.007986   19.53 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.85 on 35490 degrees of freedom
##   (515 observations deleted due to missingness)
## Multiple R-squared:  0.0809, Adjusted R-squared:  0.0807
## F-statistic:  347 on 9 and 35490 DF,  p-value: <0.0000000000000002
```

Ok, let's interpret these results through a series of Questions:

1. How well does this set of explanatory variables explain differences in self-reported happiness?

We can look at the R^2 for that. Only 8% of the individual variability in self-reported happiness is explained by the variables included.

2. According to the model, what is the effect of gender on happiness?

To answer this question it is useful to consider the mathematical form of the model we have estimated: $happiness = \beta_0 + \beta_1 * female + \beta_2 * age + \beta_{3,...,8} * sclmeet_{1,...,6} + \beta_9 * hhmb$

The effect of gender is picked up by the β_1 estimate, which is denoted as 'gndrFemale' in the summary table of our regression model above. Therefore, the answer is 0.04, which means that on average, and after controlling for household members, frequency of contacting others, and age, women reported happiness is 0.04 points higher in a scale of 1 to 10.

3. According to the model, what is the association between the number of household members and happiness?

0.16; this means that on average, and after controlling for the set of explanatory variables, for every additional member living in a household, individuals report to be 0.16 points happier in a scale of 1 to 10.

4. According to the model, how much happier are individuals living in households of three people (including themselves) than those who leave alone?

Again, to answer this question it is useful to think of the mathematical form of the model. Specifically, we want to use this part, $\beta_9 * hhmb$. We know that for those how live alone $hhmb = 1$, and for households of three we have $hhmb = 3$, and we also know from the model results that $\beta_9 = 0.16$, hence, the answer is 0.32 points happier, which I calculated as $0.16 * (3 - 1)$.

5. And how much happier would be those leaving in households of thirteen people than those in households of three.

1.92 points happier (derived as follows, 0.16*(13-1)). If you find this result surprising then you are onto something. So far we have assumed that all the associations between explanatory variables and the response variable are linear, which means that the marginal effect for every additional unit of a given explanatory variable is constant across the range of that variable. That, however, is often not the case. We could expect that at a given point living with more people actually makes individuals unhappy as they might experience overcrowded conditions. In Workshop 4 we will see how to model such non-linear effects.

6. Would you take the association between number of household members and happiness as a causal effect of the former on the latter?

We need to be careful when interpreting results from observational (i.e. non-experimental) data causally. Income, employment status, marital status, area of residence, etc. could be factors affecting both happiness and number of household members, if so, when these factors are not being controlled for, they could be biasing the effect of number of household members and happiness estimated in our model.

Even if we are controlling for such predictors of happiness, it is possible that the association between household members and happiness is the result of reverse causality. That is, we cannot rule out that happiness is also a predictor of the number of people we decide to live with. Notice how this cannot be the case when we consider the association between age and happiness, which can only go in one direction. A similar argument could be made regarding gender if we assume that changes in gender tend to be rare (which should be the case in samples describing the general population). We will learn more about causal effects, and about when should we be suspicious of statements implying *correlation = causation*, in Workshop 3.

7. Which has a stronger effect age or hhmmb? And how much so?

A priori it seems hhmmb, but it is not straightforward since each one is measured in a different scale. One way to resolve this issue would be by re-estimating the model after *standardising* the two variables. To do so you need to calculate new variables where you subtract their mean and divide by their standard deviation. Then, the regression coefficients for all standardised variables are interpreted the same way: the change in the Y variable (happiness) following a one standard deviation increase in the explanatory variable.

```
ess$agea_std = (ess$agea - mean(ess$agea, na.rm=TRUE))/sd(ess$agea, na.rm=TRUE)
ess$hhmmb_std = (ess$hhmmb - mean(ess$hhmmb, na.rm=TRUE))/sd(ess$hhmmb, na.rm=TRUE)
linear2 = lm(happyrec ~ gndr + agea_std + sclmeet + hhmmb_std, data=ess)
summary(linear2)
#This is if you only want to report the estimates of the regression coefficients.
linear2$coefficients
linear2$coefficients[3] #This is to retrieve the coefficient for age_std.
linear2$coefficients[10] #And for hhmmb_std.
```

We can now confirm that the effect of hhmb appears about five times stronger than that of age.

8. How could you test whether the effect of age on happiness varies by gender?

You could test that through an *interaction* between age and gender.

```
linear3 = lm(happyrec ~ gndr*agea + sclmeet + hhmmb, data=ess)
summary(linear3)
linear3$coefficients[3] #This is now the effect of age but only for men
linear3$coefficients[11] #And this is the effect of age for women
```

After including the interaction between gender and age we have separate effects of age for men and women. We can see that the negative association of age on happiness is only present in women; for men age appears to have no effect on happiness.

9. If you wanted to explore the effect of age on hhmmb, would you still use a linear model?

You can still use a linear model, however, given how right-skewed hhmmb is you will find that your residuals might not be normally distributed. We can adjust this problem by 'normalising' hhmmb, which could be approximated using a *log-transformation*.

```
hist(ess$hhmmb)   #Strongly right-skewed.
linear4 = lm(hhmmb~ gndr + agea, data=ess)
summary(linear4)
hist(linear4$residuals)  #The residual's distribution is only slightly right-skewed.
ess$hhmmb_log = log(ess$hhmmb)
hist(ess$hhmmb_log)  #Still not entirely normally distributed but closer than before.
linear5 = lm(hhmmb_log ~ gndr + agea, data=ess)
summary(linear5)
hist(linear5$residuals)  #The right-skewed tail is not there anymore.
```

10. Estimate the following model $log(hhmmb) = \beta_0 + \beta_1 * female + \beta_2 * age$ and report the difference in the number of household members between men aged 18 and men aged 28.

Remember that in a model when the response variables has been transformed, you need to back-transform the model estimates if you want to interpret them in terms of the original scale of the response variable.

```
linear5$coefficients[3]  #This is to select the regression coefficient for age in linear5.
linear5$coefficients[1]  #This is to select the intercept in that same model.
exp(linear5$coefficients[1]+linear5$coefficients[3]*18)  #The estimated number of
#household members for a man aged 18.
exp(linear5$coefficients[1]+linear5$coefficients[3]*28)  #The estimated happiness for a
#man aged 28.
```

11. What type of model would you use to predict the probability of a woman living alone at age 68?

If you see hhmmb as a binary (whether living alone or not), you could model that using *logistic regression*.

```
table(ess$hhmmb, useNA = "ifany")
ess$alone = ifelse(ess$hhmmb==1,1,
            ifelse(is.na(ess$hhmmb), NA, 0))
table(ess$alone, useNA="ifany")  #This is to check that the transformation went ok.
logit1 = glm(alone ~ gndr + agea, family="binomial", data=ess)
summary(logit1)
```

12. According to the last model, what are the odds of women living alone compared to men, and what is the probability of a woman aged 68 living alone?

```
exp(logit1$coefficients[2])  #The odds ratio of living alone for a woman compared to
#a man, holding age constant.
```

To transform log(odds) (the scale used in logistic regression) for a given reference category into probabilities you can use the following formula: $prob = odds/(1 + odds)$.

```
odds = exp(logit1$coefficients[1] + logit1$coefficients[2] + logit1$coefficients[3]*68)
#The odds of a woman aged 68 living alone vs living with others.
odds/(1+odds)  #The probability of a woman aged 68 living alone.
```

If you have any questions regarding what we have covered just ask. If you did manage to follow everything then move on to review the content of the module included below, and make sure that you are able to download and install all the libraries listed there.

Syllabus

Workshop 2. Selecting explanatory variables

Through Workshops 2, 3 and 4, we will learn some useful principles that we ought to consider when deciding how to choose the set of explanatory variables to be used in a regression model. In Workshop 2 we will learn a key consideration that should preempt all others when it comes to decide what explanatory variables should be included in our model. Namely, we need to identify what is the purpose of our model. We will see how we should differentiate between models - and research questions - seeking to *predict* from those seeking to *explain*.

When our goal is to predict then we should include any variable that can help us improve predictions of the outcome of interest. We will see how unsupervised variable selection (basically machine learning) procedures like *stepwise regression* can help us in deciding which is the set of variables that best predict our outcome.

When our goal is to *explain*, i.e. to estimate a causal effect, then we should follow a supervised modelling process. This means that we (the researcher) should decide what variables to include in the model. The selection of variables should be undertaken according to theoretical insights, and a few technical considerations that we will learn through the module. One of those is the problem of multicollinearity, which we will explore in this workshop. Multicollinearity could take place when we include too many variables in our model, and/or a few variables that are highly correlated. To detect whether multicollinearity is present in our models, in the second part of the practical we will use the *Variance Inflation Factor (VIF)*.

To see the difference between these two modelling strategies we will undertake exercises seeking to: i) predict custodial sentences in the Crown Court; and ii) test whether aggravating factors have a stronger effect in determining sentence severity than mitigating factors.

To be able to use the commands associated with these new techniques, plus a few other data manipulation tools (such as the %>% operator or the partition of a sample), we need to make sure that we can install and load the following libraries: **MASS**, **caret**, **regclass** and **dplyr**. You can do so using the *Packages* menu in RStudio (bottom-right corner) or directly through the R Console using the command *install.packages("name of the library")*. Once installed, check that the libraries are ready using *library("name of the library")*.

Workshop 3. Path analysis and the causal framework

This week we will learn how to classify explanatory variables based on whether they act as confounders, mediators or colliders. We will see how, in order to be able to estimate causal effects, we should identify and control for as many confounders as we possibly can, but also how we should not do this indiscriminately, since there are variables (such as *colliders*) we should never control for. In addition, we will learn two new specific techniques: *DAGs* (a type of causal diagrams), and *path analysis*. The former will help us present our causal assumptions more clearly, while the latter allows us to disentangle direct from indirect effects.

These new modelling considerations and technqiques will be put in practice through three exercises. First we will test Tyler's (1990) procedural justice model by considering how perceptions of police fairness affect crime directly, and indirectly through its effect on perceptions of police legitimacy. In the second and third exercise we will explore the gender pay gap. We will do so using a dataset of academic salaries from one american university, and the UK Labour Force Survey. Emphasis will be placed on testing not only potential gender disparities in salary, but also in understanding the indirect mechanisms through which such differences might be taking place.

To be able to use the new commands explored in this workshop (and to be able to load a dataset on academic salaries), we need to install and load the following libraries: **ggdag**, **mediation**, and **car**.

Workshop 4. Non-linear effects

In this workshop we will see how we can explore non-linear associations between variables, i.e. associations that can chage their strength and sign (e.g. a positive relationship turning negative), across the range of

values captured in our variables. To do so we will practice two new techniques *polynomial regression* and *loess* curves.

The workshop is composed of three exercises. In the first exercise we will use the study on academic salaries to explore the research question: Do salaries increase with the number of years since the PhD was obtained? In the second exercise we will use the Labour Force Survey to explore the relationship between age and salaries in the UK. Specifically we will seek to answer the following research question: Are salaries directly proportional to years of experience?

To be able to use *loess* curves in this workshop, we need to install and load **ggplot2**.

Workshop 5. Time-series

In the first exercise we will go step by step detailing all the features that should be considered when describing and modelling time-series. We will learn how to plot and decompose time-series into its different elements (seasonality, trend, residuals), and how to build *ARIMA models* by carefully considering each of the model's components. For this exercise we will use data from a bike sharing company, capturing daily count of bikes rented. We will see different exploratory techniques to learn from this particular time-series. We will then use that knowledge to model it and forecast future values.

In the second exercise we will also estimate ARIMA models, only now we will follow a *data driven* approach, i.e. unsupervised. We will use these models to assess the impact of the new sentencing guidelines. Specifically, we will test whether the guidelines can be blamed for the increase in sentence severity that has been observed in England and Wales over the last decades.

To run time-series analysis we will use the library **tseries**, and to make forecasts based on our time-series models we will use **forecast**.

Workshop 6. Data reduction techniques

This week we will learn about two data reduction techniques – *Principal Components Analysis* (PCA) and *Cluster Analysis*. Data reduction is the transformation of certain aspects of data into a simplified form. In PCA this is done by summarising and combining variables. In short, this means calculating a set of new variables (or principal components) which are fewer in number than the original variables but which retain most of the information. The principal components are less interpretable than the initial variables, but they create an easier way to explore and visualise data. Cluster analysis is a data reduction technique in that it summarises observations of data by aggregating them into groups. There are many methods of clustering, but this lecture will focus on the method of *k-means clustering*, which is a form of 'unsupervised learning'. It is used to group similar observations in a dataset together which can reveal underlying patterns.

There will be two exercises in the workshop, each of which will focus on one of the techniques introduced. Both will use a dataset containing US State Violent Crime Rates, and we will use this as a way to illustrate the kinds of insights that can be gained. Exercise 1 will use PCA to determine the number of meaningful components in the data and the extent to which they can explain the variability between states. In Exercise 2, we will develop a classification of US states in terms of their violent crime rates by using a K-means cluster analysis. We will explore the characteristics of each group and consider how meaningful the classification is.

The libraries that are required for this week's practical are **tidyverse**, **gridExtra**, **cluster**, **factoextra**, and **ggpubr**.

Workshop 7. Data quality

Here we will design and apply various adjustments to improve the validity of our estimations in the presence of missing data and other sampling issues leading to selection bias. Specifically, we will use *probability weights* and *imputation methods*, to adjust datasets derived from different surveys. These are: the European Social Survey, the Crown Court Sentencing Survey, and the Leeds Parks Survey.

The workshop is composed of two exercises. In the first exercise we will practice *mean*, *regression* and *multiple imputation* to adjust for problems of item-missingness in the Crown Court Sentencing Survey. As we did in Week 2 we will try to estimate the relative importance of different sentencing guidelines' factors in determining custodial sentences, but this time we will be able to use the degree of offence seriousness; the most important factor according to the guidelines, which has been normally discarded from studies using this survey because it is severely affected by non-response. In the third exercise you are requested to calculate probability weights for the Leeds Parks Survey (a survey created using non-probability sampling) using data from the Census, and then provide an adjusted estimate for the proportion of Leeds residents who seldom or never visit a park.

In this workshop we will use some new libraries: **survey**, to apply probability weights, and **mice**, to undertake multiple imputation.

Workshop 8. Hierarchical data

In this workshop we will work with datasets where cases are not independently collected, but nested within clusters such as courts (where sentences are imposed), or countries (where survey participants live). We will learn the different approaches that can be used to adjust for this hierarchical structure (i.e. taking within cluster correlations as a nuisance), and we will also learn how to interpret such hierarchical structures using multilevel modelling.

The workshop is composed of a single fully guided exercise. In the first part we will assess how to take into account the fact that survey participants in the European Social Survey are clustered within countries. To do we will explore the *sandwich estimator* and *fixed effects* models. As we learn how to adjust for within cluster correlations we will also try to estimate the association between household income and trust in others. In the second part of the exercise we will employ *multilevel models* to explore two research questions: Is the variability in trust between countries larger than between individuals within the same countries? And, is the association between income and trust uniform across countries?

To apply the *sandwich estimator* you will need to install **sandwich**, to specify and test *multilevel models* we will use **lme4** and **lmtest**. In addition, we will also learn how to report our model results visually using **sjPlot** and **glmmTMB**.

Workshop 9. Longitudinal data

In this workshop we will practice using one modelling approach for longitudinal data: *growth curve models*. We will employ it to model individual trajectories across time, assessing whether they converge or diverge with the average trajectory in the population.

This technique will be explored using sentencing data from the Czech Republic. We will use *growth curve models* to explore changes in the sentencing practice across judges as they become more experienced. Most of the exercises is fully guided, but in the last part you are requested to specify a random slopes model (such as those used in Workshop 8), to assess whether between judges disparities change as judges become more experienced.

Workshop 10. Crime mapping

R Libraries: **leaflet**, **leaflet.extras**, **sf**, **tmap**

Workshop 11. Agent-based modelling

Netlogo – freely available for Windows/Mac and Unix from here: https://ccl.northwestern.edu/netlogo/download.shtml

Preparation for next week's workshop

From now on we are going to try and have a 'flipped-lecture' format. This means that to make the most of our workshops you are required to prepare them in advance, so we can use our contact time more efficiently, for example, trying to resolve any issues that you might have encountered or doubts that arise as you prepare the workshop. Next week's workshop is composed of a guided and an unguided exercise. The instructions are available on Minerva. At a minimum you are requested to replicate the procedures followed in the guided exercise (exercise 1). If you want to go even further in your preparations you are also highly encouraged to try resolving the unguided exercise (exercise 2). I have left some hints in the instructions to make it a bit less challenging, but if you feel that it is still too hard do not worry as we will see it together next week and I will upload the solutions on Minerva after we complete that session.

See you all again next week.