# Lab 8: Measuring latent variables in the social world

DATA5207: Data Analysis in the Social Sciences

Dr Shaun Ratcliff

This lab will be run in two parts. First we will walk you through the use of *factor analysis* in *R*. Then you will use the output from this model to complete your assessable group project from this morning.

In the lecture, we explored the use of different methodologies to understand measures of similarities used in the social sciences (also known as data reduction techniques). These methods are usually adopted in social science research to reduce the number of variables included in a model, or to detect structure in the associations between different variables. In this lab, we will primarily focus on the first of these. This strategy is particularly important for measuring characteristics of individuals for which a single question would often not provide an accurate result; due to measurement error, social desirability bias, or because what we are trying to measure is a latent trait where no one question could possibly capture the true picture.

The most common method used to do this in the social sciences is factor analysis. While factor analysis expresses the underlying common factors for an entire group of variables, it also helps researchers differentiate these factors by grouping variables into different dimensions or factors, each of which is ideally uncorrelated with the others.

During this lab, we will walk through the process of fitting a factor analysis to survey data on quality of life. The assessable component of the lab will require you to undertake an analysis of the measure we create using the other methods you have learned in this subject.

## Learning to use factor analysis

Today we will look at factor analysis as a tool to understand quality of life. This is an appropriate method for this kind of research, as there are a number of variables in our dataset that provide information on the outcome we are interested in, and we can better understand the problem if we decompose the information they contain into a single measure, which we can then study further.

We are using the 2018 Australian World Values Survey for this research. This is a survey of Australians, run in conjunction with other similar surveys conducted around the world. It asks respondents more questions than most other surveys, and therefore provides a useful datasource for researchers. A copy of the dataset, code book and questionnaire is available in the zipped folder in the module for this class on canvas. These data come from the Australian Data Archive.

### Examining our data

Using these data, we will examine how different personal characteristics might be associated with higher and lower quality of life. We will do this by fitting a factor analysis to five variables which cover reported happiness, life satisfaction, health, and financial situation. These variables are:

- *Q46.* Taking all things together, would you say you are… (variable V10)
- *Q47.* All in all, how would you describe your state of health these days? Would you say it is… (variable V11)

- *Q48.* Some people feel they have completely free choice and control over their lives, while other people feel that what they do has no real effect on what happens to them. Please use this scale where 1 means "no choice at all" and 10 means "a great deal of choice" to indicate how much freedom of choice and control you feel you have over the way your life turns out. (variable V55)
- *Q49.* All things considered, how satisfied are you with your life as a whole these days? (variable V23)
- *Q50.* How satisfied are you with the financial situation of your household? (variable V59)

Using a factor analysis, we will create a measure for quality of life.

We begin by making sure these are coded the correct way (greater quality of life being the positive score), and we standardise the scores.

Begin by loading your data:

```
wvs.dat <- read.csv("Data/wvs_data.csv")
```

We then recode our variables. As we do this we provide the variables with more intuitive names, to make our work easier. This can be done with the code:

```
library(dplyr)

wvs.dat$happiness <- as.numeric(as.character(recode(wvs.dat$V10,
                                    '1' = '4',
                                     '2' = '3',
                                     '3' = '2',
                                     '4' = '1',
                                    '-2' = NULL)))

wvs.dat$health <- as.numeric(as.character(recode(wvs.dat$V11,
                                     '1' = '5',
                                     '2' = '4',
                                     '3' = '3',
                                     '4' = '2',
                                     '5' = '1',
                                    '-2' = NULL)))

wvs.dat$finances <- recode(wvs.dat$V59,
                                    '-2' = NULL,
                                    .default = wvs.dat$V59)

wvs.dat$satisfaction <- recode(wvs.dat$V23,
                                    '-2' = NULL,
                                    .default = wvs.dat$V23)

wvs.dat$freedom <- recode(wvs.dat$V55,
                                    '-2' = NULL,
                                    .default = wvs.dat$V55)
```

We use the `recode()` function from `dplyr` to both switch around those variables that are coded incorrectly (where the higher value is associated with lower quality of life) and where missing values are included as a

numeric value (such as -2 for `V10`), which will result in these values being included in our factor analysis and adding misleading information into our model. Here we are using the `recode()` function from `dplyr`.

We are almost ready to fit a factor analysis to these data.

To run our factor analysis, we use the `psych` package, which we need to download from CRAN using the `install.packages()` function.

Once you have done this, we load the package:

```
library(psych)
```

We use the `fa()` function, also from the `psych` package, to run our factor analysis. More information on this function can be found here. We fit the model with this syntax:

```
fa.fit <- fa(wvs.dat[,c("happiness",
                        "health",
                        "finances",
                        "satisfaction",
                        "freedom")],
             nfactors=1)
```

Within this function we specify our variables, the number of factors we wish to calculate (in our case, just one).

It is not shown here, but there are several rotation methods available with factor analysis. These can be characterised as orthogonal, which do not allow the latent dimensions to be correlated, and oblique, which do allow correlation. Varimax is a popular method (especially in the social sciences) for orthogonal rotation in factor and principal components analysis, so in instances where you have multiple dimensions, they will remain uncorrelated. We have not included this here, as we have a single dimensional factor analysis, but good to know for future work you might do where there are two or more factors. If we had wanted to include a Varimax rotation, we could do so by adding `rotate="varimax"` after the `nfactors` command (separated by a comma, of course).

The output from your model should look like this:

```
## Factor Analysis using method =  minres
## Call: fa(r = wvs.dat[, c("happiness", "health", "finances", "satisfaction",
##     "freedom")], nfactors = 1)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                MR1   h2   u2 com
## happiness    0.64 0.41 0.59   1
## health       0.52 0.28 0.72   1
## finances     0.63 0.39 0.61   1
## satisfaction 0.93 0.86 0.14   1
## freedom      0.71 0.50 0.50   1
##
##                  MR1
## SS loadings     2.43
```

```
## Proportion Var 0.49
##
## Mean item complexity =  1
## Test of the hypothesis that 1 factor is sufficient.
##
## The degrees of freedom for the null model are  10  and the objective function was  1.73 with Chi Squa
## The degrees of freedom for the model are 5  and the objective function was  0.02
##
## The root mean square of the residuals (RMSR) is  0.02
## The df corrected root mean square of the residuals is  0.04
##
## The harmonic number of observations is  1788 with the empirical chi square  22.35  with prob <  0.000
## The total number of observations was  1813  with Likelihood Chi Square =  33.28  with prob <  3.3e-0
##
## Tucker Lewis Index of factoring reliability =  0.982
## RMSEA index =  0.056  and the 90 % confidence intervals are  0.039 0.075
## BIC =  -4.23
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                  MR1
## Correlation of (regression) scores with factors   0.95
## Multiple R square of scores with factors          0.90
## Minimum correlation of possible factor scores     0.79
```

There is a lot here, and we do not expect you to understand it all straight away. For this exercise, we will focus on the first table.

The first column of the top table in our output, labelled `MR1`, shows how each item loads onto the latent trait for quality of life. A positive loading means that a higher response on this variable indicates a higher score on the combined measure of life quality created by this analysis. A negative loading (if there was one) would suggest that a higher response predicts lower quality of life.

It is only this simple to interpret the finding because we recoded our variables. If we had not, we would need to remember the way each variable was coded to infer the meaning of these results. You can see what I mean by replacing `happiness` and `health` with the original variables and re-running the factor analysis.

Once you have done this, return to the original model. There are three other columns. The second column, `h2`, represents the communalities of the variables. These communalities are the total amount of common variance between the variable and the factor(s). Higher communalities are better. If communalities for a particular variable are low, then that variable may struggle to load significantly on any factor.

The communalities for the $i$th variable are computed by taking the sum of the squared loadings for that variable. To compute the communality for `happiness`, for instance, we square the factor loading for this variable (if there was more than one dimension, we would square the loading for each then sum the results). This gives us 0.41. You can think of these values as the equivalent of multiple $R\hat{}2$ values from regression models, predicting the variables of interest for your factors. The communality for a given variable can be interpreted as the proportion of variation in that variable explained by the factor(s). If we perform multiple regression of `happiness` against the factor score, we obtain an $R\hat{}2$ of 0.41, indicating that about 41 per cent of the variation in happiness (as measured by this variable) is explained by the factor score.

One assessment of how well this model is doing can be obtained from the communalities. What you want to see is values that are close to one. This would indicate that the model explains most of the variation for those variables. In this case, the model does better for some variables than it does for others. Our results suggest the factor analysis does the best job of explaining variation in `satisfaction`. This makes sense, as we're trying to measure quality of life, and self-reported life satisfaction is likely the closest variable to this in the model.

The third column, `u2`, has nothing to do with the band, but is rather the measure of uniqueness. This is the proportion of a variable's variance that is not shared with a factor structure. Unique variance is composed of specific and error variance. The existence of uniquenesses is what distinguishes factor analysis from principal components analysis. If the variables in the model are thought to represent a 'true' or latent part of some phenomena, then factor analysis provides an estimate of the correlations with the latent factor(s) representing the data. If we believe them to be measured without error, then principal components provides the most parsimonious description of the data. When we look at our table factor loading table, the output in the third column is the direct inverse of the second; the more the variance of a item is explained by the factor score(s), the less the proportion a variable's variance will not shared with a factor structure.

The last column, `com`, is the complexity of the factor loadings for that variable (see below). This will be '1' when you only have a single dimension, but will (usually) increase as you increase the number of dimensions in your model.

Just below the first table, we can see that our latent dimension accounted for around 48 per cent of the variance in responses. If we had multiple dimensions we would be provided with a statistic for each dimension and the cumultive variance explained.

We can take the factor scores from this analysis — the latent quality of life estimated by our factor analysis — and save this back into our existing dataset using the code:

```
wvs.dat$life.quality <- fa.fit$scores
```

This saves the factor scores as a new variable called `life.quality`. We can view a sample of these using the `head()` function:

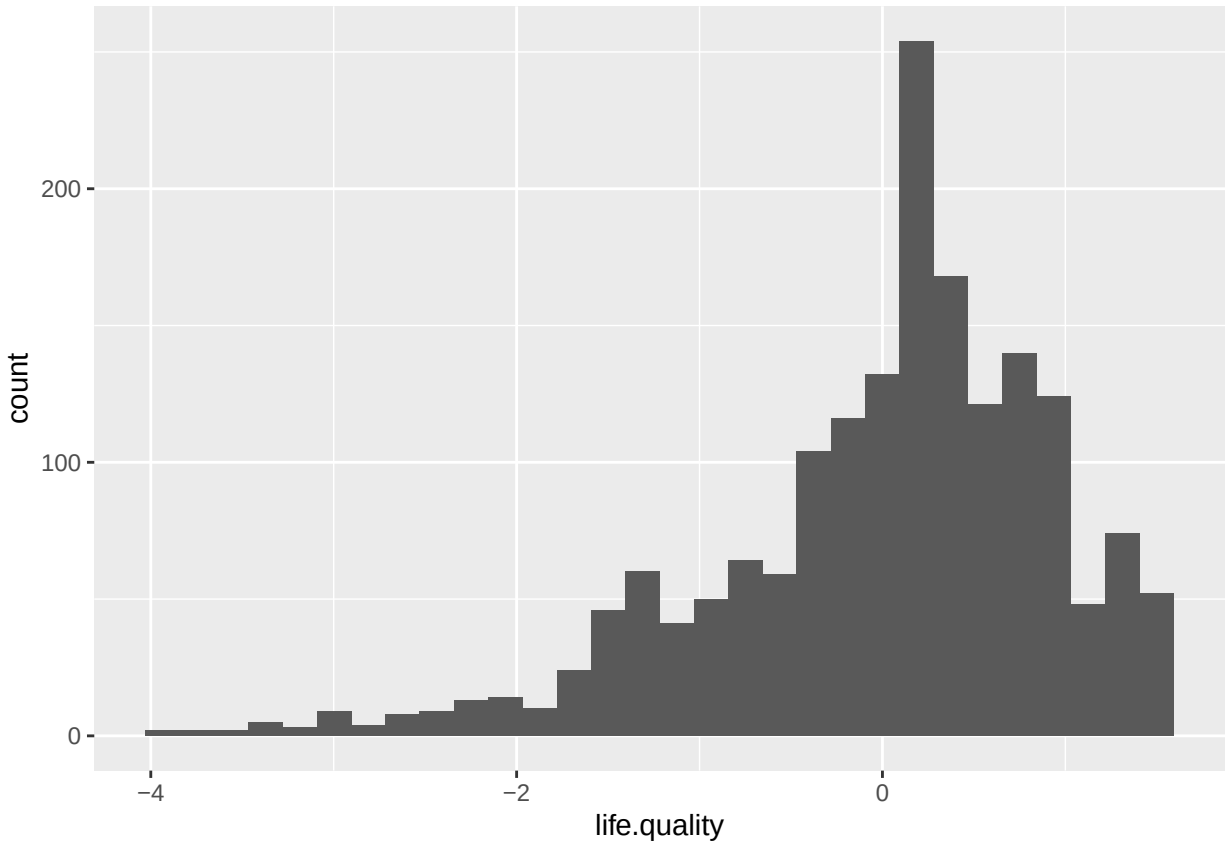
```
head(wvs.dat$life.quality, 10)
```

```
##               MR1
##  [1,]  0.72256548
##  [2,] -0.74706729
##  [3,]          NA
##  [4,]  0.26937731
##  [5,]  1.28122093
##  [6,]  0.12100610
##  [7,]          NA
##  [8,]  0.08338665
##  [9,]          NA
## [10,] -1.25791923
```

As you can see, these are approximately standardised, with a mean of zero and standard deviaiton of (almost) one. We can then use this to analyse the association between quality of life and different individual characteristics that are also available in this dataset.

```
library(ggplot2)

ggplot(wvs.dat, aes(life.quality)) +
  geom_histogram()
```




## Assessable group work

Now you are ready to finish the assessable group exercise we began this morning.

For this section of the project, you will undertake three main steps:

1. Calculate and plot the relationship between your dependent variable (the quality of life measure we calculated above using factor analysis) and the independent variables you selected this morning using descriptive statistics.
2. Run linear regressions on the quality of life measure. Use appropriate measures to test your theories.
3. Write up your key findings.

If you have any questions, do not hesitate to ask us for help. During the labs, the teaching team will be available to talk you through the project on Zoom. We will be moving through the breakout groups and can also be called to assist you. Outside of the sessions today, you can also post questions on the Ed discussion board.

We cannot do the work for you — this is an assessment – but we provide some advice.

Good luck with the exercise!