# Many Models in R

## Transcript video 1

Full resource: <https://www.ncrm.ac.uk/resources/online/all/?id=20840>

Hello, this is the first in a series of three videos on running many models within R.

In this video, we're going to look at how to create a grid of specifications for the different models that we want to run.

We're going to be using the tidyverse package to do this. The tidyverse package is actually a package of packages. It includes several popular packages such as dplyr and ggplot2, but the tidyverse is also a design philosophy. As a result, the packages within the tidyverse share common features and they work very well together. If you don't know much about the tidyverse, I've included some links to resources with this tutorial.

I've also created a worksheet which you can work through alongside these videos. The worksheet includes fully annotated code.

Let's quickly load the tidyverse package into R using the library tidyverse command. If you've not already installed it, you'll have to install this prior to using the library tidyverse command. We're going to be using a simulated data set in this tutorial. The data set contains data from four pretend cohort studies representing four generations.

Baby boomers, Gen X, Millennials, and Gen Z. I'm just going to download this data from the internet, loading it into R and assigning it to the name df. I've put the data on the OSF so you can download it yourself. This data set contains repeat measurements of BMI at each decadal sweep in each of the four studies. Now you'll notice I've used the pipe operator here, which is used extensively within the tidyverse. It allows you to chain commands together and it can be read, do this, then this.

So here we've taken the URL and then we've piped it into the read RDS command to download that data set, which is an RDS format, into R and assigned it to the variable df. Let's have a quick look at df using the structure function str. Here we see that df is a table with the dimensions 80,000 by 11.

A table is a type of data frame, which is a rectangular data structure like a spreadsheet or a data set in Stata. It has rows, which are observations, and columns, which are variables. df is a long format data set where each individual has a measurement of BMI for each follow up that the survey was run for.

So for instance, for the baby boomers, we've got observations for each cohort member from follow ups at age 15 to 65. We're going to be using this data set to examine the association between cognitive ability and later BMI. Let's have a quick look at the raw data sampling 20 observations at random using the sample n function.

So let's just make this a bit bigger. Okay, you'll see that the 11 variables are cohort ID, and we've got FUP, which is the age of follow up. So 15 to 65 for baby boomers. We've got age, which is the age of BMI measurement BMI, then we have three measures of cognitive ability. We've got nonverbal ability, verbal ability and vocabulary. We also each of these, by the way, was measured in adolescence. We also have three further variables that we use as control variables in our regression models.

So one is an indicator for whether the person is female, one or zero, then we've got two that aren't printed to the console family class, which says whether someone was from a non manual or a manual socioeconomic background, and then childhood BMI, which was measured prior to the measure of cognitive ability. Now, the current data are perfect for adopting a many model approach.

We have data from four different cohorts, each of which measures BMI at multiple different ages. We also have three different ways of measuring cognitive ability. Now, the field of cognitive epidemiology has found widespread associations between cognitive ability and multiple measures of health. Dominant theories in the field would predict that cognitive ability should have a causal effect on BMI and predicts that an association between the two should be seen regardless of the age of follow-up, the year one was born or the measure of cognitive ability that was used. It is therefore worth running a many models approach with combinations of cohort, age of follow-up and measure of cognitive ability to examine how robust associations are. These theories would also predict that we should find an association regardless of whether we control for childhood BMI or not.

So we'll add this to our analysis too. Our specification grid will contain all the information we need to run a particular model, that is the specific cohort and follow-up to take the data from, the particular set of control variables to use, and the particular measure of cognitive ability to examine associations with BMI using. We'll first create a few objects that we're going to use to construct the specification grid. First, we're going to create a list that contains the sets of control variables. This list will have two elements. One will just define a basic model with controls for age, sex and family class, and the second will add a further control for childhood BMI.

So let's just write that out now. I can copy and paste this across. So the basic model just has those three covariates, and the second one will have those three covariates plus child's BMI. And we've called these models basic and child's underscore BMI. Let's just quickly run that and save it to name modcovars. Okay, and then we can just look what modcovars looks like. So it's a list with two elements, basic and child BMI, that just contains those strings.

So modcovars is a list, and we can access elements of modcovars using the double square bracket notation, and putting within those square brackets the name of the element we want to pick out. So if we want to get the set of variables for the basic model, we can just write basic in quotation marks with the name of the element. Basic in quotation marks within the double brackets, and that will get us back to the string.

So here we go. This is the sort of code we'll be using later to construct our models. Next, we want to get the set of cognitive ability variables. We'll use the stress subject function to subset the names of DF, the column names of DF, to those beginning with cog, all of our cognitive ability variables to cog underscore, and then something after that. So the syntax is stress subject, then the string that we want to subset. In this case, it's the names, the column names from DF. And then we just use an irregular expression. This tilde sign gives us back any string that contains, sorry, that starts with cog underscore.

So we get back cog non-verbal, cog verbal, and cog vocab. Let's just save this as cog vars. Okay, so as mentioned, we want to run separate models for each cognitive ability measure, each set of control variables, and each cohort and follow-up. The cohorts differ in the set of follow-ups used to measure BMI, so we're going to be using the distinct function from the tidyverse to get the combinations of cohort and FUP that are actually observed in the DF data frame. So DF, pipe this into distinct, and then we just name the columns that we want to get the distinct values of. So we return back a table of 16 rows, which has the different combinations of cohort and FUP that are observed in the data set.

So we have the years 15 to 65 for baby boomers, 15 to 55 for Gen X, 15 to 35 for millennials, and 15 to 25 for Gen Z. This is just two of the things that we want to loop over. We also want to add two new variables for cog var and the covariates that we're going to use. So to combine that information in, to get the full set of specifications, we're going to use the expand grid function, which will just produce a table with all the combinations of the inputs that have been provided to it. So we can pipe the result from distinct straight into expand grid. Then we'll just add cog var equals cog vars, and co vars equals names of mod co vars. And now we get back a table that is 96 rows long. We've got each combination of cohort, FUP, cog var. You can see we've got cog nonverbal, cog verbal, and cog vocab. And then we've also got each combination of co vars as well. So basically child BMI.

Okay. So now we have these 96 rows. We can see that, for instance, the first row would be a specified model using data at age 15 for baby boomers, where cognitive ability is measured using a nonverbal measure of cognitive ability. And the covariates that we use are just from the basic model. In the second model, we would use the same data, same measure of cognitive ability, but we would be adding covariates for child BMI as well.

Before saving this specification grid to an object, I just want to do one last thing, which is to create a new column to identify each row. And we're going to call this new column spec ID. This will just be equal to the row number. So one for the first row, two for the second, and so on.

There we go. We're going to use this spec ID to pass it to a generic function that allows us to select a row. We'll do this rather than pass all the information from the other four columns in separate arguments, which can get cumbersome. Let's just save that to mod specs. And then just quickly check mod specs is saved correctly. Great. Okay.

So we've got, that's a basic way of getting a specification grid. In a few lines of code, we've been able to specify 96 models where we're going to loop over the cohort, the follow-up, the cognitive ability variable measured and the covariance that we're going to add to the regression. Thank you.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK