Everybody, welcome.

This video is about the factor analysis demo.

So we'll load the standard packages plus an additional one called psych

which contains some factor analysis tools that we'll be discussing today.

This data had been collected

about five socioeconomic variables for 12 census tracts in Los Angeles area,

and the variables are total population,

median years of schooling in a person in that population,

total unemployment, miscellaneous professional services,

this is employment, and median house value.

We'll perform a factor analysis on these data and talk about

the different rotations possible.

We'll load these data from the dataset that's been provided.

Here's what they look like but let's look at the plot.

Not many data points but we do see that there are certain variables that are

strongly correlated with each other but let's take a look at what that means.

Let's use the factanal function from the psych package.

It's pretty straightforward to use.

We just say here's the dataset and here's the number of

factors and we're told that the one factor has the following loadings,

the following coefficients on the latent unobserved factor.

What do we see? Now notice that first of all,

all of these are standardised,

so we can actually look at them on a 0,1 scale.

That's rather convenient and we have a hypothesis actually factanal.

I think it's in the stats package.

We have one factor which has high loadings on population and

employment and to some extent

on service but not on others whereas we

see there is a lot of leftover very niche and for school,

services and house prices.

We also have a hypothesis here that one factor is sufficient.

It's a Chi-square test that's described in the slides

and we see that based on the Chi-squared test,

the p-value is really small so there is a lot of

evidence that additional factors are needed so let's add another factor.

Now, we see a few things happen.

One is that some of the loadings are really close to zero now,

and that's good because generally we want to separate

out the variables based on the factors that influence them.

The uniquenesses have also gone down quite a bit and so what do we see?

Well, we see that there is a factor one and this is in no particular order,

covers school,

services and house prices.

The second factor deals with population,

employment and to a lesser extent services.

So what does that tell us?

Well, the first factor is probably to do with the wealth of the community,

how wealthy or how rich the community is or how poor it is because we see that,

for example, it corresponds to relatively high schooling

and relatively high proportion in the service industry and high house prices,

but mostly house prices and schooling whereas

the second factor might just be the population and the total number of employed.

Well, that's going to be a function of

the population and how many people are employed in the services.
Well, again, that's going to be bigger with the population.
We look at the Chi-square test,
the p-value is 0.136 so there's
not enough evidence to believe that we need additional factors.
Now, we don't have
a way of fitting model with three factors when we
start using this package and the reason for
that is that there's only so much information in these variables,
these observations only so many elements of the covariance matrix that can
model and the numbers given here.
By the time we actually start incorporating
these many loadings and these many uniquenesses,
well, we might run out of degrees of freedom.
Now, we can also use the psych package.
Psych package has different defaults so it uses a rotation different from varimax,
arguably a better one.
It's a much bigger topic which we won't be getting
into and it also use
a different method by default where we're going to use maximum likelihood,
so these are the settings to try to reproduce the other fit.
In fact, even though the formatting of the outputs is a bit different,
we can see that the values are very very similar.
Then we also have a similar Chi-squared statistic.
In fact, this one is the total number it corresponds
to this test although it uses a slightly different formulation.
But as you can see,

the p-value is the same and this conclusion is the same.

There are also other diagnostics which again if you are

interested in factor analysis there are additional sources about this.

We can also make a plot of these factor loadings for the different variables.

It's not very informative although

it can help you visualise that two of them are small.

Now let's try it with two factors.

Again, we get the lack of fit test here,

which now is no longer significant.

Again, we have our loadings and weaknesses.

They are basically the same as before.

Now, this plot,

actually plots first loadings and your second loadings and you can give us an idea

of what variables belong to which factor.

Here we can see that variables one and three belong

to the first factor and variables five and six to the second factor,

and variable number four belongs to both or are shared between them.

This also works.

You can in fact fit this,

but unfortunately we can no longer do any tests on

it and this is what the plot looks like.

There is also useful little function called fa.parallel,

which can be used to quickly estimate a good number of factors.

It's kind of using the Kaiser's rule for the PCA and here it is.

This suggests that we want probably two factors.

Now one thing, notice that

this method doesn't have to specify its rotation

and the reason for that is all possible rotations have the same goodness of fit.

Now, you can take a look at the contents of these objects if you want

to continue to use them for

the purposes and to do that you just use the unclass function.

I don't show the results here,

but as always I recommend you try running

the things in the demonstrations yourselves

just to see if it works and to explore a bit.

That's the demo for the factor analysis.