Data analysis meeting with SU, SI and SS

Joakim Frögren

2021-03-04

Initially, SU wanted to know the purpose of why the analyzes were done at all. I explained that it was about the fact that in the study it was a lot about presenting descriptive data, but that in addition I also want to compare those who want to participate with those who did not want to participate in a more detailed way. Thus, the bivariate analyzes in some cases constitute results in themselves, while for just *willingness* they constitute a sub-step that makes it possible to proceed with the analysis to a multivariable analysis. SU emphasized that a regression model can be made for many different reasons and that it was important to clearly explain why it was considered appropriate in this particular case.

SU agreed with SS that for *awareness* and *previous active involvement* I should use Chi square, alternatively in cases where it was a question of ordinal and nominal scale with more answer options than two (or three?) I would use Mann-Whitney or Wilcoxsons non-parametric tests.[1]

An alternative to the current presentation of data in Tables 2 and 3 would instead be to present, for example, level of education in a way that describes that "Among those who are willing, how many have elementary school as their highest education level, high school,etc.?" The way in which it is best to present the result in this case is connected with how I intend to conduct the discussion and which result then best illustrates the arguments I wish to present.

SS pointed out that it is not clear that healthier people necessarily would be more aware/ or willing. Rather, it could be argued they could presumably be motivated by their health issues and thus better aware and more willing.

Furthermore, Table 4 was considered redundant and I was asked to remove it. Figure 1 was considered good as it was, but I was asked to create a similar figure for the contents of Table 5 and thus replace that table with a figure.

When it came to willingness in Table 6, SU agreed with SS that instead of doing a Chi square test, I should perform a series of bivariate logistical regressions, variable by variable. Then, before the multivariate logistic regression, I should decide on a cut-off value, usually at .1 or .2, but here I was asked to, based on previous similar studies, suggest which p-value I considered appropriate to take the variable further to that analysis. However, this p-value should not be

[1] This means in practice using Mann-Whitney or Wilcoxson on all variables except age and sex

presented in any table, but the only table reported for willingness is the one based on the next multivariable logistic regression, and OR should be included instead of the p-value.

Furthermore, SU agreed with SS that there was nothing that justified a division into age categories.

However, when it came to Level of education, SU thought that the number of respondents who had 'Research education' was so small that that group could advantageously be incorporated into the category 'College 3 years or more.'

Regarding self-rated health, SI and SS pointed out that it was unclear whether the current response options really corresponded to the actual instrument. Here I was asked to ensure that this was the case and to use these response options and no others. These options were also used for self-rated economy, so even there I needed to ensure they were correctly formulated.

Furthermore, it was decided that I will bring in frailty as an additional variable here, motivating this decision by stating that they were both initially considered relevant to include. To check how to frame frailty in this context, SI recommended me to - rather than contact Synneve, as MK had suggested - go to her papers and see for myself how frailty is been defined, brought up and framed there.

As a general rule, a subgroup in that one brings to a multivariate analysis should not contain less than 10 respondents. Thus, self-rated health is at the border with exactly 10 respondents in one of the subgroups. For multivariate analysis I should choose the group that is big enough as = 0, because the confidence interval in the subsequent groups will follow from that. SU explained that one can in principle choose any of the categories, but one should also be able to argue theoretically why a certain category was chosen. I explained that I had planned to use the first response option as = 0, but that it then made sense to transform the five categories related to self-rated health and self-rated economy into three. this would also make the table for the next multivariate logistic regression less extensive and easier to interpret. Thus, keeping the middle category intact but merge the first and second, as well as the fourth and fifth was decided as a good way forward. SU also emphasized that I need to have the same groups for univariable and multivariable analysis. However, for the analysis of awareness and previous active involvement, which will be kept on a descriptive level, I can keep the five categories.

Next, when it comes to checking the interdependence between the independent variables, SU recommended VIF as a suitable alternative to the conventional correlation matrix. The VIF value should be above .70 or .80 in order to be a problem correlation-wise.

When it comes to the multivariable logistical regression, it emerged that I initially thought correctly when I marked the categorical variables as categorical and that this is how I should do so that they are not handled as continuous, which they otherwise are. Furthermore, SU ageed with SS that it made sense to enter all independent variables except previous active involvement as 'Block 1' and previous active involvement as 'Block 2.' However, when I describe what I have done, I should not employ the term 'block,' but rather write: "In the first model we used age , sex etc as the independent level.. then there was a second model were we included previous active involvmenet as a confounder."