**Research Replicability and Workflow Management (RRWM)**

**Replication report**

**Attempt coding Ogochukwu’s project** **by Nicole Antunes Rezende**

Research question:

Is there any significant association between total income and sex, age at immigration, and level of education of Canadian Immigrants from 2000 to 2015 who immigrated at the age of 20 years and older?

Dataset: Census 2016.csv

In green we have the instructions given and in black we have Nicole’s comments.

1. Install necessary packages and load the libraries

Nicole’s comment: The first instruction was “Install necessary packages and load the libraries”. It was impossible to imagine which are the necessary packages for this project without previous instructions.

2. Clean up data by:

a. Loading the CSV file

b. Select only TotInc, Sex, AGEIMM, HDGREE, YRIMM (variables) and discard others

c. Filter AGEIMM and select range 5 to 13 [meaning 20 years and above]

Nicole’s comment: It seems that we followed the same steps for a. b. c. e. and h.

d. Remove cases with TotInc less than 1000 and those labelled 88888888 and 99999999

Nicole’s comment: Although in the instructions only “less than1000 and those labelled 88888888 and 99999999” were mentioned, I noticed that Ogochukwu used 33 in his code, which was not clear why and was not part of the instruction.

e. Remove cases with HDGREE = 88 or 99

f. Convert all variables in the dataset to numeric

Nicole’s comment: Not clear for me this instruction as using the function “glimpse”, I see that the variables are dbl (double), which is one of the main numeric types in R. I used the function “mutate”, but anything changed as all the variables were already dbl. So, I’m not sure what was the result wanted with this instruction.

g. Re-code Sex as Female =1, Male = 2

Nicole’s comment: I don’t understand this instruction “Re-code Sex as Female =1, Male = 2”, as in the data set Female is already 1 and male is already 2. After looking at the solution, I suppose the goal was to label this variable instead of recode (?).

h. Creating HDGREE.R variable such that 1 to 8 = below bachelor, 9 = bachelor, and 10 to 13 = above bachelor

3. Run descriptive statistics to create frequency and percentage summary table for Sex and Education levels.

Nicole’s comment: When I read this instruction, I understand that we are talking about a cross table (sex and educational level). After looking at the solution, I noticed that he was intending to have one frequency/percentage table for sex and another frequency/percentage table for educational level.

4. Run a multivariate regression analysis with TotInc as the dependent variable and Sex, age at immigration, and education qualification (HDGREE.R) as independent variables.

Nicole’s comment: I used the function “lm” to run a multivariate regression. It’s not clear how it was run in the solution (I see the table framework code: “Fit the linear model”.

We are not answering all aspects of the research question, as we are not selecting the period (from 2010 and 2015).

Research question: “Is there any significant association between total income and sex, age at immigration, and level of education of Canadian Immigrants from 2000 to 2015 who immigrated at the age of 20 years and older?”

5. Write the 2 or 3 tabular output in any file format.

Nicole’s comment: I was not able to save the outputs in word, pdf or htlm. I have tried some options (see at the end of the document “Attempt coding Ogochukwu's project by Nicole”), but it did not work. I put some screenshots of my RStudio output in the document “Output\_Ogochukwu's project by Nicole”.

Final comment: It is important to highlight that our samples ended up being very different. As I mentioned in my comment on instruction number 4, before running the multivariate regression, I realized we were not fully addressing all aspects of the research question, as we are not selecting the correct time period (from 2010 to 2015). Although there were no instructions to restrict the period (from 2010 to 2015), I noticed in the solution that Ogochukwu have selected only YRIMM from 2000 to 2015. This could be the reason why our samples ended up being very different – 116 346 individuals in mine compared to 55 616 in his project.

As we could expect after the detailed comments listed above, the results of our tables and regressions are quite different. This exercise was very interesting, and I think I have learned a lot while coding from the instructions and then comparing my final code with the solution.