

Research Replicability and Workflow Management

Data Activity | Program

Goodnews' program

Initial version with associated do file (after iteration through "coder" role)

1. Data cleaning—Clean the GSS (General Social Survey), Cycle 31, 2017 [Canada] family file:
 - a) Select and keep only the dependent and independent variables of interest
 - b) Drop anyone aged over 64 from the dataset, since we are focusing on working-age individuals (15-64-year-olds) and the minimum age of the respondents is 15.
 - c) Clean the dependent and independent variables, dropping observations with missing or unusable values, recoding default variables to combine response options and generate new variables where applicable.
2. Create a table of summary/descriptive statistics.
3. Create a table of multivariate analysis.

Detailed version with associated do file (after iteration through "coder" role)

1. Data cleaning—Clean the GSS (General Social Survey), Cycle 31, 2017 [Canada] family file— "RRWM-1-DoFile-Goodnews_Oshiolebe.do":
 - a) Select only these seven (7) variables of interest:
 - i. Dependent variable is SRH_110 (Self-rated health).
 - ii. Independent variables, single focal IV written first, are MARSTAT (Marital status), AGECE (Age), SEX (Sex), LMAM_01 (Worked at a job or business last week), EHG3_01B (Education – Highest certificate, diploma or degree), and TTLINCG2 (Income of respondent - Total (before tax))
 - b) Drop anyone aged over 64 from the dataset, since we are focusing on working-age individuals (15-64-year-olds) and the minimum age of the respondents is 15.
 - c) Clean the dependent and independent variables, dropping observations with missing or unusable values, recoding default variables to combine response options and generate new variables where applicable; and using the data's codebook, define and apply labels where numbers represent subcategories (that is, do this for variables that were not recoded so that their tabulations will not display values instead of labels).
 - i. Tabulate it and drop observations with unusable values from SRH_110: "Don't know" [7], "Refusal" [8], and "Not stated" [9]
 - ii. Recode SRH_110 such that a new category "Not good" [4] combines both "Fair" [4] and "Poor" [5] and generate a new dependent variable SRH2_110 (Self-rated health cleaned) to be used going forward.
 - d) Clean the independent variables.
 - i. For MARSTAT, tabulate it and drop observations with "Don't know" [97] and "Refusal" [98]; and recode it such that a new category "Partnered" [1] combines both "Married" [1] and "Living common-law" [2] and another new category "Other" [3] combines "Widowed" [3], "Separated" [4], "Divorced" [5]; and "Single, never married" is renumbered [2] instead of [6]; and generate a new marital status variable MARSTAT2 (Marital status cleaned) to be used going forward.
 - ii. For AGECE, tabulate it to verify that the minimum is 15, the maximum is 64, and there is no missing value; and recode it from a discrete variable into an ordinal categorical one with five age groups, generating a new variable AGEGRP (Age group): [1] "15-24 (Youth)", [2] "25-34 (Young adults)", [3] "35-44 (Middle-aged adults)", [4] "45-54 (Senior adults)", [5] "55-64 (Pre-retirement adults)".
 - iii. For SEX, tabulate it and confirm no missing values, noting that 1=Male, 2=Female (default).

- iv. For LMAM_01, tabulate it and confirm no missing values, then drop “Don’t know” [7] and “Refusal” [8] observations.
 - v. For EHG3_01B, tabulate it and confirm no missing values, then drop “Don’t know” [97], “Refusal” [98], and “Not stated” [99] observations.
 - vi. For TTLINCG2, tabulate it and confirm no missing values.
2. Create a table of summary/descriptive statistics
- a) Create frequency and summary tables using these variables (dependent variable listed first): SRH2_110, MARSTAT2, AGEGRP, SEX, LMAM_01, EHG3_01B, TTLINCG2
 - b) Since all variables are categorical, create descriptive summary statistics by transforming them into dummies (using *i.var*) so that their outputted mean values will be their subcategorical percentages in fractions (e.g., 0.5=50%) and hidden reference categories will make up the remaining percentages inside 100%.
 - c) Save the summary statistics (with dummy variables) as a separate file in .jpg.
3. Create a table of multivariate analysis, focusing on answering these questions: i) What sociodemographic factors determine the quality of self-rated health among working-age Canadians? ii) Does being in a partnership have a greater impact on health than being single or never married? Iii) Is marital status a more significant determinant of health than educational attainment?
- a) Use same variables as in 2 (a) above for the **bivariate analysis**: create contingency tables, showing and testing (Chi2) associations between the DV and each IV.
 - b) Save the contingency tables with chi-square tests as six separate files in .jpg.
 - c) **Multivariate analysis**: Use the same variables as before to build two separate ordinal logistic regression models. *In the first model*, regress the dependent variable against only the marital status and educational attainment independent variables. *In the second model*, regress the dependent variable against all the independent variables. *For both models*, explicitly declare all categorical variables so that Stata automatically creates dummy variables.
 - d) Display the results as odds ratios, ignoring any weights in the analysis.
 - e) Allow Stata to automatically select the reference categories.
 - f) Save the two regression output tables as two separate files, that is, .jpg images. One image should contain the results from the first model (with two IVs), and the other should contain the results from the second model (with all IVs).
 - g) (Optional) Create a publication-quality table from stored regression estimates (that is, in a journal-like, customizable format), showing the major outputs of the two models in one table.

Note: There will be **nine tables** (minimum, excluding optional output) in all. It is okay to create all tables simply as cropped Stata outputs instead of outputting them using commands.