**Appendix 10.2: Planning how to create the variables you need from the variables you have**

Before you analyze data for your research question, think about (and discuss with your professor or research mentor) whether you need to create any new variables in order to address that research question, working from the variables available to you in an existing data set. Examples of new variables you might need to create for an analysis include:

* categorical versions of continuous variables, such as age group from age in years
* simplified (collapsed) categorical variables, such as a 3 racial/ethnic groups from a detailed list of many racial/ethnic groups;
* a binary indicator such as low birth weight (“yes” versus “no”) from a continuous measure of birth weight;
* a new continuous variable such as Body Mass Index from height and weight

This planning step should be done *before* you point and click (or write syntax) to create one or more new variables in your electronic database.

Create a Word document that includes the following information:

1. The hypothesis for your research project in which you name the independent and dependent variables needed to answer that question (in the form you wish to analyze them).
2. List the variables in the original data set that measure the concepts you plan to study, along with the following information about each of these variables, which should be available in the documentation for the data set:
	1. Variable name (acronym);
	2. Variable label – a short phrase describing what the variable measures;
	3. Units or categories.
	4. Missing value codes, if any.

**NOTE:** The names, labels, units, coding, and missing values of the original variables must match those in the actual dataset that you will use for your statistical analysis. Do not make up hypothetical versions!

**For each new variable you will create, specify:**

1. A variable name (acronym). It should convey the content (meaning) of the new variable, and if dates or survey rounds pertain, an abbreviation for that information.

1. A label – a short phrase describing what that variable measures, including units where pertinent.
2. Missing value codes. If there were missing values on the original variable(s), those should be mapped into missing values on the new variable.
3. Explain the logic and math that will be used to create the new variables. Do not write the syntax or describe the point-and-click steps you will use in your software program. To learn what formulas, cutoffs or standards pertain to your topic, review the literature in your field.
4. For new *continuous* variables,
	1. Write the mathematical formula by which that new variable is created from one or more variables available in the original dataset.
	2. Specify the units of the original variable(s) and the new variable.

**(Continued on next page)**

1. For new *categorical* variables,
	1. Fill out the attached planning grid, making sure to show how every possible value of the original variable maps into a value of the new variable.
	2. List the:
		1. Code (numeric value) that the new variable will take on for each value or set of values of the original variable;
		2. Value label (descriptive phrase) for each value (category) of the new variable.

**HINT:** It often helps to write the value labels for categories of the new variables first, then give each category a numeric code, and finally list the values of the original variable that map into each category.

1. Provide a bibliographical citation for each reference source you used to identify formulas or classification criteria, and cite that reference next to the pertinent step in the planning process.

**Planning template for new categorical variables**

|  |  |
| --- | --- |
| **NAME of original variable \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_****LABEL for original variable \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | * **NAME of new variable \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**
* **LABEL for new variable \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**
 |
| **Values of original variable** | **Values (codes) of new variable** | **Value labels** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Citation(s) for source(s) of information about how to create the new variable, e.g., formula for calculation of a continuous variable, or cutoffs for classification of new categorical variable.

**Example answer:**

**Hypothesis**: That people from families with more than 6 people are more likely to be poor.

**Independent variable:**

* Original variables:

Variable name Variable label

* 1. NADULTS number of adults in family
	2. NKIDS number of children in family
* First new variable: Variable name Variable label

FAMSIZE total number of persons in family

Formula: FAMSIZE = NADULTS + NKIDS.

Create using “compute” from the transform menu in SPSS. [Note: Don’t describe that step; I put it here to clarify for those of you who use that software.]

* Second new variable: FAMGT6, created from FAMSIZE (total # of people in family).

Categorical version:

|  |  |
| --- | --- |
| **NAME of original variable \_FAMSIZE\_\_\_\_\_\_\_\_\_\_\_\_\_** | * **NAME of new variable \_FAMGT6\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Label for new variable =** more than 6 people in the family |
| If FAMSIZE= | * FAMGT6
 |
| **Values of original variable** | **Codes for new variable** | **Value labels for new variable** |
| <=6 | 0  | No |
| >6 | 1  | Yes |
| Missing | 9  | Missing |

**Dependent variable:**

* Original version from dataset: INC08 (2008 annual income in $; 999999 = missing)
* New version to be used in analysis: POOR2008

|  |  |
| --- | --- |
| **NAME of original variable: INC08****LABEL for original variable:** 2008 annual income in $ | * **NAME of new variable \_\_POOR08**

**Label for new variable =** family poverty status in 2008 |
| If INC08 = | POOR08= |
| **Values of original variable** | **Codes for new variable** | **Value labels for new variable** |
| <20,000 | 1  | Poor |
| >=20,000 | 2  | Non-poor |
| Missing (999999) | 9 | Missing  |

Create using “recode into new variable” in SPSS. [Note: Don’t describe that step; I put it here to clarify for those of you who use that software.]

Definition of “poor” as annual income <$20,000 from

Jones et al. 2000. Title, journal, volume, page #s.

[This is an example of providing the citation from which you obtained the cutoffs, formula, or logic by which you created your new variables from those in the original data set.]