

Hwk #2: Data Management & Cleaning

Biostat 202: Opportunities and Challenges of “Big Data”

Due Wednesday, August 10th by Midnight

For this homework we will also use the data from the National Electronic Injury Surveillance System (NEISS) available at <http://www.cpsc.gov/en/Research--Statistics/NEISS-Injury-Data/> We will be interested in examining age, and injury location (body_part) along with the disposition variable you created in Hwk #1.

Open your stream from Hwk #1 and the Excel file included in Hwk #2.

Now that you have already audited the data, we will do additional cleaning and recoding.

1. (4 points). Examine the race_other variable in the Excel file, If you were to use this variable in your analyses suggest how it could be recoded given its inconsistencies.
2. (6 points). Based on your 'Audit' node from Hwk #1 and an examination of the codebook coding for age, what is wrong with the age variable and how would you recode the information? Recode the variable into a usable variable for analyses. How have you changed the '000' used for 'not recorded' vs. the '2xx' variables used for less than 2 years old? You can use the 'Filler' node in the Field Ops tab to change the codes 200 and above to a number less than 2.
3. (2 points). Should you replace the 'not recorded' observations with missing value imputation here or replace them with a specific missing value code? In your 'Type' node and, if keeping the missing data as missing or 'not recorded,' create a missing value code for this that matches your recoding from #3 in the 'Type' node.
4. (5 points). Recode body_part into two Flag variables to identify Head (=75) or Other as the body_part injured, and another variable to identify Toe (=93) or Other as the body_part injured. Note that you can use 'Default value' as 0 for unspecified values.
5. (3 points). Use the 'Filter' node to eliminate variables you will not use in an analysis. Which variables have you eliminated?
6. (2 points). Audit your new dataset and examine the quality and that the recodes worked. In general, what is the quality of your recoded and existing variables?
7. (4 points). Run a 'Means' node from the Output tab for age by your new body_part variables and admitted or not (by selectively changing the target to examine age by admitted or Head or Toe). What is the mean age for head injuries, toe injuries, and hospital admissions. What does this tell you? Tabulate admitted by Head or Toe. Which body_part is more likely to require admission?
8. (3 points). Predict Hospital admission based on your new body_part variables, your new age variable, and any others that you wish to include in a Logistic model and Neural Net. Which model had higher prediction & how good were the predictions? Did the model predict Head or Toe injuries better?

Visual Aids:

In the 'Type' node you can change '0' to be a missing value by choosing 'Specify' in the Missing field:

The 'age Values' dialog box is shown. It has a title bar with standard window controls. The 'Measurement' dropdown is set to 'Continuous'. The 'Storage' dropdown is set to 'Real'. The 'Depth' is set to '0'. There is a 'Model Field...' button. Under 'Values', the 'Specify values and labels' radio button is selected. The 'Lower' field contains '1.0' and the 'Upper' field contains '107.0'. There is a 'Labels...' button. Below that, the 'Extend values from data' checkbox is unchecked. The 'Check values' dropdown is set to 'None'. The 'Define blanks' checkbox is checked. Below it, the 'Missing values' list is empty. There are 'Range' and 'Null' checkboxes, both checked, with 'White space' unchecked. The 'Range' field contains '0.0' and the 'to' field contains '0.0'. There is a 'Description' field at the bottom. At the bottom right are 'OK', 'Cancel', and 'Help' buttons.

Recoding body_part to Head:

The 'Reclassify field' dialog box is shown. It has a title bar with standard window controls. The 'Reclassify into:' dropdown is set to 'New field'. The 'Reclassify field:' dropdown is set to 'body_part'. The 'New field name:' field contains 'Head'. Under 'Reclassify values:', there are four buttons: 'Get', 'Copy', 'Clear new', and 'Auto...'. Below these is a table with two columns: 'Original value' and 'New value'. The table has two rows: the first row has '75' in the 'Original value' column and '1' in the 'New value' column; the second row has an empty 'Original value' cell and '0' in the 'New value' column. To the right of the table are three buttons: a 'Fill' button, a 'Delete' button, and a 'Move' button. Below the table, the 'For unspecified values use:' section has two radio buttons: 'Original value' (unchecked) and 'Default value' (checked). The 'Default value' field contains '0'. At the bottom are 'OK', 'Cancel', 'Apply', and 'Reset' buttons.

Original value	New value
75	1
	0

Using the 'Filler' node to recode ages ≥ 200 to 1

The 'Filler' dialog box is shown. It has a title bar with standard window controls. There are 'Preview', 'Settings', and 'Annotations' tabs. The 'Settings' tab is selected. Under 'Fill in fields:', the 'age' field is listed. Below that, the 'Replace:' dropdown is set to 'Based on condition'. Under 'Condition:', there is a list box with two items: '1 age >= 200' and '2'. Below that, the 'Replace with:' section has a list box with two items: '1 1' and '2'. At the bottom are 'OK', 'Cancel', 'Apply', and 'Reset' buttons.