

Hwk #1: Introduction to IBM SPSS Modeler

Biostat 202: Opportunities and Challenges of “Big Data”

Brief Solutions

For this homework we will use the data from the National Electronic Injury Surveillance System (NEISS) available at <http://www.cpsc.gov/en/Research--Statistics/NEISS-Injury-Data/>

The database was designed by the Consumer Product and Safety Commission (CPSC) to monitor consumer products that can cause injuries (think hoverboard). It is a national sample of hospitals in the U.S. and records emergency department visits that involve any sort of injury related to a consumer product. The data from 2015 alone consists of approximately 360,000 records and contains the following variables: CPSC Case #, trmt_date, PSU (primary sampling unit), weight, age, stratum (sampling stratum), sex, race, race_other, diag, diag_other, body_part, disposition, location, fmv, prod1, prod2, narr1, narr2

The disposition variable is coded as 1= treated and released, 2 = transferred, 4 = admitted to hospital, 5 = held for observation, 6= left emergency department, 8 = died, 9 = not recorded.

We will be interested in predicting if the patient was admitted (code = 4) or died (code = 8).

1. (2 points). Read the dataset in from the course website. It is called NEISS_data_2015.csv.
2. (5 points). Fix whatever errors you encountered when reading in the data. What did you have to fix and why? **There are commas in the narrative fields and it has trouble reading the dates. Uncheck the “Automatically recognize dates and times” and under “Quotes” chose “Pair and discard.”**
3. (5 points). Use a “Reclassify” node to create a variable called admit that is equal to 1 for disposition equal to 4 or 8 and is 0 otherwise.
4. (5 points). Use a “Type” node to specify admit as the target, CPSC Case # as the ID variable and correctly code the remaining variables. Note that diag, body_part, location, and prod1/prod2 are all numeric codes for different diagnoses, body parts, locations where injuries occurred, and products (up to two entered) but because they take on so many values, Modeler thinks they are continuous. Fmv is coded for visits that involve fire with 1=fire department involved, 2=fire department not involved and 3 = unknown. If you are curious the codebook is on the course website (NEISS Coding Manual.pdf). Prod1 and prod2 have *many* different values and the default is to limit it to 250. You will need to change that by unchecking the box for “Maximum members for nominal fields” in the “Stream Properties” (under the File menu).
5. (12 points). Create a “Data Audit” node to answer the following questions:
 - a. At what rate was the disposition admission or death? **7.47%**
 - b. Which variables have significant amounts of missing data? **Treatment date, race_other, diag_other, prod2, narr2.**

- c. Are there data quality issues? If so, what is the nature of that issue? Age needs recoding. Weight is not the weight of the person. Race_other has many misspellings.
- d. Which variable(s) seem to be related to being admitted or dying? Age, body_part, and perhaps race.