Homework 2 Take Home Points

Hwk #2: Data Management & Cleaning
Biostat 202: Opportunities and Challenges of "Big Data"
Due Wednesday, August 9th by Midnight

For this homework we will also use the data from the National Electronic Injury Surveillance System (NEISS) that you used for Hwk #1 (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. I have included the data as an Excel file 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.

There are a couple of issues with the race_other variable. First, if you look at the NEISS coding manual, you can see that race_other is only coded if a certain value is there for the race variable. Second, there are inconsistencies about how this is coded springing from the fact that this is coded as a string and not as a numeric variable. Because correct coding of text strings are higher operator dependent, it is subject to much variability as you can see with the frequent misspellings. Third, there are some entries in there such as Somalia, which doesn't quite make sense as a "race" category. With regards to how to recode this variable, there are several options, but the easiest would be to recode it into a numeric categorical variable, e.g. 1 = Hispanic, 2 = multiracial, . (or blank) = missing. Because one has to necessarily deal with misspellings, there are ways that you can specify certain values to be imputed. For example, because Hispanic contains the most number of entries, you can perform a condition statement such as "if field contains 'His', then set race_other = 1". The negative conditional (i.e. if race_other is not equal to 1 already, and there is an entry there, set value as 2 – these people are most likely multiracial). Then we can keep everything else as missing.

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. [An example is given below]
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 #2 in the 'Type' node.

In this question, it is crucial to understand when to use missing value imputation vs a missing value code. SPSS automatically enters a BLANK value for missing values once you specify them to be missing. The code is automatically generated by the program and you don't necessarily need to add an additional "code". In the NEISS data, you will notice that there are only 37 missing values in the age category compared to the 300k+ observations. The advantage of using a missing value code is that you use the data in its purest, unadulterated form. Your statistical analysis of the data will be dependent upon all the observations that do not have missing data. If your data are missing at random, keeping the missing value code will not bias your analysis in any way.

However, imagine that you have 100 entries and 37 of those are considered missing. This is a relative large percentage of missing values and one can imagine that potential for bias is much higher. Consider the case where age is missing because most of them are elderly people that forget how old they are when they show up to the emergency room. By removing these entries from the dataset, we have unintentionally skewed the data towards a younger patient population introducing bias to our data. In this case, by imputing a value for age can decrease the bias, especially if we know that the average age of the missing values should be higher compared to the population from which do not have missing age information.

4. (5 points). Recode body_part into two Flag variables to identify Head (body_part=75) or Other as the body_part injured, and another variable to identify Toe (body_part=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 (i.e.; text fields). Which variables have you eliminated?

Variable selection is also an important concept to understand, one that can also affect the analysis of your data. What constitutes good predictive information that can be entered into the model? Preferably, a variable that you know may affect the outcome in some way or another. If you are certain that the variable will not affect the outcome (e.g. a randomly assigned case ID number), then this should not be included into the model. The treatment date variable is a tricky one to think about. A certain date is unlikely to be predictive of admission/death, but perhaps the month or season of ED visit may be more predictive. You can imagine that in the winter time, people are more likely to be admitted from the ED because of severe flu. In this situation, treatment date would have to be recoded in order for you to ascertain any differences in admission based upon season or month.

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?

This question is highly dependent upon what you did with the 37 missing values for age. If you kept them as 0, then there would not be any missing data in that field. If you used a missing value code, you would have generated 37 missing values. And if you imputed a value to that field, you should again have not had any missing data in this field. Depending upon which variables you selected in question 5, this may have uncovered other data quality issues.

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?

The mean age for toe injuries was 29, mean age for head injuries was 30, and mean age for hospital admission was 55. The matrix node should have been used to tabulate admit by head or toe injury. Approximately 10% of head injuries were admitted compared to 1% of toe injuries. From this, you should have seen that head injuries were 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?*

This question is dependent upon your predictors that you selected for you model in question 5. In general, you should have seen that both logistic regression and neural network were approximately the same in accuracy

(~90-92%), meaning that our simple model is quite accurate in terms of predicting admission/death based upon the variables selected. You may find that if you incorporated variables such as treatment date or case ID, the accuracy of the model falls. This is because you may be controlling for a collider in your model (more on this in Epi 203!). Hence, it is important to remember to only include variables into your model that you think directly affect either the exposure and/or outcome of your model. The question regarding whether head or toe injuries predicting admission is slightly trickier. There are a couple of ways to do this including separating out the analysis of the model based upon toe injury or head injury. An alternative way (and perhaps a more straight forward way) is to look at the model output of the logistic and neural network. The neural network output is slightly easier to interpret and you should see that toe injury is actually a more important predictor compared to head (higher up on the list of predictors). The actual neural network is difficult to interpret and is beyond the scope of the course though. If we look at the model output for the logistic regression, the output shows exp(B) which provides us with the odds ratio from the model (see homework 2 guidance paper). An odds ratio is the probability of A happening divided by the probability of B happening – P(A)/P(B). Odds ratios also have the property that they are symmetrical. If you look at the output, you will see next to head = 0 that the exp(B) = 0.7 while for toe = 0, the exp(B) = 5. What this means is that if you don't have a head injury (head = 0) compared to patients that have a head injury (head = 1), you have a 0.7 times the odds of being admitted to the hospital. Because it is symmetric, we can easily say that if you have a head injury (head = 1), then you have a 1/0.7 = 1.3 times the odds of being admitted to the hospital compared to if you don't have a head injury (head = 0). Odds ratios always must have comparison group. Similarly, you can see that toe injuries (toe = 1) have a 0.2 times the odds of being admitted to the hospital compared to if you don't have a toe injury.

Recall that if you have an odds ratio of 1, you are equally likely to be admitted to the hospital if you have the predictors compared to if you don't have the predictor. The further you are away from 1, the "stronger" the association. Hence, since 0.2 is further away from 1 compared to 1.3, you can say that toe injuries are a better predictor of admission compared to head injuries. Alternatively, you can say that the predictive value for not being admitted to the hospital if you have a toe injury is stronger than the predictive value for being admitted to the hospital if you have a head injury.

*Reminder: you want to partition your data to have a training & testing subset here