Lab #1: Introduction to IBM SPSS Modeler
Biostat 202: Opportunities and Challenges of "Big Data"
Lab notes

6. The menu will display the first few lines of the dataset. Click on the "Preview" button to see how it did at reading in the data. What do you notice about the first 3 variables? **Q1. Why do you think Modeler was unable to read in the dataset using its default values?** **The commas in the name field are causing problems.**
7.
8. **Q2. For passenger number 6, what do you think happened to the value of "age"?** **Age is missing in the original dataset.**
9.
10.
11.
12.
13.
14.
15. Modeler will display each of the variables in a histogram (for continuous measurements) or a bar chart (for categorical measurements) with color coding for survival or not (since we have designated that as the target variable). Double-click on the Fare variable figure to enlarge the histogram. Note that it is highly skewed to the right, which we will consider below. **Q3. Looking at the figure, which fares are likely to fare better with regard to survival?** **Higher fares since those bars in the histogram are predominantly red (for survivors).**
16. Close the histogram window and click on the "Quality" tab in the output window from the "Data Audit" node. Modeler marks some of these as outliers (you can control the definition of an outlier in the "Quality" tab in the "Data Audit" node). **Q4. Which variables have missing data?** **Age and Embarked.**
17.
18.
19. Disconnect the "Data Audit" node from the "Type" node and attach it to the "Derive" node. You can do this with the F3 key, or by right-clicking on it and selecting "Disconnect." Or, if all else fails, delete it and add a new "Data Audit" node. Now check the effect of the transformation on the skewness and data quality. **Q5. Is the log_fare variable less skewed? Does it have fewer outliers?** **Yes – less skewed. And fewer outliers (3 versus 17).**
20.
21.
22.
23.

24. Next we will calculate some simple descriptive statistics. Let's create a table that gives the average value of age and the logarithm of fare separately for those that did and did not survive. Under the "Output" tab connect a "Means" node to the "SetToFlag" node. Open up the node and select "Survived" as the "Grouping Field" and select log_fare, Age and Pclass for the "Test Field" and click on "Run". **Q6. Do there seem to be differences between those that did and did not survive? Do they make sense? Yes – survivors paid more, were younger (e.g., children), and were in better classes (lower = better).**
25.
26.
27.
28.
29. Next we will calculate some summary statistics to see how well the model works. Add an "Analysis" node (under the "Output" tab) to the model nugget node, open it and click on "Run." **Q7. How well did it work? Did it do better on the training or test dataset? It did only OK, getting 62% of the predictions correct on the test data. The training data was slightly optimistic at 65%.**
30. Next try another method to see if you can do better than the logistic regression (so add "Modeling" and "Analysis" nodes). Good bets would be some of the modern machine learning methods such a Classification and Regression Tree ("C&R Tree"), C5.0, a neural net, or a support vector machine (SVM). More details on some of these algorithms later in the course. **Q8. What additional method did you try? Was it better, worse or about the same as the logistic regression? Can try a large number of methods, e.g., classification and regression trees (C&R Tree), neural net or support vector machine (SVM). Another alternative would be C5.0, which did better at 78%.**