Class Exercises

1. Using the concept of overfitting, explain why when a model is fit to training data, zero error with those data is not necessarily good.

* **The model may be overfitted for that particular dataset, yet in practice it may be inaccurate.**

1. Two models are applied to a dataset that has been partitioned. Model A is considerably more accurate than model B on the training data, but slightly less accurate than model B on the validation data. Which model are you more likely to consider for final deployment?

* **Model B because it is more accurate on the validation data.**

1. The dataset *ToyotaCorolla.jmp* contains data on used cars on sale during the late summer of 2004 in the Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications.

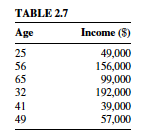
a. Explore the data using the data visualization (e.g., Graph > Scatterplot Matrix and Graph > Graph Builder) capabilities of JMP. Which of the pairs among the variables seem to be correlated? (Refer to the guides and videos at jmp.com/learn, under Graphical Displays and Summaries, for basic information on how to use these platforms.)

b. We plan to analyze the data using various data mining techniques described in future chapters. Prepare the dataset for data mining techniques of supervised learning by creating partitions using the JMP Pro Make Validation Column utility (from the Cols menu). Use the following partitioning percentages: training (50%), validation (30%), and test (20%). Describe the roles that these partitions will play in modeling.

1. A dataset has 1000 records and 50 variables with 5% of the values missing, spread randomly throughout the records and variables. An analyst decides to remove records that have missing values. About how many records would you expect would be removed?

-**The probability of an entire row having no missing values is (0.95)^50 = 0.0769. Therefore, the expected number of values is 7.69% of 1,000 or 70/1000. Theses are terrible odds and would greatly decrease your dataset size.**

1. Standardize (normalize) the data below showing calculations. Confirm your results in JMP.



* **Z Transform or 0-1 Transform.**