CPH 636 — Spring 2009 — Dr. Charnigo

Written Assignment 2

Written Assignment 2 is due on Monday 16 February at the end of lecture. You are encouraged to work in groups of two or three, though you may work individually if you prefer. *If you work in groups of two or three, please be sure that each group member is able to run the SAS macros, as each group member must be able to use the SAS macros for the noncollaborative final project.*

The data set in {rhythm.sas7bdat} is from {http://archive.ics.uci.edu/ml/datasets/Arrhythmia}. See {http://archive.ics.uci.edu/ml/machine-learning-databases/arrhythmia/arrhythmia.names} for information about the variables. For simplicity I retained only 11 variables, which I named as follows: Age, Female, Height, Weight, QRSDur, PRInt, QTInt, TInt, PInt, HeartRate, and ArrhythmiaType. Also, I created a recoded version of ArrhythmiaType called ArrhythmiaAny that equals 1 for patients with ArrhythmiaType > 1 and 0 for patients with ArrhythmiaType $= 1$.

The data sets in {rhythmtrain.sas7bdat}, {rhythmvalid.sas7bdat}, and {rhythmtest.sas7bdat} contain training, validation, and testing subsets that I generated using the RANSPLIT macro.

Please use the full data set for the first two exercises and the training subset for the last exercise.

[40] 1. Apply the UNIVAR macro to Age, Height, Weight, QRSDur, PRInt, QTInt, TInt, PInt, and HeartRate. *Do not hand in output from the UNIVAR macro except that which may be necessary to defend your answer to part b. Items a through c can be answered with a single well-constructed table.*

[10] a. Report the sample mean and sample standard deviation for each variable to which you have applied the UNIVAR macro.

[10] b. Comment on the shape of the distribution for each variable.

[10] c. For each variable identify any observations with outlying values as determined by cutoffs of $Q_3 + 1.5IQR$ and $Q_1 - 1.5IQR$. If any of the outlying values are so far out of line that you think they may be mistakes (there may or may not be any such outlying values), please make note of them.

[10] d. How many observations have outlying values on at least one of the variables? What do you think would happen if there were 90 variables instead of nine? What concept from Lecture 1 or Lecture 2 is suggested by your answer?

[10] 2. Apply the FREQ macro to ArrhythmiaAny while stratifying by Female. Report the numbers and percentages of observations in each of the $4 = 2 \times 2$ groups determined by these variables. *Present one graphical display that illustrates these numbers and percentages, but do not hand in any other output from the FREQ macro.*

[50] 3. Apply the REGDIAG macro to perform multiple linear regression on the training subset. Use HeartRate as the response variable. Use Age, Female, Height, Weight, QRSDur, PRInt, QTInt, TInt, PInt, and ArrhythmiaAny as explanatory variables. *Before completing the items below, you may remove any observations with outlying values so far out of line that you think they may be mistakes. If you do so, please state explicitly which observations are removed. Do not hand in output from the REGDIAG macro except that which may be necessary to respond to parts c and d.*

[10] a. Which explanatory variables have the strongest (pairwise) correlations with the response variable? Are there any explanatory variables for which you feel uneasy reporting correlations? Why or why not?

[10] b. Report and interpret R^2. Also, report adjusted R^2 and predicted R^2. Which is greater, R^2 or predicted R^2? Does that surprise you? Why or why not?

[10] c. Furnish the graphic that plots observed responses against fitted values. Comment on any pattern that you may see.

[10] d. Furnish the graphic that reflects response variability and residual variability. Comment on any pattern that you may see.

[10] e. Suppose that I had asked you to use ArrhythmiaType rather than ArrhythmiaAny as an explanatory variable. Even though there are 16 possible values for ArrhythmiaType, treating ArrhythmiaType as if it were continuous would be a serious mistake. Explain why. Then describe what you would have done instead.