CPH 636 — Spring 2009 — Dr. Charnigo

Written Assignment 3

Written Assignment 3 is due on Monday 09 March 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.*

[35] 1. Continue inspecting the output that you obtained from exercise 3 of Written Assignment 2:

[05] a. Are there any explanatory variables for which VIF plots suggest a collinearity problem? If so, which ones?

[05] b. Based on the normal probability plot of student-residuals and the plot of residuals against predicted values, do the assumptions of normality and constant variance seem reasonable? Briefly explain.

[05] c. Use $C(p)$ to identify which model seems best among those for which the REGDIAG macro provided output in the section on "2 best models within each subset". Call this "Model I". To get the coefficient estimates for Model I, you will need to run the REGDIAG macro again using only the variables in Model I.

[05] d. Repeat item c using the AIC. Let the model so identified be called "Model II". Model II may or may not be the same as Model I.

[05] e. Repeat item c using the SBC. Let the model so identified be called "Model III".

[10] f. Among the explanatory variables appearing in Model I, Model II, and/or Model III, for which variables are the augmented partial residual plots supportive of quadratic terms? Let Models IV, V, and VI be defined as analogues to Models I, II, and III that include quadratic terms for the variables that you identified. Your classmates may have a different Model IV, Model V, or Model VI based on differences in subjective visual impressions of augmented partial residual plots (as well as differences in which observations were deemed mistakes and removed in exercise 3 of Written Assignment 2).

Remark. I could have also added an item g, "Among the explanatory variables appearing in Model I, Model II, and/or Model III, for which pairs of variables are the interaction plots supportive of interaction terms? Let Models VII, VIII, and IX be defined as analogues to Models I, II, and III that include interaction terms for the pairs of variables that you identified. Let Models X, XI, and XII be defined as analogues to Models IV, V, and VI that include interaction terms for the pairs of variables that you identified." I have refrained from adding an item g only to keep this assignment to a reasonable length, not because interaction terms are unimportant. Indeed, when you undertake your final project, you will do well to consider interaction terms.

[10] 2. Use appropriately modified code from {SASMacroInstr4.txt} to assess R^2 on the *validation* subset for each of Models I through VI. Which model do you choose on this basis?

[05] 3. For the model chosen in exercise 2, use appropriately modified code from {SASMacroInstr4.txt} to assess R^2 on the *test* subset. This is an estimate of the fraction of variability in the response explained by the model *in the population* from which the sample was drawn.

[30] 4. Now consider the *classification problem* in which ArrhythmiaAny (or some misspelling thereof) is the response variable and candidate explanatory variables are Age, Female, Height, Weight, QRSDur, PRInt, QTInt, TInt, PInt, and HeartRate.

[05] a. Use the AIC to identify which model seems best among those for which the LOGISTIC macro provides output in the section on “model selection - AIC and SC statistics”. Call this “Model I”.

[10] b. For Model I, what cutoff or threshold for the estimated risk (of arrhythmia) should be used if we wish to maximize the correct classification rate? What if we wish to maximize the specificity subject to the constraint that the sensitivity exceed 0.90?

[05] c. Repeat item a using the SC. Let the model so identified be called “Model II”.

[10] d. Among the explanatory variables appearing in Model I and/or Model II, for which variables are the logit and/or partial delta logit plots supportive of quadratic terms? Let Models III and IV be defined as analogues to Models I and II that include quadratic terms for the variables that you identified.

[10] 5. Use appropriately modified code from {SASMacroInstr5.txt} to find the best attainable correct classification rate on the *validation* subset for each of Models I through IV. Which model do you choose on this basis?

[10] 6. For the model chosen in exercise 5, use appropriately modified code from {SASMacroInstr5.txt} to find the best attainable correct classification rate on the *test* subset. What is the best attainable specificity subject to the constraint that the sensitivity exceed 0.90?