FINAL PROJECT

COMPUTING FOR RESEARCH

Spring 2012

DUE APRIL 30, 5PM**. Submit via email to garrettm@musc.edu**

Use Sweave and R to complete the questions below. Turn in both (1) the final report in either postscript (PS) or PDF format and (2) your Sweave file. You may contact the instructor with any questions. Note that there is not a ‘correct’ model. This project is designed to show your facility with using R, latex and Sweave and reasonable statistical logic.

Use the ICU data and consider the multiple logistic regression model of STA on AGE, cancer part of the present problem (CAN), CPR prior to ICU admission (CPR), infection probable at ICU admission (INF), and race (RACE).

1. Perform any data exploration that you feel is necessary.
2. Write down an expression for the logistic regression model of STA on AGE, CAN, CPR, INF, and RACE. (note that RACE has three levels.) Write down the equation for the logistic regression model.
3. Obtain the maximum likelihood estimates of the parameters of the logistic regression model, their standard errors and p-values. Present them in a table and interpret the results.
4. Fit a reduced model based on statistical significance of predictors as determined in 3 (and state what your threshold for significance is) and any other justification (such as data exploration findings, collinearity, or model diagnostics). Determine a “final” model.
5. Based on the model in 4, present estimated odds ratios and 95% confidence intervals in a table. Interpret your findings in words.
6. Create (and include) an ROC curve for the model selected in 5.\*\*\*
   1. what is the AUC of the curve?
   2. does your model do a good job of predicting STA?
   3. Choose a probability cutoff based on the sensitivity and specificity produced by your model. What is the probability cutoff?
7. Comment on the predictive ability of the model and also on the difference between a model with significant p-values for the predictors versus a high AUC.

\*\*\* To create an ROC curve, you can use the R library ROCR. Here is an example of using commands with the ROCR library that should help you figure out the syntax of the R commands prediction and performance..

library(ROCR)

reg1 <- glm(cap.inv ~ gleason + log(psa) + detectable + factor(dpros),

family=binomial)

pred1 <- prediction(reg1$fitted.values, reg1$y)

perf1 <- performance(pred1,"tpr","fpr")

auc1 <- performance(pred1,"auc")@y.values[[1]]

plot(perf1, lwd=2, col=2)

abline(0,1)

legend(0.5,0.4, c("AUC=0.83"), lwd=2, col=2)