November 19, 2020 Name:_____

CS 624, Fall 2020

Quiz 2

100 total points

Use R to perform all necessary calculations. Attach your code and output. Give interpretation and discuss all relevant statistical measures.

Problem 1. (60 points) We are interested in finding the important predictors of behavioral type (Aggressive/Passive), assess their adjusted effect sizes (in direction and magnitude) and use the best logistic regression model for interpretation and prediction. We want to analyze the *Western Collaborative Group Study* dataset available in the *faraway* package in R. Variable description can be obtained via the ?wcgs command. The list below contains steps that might help you with your analysis.

a) (10 points) Remove the redundant or unnecessary variables. Check the variable types and recode if needed.

b) (10 points) Automatically build the best logistic regression model with main effects only according to the AIC minimization criterion. Give interpretation of the logistic regression coefficients.

c) (10 points) Obtain the ROC curve associated with the model obtained in part b. Find the AUC and comment.

d) (10 points) Find the optimal threshold that simultaneously maximizes specificity and sensitivity.

e) (10 points) Automatically build the best logistic regression model with main effects, 2-way and 3-way interactions. Compare the three models.

f) (10 points) Write a short paragraph summarizing all findings.

Problem 1. (40 points) We are interested in finding the important predictors of the number of damage incidents of ships, assess their adjusted effect sizes (in direction and magnitude) and use the best poisson regression model for interpretation and prediction. We want to analyze the *ships.dat* dataset available at *https://data.princeton.edu/wws509/datasets/#ship*. Variable description is provided at the site. The list below contains steps that might help you with your analysis.

a) (10 points) Build the best poisson regression model with main effects only. Give interpretation of the poisson regression coefficients. Since the ships have been observed over different time durations, an offset variable log(months) has to be introduced to the model to rescale the number of damage events to a common time duration. The code should look like:

glm(damage ~ type, offset(log(months)),data=d, family=poisson)

b) (10 points) Use the residual deviance to assess the goodness-of-fit.

c) (10 points) Use the test for overdispersion and comment.

d) (10 points) Refit the model using quasipoission family. Compare the two models.