October 22, 2020 Name:_____

CS 624 Biostatistics, Fall 2020

Quiz 1

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. (50 points) We are interested in finding the important predictors of fuel efficiency of cars, assess their adjusted effect sizes (in direction and magnitude) and use the best linear regression model for interpretation and prediction. We want to analyze the mpg dataset available in the ggplot2 package in R. that contains 234 observations on 11 variables. Description of the variables is available at https://ggplot2.tidyverse.org/reference/mpg.html. The list below contains steps that might help you with your analysis.

a) (5 points) Read the data into an appropriate data structure in R. Use the head() function to check if the first six rows of data are read properly. Average the cty and hwy to create a single outcome variable mpg.

b) (5 points) Use the dim(), names(), summary(), str(), hist(), boxplot() functions to obtain the dimensions of the dataset, the names of all variables, their summary statistics and histograms/barplots.

c) (5 points) Investigate if the variables have the correct type, outliers/sparse categories, engineer new variables if needed.

d) (5 points) Install the R package “MASS”. Use the stepAIC() function to build the best main effect model according to the AIC minimization criterion.

e) (5 points) Assess the goodness-of-fit using plots and analytical measures.

f) (5 points) Based on the result from part d), write a paragraph discussing the predictors of fuel efficiency and their adjusted effect sizes.

g) (5 points) Use the stepAIC() function to build the best model that includes 2-way interactions according to the AIC minimization criterion.

h) (5 points) Assess the goodness-of-fit using plots and analytical measures.

i) (5 points) Based on the result from part g), write a paragraph discussing the predictors of bmi and their adjusted effect sizes.

j) (5 points) Compare the two models and comment on which one you prefer.

Problem 2. (30 points) We know that the leverages h_i, $i = 1, 2, \ldots, n$ associated with a linear regression model $Y = X\beta + \varepsilon$ are the values on the main diagonal of the hat matrix $H = X(X^T X)^{-1} X^T$.

a) (10 points) Show that H is symmetric ($H = H^T$) and idempotent ($H = H^2$).

b) (10 points) Find the average of all leverages.

c) (10 points) Show that all leverages do not exceed one.