December 10, 2020

Final Exam

CS 624 Biostatistics, Fall 2020

Due on December 18, 2020

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. (20 points) We are interested in finding the important predictors of accumulated wealth at the time of retirement, 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 *Pension.txt* dataset (available on *Canvas*) that contains 194 observations on 17 variables: pyears - years of employment, prftshr - indicator for profit sharing company, choice - indicator for company giving a choice to contribute, female, married, age, educ - years of education, finc25, finc35, finc50, finc75, finc100, finc101- indicators for 25, 35, 50, 75, 100 and 101 levels of retirement contribution, wealth89 - wealth in thousands of dollars, race, stckin89 - percent of the portfolio in stock, irain89 - percent of the portfolio in IRA. Perform all necessary data analysis steps and write a section summarizing the findings.

Problem 2. (20 points) We are interested in finding the important predictors of online customers booking a room at a hotel, 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 *Travel.txt* dataset (available on *Canvas*) that contains 20,000 observations on 26 variables (description of all variables is presented in the *Data_Dictionary_Travel* file available on *Canvas*). Perform all necessary data analysis steps and write a section summarizing the findings.

Problem 3. (20 points) We are interested in finding the important predictors of number of non-spinal bone fractures in women with low bone densities, 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 *FITglm2.txt* dataset (available on *Canvas*, use the command read.delim("…\\FITglm2.txt",sep="\t")) that contains 6,459 observations on 18 variables: alloc - id, ra_age –age in years, frx – indicator for spinal fractures, nosp – indicator for non-spinal fractures, numnosp – number of non-spinal fractures (outcome variable), trt01 – indicator for treatment, p3_weigh – weight over 100 pounds, htotbmd – bone mass density 1, nbmd – bone mass density 2, trialyrs – duration of follow-up, riskcat4 – risk category, tneck –

bone density at the neck, bmd25 – indicator for osteoporosis (based on tneck values), hplac – indicator for high placebo dose, htrt – indicator for high dose treatment, lplac – indicator for low placebo dose, ltrt- indicator for lose dose treatment, rtgroup – risk group for falling. Perform all necessary data analysis steps and write a section summarizing the findings.

Problem 4. (20 points) We are interested in finding the predictors of three types of wines. The dataset *wine* is a part of the R *rattle.data* package that also provides the necessary variable descriptions (?wine). Perform all necessary data analysis steps and write a section summarizing the findings.

Problem 5. (20 points) We are interested in finding the predictors of the mortality risk for lung cancer patients. The dataset *lung* is a part of the R *survival* package that also provides the necessary variable descriptions (?lung). Perform all necessary data analysis steps and write a section summarizing the findings.