Assignment 3 – Biostatistics 3 – 2021

Instructions: Choose 1 (one) of the following data sets / problems to analyse. Submit your write up and code file. Your analysis should focus fairly narrowly on the question mentioned in text, and you should restrict yourself to no more than 3 pages, including any tables and figures. You should work on your analysis and write up on your own, but you are welcome to ask questions about coding in the forums, or any questions about clarification of the variables in the data sets. All data sets can be found in data on Vula.

Problem 1:
1. Data set: WHAS Analysis: Survival

Fit Cox proportional hazards AND at least **2 parametric survival models** to the WHAS data set. Estimate the effects of age, sex, length of hospital stay (LENSTAY), grouped cohort year (YRGRP), and left heart failure complications (CHF) on long-term survival following hospitalization for an acute myocardial infarction in the WHAS dataset. Use LENFOL as survival time (days) and FSTAT as the censoring variable (1 = death observed). Report hazard ratios and 95% confidence intervals. Additionally, test the null hypothesis that men and women have equal survival curves using a log-rank test. Compare models, which model is best?

Problem 2:
2. Data set: city-data-gee.csv Analysis: Clustered data / interaction

We are interested in the association between smoke exposure on respiratory problems in children (in this case wheeze).
Resp - 1 = wheeze, 0 = doesn't.
Id - child id number – should be coded as a factor.
Age - how many years older than 9 the child is. [ie this is a centred age variable]
smoke - did their mother smoke in their first year of life?

Using GEE or mixed model approach, fit a model including both fixed effects (age, smoke) as well as the interaction between age and smoking. Try to get "Wald-type" tests of associations. In R you can get this via geeglm::anova; in Stata via xtgee postestimation test. What are your conclusions? **Fully interpret the age and smoking effects in the prescence of interaction.**

Problem 3:
3. Data set: fatal-train.csv Analysis: count – exposure
Numbers of fata rail accidents and millions of main line train km annually for Britain 1946-2003, rail was privatised in 1994. Fit Poisson, and negative binomial models to this data. **Is the rate of fatalities changing? Did privatisation have an impact on the rate fatalities?**

Problem 4:
4. Data set: nhanes-adult Analysis: multinomial
What are the factors associated with self reported good health?

HealthGen: Self-reported rating of participant's health in general. Excellent, Vgood, Good, Fair, or Poor.
Age: Age at time of screening (in years). Participants 80 or older were recorded as 80.
PhysActive: Participant does moderate to vigorous-intensity sports, fitness or recreational activities
Poverty: 0-5 – ratio measure of poverty (family income:standard), smaller numbers indicate worse level of poverty
BMI: BMI at time of screening