Assignment 5

Note: For this problem, we will use publicly available training data (the Synthetic National Early Warning Scores Data) from the UK NHS. Follow these steps to load the data:

- Run the commands install.packages("NHSRdatasets") and library(NHSRdatasets)
- Then load and store the dataset called "synthetic_news_data". Use ?synthetic_newd_data to get the data dictionary.

Problem 1: Regression. We will examine the relationship between patient demographics and a predicted "national early warning score" used to prioritize patients. The dependent variable, NEWS, ranges from 0 (no risk) to 12 (urgent risk). We will use the following independent variables: patient age, patient gender, patient alertness ("alert"), patient blood pressure ("syst" and "dias") and patient temperature ("temp").

a. First regress NEWS only on patient age. Report the coefficients, standard errors, confidence intervals, p-values, R^2, and sample size in a well-formatted regression table. Interpret the table, noting the economic and statistical significance of the relationship. What is the associated change in NEWS from a 10-year increase in patient age?

b. Now run the full regression and update the table (or add a column to the initial table, if you like). Answer the same questions above – how have they changed?

c. What other regressors are significant? How might we interpret the associations (in their contexts)? How do we interpret the associations for the insignificant variables?

d. Do you recommend a nonlinear transformation for any of the variables included? If so, defend your choice and repeat the regression with the appropriate transformations. Interpret how your results have changed.

e. Rather than including patient blood pressure ("syst" and "dias") and patient temperature ("temp") in levels, create dummy variables for high blood pressure (syst >= 130 or dias >= 80) and high temperature (temp > 38). Update the regression using these dummy variables; how have things changed?

f. Finally, include an interaction term between patient age and your high blood pressure dummy. What are you measuring with this interaction, and why might it be meaningful? Interpret the results of this coefficient in an updated regression.

g. Draw a DAG that suggests relationships between your variables and justify it with text. What does your DAG tell you about interpreting any regression coefficients causally? Why are establishing causal relationships difficult given this data and regression equation?

h. What choice of standard errors would you make in this regression (i.e., homoskedastic, heteroskedasticity-robust or clustered SEs)? Update the (latest) regression and interpret any changes.