

SMDE FIRST ASSIGNMENT (40% OF THE FINAL MARK)

THIRD QUESTION: REGRESSION ANALYSIS (25% OF THE FIRST ASSIGNMENT).

Continue on working on the Kaggle data set ([Technology Usage, Stress and Wellness](#)) . You need to use the variables created in the previous questions to answer following questions.

We aim to fit a linear regression model to predict mental health score.

- a) Consider the numerical variables in the data set and find the best simple linear regression model to predict the mental health score. Test the assumptions and use transformations if it is required. Explain why the model you find is the best simple linear regression model and interpret the model coefficients. (2.5p)
- b) Fit a multiple linear regression model with the most important two (numerical) variables. Use transformations if it is needed and test all the assumptions. Then compare this model to the simple linear regression model that you fit in (a). Which one is a better model? Why? (2.5p)
- c) Now we will add a factor to the regression model you have chosen in section (b). First, add generation (age_cat) as a factor. Interpret the model coefficients and overall summary of the model. Then add the stress level factor (stress_cat) to the model that you had chosen in (b). Interpret the model coefficients and overall summary of this model similar to the previous case. (3p)
- d) Compare the model in section (b) with the two models that had constructed with an additional factor in (c). Which one would you choose? Why? (1p)
- e) Test the validity of the final model that you choose. (1p)