Exercise 5 (Regression models)

We keep using the NHANES data from your R Data Project.

Task 1 (linear regression)

1. How strong is the relationship between BMI and systolic blood pressure? Is it significant? How much of the variation in systolic blood pressure can be explained (in a statistical sense) by variation in BMI?

For the following exercise, create a binary age (younger: $age \leq 50$ and older: $age > 50$) and a binary smoking variable.

2. Does the relationship between systolic blood pressure and BMI change when you adjust for age (categorized)? Interpret the coefficients of the resulting model (when you mean-center BMI before fitting the model, you can also interpret the intercept). Would you say that BMI has a clinically relevant impact on blood pressure, according to your model?
3. Try to find a better model to predict systolic blood pressure by including more covariates. Select a number of candidate covariates which in your opinion may be related to systolic blood pressure, and then choose a model selection strategy and a criterion/test for comparing models. Describe the model with the best fit according to your search, and interpret the model coefficients.

Task 2 (logistic regression)

1. Analyze the relationship between lifetime diagnosis of cancer and exposure to pollutants, using the categorized age variable (Note: No information on pollutant exposure was collected from participants aged 80+, so these cannot be included in the analysis). Does the adjustment for age change the picture? Interpret the model coefficients including the intercept.

2. Try to find a good model of cancer diagnosis, describe and interpret it as you did for systolic blood pressure.