Lab: Linear regression for machine learning

Learning objectives.

- Predict a quantitative outcome/label with linear regression in `R`.
- Assess prediction performance model using a test set.
    a. Read in (`read.table`) the Brain weight dataset. Examine (`head`) the dataset.
    b. Convert Sex and Age to factor variables so that `lm` can be properly dealt with them.
    c. Split the data into training (70%) and test (30%) sets. (set the seed with for example `set.seed(2024)` for reproducibility)
    d. Fit a linear regression model with brain weight as the outcome and head size and Sex as predictors. What is the interpretation of the coefficients for head size and Sex? (not a ML question, just checking!) Compute the training and test RMSE and R^2.
    e. Fit now a linear regression model with brain weight as the outcome and head size, Sex, and Age as predictors. What is the interpretation of the coefficients for Sex and Age? Compute the training and test RMSE and R^2. Does adding Age improves *prediction performance* over the model with Sex and head size alone?
    f. Explore whether fitting a linear regression model with separate intercepts and separate slopes for $20 \leq$ Age < 46 and Age ≥ 46 improves prediction performance over the model `Brain.weight ~ Age + Brain.size` (hint: you can specify an interaction between Age and Head size by including `Head.Size:Age` in the model formula.
    g. Bonus points! Compare your results from f. to fitting two separate models: `Brain.weight ~ Brain.size` for individuals $20 \leq$ Age < 46 and `Brain.weight ~ Brain.size`` for individuals Age ≥ 46. Is this equivalent to the single model you fitted in e.? Explain (hint: think about the residual sum of squares being minimized in each case to obtain the model coefficients).