Lab 04: Regression Criticism

**Due date:** Thursday, March 14, 2025 submitted as Word document to Canvas ***Lab04*** link

This lab counts 9 % toward your total grade.

**Objectives:** In this lab, you will practice your skills in

1. Explore multiple regression
2. Residual plot
3. Multicollinearity
4. Outlier detection

**Format of answer:** Submit your answers as a **Word document** with graphs and verbal descriptions, properly labeled in the task sequence, with answers in red text and only relevant content included

# Task 1: Multiple regression (9 pts)

We will use the Boston dataset from the **MASS** package, which contains **506 observations and 15 variables** related to housing prices in the Boston suburbs. Data information please check this [link](https://search.r-project.org/CRAN/refmans/ISLR2/html/Boston.html) (Provide R code for each task and explain the result).

Total\_area is a simulated variable that is calculated based on **rm**.

1. Construct a multiple regression model using **medv** (median home value) as the dependent variable, with all other variables **except lstat** as independent variable (a total of 14 variables). (1 pt)
2. Use Variance Inflation Factor (VIF) to detect multicollinearity among the predictors. (1 pt)
3. If multicollinearity is detected, remove **one highly collinear variable** based on domain knowledge and reasoning. Fit a new regression model (Model 2) without the problematic variable. (2 pts)
4. Compare Model 1 (from Task1. a) and Model 2 (from Task1. c) using: Standard errors and t-statistics. Please illustrate how multicollinearity impacts model estimates by analyzing changes in coefficient significance. (2 pts)
5. Check whether residuals exhibit constant variance (homoscedasticity) in Model 2. If heteroscedasticity is detected, suggest one potential remedies. (1 pt)
6. Create Model 3 by adding **lstat** to Model 2. Use **ANOVA** to compare Model 2 (without lstat) and Model 3 (with lstat) to determine if lstat significantly improves model performance. (1 pt)
7. Based on Model 3, using **car::influenceIndexPlot()** to investigate the presence of outliers or influential cases in the dataset. Explain the results and analyze the possibility that these observations might be influential. (1 pt)