Linear Regression Exercise 1

Part 1

Given a matrix of explanatory variables `X` and a vector with response variable `Y`, write a function that calculates the coefficients of an OLS model. The function should output a list with three elements:

- `beta_hat`: A vector of the coefficient estimates
- `v_beta_hat`: A matrix with the estimated variance of `beta_hat`
- `S2`: A scalar estimate of σ^2

For this function, you may assume that the regression matrix `X` has full rank.

Part 2

Load the Boston Housing Dataset contained in `HousingData.csv`. There are 14 variables:

- `CRIM`: per capita crime rate by town
- `ZN`: proportion of residential land zoned for lots over 25,000 square feet
- `INDUS`: proportion of non-retail business acres per town
- `CHAS`: Does tract bound the Charles River?
- `NOX`: Nitric oxides concentration
- `RM`: Average number of rooms per dwelling
- `AGE`: Proportion of owner-occupied units built prior to 1940
- `DIS`: Weighted distances to five Boston employment centers
- `RAD`: Index of accessibility to radial highways
- `TAX`: Full-value property tax rate per $10,000
- `PTRATIO`: Pupil-teacher ratio by town
- `B`: A measure of the proportion of Black residents by town
- `LSTAT`: Percentage lower status of the population
- `MEDV`: Median value of owner-occupied homes in $1000s

Use the regression function you defined previously to estimate the coefficients of a regression model using the average number of rooms per dwelling (`RM`) as the explanatory variable and the median value of owner-occupied homes (`MEDV`) as the response variable. Then, compare the results of your function to the results you receive using the `lm()` function.

Part 3

Economists typically model real estate prices as a function of the amenities provided by the house (e.g. number of rooms, age, distance to workplace, education quality, etc.). In this section, we focus on the effect of education on real estate prices. We assume that a higher pupil-teacher ratio usually indicates lower funding for education. Notably, in the given dataset, there are two conflicting effects on home values:

- A lower pupil-teacher ratio indicates higher funding for education, leading to higher home values
- Higher funding for education often requires higher property taxes, which likely leads to lower home values

Using your regression function defined above, fit a regression model to quantify the associations between pupil-teacher ratio (`PTRATIO`), property taxes (`TAX`), and home values (`MEDV`). Compare the results of this model to the results you receive using the `lm()` function.