Assignment 16: Energy Efficiency Analysis with Single Linear Regression

62 Points scaled to 20 Points

Introduction

This assignment explores single linear regression to predict the heating load for buildings based on characteristics of the buildings. Specifically, you will explore regression models in which building heat load is predicted using building relative compactness, wall area, and roof area. These data ("energy_efficienty_data.csv") were obtained from Kaggle: https://www.kaggle.com/datasets/ujjwalchowdhury/energy-efficiency-data-set.

Objectives

- *Prepare data for analysis and prediction using single linear regression*
- *Explore correlation between continuous variables*
- *Create and compare linear regression models*
- *Interpret regression coefficients and R-squared and RMSE metrics*

Deliverables

- *Jupyter Notebook (Python) or R Markdown file (R) with all code and results embedded. Files can be rendered to HTML webpages if your instructor requires this. Questions should be stated and answered within Markdown cells.*

Background Questions

Question 1: Provide a general single linear regression equation and explain all the terms. (5 Points)

Question 2. Explain how the best fitting line is determined using the ordinary least squares method. (5 Points)

Question 2. What is a residual? How are residuals calculated? (5 Points)

Question 3. Explain how R-squared is calculated. (5 Points)

Question 4. What are the units of R-squared? (5 Points)

Question 5. Explain how Root Mean Square Error (RMSE) is calculated. What are the units of measurement for RMSE? (5 Points)

Tasks and Questions

This assignment can be conducted using either Python or R, whichever you prefer or whichever you instructor requires. Generate code to perform the following analyses and answer the stated questions.

Task 1: Extract out only the "Heating_Load", "Relative_Compactness", "Wall_Area", and "Roof_Area" attributes. (2 Points)

Task 2: Create scatterplots to compare the dependent variable ("Heat_Load") with each of the three independent variables separately. (5 Points)

Task 3. Obtain Pearson correlation coefficients for "Heat_Load" and each of the three independent variables of interest. (5 Points)

Question 6. Discuss the relationships between the dependent variable and each independent variable based on the scatterplots and correlation coefficients obtained. Does each independent variable appear to be correlated with "Heat_Load"? If so, is the correlation positive or negative? Is the correlation linear? (5 Points)

Task 3. Create three separate single linear regression models in which "Heat_Load" is predicted using one of the available predictor variables. (5 Points)

Task 4. Calculate or obtain the R-squared and RMSE metrics for each of the three models. Note that you do not need to separate the data into separate training and testing sets in this assignment. So, you can just predict back to the data used to create the models. (5 Points)

Question 7. Compare the model fit using R-squared and RMSE. Do the model results agree with the results obtained using the scatterplots and Pearson correlation coefficients? In other words, did the variable with the highest linear correlation yield the best performing model? (5 Points)