Practice 3.1 Regression

Statement

Description

The dataset **insurance.csv** contains hypothetical medical expenses for patients in the United States. This data was created using demographic statistics from the US Census Bureau, and thus, approximately reflect real-world conditions.

The goal of this analysis is to use patient data to estimate the average medical care expenses for such population segments. These estimates can be used to create actuarial tables that set the price of yearly premiums higher or lower, depending on the expected treatment costs.

Attribute Information:

1. **age**: An integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
2. **sex**: The policy holder's gender, either male or female.
3. **bmi**: The body mass index (BMI), which provides a sense of how over- or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
4. **children**: An integer indicating the number of children/dependents covered by the insurance plan.
5. **smoker**: A yes or no categorical variable that indicates whether the insured regularly smokes tobacco.
6. **region**: The beneficiary's place of residence in the US, divided into four geographic regions: northeast, southeast, southwest, or northwest.
7. **Charges**: total medical expenses charged to the plan for the calendar year.

Tasks

1. Load the data set: **insurance.csv**
2. Identify a linear regression model to estimate the variable “charges”.