ASSIGNMENT

Predicting Medical Insurance Charges Using Linear Regression

Course: Regression Analysis / Applied Statistics

Assignment Type: Individual Assignment

Total Marks: 100

Submission Deadline: 2rd March 2026

Submission Format: PDF Report (E-Learning) + Code File (.ipynb) (GitHub repository)

1. Background and Context

Health insurance companies use statistical models to estimate medical charges based on demographic and lifestyle characteristics of clients. Accurate prediction models help companies:

- Set fair premiums
- Manage financial risk
- Identify high-risk groups
- Improve pricing strategies

In this assignment, you will act as a data analyst working for an insurance company. You are required to develop and evaluate a **Linear Regression model** to predict medical insurance charges using real-world data.

2. Dataset

Dataset Name: Medical Cost Personal Dataset

Source: Provided (excel)

You are required to download the dataset directly from E-Learning.

Variables in the Dataset

- age – Age of beneficiary
- sex – Gender
- bmi – Body Mass Index
- children – Number of dependents
- smoker – Smoking status
- region – Residential region
- charges – Medical insurance cost (**Target Variable**)

ASSIGNMENT REQUIREMENTS

You must document all steps clearly. Code alone is NOT sufficient. Explanations and interpretations are mandatory.

PART A: Data Acquisition and Understanding (5 Marks)

1. Download the dataset and import it into Python or R.
2. Display:
   - First 5 observations
   - Number of observations and variables
   - Data types of each variable
3. Briefly describe:
   - What the dataset represents
   - Which variable is the dependent variable
   - Which variables are independent variables

Guidance: Use .info() and .describe() (Python).

Part B: Data Cleaning and Preprocessing (20 Marks)

You must clearly explain each step taken.

1. Missing Values

- Check for missing values.
- If present, explain how you handled them.

2. Duplicate Records

- Check for duplicates.
- Remove or justify keeping them.

3. Outliers

- Use boxplots or statistical methods (e.g., IQR).
- Identify extreme values in BMI and charges.
- Explain whether you removed or retained them and why.

4. Encoding Categorical Variables

- Convert categorical variables into numerical format.
- Clearly explain the encoding method used (e.g., dummy variables).

5. Feature Scaling

- State whether scaling is necessary for linear regression.
- Justify your answer.

Marks will be awarded for correct reasoning, not just execution.

Part C: Exploratory Data Analysis (EDA) (15 Marks)

Include visualizations and interpretation.

1. Summary Statistics

- Present descriptive statistics for numeric variables.

2. Distribution of Charges

- Plot histogram of charges.
- Comment on skewness.

3. Relationship Analysis

Produce and interpret:

- Scatter plot: Age vs Charges
- Scatter plot: BMI vs Charges
- Boxplot: Charges by Smoker Status
- Correlation matrix (numeric variables)

Guiding Questions:

- Which variable appears most strongly related to charges?
- Do smokers pay more than non-smokers?
- Are relationships approximately linear?

Interpretations must accompany all graphs.

Part D: Linear Regression Modeling (25 Marks)

Section 1: Simple Linear Regression (10 Marks)

1. Select ONE independent variable.
2. Fit a simple linear regression model.
3. Write the regression equation in mathematical form.
4. Interpret:

  - Intercept
  - Slope coefficient
  - R^2

Explain what the slope means in practical terms.

Section 2: Multiple Linear Regression (15 Marks)

1. Fit a multiple linear regression model using all relevant predictors.
2. Present:
   - Coefficients
   - Standard errors
   - p-values
   - R^2
   - Adjusted R^2
3. Write the full regression equation.

Interpret:

- Which variables are statistically significant?
- Which factor has the strongest impact?
- How does smoking affect medical charges?

Part E: Model Evaluation and Assumptions (15 Marks)

1. Split data into training (70%) and testing (30%).
2. Compute:
   - RMSE
   - MAE
   - R^2 on test data
3. Check regression assumptions:

- Linearity
- Homoscedasticity (residual plot)
- Normality of residuals (histogram or Q-Q plot)
- Multicollinearity (VIF)

Discuss whether the assumptions are satisfied.

Part F: Interpretation and Business Recommendations (10 Marks)

Answer the following:

1. Which factor increases medical insurance charges the most?
2. If BMI increases by 1 unit, what is the expected change in charges?
3. Provide three practical recommendations for the insurance company.

Recommendations must be supported by your statistical findings.

SECTION G: Report Writing and Presentation (10 Marks)

Your report must include:

1. Title Page
2. Introduction
3. Methodology
4. Results
5. Discussion
6. Conclusion
7. References

Formatting Requirements:

- Maximum 8 pages
- 12-point font
- 1.5 line spacing
- All figures properly labeled
- Clear tables with titles

Marks will be awarded for clarity, structure, and professionalism.

MARKING RUBRIC

Section	**Marks**
Data Acquisition	5
Data Cleaning	20
EDA	15
Regression Modeling	25
Model Evaluation	15
Interpretation	10
Report Quality	10
Total	**100**