Comprehensive Linear Regression Analysis: Insurance Charges Prediction

Date: October 26, 2023

1. Introduction

This report details a comprehensive linear regression analysis performed on the 'insurance.csv' dataset to predict insurance charges. The analysis includes data loading and inspection, exploratory data analysis, data preprocessing, implementation and evaluation of simple linear regression models (for age, smoker, BMI), and a multiple linear regression model. Visualizations are provided to illustrate model performance and insights.

2. Data Loading and Inspection

The 'insurance.csv' dataset, consisting of 1338 entries and 7 columns, was successfully loaded. The dataset includes numerical features like `age`, `bmi`, `children`, and `charges`, and categorical features such as `sex`, `smoker`, and `region`. No missing values were observed, and data types were appropriate for analysis.

3. Exploratory Data Analysis (EDA) and Feature Selection

Distributions of all features were visualized. Key observations included: `age` and `bmi` showing somewhat normal distributions, `children` skewed, and `charges` heavily right-skewed. `smoker` status showed a very strong relationship with `charges`. For linear regression, `age`, `bmi`, and `smoker` were identified as key features to explore, with `smoker` requiring one-hot encoding.

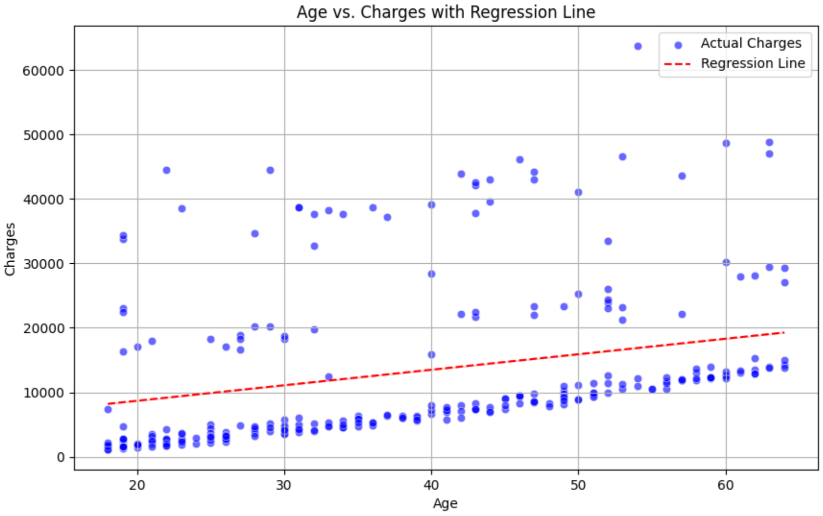
4. Simple Linear Regression Model: Age

This model uses `age` as the sole independent variable to predict `charges`.

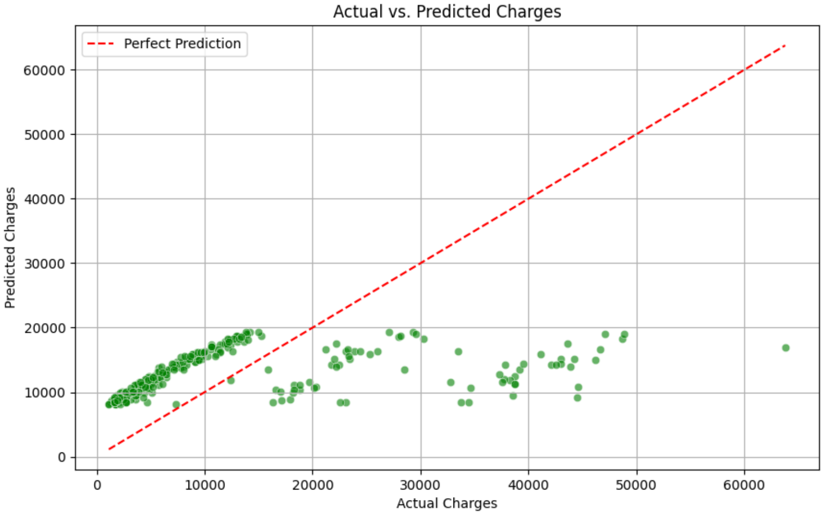
Model Intercept: 3876.93**Model Coefficient (for age)**: 240.60

Interpretation: For every one-year increase in age, predicted charges increase by approximately $240.60.

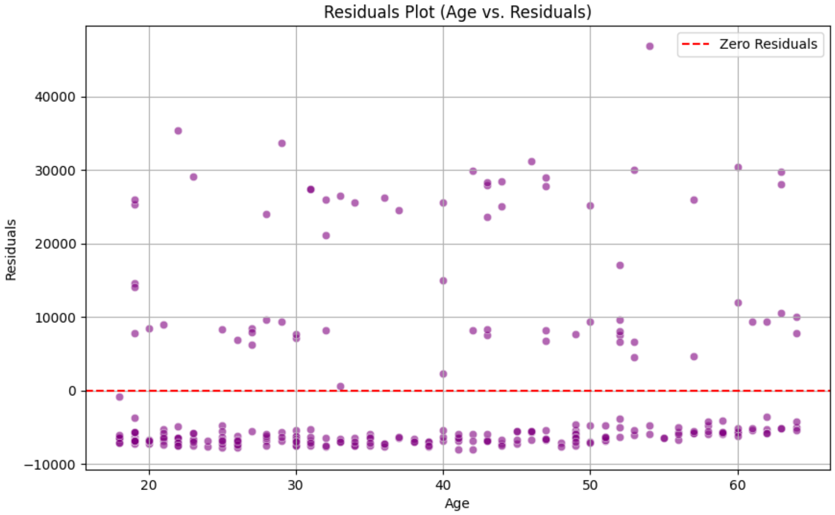
Performance: R-squared: 0.1241, MAE: $9173.26, MSE: $135983957.48. The low R-squared indicates `age` alone is a weak predictor.



Age vs. Charges with Regression Line



Actual vs. Predicted Charges (Age Model)



Residuals Plot (Age Model)

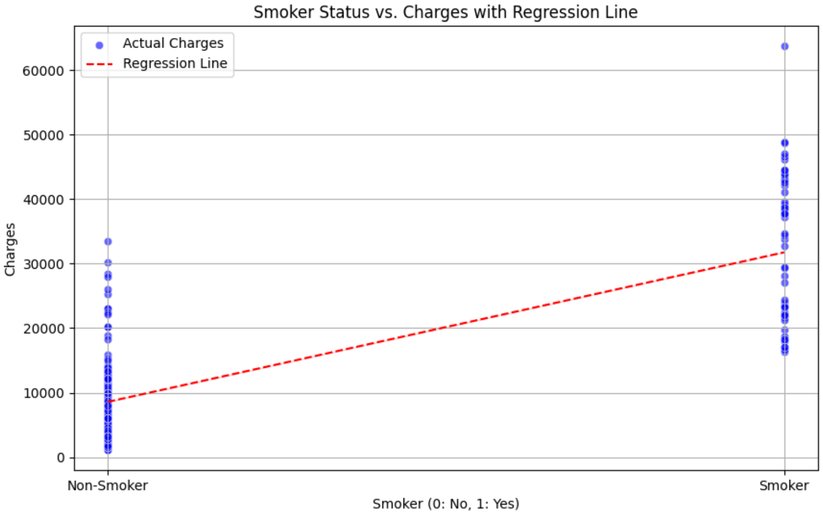
5. Simple Linear Regression Model: Smoker

This model uses one-hot encoded `smoker_yes` (1 for smoker, 0 for non-smoker) to predict `charges`.

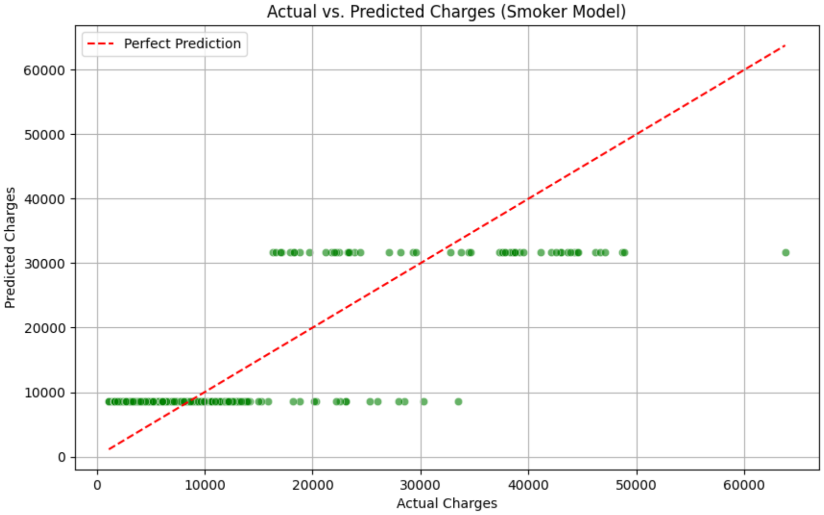
Model Intercept: 8578.32**Model Coefficient (for smoker_yes)**: 23188.69

Interpretation: Non-smokers have an estimated charge of $8578.32. Smokers incur an additional $23188.69 compared to non-smokers.

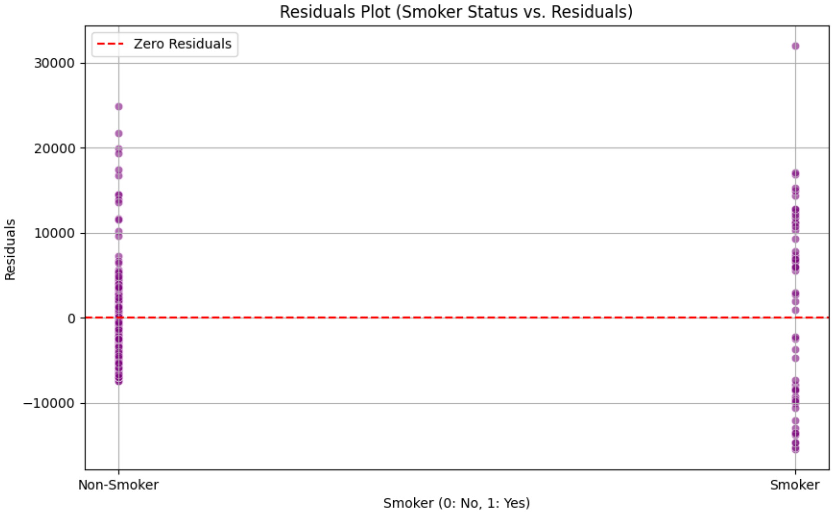
Performance: R-squared: 0.6602, MAE: $5625.81, MSE: $52745964.73. This model shows a significantly better fit than the age model.



Smoker Status vs. Charges with Regression Line



Actual vs. Predicted Charges (Smoker Model)



Residuals Plot (Smoker Model)

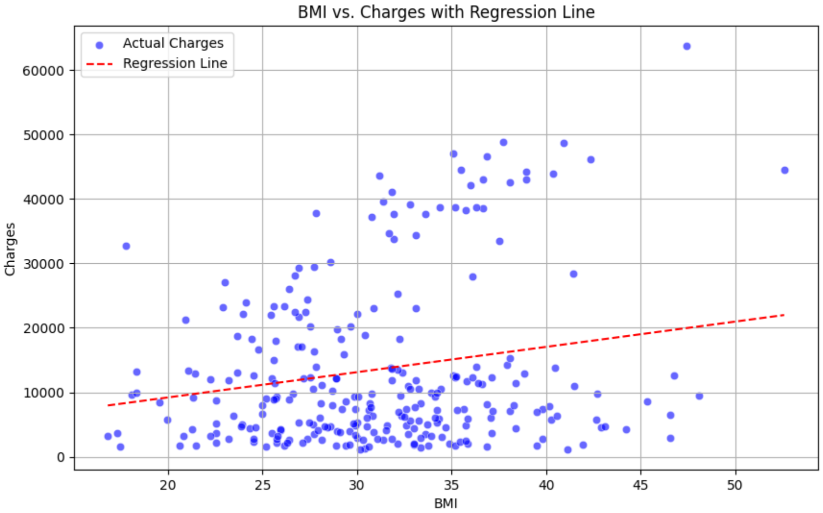
6. Simple Linear Regression Model: BMI

This model uses `bmi` as the sole independent variable to predict `charges`.

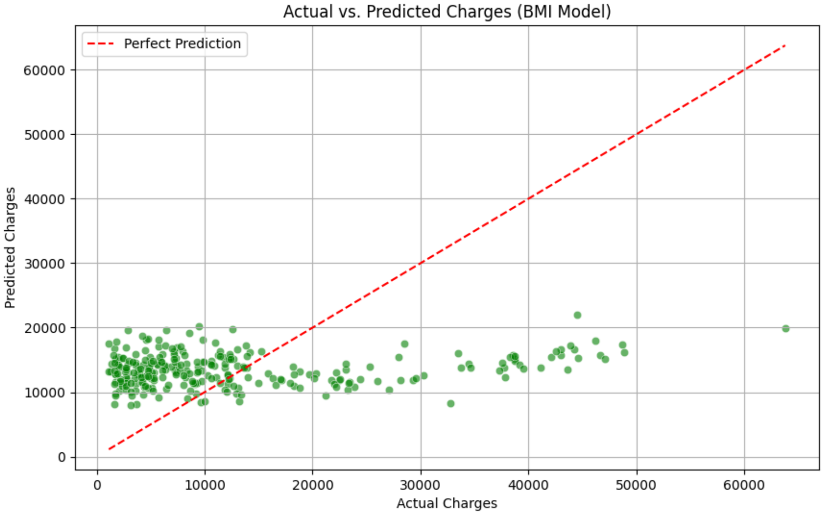
Model Intercept: 1353.07**Model Coefficient (for bmi)**: 392.44

Interpretation: For every one-unit increase in BMI, predicted charges increase by approximately $392.44.

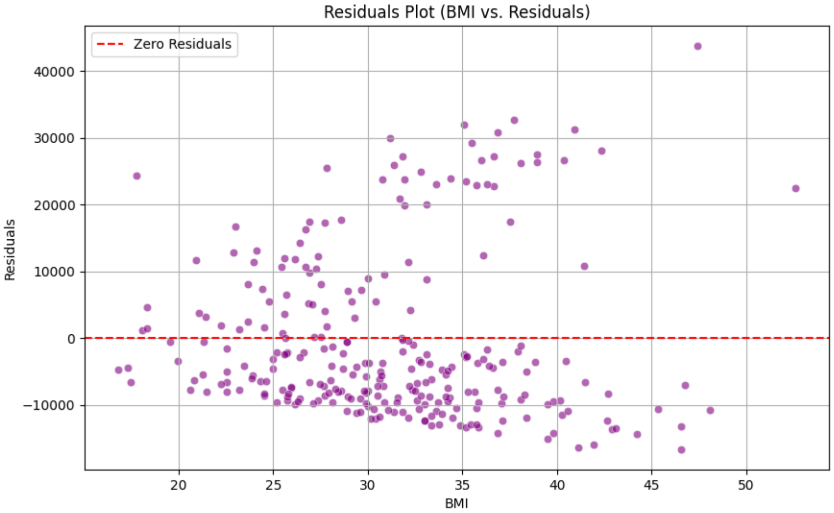
Performance: R-squared: 0.0397, MAE: $9784.65, MSE: $149085057.04. The low R-squared indicates `bmi` alone is a weak predictor.



BMI vs. Charges with Regression Line



Actual vs. Predicted Charges (BMI Model)



Residuals Plot (BMI Model)

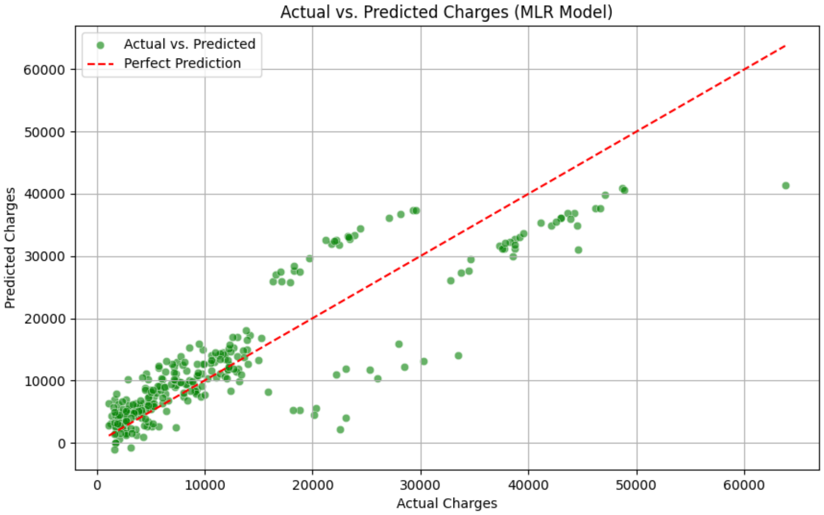
7. Multiple Linear Regression Model (Age, BMI, Smoker)

This model incorporates `age`, `bmi`, and `smoker_yes` as independent variables.

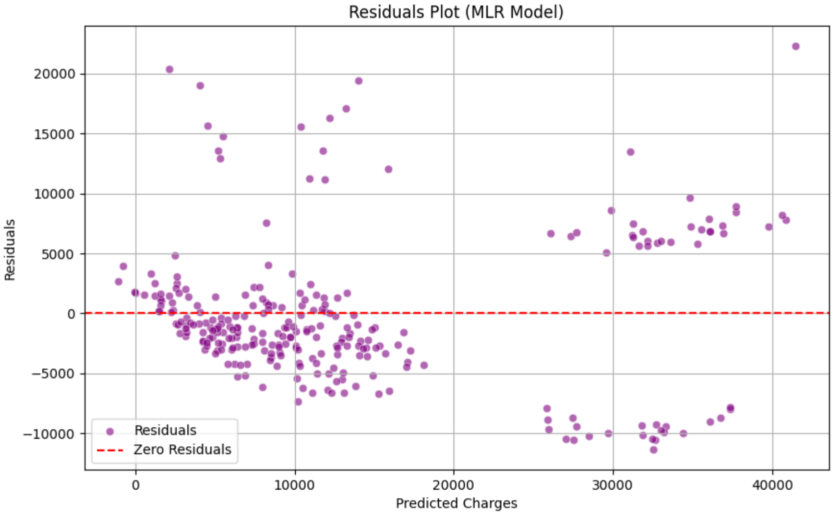
Model Intercept: -11707.80**Model Coefficients**: Age: 259.41, BMI: 326.45, Smoker_yes: 23675.37

Interpretation: Each additional year of age increases charges by $259.41. Each additional BMI unit increases charges by $326.45. Being a smoker increases charges by $23675.37 compared to non-smokers (holding other factors constant).

Performance: R-squared: 0.7777, MAE: $4260.56, MSE: $34512843.88. This model shows a substantial improvement over the simple linear regression models.



Actual vs. Predicted Charges (MLR Model)



Residuals Plot (MLR Model)

8. Comparative Analysis of Models

A comparison of the R-squared values across models highlights the significant impact of including `smoker` status in the prediction:

Age SLR: R-squared = 0.1241**BMI SLR**: R-squared = 0.0397**Smoker SLR**: R-squared = 0.6602**Multiple LR (Age, BMI, Smoker)**: R-squared = 0.7777

The 'Smoker' simple linear regression model significantly outperforms both 'Age' and 'BMI' simple linear regression models, demonstrating that `smoker` status is the strongest individual predictor among the features examined. The multiple linear regression model, combining 'age', 'bmi', and 'smoker', achieved the highest R-squared and lowest MAE/MSE, indicating it's the most effective model for predicting charges among those built.

9. Limitations and Future Work

The simple linear regression models, particularly for 'age' and 'bmi', showed limited predictive power. While the 'smoker' variable dramatically improved performance, the residuals plots for all models indicated remaining unexplained variance and potential violations of linearity assumptions, suggesting that a purely linear relationship might not fully capture the complexities of insurance charges. Future work should explore:

Including more categorical variables like `sex` and `region` (one-hot encoded) and `children` in the multiple linear regression model.Investigating interaction terms between variables (e.g., age * smoker, bmi * smoker).Considering non-linear models or transformation of variables to better address patterns in the residuals.Analyzing outliers and their impact on model performance.