**Part I: Research Question**

I have chosen a dataset with 7 columns which stores the age, gender, BMI, number of children, smoking habits, region of residence and the insurance charges that a company would charge for each person.

1. Summarizing the Research Question:

Research Question: "How do age, BMI, number of children, smoking status, and region affect insurance charges for individuals?"

Explanation: In this context, insurance companies might want to predict the medical charges they expect to incur based on customer demographics and lifestyle choices. This research question can help insurers understand which factors most significantly influence the cost of healthcare and thus help in pricing their insurance products more effectively.

2. Define the Goals of the Data Analysis:

Primary Goal: To build a predictive model using multiple linear regression to estimate insurance charges based on customer characteristics such as age, BMI, number of children, smoking status, and region.

Secondary Goal: To identify which factors (e.g., smoking status, BMI) are the most significant drivers of high medical charges. This insight can help in segmenting risk categories for different customers and potentially guide the creation of targeted wellness programs or insurance plans.

Justification:

The dataset provides enough relevant variables to model insurance charges as the dependent variable (charges) and includes a mix of demographic and lifestyle-related independent variables (age, sex, bmi, children, smoker, and region). By using multiple linear regression, we can estimate the relationship between these variables and insurance costs.

**Part II: Method justification**

1. Four Assumptions of a Multiple Linear Regression Model:

Multiple linear regression relies on several key assumptions to provide valid results. These are the four most important assumptions:

Linearity: The relationship between the independent variables and the dependent variable (insurance charges) must be linear. This means that changes in the predictors (age, BMI, smoker, etc.) should result in proportional changes in the dependent variable.

Independence: Observations should be independent of each other. This means that the data points (insurance charges of different individuals) should not influence one another.

Homoscedasticity: The residuals (differences between observed and predicted values) should have constant variance across all levels of the independent variables. If the variance of the residuals is not constant (i.e., heteroscedasticity), it could indicate model misspecification.

Normality of Residuals: The residuals should be approximately normally distributed. This assumption is necessary for hypothesis testing and generating reliable confidence intervals.

2. Two Benefits of Using Python or R in Support of Various Phases of the Analysis:

Both Python and R are highly efficient for performing data analysis, particularly for tasks like multiple linear regression. Here are two major benefits:

Data Manipulation and Cleaning: Python (using pandas) and R (using packages like dplyr and tidyverse) offer powerful tools to clean, preprocess, and manipulate data. You can easily handle missing values, convert categorical data into numerical formats, and generate summary statistics.

Visualization and Interpretation: Both Python (via matplotlib and seaborn) and R (via ggplot2) excel at creating visualizations. This is essential for diagnosing regression models by visualizing residuals, checking assumptions (like linearity or homoscedasticity), and understanding variable distributions through plots.

3. Why Multiple Linear Regression is an Appropriate Technique for Analyzing the Research Question:

Multiple linear regression is well-suited for the research question, "How do age, BMI, number of children, smoking status, and region affect insurance charges?" Here’s why:

Predicting a Continuous Outcome: The dependent variable (insurance charges) is continuous, which is precisely the type of outcome multiple linear regression is designed to predict. By using this technique, you can estimate the average change in insurance charges for each unit change in the independent variables.

Multiple Predictors: The dataset includes multiple independent variables (age, BMI, children, smoker, and region), each of which may influence the dependent variable. Multiple linear regression allows you to evaluate the combined effects of these variables and identify which factors have the most significant impact.

Interpretability: One of the strengths of linear regression is its interpretability. The resulting model will provide coefficients that show how each variable influences insurance charges, making it easy to explain the findings to stakeholders.

**Part III: Data Preparation**

1. Data Cleaning Goals and Steps:

Data Cleaning Goals: The primary goal is to prepare the dataset for multiple linear regression analysis by ensuring that all variables are in the correct format and there are no missing values or outliers that could distort the model.

Steps:

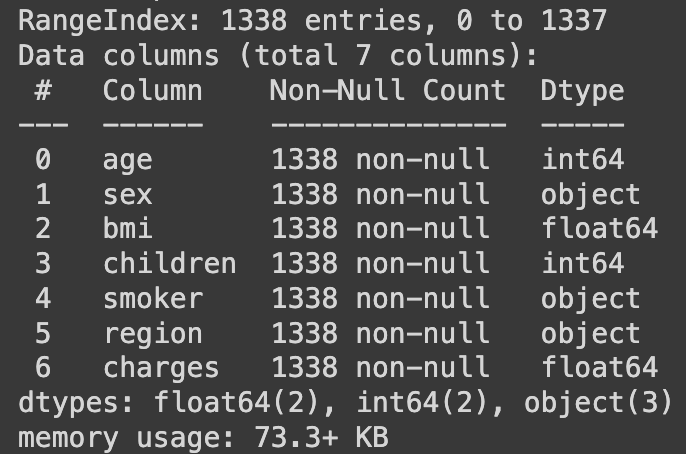
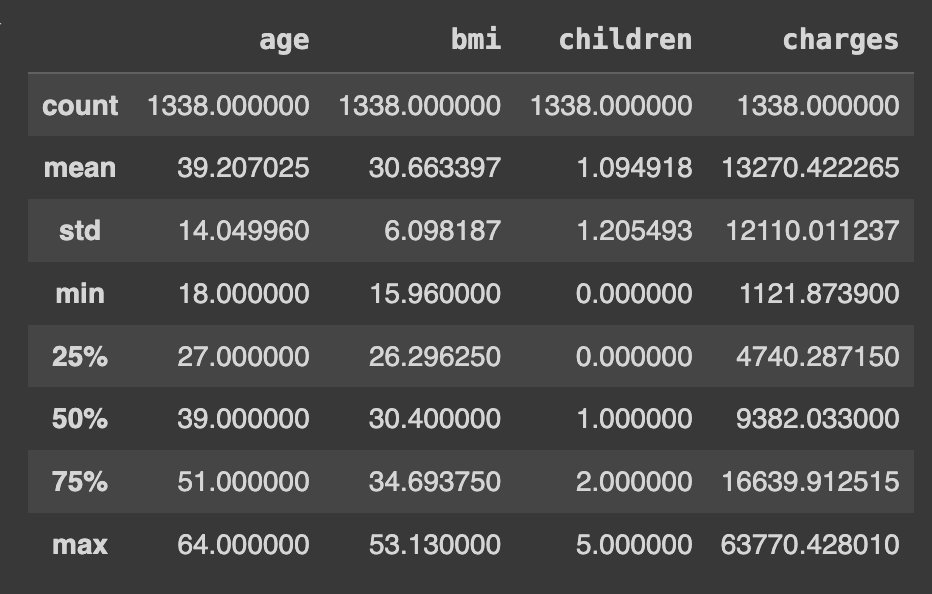
* Handling Missing Values: We first have to check for any missing values in the dataset.

Since the summary shows no missing values, we can proceed to the next step.

* Convert Categorical Variables: Variables like sex, smoker, and region are categorical and need to be converted to numerical format. I have used label encoding for this.

2. Summary Statistics for Dependent and Independent Variables:

Dependent Variable: charges (insurance charges).

Independent Variables: age, bmi, children, sex, smoker, and region.

3. Univariate and bivariate plots have been generated in the notebook.

4. Data Transformation Goals and Steps:

Had to ensure all variables are in the correct format, i.e. converting categorical features to numerical features and normalizing or scaling variables when needed.

5. Final transformed and cleaned dataset has been downloaded as a separate csv file.

**Part IV: Model Comparison and Analysis**

D

1. Initial Multiple Linear Regression Model

I began by constructing an initial multiple linear regression model using all the independent variables identified in Part C2: age, sex, bmi, children, smoker, and region. The goal of this model was to predict charges (insurance costs) based on the given predictors.

Adjusted R² score: The model’s adjusted R² score was 0.7879, indicating that approximately 78.79% of the variance in insurance charges can be explained by the selected independent variables. Although this is a reasonable fit, I aimed to improve the model’s performance through feature selection to focus on the most relevant predictors.

2. Feature Selection Using Backward Elimination

To reduce the model, I employed backward elimination as the feature selection method. This technique starts with all predictors and iteratively removes the least significant variables based on their p-values (using a threshold of 0.05). The significance of each predictor was evaluated, and variables with high p-values, which did not significantly contribute to the model, were removed.

In addition to p-values, I used the adjusted R² score as a model evaluation metric to ensure that removing variables did not reduce the explanatory power of the model. The adjusted R² accounts for the number of predictors in the model, making it more reliable for comparison between models with different numbers of variables.

Through this process, the variables sex and children were identified as statistically insignificant and were subsequently removed.

3. Reduced Linear Regression Model

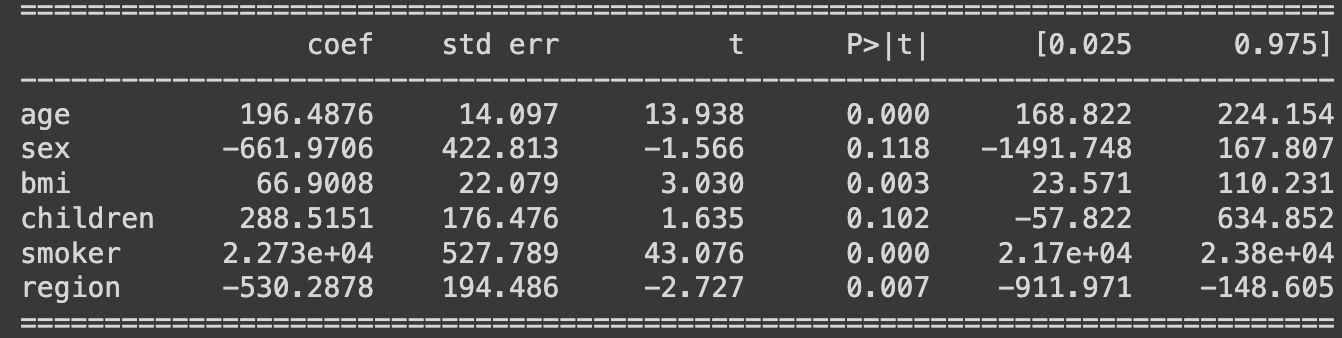
After applying backward elimination, the final reduced model included the following independent variables: age, bmi, smoker, and region. These variables were retained as they had significant p-values and were shown to contribute meaningfully to the prediction of insurance charges.

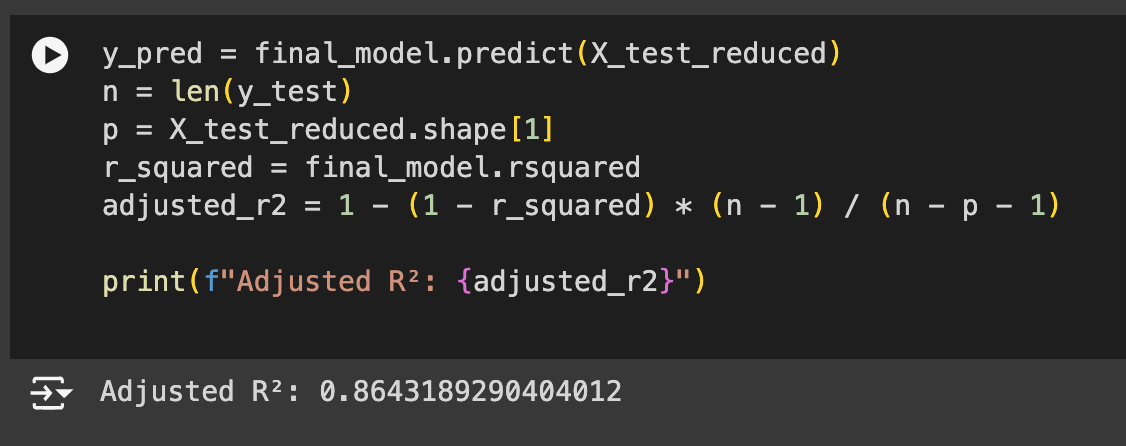
New Adjusted R² score: The reduced model achieved an adjusted R² score of 0.8643, which is significantly higher than the initial model’s score of 0.7879. This improvement suggests that the reduced model not only simplifies the prediction process by using fewer variables but also enhances the model’s ability to explain the variance in insurance charges.

This result indicates that age, bmi, smoker status, and region are the most important predictors of insurance charges, with factors like gender and number of children contributing little to the model’s explanatory power. The reduced model performs better, is more interpretable, and avoids overfitting by excluding less relevant variables.



R2 score for initial model





R2 score for reduced model

E

1. Data Analysis Process: Comparing the Initial and Reduced Multiple Linear Regression Models

In this analysis, I compared the initial multiple linear regression model, which included all the independent variables, with the reduced model, which was refined using backward elimination based on p-values and the adjusted R² score.

Initial Model:

Independent Variables: age, sex, bmi, children, smoker, region

Adjusted R²: The initial model achieved an adjusted R² score of 0.7879, meaning that approximately 78.79% of the variability in the dependent variable (charges) could be explained by the combination of all independent variables. While the model fit was reasonable, not all variables contributed significantly to the prediction, as indicated by their high p-values.

Reduced Model:

Independent Variables: age, bmi, smoker, region

Adjusted R²: After removing statistically insignificant variables (sex and children), the reduced model produced an adjusted R² score of 0.8643, representing a significant improvement. The higher adjusted R² indicates that the reduced model better explains the variance in charges, despite using fewer predictors.

Comparison:

The adjusted R² score serves as the key model evaluation metric in this analysis. It adjusts the R² value for the number of predictors, making it a useful tool to compare models with different numbers of variables. The increase from 0.7879 to 0.8643 demonstrates that the reduced model, with only the most significant predictors, offers better explanatory power and generalizability.

Model evaluation metric: Adjusted R2 score

R² score (coefficient of determination) is a measure of how well the model fits the data. It ranges from 0 to 1:

R² = 1: The model perfectly explains the variance in the target variable.

R² = 0: The model does not explain any variance, and it’s as good as using the mean of the target variable as the predictor.

In general, the higher the R², the better the model explains the variance in the target variable. The higher the R² score, the better the model fits the data.

However, there are important caveats:

* R² Doesn’t Account for Overfitting

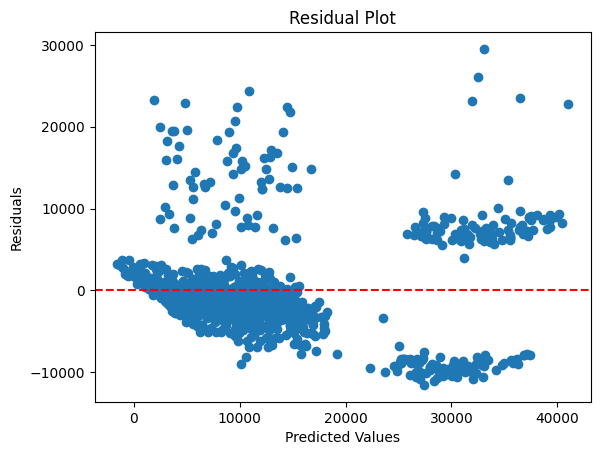
Overfitting happens when the model is too complex and fits the noise in the data rather than the underlying trend. In this case, R² can be high, but the model may perform poorly on new, unseen data (it won’t generalize well).

* Adjusted R² Is More Reliable for Multiple Predictors

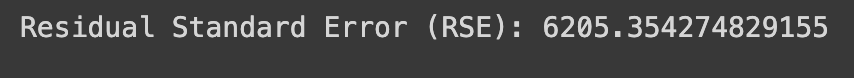
In models with multiple independent variables, Adjusted R² is often preferred because it adjusts for the number of predictors.

Adjusted R² accounts for the fact that adding more features will usually increase the R², even if those features are not truly improving the model. It penalizes the addition of irrelevant features, making it more reliable for comparing models with different numbers of predictors.

2. A residual plot helps visualize the difference between the actual target values and the predicted values. Ideally, the residuals should be randomly scattered around zero (no clear pattern), indicating a good fit.

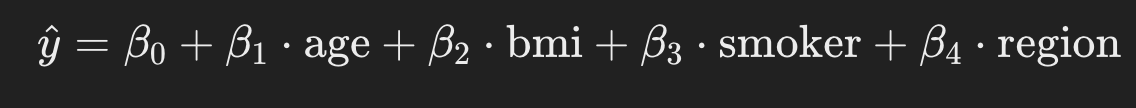


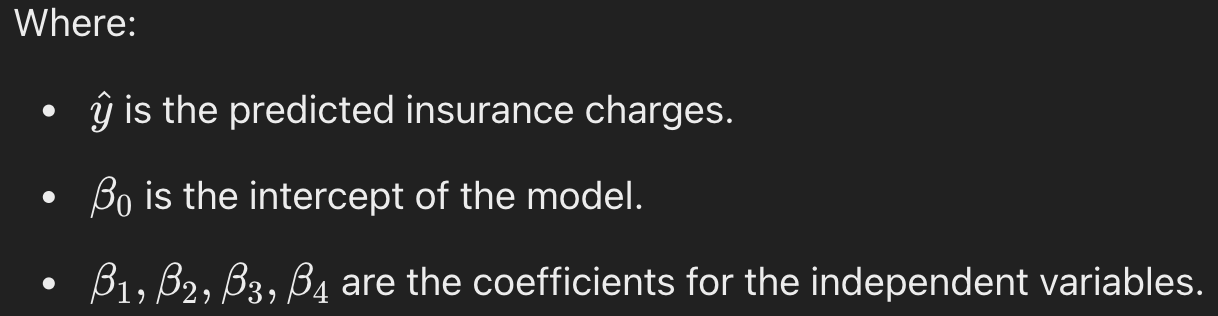
The Residual Standard Error (RSE) measures the typical size of the residuals (prediction errors). It provides an estimate of the standard deviation of the residuals and helps gauge how well the model fits the data.

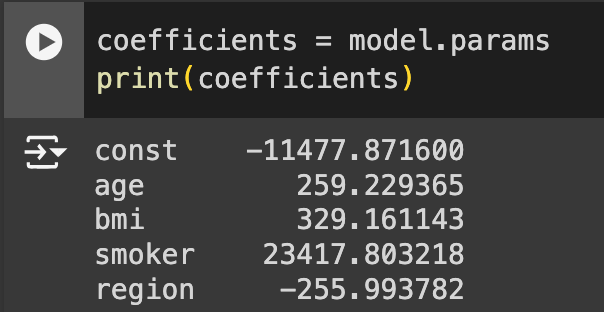


**Part V: Data Summary and Implications**

1. Discussion of Data Analysis Results

Regression Equation for the Reduced Model: The regression equation for the reduced multiple linear regression model is as follows





For example, if the coefficient for age is almost 260, it means that for each additional year in age, the insurance charges increase by 260 units, assuming all other variables are held constant.

Interpretation of the Coefficients:

Age: The coefficient for age represents the increase in insurance charges for each additional year of age. A positive coefficient indicates that as a person gets older, their expected insurance charges rise, which is intuitive as healthcare costs often increase with age.

BMI: The coefficient for bmi indicates how much the insurance charges increase per unit increase in body mass index. A higher BMI typically correlates with higher health risks, hence higher medical costs.

Smoker: Since this is a binary variable (e.g., 1 for smoker, 0 for non-smoker), the coefficient for smoker tells us how much more a smoker is expected to pay in insurance charges compared to a non-smoker. As the coefficient is around 23417, it suggests that smokers incur significantly higher medical costs.

Region: The region variable captures differences in charges based on geographic location.

Statistical and Practical Significance:

Statistical Significance: The p-values associated with the variables in the reduced model were significant, indicating that these predictors (age, BMI, smoker status, and region) have a statistically significant effect on insurance charges. The high adjusted R² score of 0.8643 further confirms the model’s strong fit to the data.

Practical Significance: From a practical standpoint, the analysis highlights key factors (age, BMI, smoking) that substantially influence insurance charges. For instance, smoking was found to have a considerable effect on charges, making it a crucial variable for insurers to consider when pricing their products. This information is valuable for developing tailored insurance plans, risk assessments, and health initiatives.

Limitations of the Data Analysis:

* Model Assumptions: While the regression model assumptions (linearity, homoscedasticity, normality of residuals) were checked, the model may still suffer from potential biases such as omitted variable bias if relevant factors like pre-existing conditions or lifestyle habits are not included.
* Categorical Variable Granularity: Variables like region are categorical but lack deeper geographic granularity (e.g., urban vs. rural). This could mask variations within regions.
* External Validity: The data used is specific to a particular group of individuals and may not generalize well to other populations or regions. For instance, medical costs and insurance policies vary widely across countries, and this model may only apply to specific markets.

2. Recommended Course of Action

Based on the results, I recommend the following actions:

* Targeted Premium Pricing: Insurance companies should consider charging higher premiums for smokers and individuals with high BMI, as these variables significantly affect medical costs. By identifying high-risk individuals, insurers can create personalized premium structures that more accurately reflect the anticipated costs.
* Health and Wellness Programs: The results suggest that insurers can reduce claims by incentivizing healthier lifestyles. For instance, they can offer lower premiums or discounts to individuals who quit smoking or maintain a healthy BMI through wellness programs.
* Geographic-Based Adjustments: Since region plays a role in predicting charges, insurers could refine their pricing models based on location-specific risk factors. Further research into how regional healthcare infrastructure and costs influence medical expenses would be valuable.
* Further Research: To enhance the model’s predictive accuracy, additional factors such as pre-existing conditions, exercise habits, and dietary patterns should be incorporated in future analyses.