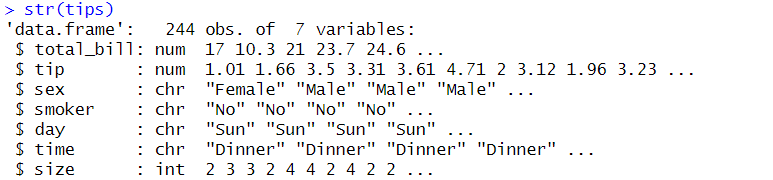
**Data Exploration:**

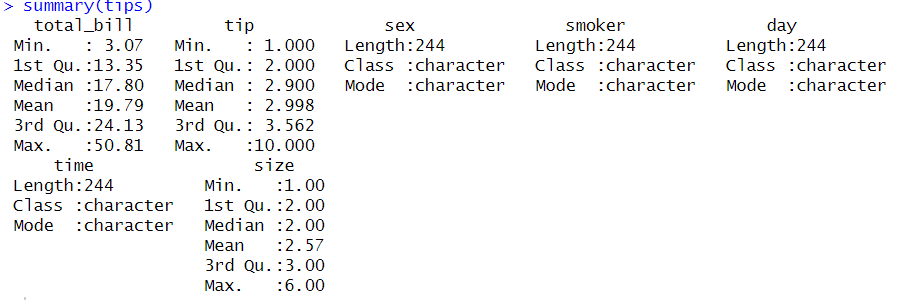
1. Load the dataset and inspect its structure.



Check for missing values and clean the data if necessary.

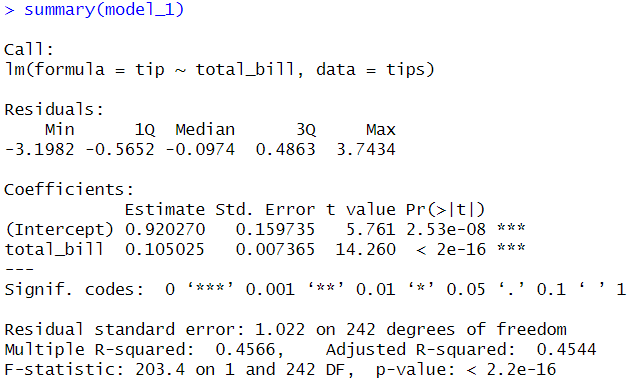


2. Generate summary statistics for all variables.



**Simple Linear Regression:**

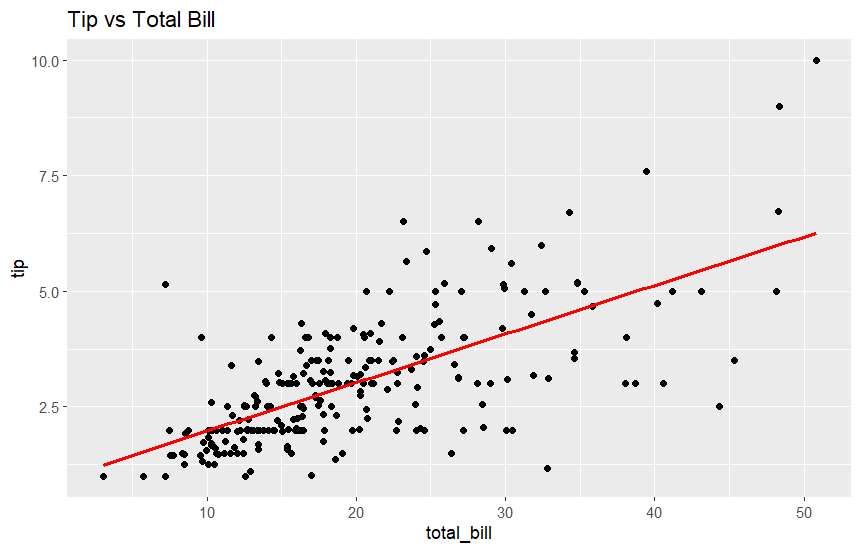
1. Perform a regression analysis with tip as the dependent variable and total\_bill as the independent variable.



2. Analyze the model output, including coefficients, R-squared, and p-values.

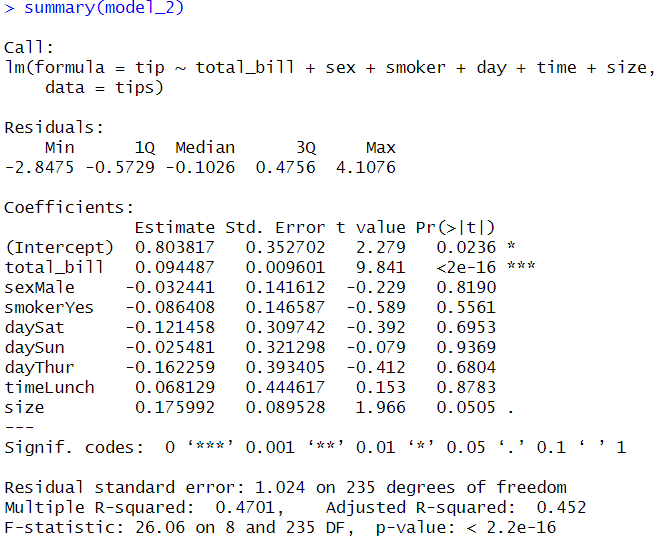
* For one USD spent on the total bill, the tip increases by 0.105 USD.
* The intercept (0.92) represents the expected tip is 0.92 USD when the total bill is $0.
* The R-squared value of 0.4566 means that about 46% of the variation in tip amounts is explained by the total bill alone.
* The relationship is statistically significant (p < 0.001), indicating a strong positive association between total\_bill and tip.

3. Visualize the regression line on a scatter plot.



**Multiple Linear Regression:**

1. Extend the model to include additional variables: sex, smoker, day, time, and size.

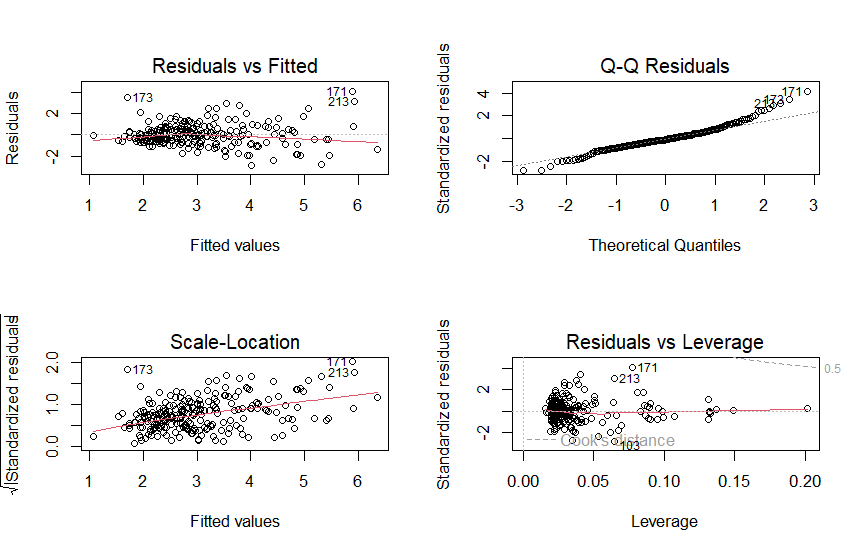


2. Evaluate the model performance and identify statistically significant predictors.

* The R-squared value of 0.4701 means that about 47,01% of the variation in tip amounts is explained by independent variable, while the 52.99% is explained by other independent variable that not in the model.
* The statistically significant predictor of tip is total\_bill and size.
* Total bill significant in 0 percent. For one USD spent on the total bill, the tip increases by 0.09 USD.
* Size is significant in 0.05 percent. For one people in the party, the tip increases by 0.176 USD.

**Model Diagnostics:**

1. Plot diagnostic graphs to check for assumptions (linearity, normality, homoscedasticity).



2. Interpret the results of the diagnostic tests.

* Linearity: As showed in that the residuals is spread following the dashed line in Residuals vs Fitted graph, the model is linear.
* Normality: As showed in Q-Q plot, the residuals linear with the dashed line.
* Homoscedasticity: The spreads of residuals slightly increases with fitted values in Residuals vs Fitted showed that the model suffers from heteroscedasticity, meaning the variance of residuals is not constant across fitted values.

**Prediction and Evaluation:**

1. Calculate the RMSE to assess model accuracy.

