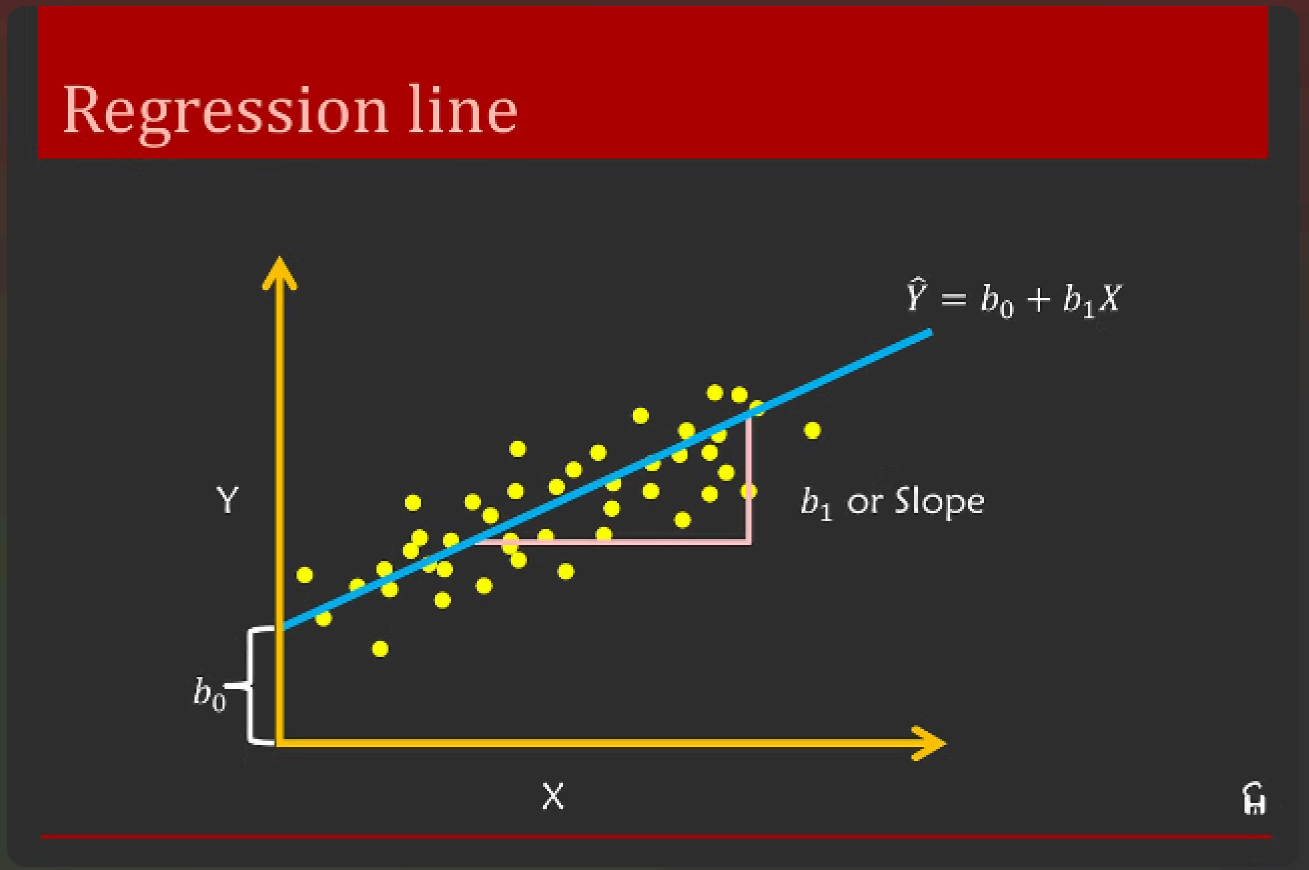
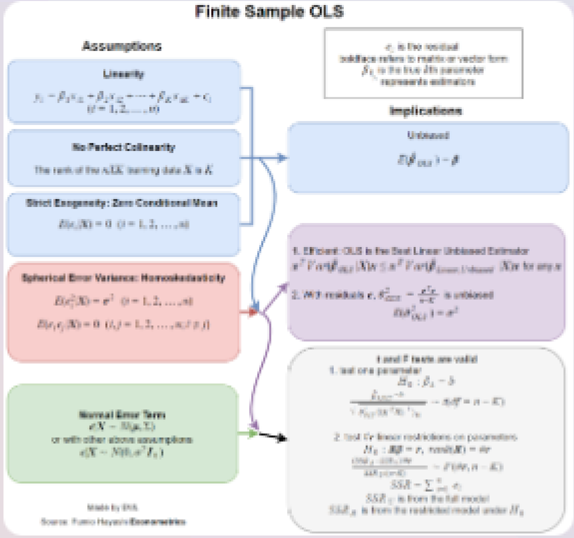
Multiple Linear Regression: Unveiling Relationships

Multiple linear regression is a statistical technique that examines the relationship between a dependent variable and two or more independent variables. It helps understand how these variables influence the outcome and predict future values.

Manoj Kumar Sahoo





Assumptions of Multiple Linear Regression

1. **Linearity**

   The relationship between the dependent variable and each independent variable should be linear. A scatter plot can help visualize this.

2. **Independence**

   The error terms should be independent of each other. This ensures that the residuals are not correlated.

3. **Homoscedasticity**

   The variance of the error terms should be constant across all values of the independent variables.

4. **Normality**

   The error terms should be normally distributed. This assumption is particularly important for hypothesis testing.

	Content	SocialMedia	Email	Consumer_Intention	var	var
1	5.00	5.00	4.00	4.00		
2	4.00	4.00	4.00	4.00		
3	5.00	5.00	5.00	4.00		
4	4.00	4.00	5.00	4.00		
5	5.00	5.00	4.00	4.00		
6	4.00	5.00	5.00	5.00		
7	5.00	4.00	5.00	5.00		
8	4.00	5.00	5.00	4.00		
9	4.00	5.00	4.00	4.00		
10	4.00	4.00	5.00	5.00		
11	5.00	5.00	4.00	4.00		
12	5.00	5.00	5.00	5.00		
13	4.00	5.00	5.00	4.00		
14	4.00	4.00	4.00	5.00		
15	4.00	5.00	5.00	4.00		
16	4.00	4.00	5.00	5.00		
17	4.00	5.00	4.00	4.00		
18	4.00	4.00	5.00	5.00		
19	4.00	5.00	4.00	4.00		
20	5.00	5.00	4.00	4.00		
21	5.00	4.00	4.00	5.00		
22	4.00	5.00	5.00	4.00		
23	5.00	4.00	4.00	5.00		
24	4.00	5.00	5.00	4.00		
25	5.00	4.00	4.00	4.00		
26	4.00	4.00	4.00	4.00		
27	4.00	5.00	4.00	4.00		
28	5.00	4.00	5.00	5.00		
29	5.00	5.00	4.00	4.00		
30	5.00	5.00	5.00	4.00		
31						

Interpreting Regression Coefficients

Coefficient	Interpretation
Intercept	The predicted value of the dependent variable when all independent variables are zero.
Slope (for each independent variable)	The change in the dependent variable for a one-unit increase in the corresponding independent variable, holding all other independent variables constant.

Understanding R-squared and Adjusted R-squared

R-squared

The proportion of the variation in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit. However, adding more variables can artificially inflate R-squared.

Adjusted R-squared

It penalizes the model for adding irrelevant variables, providing a more accurate measure of the model's explanatory power. A higher adjusted R-squared indicates a better model.

Dealing with Multicollinearity

1

Identify

Use correlation matrix or Variance Inflation Factor (VIF) to identify highly correlated independent variables.

2

Remove

Remove one of the highly correlated variables or combine them into a new variable.

3

Use Regularization

Techniques like Ridge or Lasso regression can penalize models with high correlation, reducing the impact of multicollinearity.

Selecting Relevant Features

Forward Selection

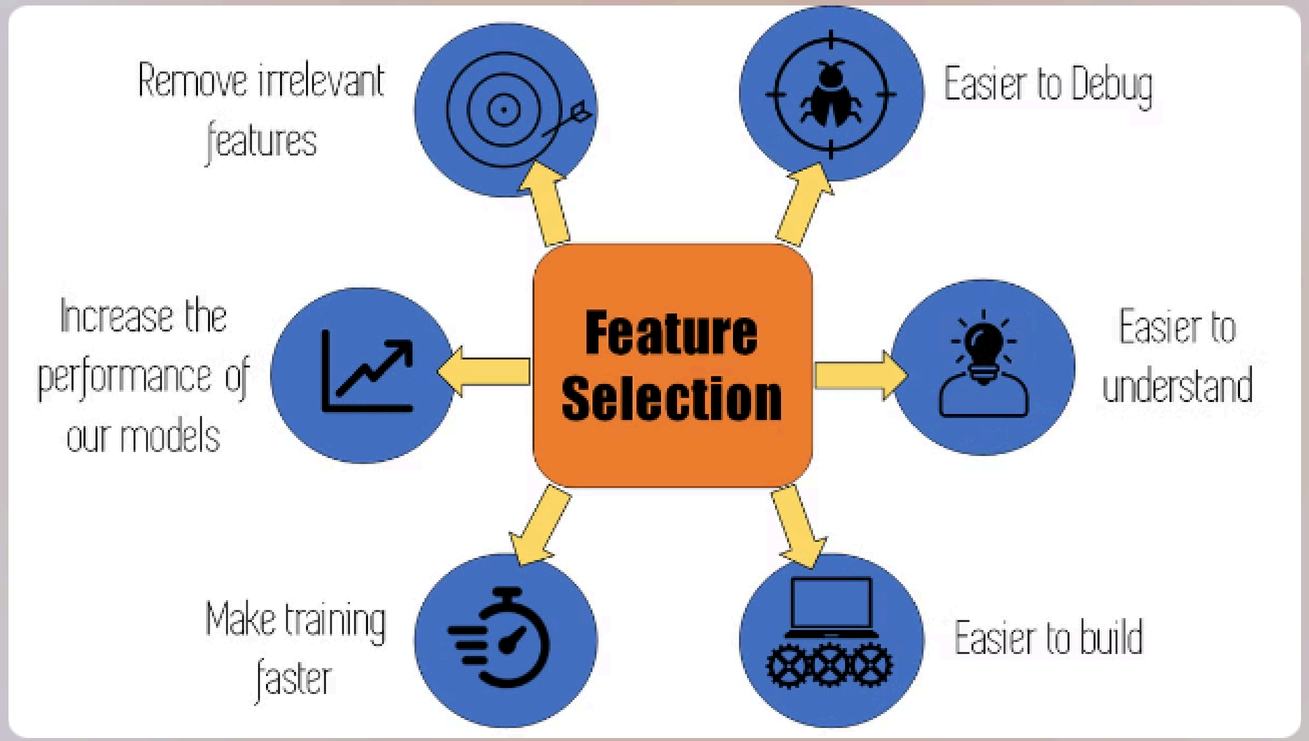
Start with an empty model and gradually add variables with the highest impact.

Backward Elimination

Start with all variables and remove those with the least significant contribution.

Stepwise Selection

Combines forward and backward selection, adding and removing variables based on their impact.



Handling Categorical Variables

Dummy Coding

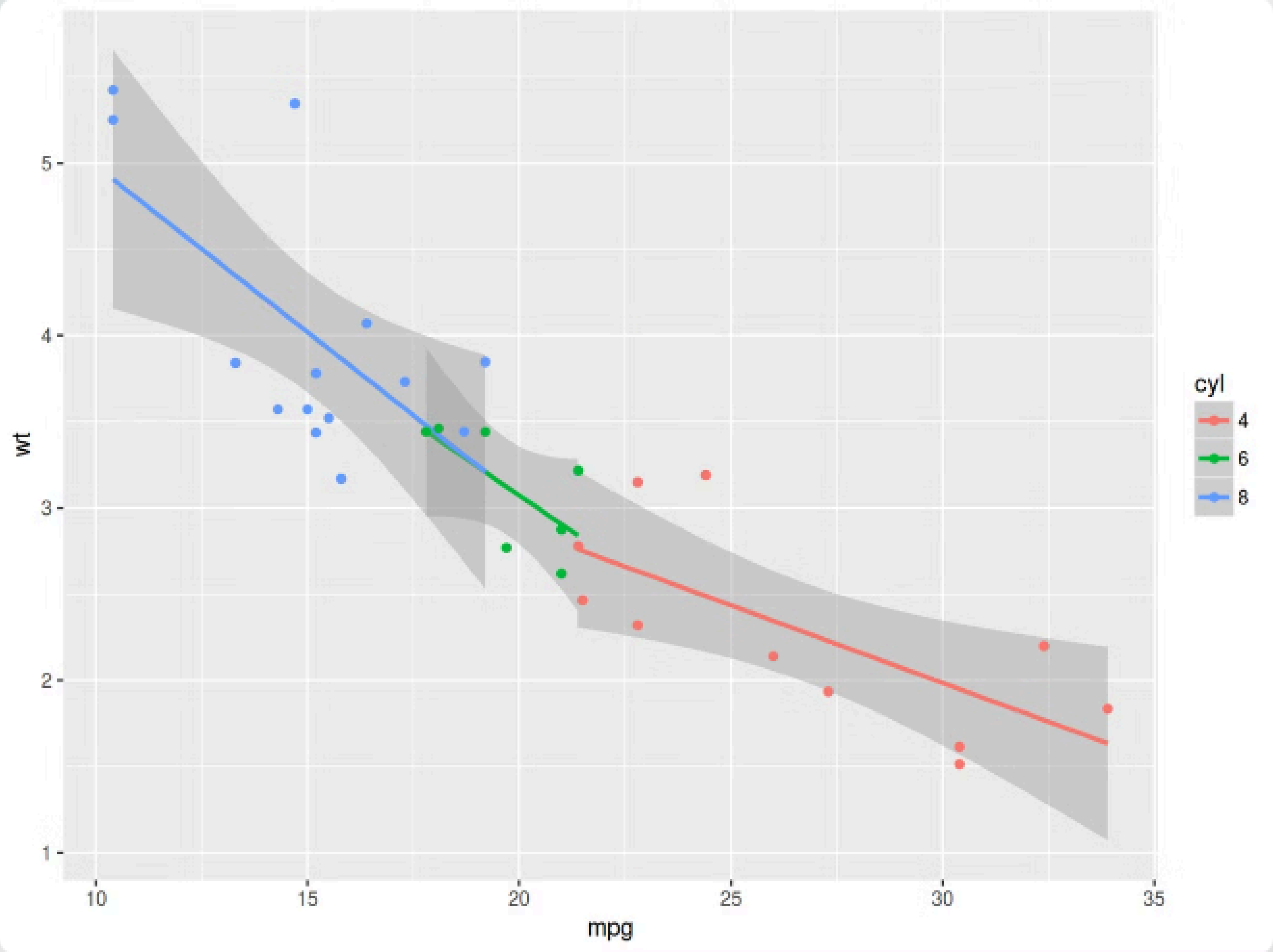
Convert categorical variables into binary variables (0 or 1) to represent different categories.

One-Hot Encoding

Create a separate variable for each category, with a value of 1 for the corresponding category and 0 otherwise.

Effect Coding

Similar to dummy coding, but uses a combination of 0 and -1 to represent categories, allowing for comparisons.



Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

Evaluating Model Performance

1

Root Mean Squared Error (RMSE)

Measures the average difference between predicted and actual values.

2

Mean Absolute Error (MAE)

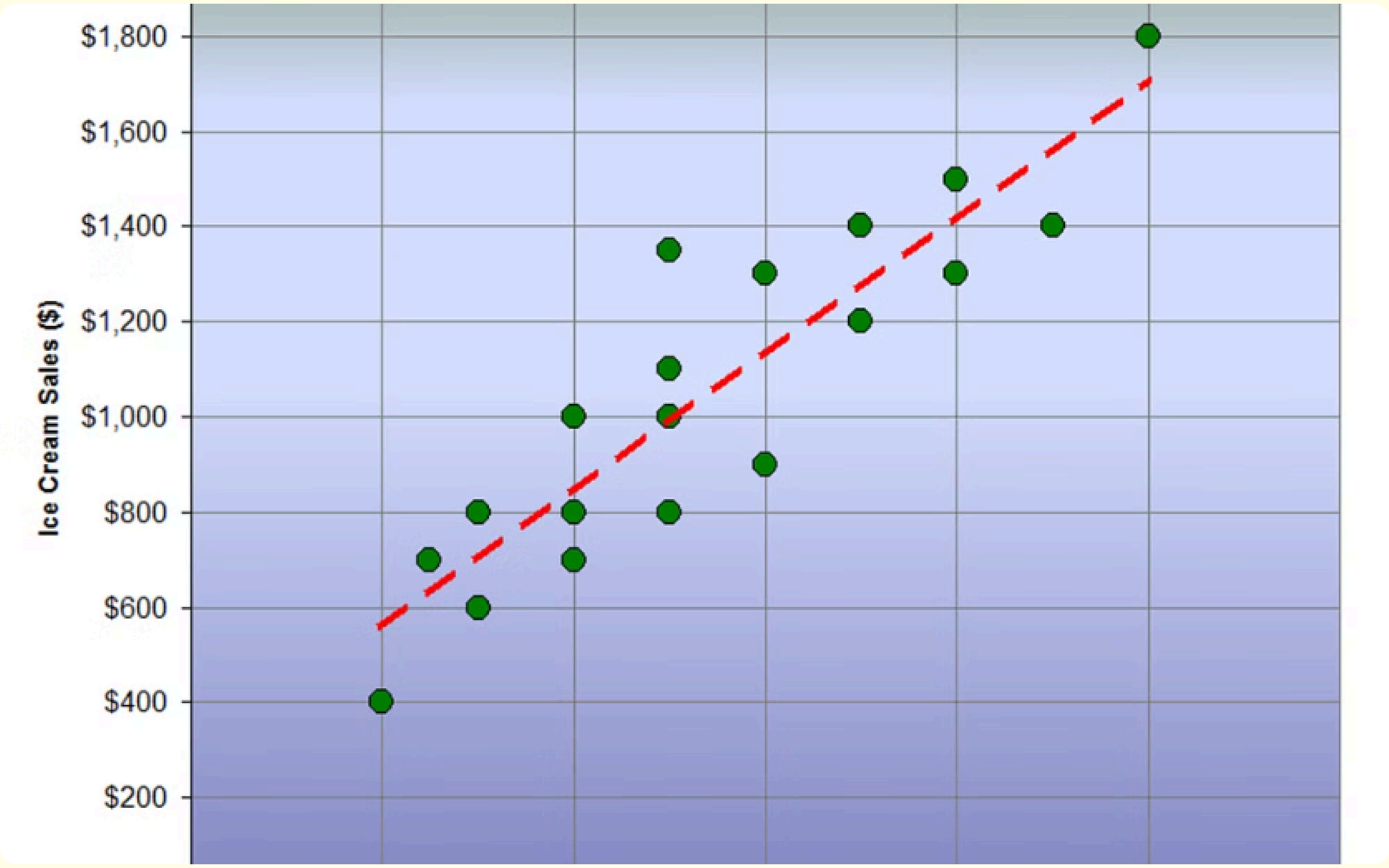
Calculates the average absolute difference between predicted and actual values.

3

R-squared

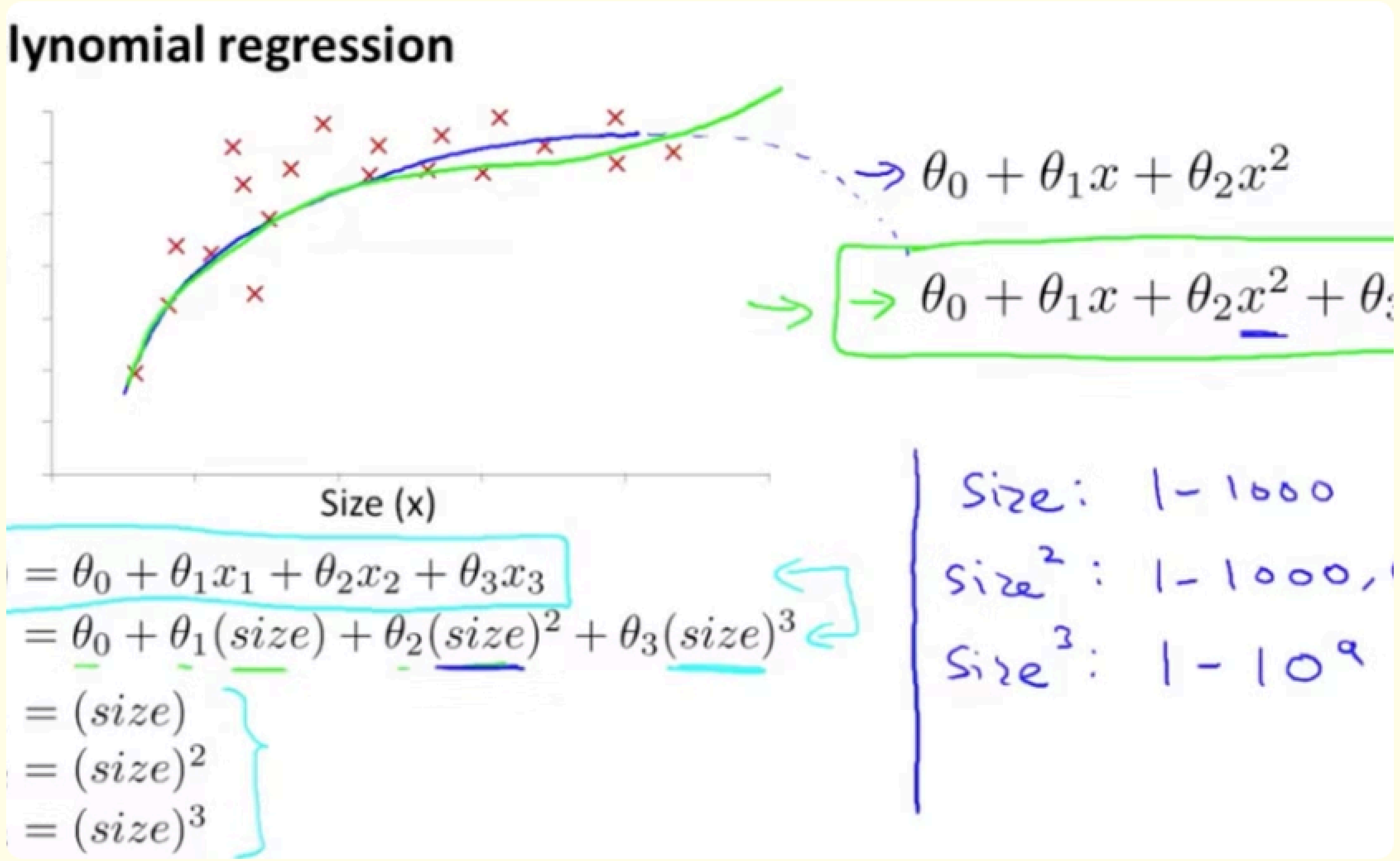
Represents the proportion of variance explained by the model.

Comparing Simple Linear Regression and Multiple Linear Regression



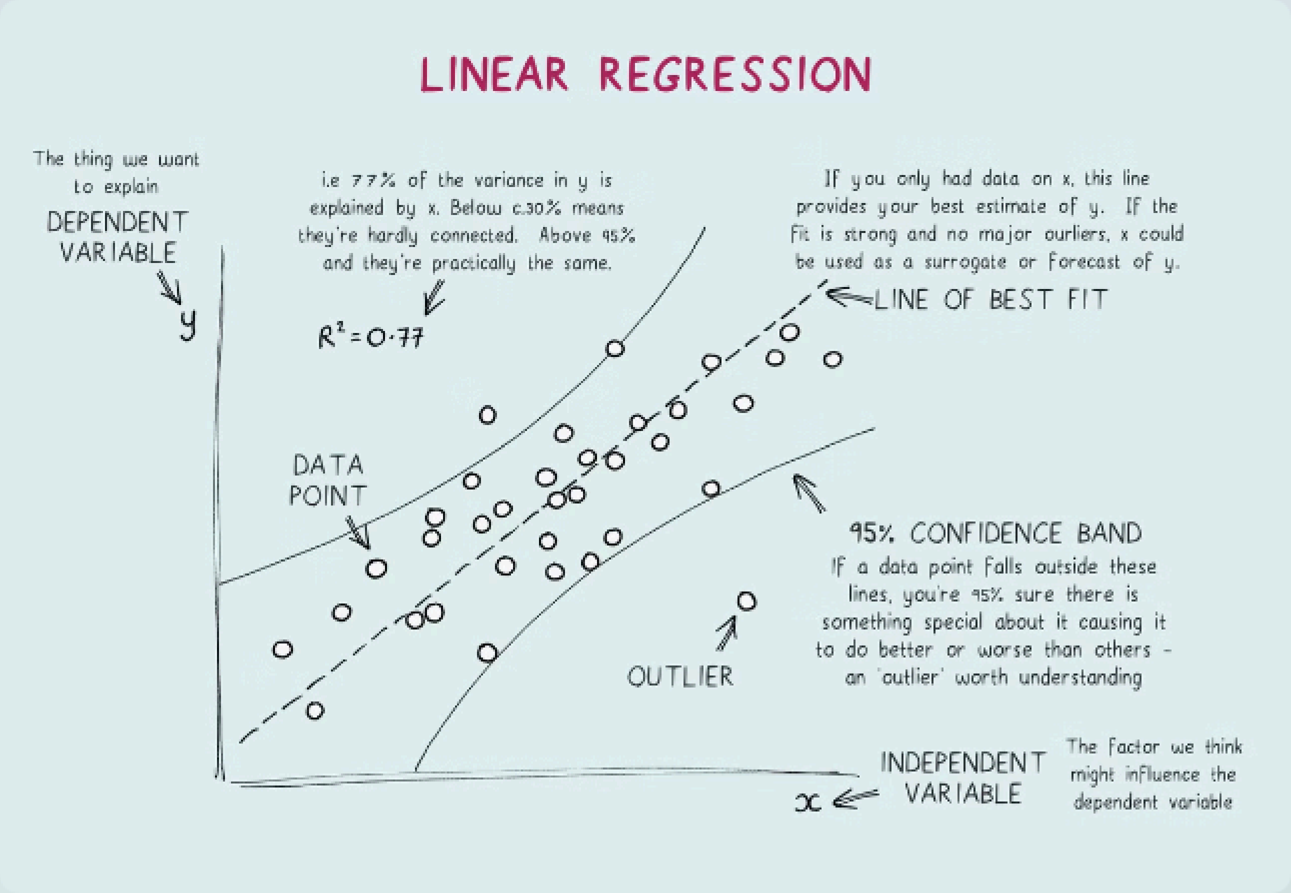
Simple Linear Regression

Examines the relationship between one dependent variable and one independent variable.



Multiple Linear Regression

Examines the relationship between one dependent variable and multiple independent variables.



Conclusion and Key Takeaways

Multiple linear regression is a powerful tool for understanding and predicting relationships between variables. By carefully considering the assumptions, interpreting coefficients, and selecting appropriate features, you can build robust and informative models.