**Data learning: Linear Model vs Regression Tree**

Example 1: Single Outcome Variable and One Covariate Data source: Simulation by R or Python programming language Data:



Step 1: Formulate your research question based on this data set.

Is advertisement budget an important factor influencing sales?

* The budget has an impact on sales.

Step 2: Define data role and data type.

|  |  |  |
| --- | --- | --- |
| Variable name | Data role | Data type |
| Advertisement budget | independent Variable | Numerical |
| Sales | dependent Variable | Numerical |

Step 3: Select an appropriate method to analyze the data e.g., statistical learning, machine learning.

**Simple linear regression** is used to analyze the data because it examines the linear relationship

between advertisement budget and sales, allowing for easy interpretation of coefficients.

**Regression tree** is used to analyze the data because it can model non-linear relationships and

identify decision splits that are important for predicting sales based on budget ranges.

Step 4: Collect data.

A simulated data set from R or Python programming language is used for this example.

* This dataset includes Advertisement Budget (X) and Sales (Y) values for analysis.

Step 5: Explore your data using numerical summary and graphs Table 1: Descriptive statistics for advertising data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable name | n | min | max | mean | median | sd | variance |
| Advertisement budget | 100 | 10.00 | 49.98 | 30.36 | 30.20 | 11.89 | 141.49 |
| Sales | 100 | 15.51 | 134.97 | 76.28 | 75.92 | 28.56 | 815.54 |

Show your graphs:

A graph with blue dots

Description automatically generated

A graph of sales

Description automatically generatedA graph of green bars

Description automatically generated

Interpret:

From the descriptive statistics:

- Advertisement Budget has a mean of 30.36 and variance of 141.49.

- Sales has a mean of 76.28 and high variance of 815.54.

- Scatter plots may indicate a positive relationship between Advertisement Budget and Sales.

**Linear model**

Step 6: Fit the model.

(Hint: t-test, linear regression model)

t-test ; Ho : B1 = 0 , H1 : B1 ≠ 0 p-value = 0.001 < alpha = 0.05

Thus, X has relationship between Y

Step 7: Check standard assumptions.

1. Linearity:

A graph with a line

Description automatically generated

2. Normality:

A screenshot of a computer program

Description automatically generated

3. Normal distribution

A screenshot of a calculator

Description automatically generated

Step 8: Evaluate model accuracy.

= 0.9729

RMSE = 1.8257

MSE = 3.3331

Step 9: Interpret the results.

- If p-value < 0.05 in the t-test, the Advertisement Budget significantly impacts Sales.

- The coefficient (B1) indicates the change in Sales for every unit increase in Advertisement Budget.

- R-squared value explains how much variance in Sales is explained by Advertisement Budget.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Show your R or Python programming language for simple linear regression.

A screen shot of a computer program

Description automatically generated

**Regression tree**

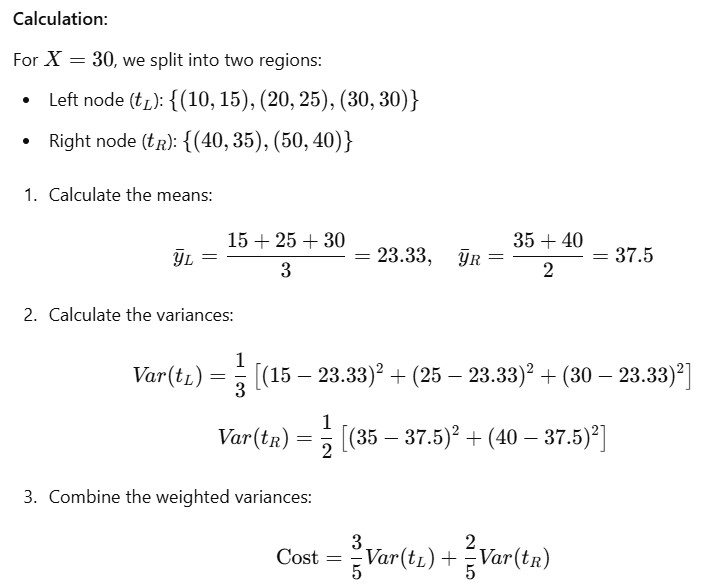
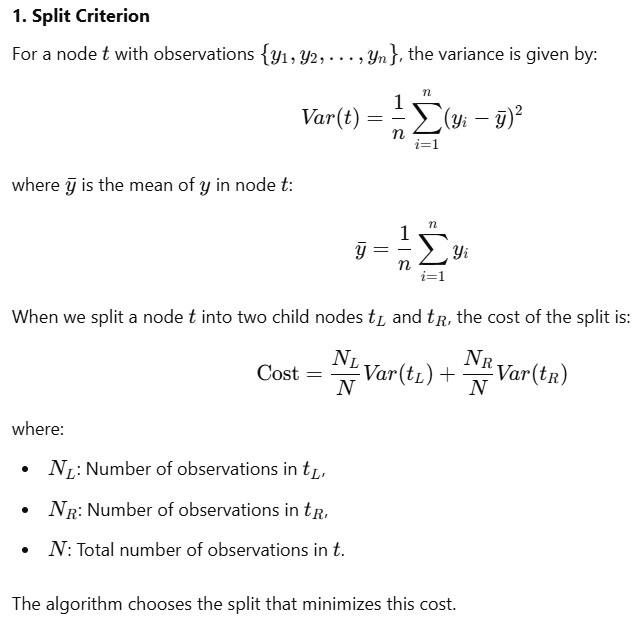
Step 6: Fit the model.

To understand the mathematical fundamentals behind regression trees, we’ll break it down into the following key components:

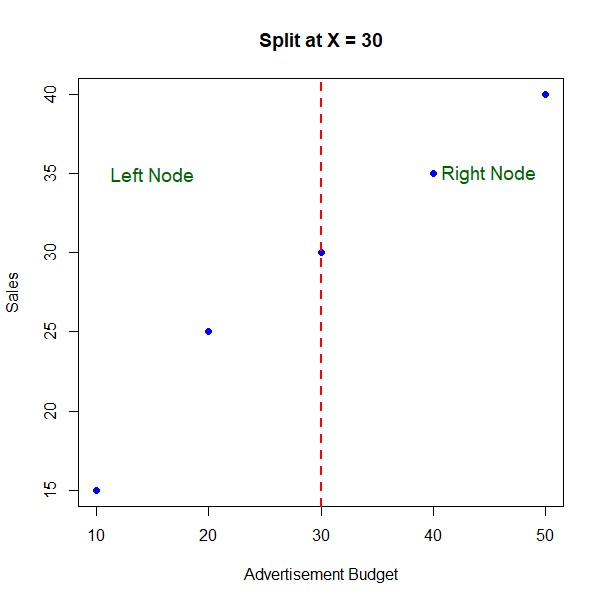
1. Splitting Criterion: Minimize Variance or Mean Squared Error (MSE)

A regression tree aims to split the data into regions (nodes) such that the variance (or sum of squared residuals) within each region is minimized.

At each step, the algorithm searches for the optimal split point to minimize the MSE.



From this data, calculate splits based on minimizing mean squared error (MSE) and understand how the tree is built with one split. The split is chosen to minimize the weighted variance in the two resulting nodes.

For Example 1 (One Covariate) with a split at X=30:

Left Node Mean: 23.33

Right Node Mean: 37.5 Variance in Left Node:

Variance in Right Node:

Split Cost (Weighted Variance):

a split at X = 10 Left Node Mean: 15

Right Node Mean: 32.5

Variance in Left Node: 0

Variance in Right Node: 31.25

Split Cost (Weighted Variance): 25

a split at X = 20 Left Node Mean: 20

Right Node Mean: 35

Variance in Left Node: 25

Variance in Right Node: 16.6667

Split Cost (Weighted Variance): 20

a split at X = 40 Left Node Mean: 26.25

Right Node Mean: 40

Variance in Left Node: 54.6785

Variance in Right Node: 0

Split Cost (Weighted Variance): 43.75

**Which split point would you select for minimizing MSE?**

Step 7: Interpret the results.

Split point we select for minimizing MSE is X=20 ; Split cost = 20

**Key Takeaways:**

**Minimizing Variance is Central:**

The primary goal is to create "pure" nodes where the outcome Y has low variance within each node.

**Recursive Partitioning:**

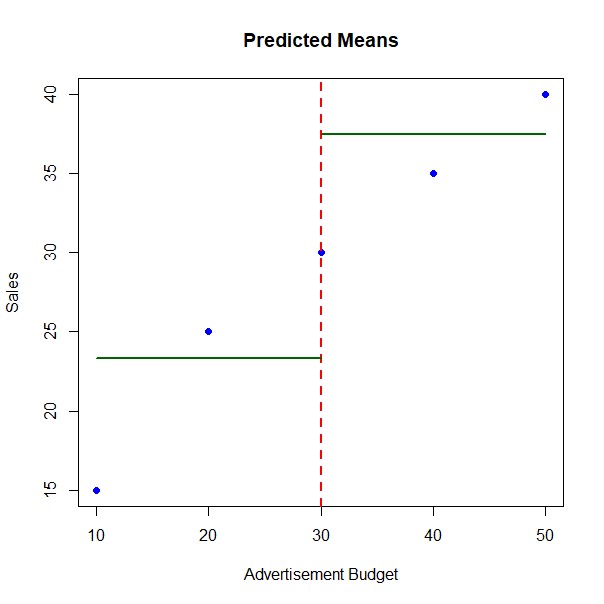
After splitting, the process is repeated for each child node, creating a hierarchical tree structure.

**Piecewise Constant Prediction:**

The tree predicts outcomes using the mean of Y within each terminal node.

**Trade-offs:**

While deeper trees can capture more complex patterns, they risk overfitting. Pruning or stopping criteria are often applied to prevent this.



**Model selection/comparison: Simple Linear Regression VS Regression Tree by comparing MSE, RMSE**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | y | Simple Linear Regression | | |  | Regression Tr | ee |
| y.hat | e=y-y.hat | e^2 | y.hat | e=y-y.hat | e^2 |
| 1 | 15 | 17 | -2 | 4 | 29 | -14 | 196 |
| 2 | 25 | 23 | 2 | 4 | 29 | -4 | 16 |
| 3 | 30 | 29 | 1 | 1 | 29 | 1 | 1 |
| 4 | 35 | 35 | 0 | 0 | 29 | 6 | 36 |
| 5 | 40 | 41 | -1 | 1 | 29 | 11 | 121 |
|  |  |  |  |  |  |  |  |

|  |  |  |
| --- | --- | --- |
| Model/Criterion | MSE | RMSE |
| SLR | 2 | 1.414214 |
| RT | 74 | 8.602325 |

**Which model would you select for prediction?**

We select SLR because RMSE of SLR less than RMSE of RT.

A screen shot of a computer

Description automatically generated\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Show your R or Python programming language for regression tree.

A screen shot of a computer

Description automatically generated