1. Introduction
2. Explaining Regression Trees
3. Testing Regression Trees with Simulations
4. Using Regression Trees housing Market Data
5. Conclusions

1 Introduction

While linear regression, probably the most used method by empirical economist can be useful in many situations, it can be useful to look at where it fails, and different methods excel. Regression trees have been used in many situations, with many extension being developed greatly improving their usefulness. In this paper I will first give a short exposition on Regression Trees, their theory and extensions and applicatios and contrast them with linear regression. I will perform easily understandable simulations comparing Regression Trees and Linear Regression. Show how Regression Trees can be used on a Dataset on the housing Market.

Linear regression performs poorly on many kinds of data, particularly those with non-linear relationships and interaction effects. What might be a better approach for these situations? Regression Trees can be useful in many situations where linear regression falls short.

Regression trees are, a powerful machine learning technique for predictive modeling. Regression trees offer an alternative approach that can be useful in many situations. We will discuss their advantages over traditional linear regression methods, cover the basics of regression trees, compare them with linear regression, address the issue of overfitting, and introduce advanced ensemble methods like Bayesian Additive Regression Trees (BART).

[Figure: Can you guess where regression trees and where linear regression will perform better?]

Hastie and Tibshirani's an Introduction to Statistical Learning serves as an eccelent introduction to Regression trees, while Tan and Roy (2019) can give a deeper dive on BART a powerful ensemble method.

The simulations show how regression trees outperform linear regression on datasets with non lenear relationships. etc.

2 Regression Trees

Regression Trees are a machine learning method, that splits the predictor space into subregions and makes predictios for each region. In doing so it doesn't make assumptions about linearity or non-interaction between different dimensions.

Regression trees split the predictor space into regions that minimize the residual sum of squares given by:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{1}$$

In each region, $\hat{y}$ simply takes on the mean of all observations in that region. Typically, we can't find the optimal regions. Instead, we use a greedy algorithm called Recursive binary splitting to find the optimal split to minimize prediction error at each stage. One will typicall continue the algorithm until some threshold is reached, such as each new split not leading to much improvment.

This model is clearly quite different from regular linear regression. Regression trees do not assume a linear relationship between predictors and the response. They also have more parameters that can be tweaked e.g. how many splits to perform.

One other major benifit of Trees over Linear Regression is thet they capture interaction effects naturally. If for example being blong typically increases pay, but only for women, a Regression Tree will ofnet naturally split along gender and then, for women only, along haircolor. Whearas in a linear Regression the researcher would have to manually include terms for all interaction effects they want to study.

2.1 Pruning

Some of the main problems with Regression Trees are that they are easily prone to overfitting and that the greedy nature of the splitting algorithm doesn't necessarily create the best possible, or even close to the best possible carving up of the predictor space.

Cost complexity pruning (or pruning) is one popular way to improve Regression Trees. It counteracts overfitting by removing non-essential splits. We grow a large Tree and then use the following formula:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_j})^2 + \alpha |T| \tag{2}$$

The process involves:

Selecting a parameter α For each α, finding the subtree that minimizes the cost Using

cross-validation to select the best α Instead of evaluating a Model on the data we trained it on we evaluate it on a separate set. Achieve a good tradeoff between bias and variance. Large α results in very small trees, small α in larger trees. The objective is to achieve a good tradeoff between bias and variance. A large α results in very small trees, while a small α results in larger trees.

2.2 Ensemble Methods

Even with pruning, trees often perform worse than other ML methods. Ensemble methods improve results by combining many regression trees. Each one contributes a small part to the overall prediction. These can be independent of previous trees (e.g., Random Forests) or grown on the residuals of the current fit (e.g., Bayesian Additive Regression Trees - BART).

BART models the response as a sum of many tree-based models plus noise:

$$Y_i = \sum_{j=1}^{m} g(X_i; T_j, M_j) + \epsilon_i \tag{3}$$

BART calculates the residuals of the current sum of Trees, then modifies one Tree to decrease the residuals. It then takes the average over all but the burn-in iterations. Unlike single trees, BART avoids overfitting by averaging the predictions of many trees and provides a probabilistic prediction, giving a measure of uncertainty.

3 Simulations

Simulations can serve as a testing ground for statistical methos, allowing for easyer repetability and eleminating many of the complications that arise when using datasets.

On two datasets

We ran simulations 400 times with regression trees having 4 terminal nodes. We compared the Mean Squared Error (MSE) for both methods. The data was generated as: $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.

Results:

- MSE for OLS model: 0.9917
- MSE for regression tree model: 1.4935

3.1 Non-linear Data

We generated non-linear data using 4 normal distributions (NW and SE = red, NE and SW = blue). The results showed:

- Classification Tree MSE: 0.3757
- Linear Classification MSE: 0.5103

Trees capture interaction effects better. For example, a large Y is only indicative of red if X is small.

3.2 Finding the Optimal Tree Size

In our simulation:

- Original Tree MSE on Test set = 198.2861
- Pruned Tree MSE on Test set = 173.861

The sweet spot balances variance and bias, minimizing overfitting. We use a training set to build the model and a test set to evaluate performance.

4 Applying Regression Trees To Housing Market Data

and so on

5 Conclusion

Regression trees are powerful tools for handling non-linear and interactive effects, often outperforming linear regression in these scenarios. They are also very easy to interpret. However, trees require pruning to combat overfitting. Ensemble methods, by averaging independent trees or fitting trees on the residuals, can significantly improve results. BART, in particular, is a sophisticated method offering good results in many scenarios. While regression trees have their strengths, it's important to choose the right tool for each specific data analysis task.

5.1 Results and short summary

and so on

This is a very important area of research and clearly there are still many open questions.

6 References

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R (Second Edition). Springer.
- Tan, Y. V., Roy, J. (2019). Bayesian additive regression trees and the General BART model. Statistics in Medicine, Band/Volume 38(25), 5048-5069.
- Chipman, H. A., George, E. I., McCulloch, R. E. (2010). BART: Bayesian additive regression trees. The Annals of Applied Statistics, Band/Volume 4(1), 266-298.
- Townshend, R. Lecture 10 - Decision Trees and Ensemble Methods | Stanford CS229: Machine Learning (Autumn 2018). https://www.youtube.com/watch?v=wr9gUreWdA, accessed 21.05.24.
- All images were made using R. Also thanks to Claude and ChatGPT for making LaTeX a lot nicer to use.