Joint Math Meeting
April 2022

Applications of Regression Tree and Linear Model Evaluation in Quantitative Trading

Matthew Cho[1] Haoyu Du[2] Tomas Escalante[3] Ashley Tran[4]

[1]Massachusetts Institute of Technology [2]University of Michigan, Ann Arbor
[3]University of Southern California [4]University of California, Irvine

Motivation

Our sponsor, Aquatic Capital Management, collected financial data over the past several years consisting of features which have been identified to be predictive of future returns. Our goal is to better understand the relationship between the target Y that is correlated with a predictive feature X, and which is influenced by an interactor variable Z. This is demonstrated by the equation:

$$Y \sim X \cdot \beta(Z) + \epsilon \tag{1}$$

Data

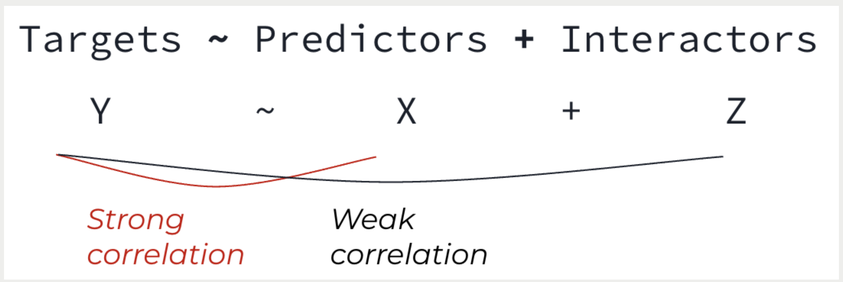
More than 3000 different stocks sampled at five-minute interval on each trading day from market start to market closing from 2007 to 2015.

- **Target (Y)**: stock returns
- **Predictors (X's)**: stock market variables strongly correlated to the targets
- **Interactors (Z's)**: weakly correlated with the targets, but are mediators of the relationship between certain target-predictor relationship

Big Idea

Goal 1 Find most significant information (feature selection)
Goal 2 Find return-indicator relationship

- Find relationships between features and targets
- Rank the most predictive features
- Feed these features to our model
- Train models



Linear Model Tests For Feature Selection

Ordinary Least Squares Regression (OLS) Consider

$$Y \sim \beta_0 + \beta_1 X + \beta_2 I_{q2} ZX + \beta_3 I_{q3} ZX + \beta_4 I_{q4} ZX \tag{2}$$

- I_{qn} is the indicator function for the nth quartile of values for the interactor variable Z
- Z variable is discretized as such because of its smaller impact on the target value
- first quartile of Z is dropped to avoid multicollinearity

Using this relation, we evaluate if a certain predictor is significantly predictive.

F-Statistic The F-test checks whether the regressors and their respective coefficients are relevant in approximating Y. The test is applied on the linear regression relationship (2) against the restricted regression relationship

$$Y \sim \beta_0 \tag{3}$$

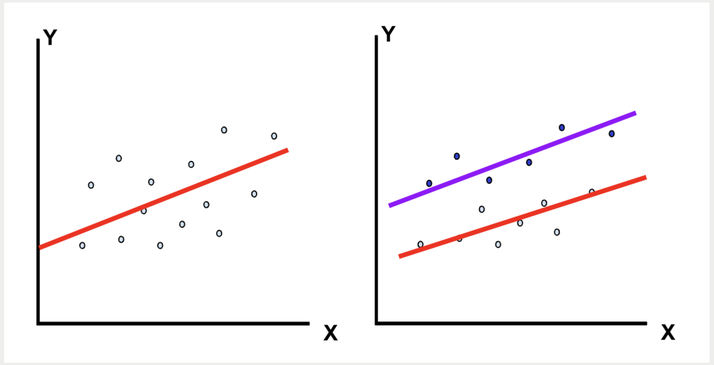
with fixed X and iterated Z.

Wald Test The Wald Test checks the significance of regression coefficients. We apply the test on (2) against the base relationship

$$Y \sim \beta_0 + \beta_1 X \tag{4}$$

which assumed X to be predictive; if a certain Z contributed significantly to the relationship between our target and X of choice, then the coefficients would not be zero, which the Wald test rigorously checks. The two different null hypotheses with the coefficients in (2) applied were:

1. $\beta_2 = \beta_3 = \beta_4 = 0$. If rejected, the regression relationship (2) with the Z quartiles is a better fit than the regression without.
2. $\beta_4 - \beta_3 = \beta_3 - \beta_2 = \beta_2$. If this null hypothesis holds, then there is a linear dependency of β on Z. The regression equation can then be understood as
$$Y \sim \beta_0 + \beta_1 (XZ)$$
where XZ is still the pointwise product of these two variables.



Model-Based Recursive Partitioning (MOB)

Model-based recursive partitioning (MOB) is an algorithm to construct a regression tree for a given parametric model $\mathcal{M}$, by recursively splitting a node whenever significant instabilities are found. The result is a tree where each node is fit into the model $\mathcal{M}$, which is OLS, and the instability is assessed using flucuation test with a Bonferroni-corrected significance level of $\alpha = 0.05$ and the nodes are split with a required minimal segment size of 20 observations. This is implemented in R as the function `mob()` in the package `party`.

Modification On Regression Tree Prediction

In the results of the MOB algorithm given the OLS model, the best-fit lines at the leaf nodes had a slope very close to zero, as we expect the noise-to-signal ratio in financial data to be very high. Thus we also experiment with predicting the sign of the return only. This is given as

$$\sum_{i=1}^{n} Y_i \cdot \text{sign}(Y_i')$$

where n is the number of observations, Y_i is the actual Y-value of the ith observation, and Y_i' is the predicted Y-value of the ith observation.

The Predictive Algorithm

- **Pre-process the data**: Subset into train and test sets
- **Top feature selection**: Statistical tests on X-Z pairs and create sets of top features data for models
- **Model evaluation**: Linear regression, PCA, `mob()`
- **Results**: Correlation matrix, average R^2 values

References

This project was jointly supported by Aquatic Capital Management and NSF Grant DMS-1925919, as a part of RIPS at UCLA IPAM.

[1] A. Zeileis, T. Hothorn, and K. Hornik, Model-based recursive partitioning, *J. Comput. Graph. Statist.* **17** (2008), 492–514.

[2] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning. Data mining, inference, and prediction* (Second edition), Springer, New York, 2009.