Metadata of the chapter that will be visualized online

Chapter Title	Trading Volatility Using Highly Accurate Symbolic Regression
Copyright Year	2015
Copyright Holder	Springer International Publishing Switzerland

Corresponding Author	Family Name	**Korns**
	Particle	
	Given Name	**Michael F.**
	Suffix	
	Organization	
	Address	2240 Village Walk Drive, Henderson, NV , 89052, USA
	Email	mkorns@korns.com

Abstract

Research efforts, directed at increasing the accuracy and dependability of Symbolic Regression (SR), have resulted in significant improvements in symbolic regression's range, accuracy, and dependability. Previous research has also demonstrated the practicability of estimating corporate forward 12 month earnings, using advanced symbolic regression. In this paper we put these prior results and techniques together to select a 100 stock semi-passive index portfolio, extracted from the Value Line Timeliness stocks (*Value Line*), which delivers consistent performance in both bull and bear decades and we will compare its performance to the Standard & Poors 100 index.

We intend to produce our 100 stock semi-passive index buy list on a weekly basis using automated forward 12 month EPS (*ftmEPS*) prediction involving the analysis of many securities, involving multiple training regressions each on hundreds of thousands of training examples. Plus the timeliness issue will require that our analytic tools be strong and thoroughly matured. The 100 stock buy list will be the foundation for a new semi-passive Value Line 100 index fund which should have great appeal to many high net worth clients, enjoy low management costs, and be easily acceptable to the compliance and regulatory authorities.

Valuation of Value Line securities via their forward 12 month price earnings ratio (*ftmPE*) is a very common securities valuation method in the industry. Obviously the *ftmPE* valuation depends heavily on the estimate of forward 12 month corporate earnings per share (*ftmEPS*). Several obvious inputs to the *ftmEPS* prediction process are the past earnings time series plus one or more analyst predictions.

Valuation via *ftmEPS* is a necessary but not a sufficient attraction for a semi-passive index fund. So we will introduce the advantages of trading volatility. Our thesis will be that emotional trading patterns tend to make markets less efficient.

The efficient market hypothesis depends upon equal access to information and rational trading patterns. Trading on insider information is illegal in most developed securities markets; however, trading when others are emotional is unregulated. In this paper we will develop a set of factors—all of which incorporate a measure of volatility indicating possible overly emotional trading patterns. The theme of our new semi-passive index fund will be "*Buy value from those who are selling in a highly emotional state*".

Keywords (separated by "-")	Value investing - Symbolic regression - Logistic regression - Genetic programming - Nonlinear regression

Chapter 21 Trading Volatility Using Highly Accurate Symbolic Regression

Michael F. Korns

AQ1

21.1 Introduction

The discipline of Symbolic Regression (SR) has matured significantly in the last few years. There is at least one commercial package on the market for several years (http://www.rmltech.com/). There is now at least one well documented commercial symbolic regression package available for Mathematica (www.evolved-analytics.com). There is at least one very well done open source symbolic regression package available for free download (http://ccsl.mae.cornell.edu/eureqa).

AQ2

In addition to our own ARC system (Korns 2013, 2014), currently used internally for massive financial data nonlinear regressions, there are a number of other mature symbolic regression packages currently used in industry including Smits et al. (2010) and Castillo et al. (2010). Plus there is an interesting work in progress by McConaghy et al. (2009).

Research efforts, directed at increasing the accuracy and dependability of Symbolic Regression (SR), have resulted in significant improvements in symbolic regression's range, accuracy, and dependability (Korns 2013, 2014). Previous research has also demonstrated the practicability of estimating corporate forward 12 month earnings, using advanced symbolic regression (Korns 2012a, b). In this paper we put these prior results and techniques together to select a 100 stock semi-passive index portfolio (*VEP100*), from the Value Line Timeliness (*Value Line*), which delivers consistent performance in both bull and bear decades.

We intend to produce our VEP100 buy list on a weekly basis using automated *ftmEPS* prediction involving the analysis of many securities, involving multiple training regressions each on hundreds of thousands of training examples. Plus the timeliness issue will require that our analytic tools be strong and thoroughly

AQ3

M.F. Korns (✉)
2240 Village Walk Drive, Henderson, NV 89052, USA
e-mail: mkorns@korns.com

A.H. Gandomi et al. (eds.), *Handbook of Genetic Programming Applications*,
DOI 10.1007/978-3-319-20883-1_21

matured. Our new VEP100 semi-passive index fund should have great appeal to many high net worth clients, enjoy low management costs, and be easily acceptable to the compliance and regulatory authorities.

Valuation of securities via their forward 12 month price earnings ratio (*ftmPE*) is a very common securities valuation method in the industry. Obviously the *ftmPE* valuation depends heavily on the estimate of forward 12 month corporate earnings per share (*ftmEPS*). Obvious inputs to the *ftmEPS* prediction process are the past earnings time series plus one or more analyst predictions.

Valuation via *ftmEPS* is a necessary but not a sufficient attraction for a semi-passive index fund. So we will introduce the advantages of trading volatility. Our thesis will be that emotional trading patterns tend to make markets less efficient.

The efficient market hypothesis assumes rational trading patterns and equal and open access to information. Trading on insider information is illegal in most developed securities markets; but, *trading when others are emotional is unregulated*. In this paper we will develop a set of factors—all of which incorporate a measure of volatility indicating possible overly emotional trading patterns. The theme of our new VEP100 semi-passive index fund will be "*Buy value from those who are selling in a highly emotional state*".

Now would be a good time to provide an overview general introduction to symbolic regression as follows.

Symbolic Regression is an approach to general nonlinear regression which is the subject of many scholarly articles in the Genetic Programming community. A broad generalization of general nonlinear regression is embodied as the class of *Generalized Linear Models* (GLMs) as described in Nelder and Wedderburn (1972). A GLM is a linear combination of **I** basis functions $\mathbf{B_i}$; i = 1,2, …, I, a dependent variable **y**, and an independent data point with **M** features $\mathbf{x} = <x_1, x_2, x_3, \ldots, x_m>$: such that

$$y = \gamma(x) = C_0 + \sum_{i=1}^{I} c_i B_i(x) + err \tag{21.1}$$

As a broad generalization, GLMs can represent any possible nonlinear formula. However the format of the GLM makes it amenable to existing linear regression theory and tools since the GLM model is linear on each of the basis functions $\mathbf{B_i}$.

For a given vector of dependent variables, Y, and a vector of independent data points, X, symbolic regression will search for a set of basis functions and coefficients which minimize *err*. In Koza (1992) the basis functions selected by symbolic regression will be formulas as in the following examples:

$$B_1 = x_3 \tag{21.2}$$

$$B_2 = x_1 + x_4 \tag{21.3}$$

$$B_3 = \mathrm{sqrt}(x_2) / \tan(x_5/4.56) \tag{21.4}$$

$$B_4 = \tanh(\cos(x_2 * .2) * \text{cube}(x_5 + \text{abs}(x_1))) \tag{21.5}$$

If we are minimizing the least squared error, ***LSE***, once a suitable set of basis functions {**B**} have been selected, we can discover the proper set of coefficients {**C**} deterministically using standard univariate or multivariate regression. The value of the GLM model is that one can use standard regression techniques and theory. Viewing the problem in this fashion, we gain an important insight. Symbolic regression does not add anything to the standard techniques of regression. The value added by symbolic regression lies in its abilities as a search technique: how quickly and how accurately can SR find an optimal set of basis functions {**B**}.

The immense size of the search space provides ample need for improved search techniques In standard Koza-style tree-based Genetic Programming (Koza 1992) the genome and the individual are the same Lisp s-expression which is usually illustrated as a tree. Of course the tree-view of an s-expression is a visual aid, since a Lisp s-expression is normally a list which is a special Lisp data structure. Without altering or restricting standard tree-based GP in any way, we can view the individuals not as trees but instead as s-expressions such as this depth 2 binary tree s-exp: (/ (+x_2 3.45) (*x_0 x_2)), or this depth 2 irregular tree s-exp: (/ (+x_2 3.45) 2.0).

In standard GP, applied to symbolic regression, the non-terminal nodes are all operators (implemented as Lisp function calls), and the terminal nodes are always either real number constants or features. The maximum depth of a GP individual is limited by the available computational resources; but, it is standard practice to limit the maximum depth of a GP individual to some manageable limit at the start of a symbolic regression run.

Given any selected maximum depth k, it is an easy process to construct a maximal binary tree s-expression U_k, which can be produced by the GP system without violating the selected maximum depth limit. As long as we are reminded that each f represents a function node while each t represents a terminal node, the construction algorithm is simple and recursive as follows.

$$U_0: \text{ t}$$

$$U_1: \text{ (f t t)}$$

$$U_2: \text{ (f (f t t) (f t t))}$$

$$U_3: \text{ (f (f (f t t) (f t t)) (f (f t t) (f t t)))}$$

$$U_k: \text{ (f } U_{k\text{-}1} U_{k\text{-}1}\text{)}$$

Any basis function produced by the standard GP system will be represented by at least one element of U_k. In fact, U_k is isomorphic to the set of all possible basis functions generated by the standard GP system.

Given this formalism of the search space, it is easy to compute the size of the search space, and it is easy to see that the search space is huge even for rather simple basis functions. For our use in this chapter the function set will be the following functions: $\mathbf{F} = \{+\ -\ *\ /$ **abs sqrt square cube cos sin tan tan h log exp max min**$\aleph\}$ (where $\aleph(a,b) = \aleph(a) = a$). The terminal set is the features $\mathbf{x_0}$ thru $\mathbf{x_m}$ and the real constant **c**, which we shall consider to be 2^{64} in size. Where $|\mathbf{F}| = 17$, $\mathbf{M}=20$, and $\mathbf{k} = 0$, the search space is $S_0 = \mathbf{M} + 2^{64} = 20 + 2^{64} = 1.84 \times 10^{19}$. Where $\mathbf{k} = 1$, the search space is $S_1 = |\mathbf{F}| * S_0 * S_0 = 5.78 \times 10^{39}$. Where $\mathbf{k} = 2$, the search space grows to $S_2 = |\mathbf{F}| * S_1 * S_1 = 5.68 \times 10^{80}$. For $\mathbf{k} = 3$, the search space grows to $S_3 = |\mathbf{F}| * S_2 * S_2 = 5.5 \times 10^{162}$. Finally if we allow three basis functions $\mathbf{B} = 3$ for financial applications, then the final size of the search space is $S_3 * S_3 * S_3 = 5.5 \times 10^{486}$.

21.2 Methodology

Creating the weekly buy list for a modern semi-passive index fund requires many fully automated multiple regressions, all of which must be run in a timely fashion, and all of which must fit together seamlessly without human intervention. Our methodology is influenced by the practical issues of applying symbolic regression to the real world investment finance environment. First there is the issue that form of each symbolic regression must be preapproved by the regulatory authorities, the compliance officer, management, and clients. Second there is the issue of adapting symbolic regression to run in a real world financial application with massive amounts of data. Third there is the issue of modifying symbolic regression, as practiced in academia, to conform to the very difficult U.S. Securities Exchange Commission regulatory compliance environment.

Weekly preparation of our VEP100 semi-passive index fund buy list will require ∼1502 fully automated regressions (*as many as there are Value Line Timeliness stocks that week*). For each of the ∼1500 Value Line Timeliness stocks, a set of pre-approved earnings estimate inputs will be fed into a multiple linear regression for each stock, resulting in an interim forward 12 month earnings per share estimate for the stock. This will require ∼1500 regressions; but, they are relatively quick multiple linear regressions. Next, a set of preapproved earnings estimate inputs plus the interim *ftmEPS* estimate produced by the linear regressions will be input to a nonlinear weighted regression on all ∼1500 stocks. This expensive nonlinear weighted regression will produce a final *ftmEPS* estimate for each of the Value Line stocks. Finally, a set of preapproved z Score factor inputs plus the interim *ftmEPS* estimate produced by the linear regressions will be input to a nonlinear logistic regression on all ∼1500 stocks. This final very expensive nonlinear logistic regression will produce a final *expected forward 12 month total return* estimate for each of the Value Line stocks.

We use only statistical best practices out-of-sample testing methodology. For each regression, a matrix of independent variables will be constructed solely from the prior 10 years of historical data—520 weeks. No forward looking data will be allowed. This is very important because it will be the subject of detailed regulatory due diligence reviews. Then the preapproved regression model will be applied to produce the dependent variable.

For the forward estimation of corporate earnings, this paper uses an historical database of the Value Line stocks with daily price and volume data, weekly analyst estimates, and quarterly financial data from January 1990 to the December 2009. The data has been assembled from reports published at the time, so the database is highly representative of what information was realistically available at the point when trading decisions were actually made. No forward looking data is included in any historical point in the database.

From all of this historical data, 20 years (*1990 thru 2009*) have been used to produce the results shown in this research. This 24 year period includes a historically significant bull market decade followed by an equally historically significant bear market decade.

Multiple vendor sources have been used in assembling the data so that single vendor bias can be eliminated. The construction of this point in time database has focused on collecting weekly consolidated data tables, collected every Friday from January 3, 1986 to the present, representing detailed point in time input to this study and cover the Value Line stocks on a weekly basis. Each stock record contains daily price and volume data, weekly analyst estimates and rankings, plus quarterly financial data as reported. The primary focus is on gross and net revenues.

Our historical database contains 1050 weeks of data between January 1990 and December 2009. In a full training and testing protocol there is a separate symbolic regression run for each of these 1050 weeks. Each SR run consists of predicting the *ftmEPS* for each of the Value Line stocks available in that week, using the 520 prior weeks as the training data set for that week. A sliding training/testing window will be constructed to follow a strict statistical out-of-sample testing protocol.

For each of the 1050 weeks, the 520 prior weeks training examples will be extracted from records in the historical trailing 10 years behind the selected record BUT *not including any data from the selected week or ahead in time*. The training dependent variable will be extracted from the historical data record exactly 52 weeks forward in time from the selected record BUT *not including any data from the selected week or ahead in time*. Thus, as a practical observation, the training will not include any records in the first 52 weeks prior to the selected record—*because that would require a training dependent variable which was not available at the time*.

For each of the 1050 weeks, the testing samples will be extracted from records in the historical trailing 10 years behind the selected record *including all data from the selected week BUT not ahead in time*. The testing dependent variable will be extracted from the historical data record exactly 52 weeks forward in time from the selected record.

Each experimental protocol will produce approximately ∼1500 linear regressions and 2 symbolic regression runs over an average of ∼780,000 ($\sim 1500 \times 520$) records for each training run and for ∼1500 records for each testing run. Ten hours will be allocated for training. Of course separate R-Square statistics will be produced for each experimental protocol. We will examine the R-Square statistics for evidence favoring the addition of swarm intelligence over the base line and for evidence favoring one swarm intelligence technique over another.

Finally we will need to adapt our methodology to conform to the rigorous United States Securities and Exchange Commission oversight and regulations on investment managers. The SEC mandates that every investment firm have a compliance officer. For any automated forward earnings prediction algorithm, *which would be used as the basis for later stock recommendations to external clients or internal portfolio managers*, the computer software code used in each prediction, the historical data used in each prediction, and each historical prediction itself, must be filed with the compliance officer in such form and manner so as to allow a surprise SEC compliance audit to reproduce each individual forward prediction exactly as it was at the original time of publication to external clients or internal portfolio managers.

Of course this means that we must provide a copy of all code, all data, and each forward prediction for each stock in each of the 1050 weeks, to our compliance officer. Once management accepts our symbolic regression system, we will also have to provide a copy of all forward predictions on an ongoing basis to the compliance officer.

Furthermore there is an additional challenge in meeting these SEC compliance details. The normal manner of operating GP, and symbolic regression systems in academia will not be acceptable in a real world compliance environment. Normally, in academia, we recognize that symbolic regression is a heuristic search process and so we perform multiple SR runs, each starting with a different random number seed. We then report based on a statistical analysis of results across multiple runs. This approach produces *different results* each time the SR system is run. In a real world compliance environment such practice would subject us to serious monetary fines and also to jail time.

The SEC compliance requirements are far from arbitrary. Once management accepts such an SR system, the weekly automated predictions will influence the flow of millions and even billions of dollars into one stock or another and the historical back testing results will be used to sell prospective external clients and internal portfolio managers on using the system's predictions going forward.

First the authorities want to make sure that as time goes forward, *in the event that the predictions begin to perform poorly*, we will not simply rerun the original predictions again and again, with a different random number seed, until we obtain better historical performance and then substitute the new better performing historical performance results in our sales material.

Second the authorities want to make sure that, *in the event our firm should own many shares of the subsequently poorly performing stock of "ABC" Corp*, that we do not simply rerun the current week's predictions again and again, with a different random number seed, until we obtain a higher ranking for "ABC" stock thus improperly influencing our external clients and internal portfolio managers to drive the price of "ABC" stock higher.

In order to meet SEC compliance regulations we have altered our symbolic regression system, used in this chapter across all experiments, to use a pseudo random number generator with a pre-specified starting seed. Multiple runs always produce *exactly the same results*.

21.3 Investing Strategies

Value investing (Graham and Dodd 2008) has produced several of the wealthiest investors in the world including Warren Buffet. Nevertheless, value investing has a host of competing strategies including momentum (Bernstein 2001) and hedging (Nicholas 2000).

One of the most difficult challenges in devising a securities investing strategy is the a priori identification of pending regime changes. For instance, momentum investing strategies were very profitable in the 1990s and not so profitable in the 2000s while value investing strategies were not so profitable in the 1990s but turned profitable in the 2000s. Long Short hedging strategies were profitable in the 1990s and early 2000s but collapsed dramatically in the late 2007 thru 2008 period. Knowing when to switch from Momentum to Value, Value to Hedging, and Hedging back to Value was critical for making consistent above average profits during the 20 year period from 1990 thru 2009.

The challenge becomes even more difficult when one adds the numerous technical and fundamental buy/sell triggers to currently popular active management investing strategies. Bollinger Bands, MACD, Earning Surprises, etc. all have complex and dramatic effects on the implementation of securities investing strategies, and all are vulnerable to regime changes. The question arises, "*Is there a simple securities investing strategy which is less vulnerable to regime changes than other strategies*?".

An idealized value investing hypothesis is put forward: "*Given perfect foresight, buying stocks with the best future earning yield* **(*Future12MoEPS/CurrentPrice*)** (*ftmEP*) *and holding for 12 months will produce above average securities investing returns*".

Of course the ideal hypothesis is *impossible to implement* because it requires perfect foresight which is, in the absence of time travel, unobtainable. Nevertheless the ideal hypothesis represents the theoretical upper limit on the profits realizable from a strategy of buying future net revenue cheaply; yet, the theoretical profits are so rich that one cannot help but ask the question, "*Are there revenue prediction models which will allow one to capture some portion of the profits from the ideal hypothesis*?".

The easiest revenue prediction model involves simply using the current year's trailing 12 month revenue as a proxy for future revenue.

Table 21.1 Returns for SP100 High ttmEP/ftmEP 100

Year	SP100 stocks	100ttmEP stocks	100 ftmEP stocks
1990	(6 %)	(17 %)	3 %
1991	24 %	40 %	111 %
1992	3 %	22 %	56 %
1993	8 %	9 %	46 %
1994	0 %	6 %	18 %
1995	36 %	22 %	49 %
1996	23 %	28 %	38 %
1997	28 %	27 %	51 %
1998	32 %	12 %	12 %
1999	31 %	38 %	22 %
2000	(13 %)	14 %	45 %
2001	(15 %)	11 %	56 %
2002	(24 %)	(15 %)	8 %
2003	24 %	52 %	67 %
2004	4 %	13 %	45 %
2005	(1 %)	17 %	43 %
2006	16 %	7 %	19 %
2007	3 %	(5 %)	20 %
2008	(37 %)	(28 %)	(17 %)
2009	19 %	43 %	120 %
CAGR%	6 %	14 %	37 %
Volatility	20 %	20 %	30 %
CAGR% 1990s	17 %	18 %	38 %
CAGR% 2000s	(4 %)	8 %	37 %

Note: Per annum total returns for each year

The data supports the conclusion that even using this current revenue proxy model buying the top one hundred stocks with the highest (***current12MoEPS/ currentPrice***) (ttmEP) and holding for 1 year produces above average securities investing profits, *as least for the Value Line stocks*, as shown in Table 21.1.

Nevertheless, buying a stock with high EP, *but whose future 12 month earnings will plummet bringing on bankruptcy*, is an obviously poor choice. So why is high EP investing so successful given that future 12 month earnings can vary significantly? Placing current earnings yield investing in this context puts a new spin on this standard *value investing* measure. In this context we are saying that current earnings yield (also known as high EP investing) works precisely to the extent that *current earnings are a reasonable predictor of future earnings*! In situations where current earnings are NOT a good predictor of future earnings, then current earnings yield investing loses its efficacy.

This agrees with our common sense understanding. For instance, given two stocks with the same high current earnings yield, where one will go bankrupt next year and the other will double its earnings next year; we would prefer the stock whose earnings will double. Implying that, *in the ideal*, current earnings are just a data point. We want to buy *future earnings* cheap!

Precisely because the per annum returns from this current revenue prediction model are far less than the returns achieved with perfect prescience, we must now look for more accurate methods of net revenue prediction.

21.3.1 Estimating Forward 12 Month EPS

Each week we will perform ~1500 linear regressions, one for each of the Value Line stocks. The preapproved linear regressions are expressed by the following Regression Query Language **RQL** (Korns 2013) expression:

$$\mathbf{regress}\,(\mathbf{x0}, \mathbf{x1}, \mathbf{x2}, \mathbf{x3}, \mathbf{x4}, \mathbf{x5}, \mathbf{x6})\ \ \mathbf{where}\ \ \{\}$$

For each of the ~1500 Value Line stocks in the current week, from each of the 520 trailing historical weeks for that stock (*see our methodology section above*) the following seven input (*independent*) variables will be collected:

1.	**Estimated12MoEPS**(x0)	Wall Street analysts 12Mo forward EPS estimate
2.	**Forward12MoEPS**(x1)	CurrentEPS + (CurrentEPS-Past1YrEPS)
3.	**Projected12MoEPS**(x2)	CurrentEPS + ((CurrentEPS-Past1QtrEPS) * 4)
4.	**EstimatedS12MoEPS**(x3)	(Wall Street analysts 12Mo forward SPS estimate) * CurrentMargin
5.	**ForwardS12MoEPS**(x4)	(CurrentSPS + (CurrentSPS-Past1YrSPS)) * CurrentMargin
6.	**ProjectedS12MoEPS**(x5)	(CurrentSPS + ((CurrentSPS-Past1QtrSPS) * 4)) * CurrentMargin
7.	**WeeksSinceLastReport**(x6)	Absolute count of weeks since last quarterly report

Each of the ~1500 linear regressions produces an *ftmEPS* estimate for each of the Value Line stocks for that week (**LRegress12MoEPS**). This regression output is then used as an input to a single preapproved nonlinear weighted regression on the following input variables:

The preapproved nonlinear weighted regression is expressed by the following Regression Query Language **RQL** (Korns 2013) expression:

1.	**Estimated12MoEPS** ($x0$)	Wall Street analysts 12Mo forward EPS estimate
2.	**Forward12MoEPS**($x1$)	CurrentEPS + (CurrentEPS-Past1YrEPS)
3.	**Projected12MoEPS**($x2$)	CurrentEPS + ((CurrentEPS-Past1QtrEPS) * 4)
4.	**EstimatedS12MoEPS**($x3$)	(Wall Street analysts 12Mo forward SPS estimate) * CurrentMargin
5.	**ForwardS12MoEPS**($x4$)	(CurrentSPS + (CurrentSPS-Past1YrSPS)) * CurrentMargin
6.	**ProjectedS12MoEPS**($x5$)	(CurrentSPS + ((CurrentSPS-Past1QtrSPS) * 4)) * CurrentMargin
7.	**WeeksSinceLastReport**($x6$)	Absolute count of weeks since last quarterly report
8.	**LRegress12MoEPS**($x7$)	Result of the linear regression for the stock in question

$$\begin{aligned}&\mathbf{model}\big(\mathbf{c0} * \mathbf{f0}\,(\mathbf{x0}, \mathbf{v0})\,, \mathbf{c1} * \mathbf{f1}\,(\mathbf{x1}, \mathbf{v1})\,, \mathbf{c2} * \mathbf{f2}\,(\mathbf{x2}, \mathbf{v2})\,,\\ &\quad \mathbf{c3} * \mathbf{f3}\,(\mathbf{x3}, \mathbf{v3})\,, \mathbf{c4} * \mathbf{f4}\,(\mathbf{x4}, \mathbf{v4})\,, \mathbf{c5} * \mathbf{f5}\,(\mathbf{x5}, \mathbf{v5})\,,\\ &\quad \mathbf{c6} * \mathbf{f6}\,(\mathbf{x6}, \mathbf{v6})\,, \mathbf{c7} * \mathbf{f7}\,(\mathbf{x7}, \mathbf{v7})\,\big)\end{aligned}$$

$$\mathbf{where}\ \Big\{\mathbf{op}\,(\aleph, +, -, \mathbf{min}, \mathbf{max})$$

$$\mathbf{c0}\,(\mathbf{0.0}, \mathbf{1.0})\ \ \mathbf{c1}\,(\mathbf{0.0}, \mathbf{1.0})\ \ \mathbf{c2}\,(\mathbf{0.0}, \mathbf{1.0})\ \ \mathbf{c3}\,(\mathbf{0.0}, \mathbf{1.0})\ \ \mathbf{c4}\,(\mathbf{0.0}, \mathbf{1.0})$$

$$\mathbf{c5}\,(\mathbf{0.0}, \mathbf{1.0})\ \ \mathbf{c6}\,(\mathbf{0.0}, \mathbf{1.0})\ \ \mathbf{c7}\,(\mathbf{0.0}, \mathbf{1.0})\Big\}$$

This nonlinear weighted regression will achieve regulatory and client preapproval because it is so intuitive and so easy to explain. Let us start with the simplest case where the functions (**f0** thru **f7**) are all noops $= \aleph$, then the final result will always be like the following example:

$$\begin{aligned}\mathbf{NLREstimated12MoEPS}\,(y) = {} & \mathbf{.34} * \mathbf{x0} + \mathbf{.16} * \mathbf{x1} + \mathbf{.81} * \mathbf{x2} + \mathbf{.54} * \mathbf{x3}\\ & + \mathbf{.26} * \mathbf{x4} + \mathbf{.72} * \mathbf{x5} + \mathbf{.59} * \mathbf{x6} + \mathbf{.21} * \mathbf{x7}\end{aligned}$$

We have eight inputs in the form of dollar values for next year's estimated EPS. Our model simply assigns a weight (*0.0<= 1.0*) to each estimate—with the added benefit that, in the past 520 weeks for all ~1500 Value Line stocks, these weights have been the most successful in predicting next year's EPS values for the Value Line stocks. Now moving on to the case where one of more of the functions (**f0** thru **f7**) are other than noops, then the final result will always be something like the following example:

$$\begin{aligned}\textbf{NLREstimated12MoEPS}\,(y) = {} & \mathbf{.34 * x0 + .16 * x1 + .81 * max\,(x2, x0)} \\ & \mathbf{+ .54 * x3 + .26 * x4 + .72 * x5 + .59 * x6} \\ & \mathbf{+ .21 * x7}\end{aligned}$$

Again we have eight inputs in the form of dollar values for next year's estimated EPS. Our model simply assigns a weight (*0.0<= 1.0*) to each possible simple combination of those estimates—with the added benefit that, in the past 520 weeks, these weights and these combinations have been the most successful in predicting next year's EPS values for the SP100 stocks.

In all cases we are simply weighting simple estimates or simple combinations of estimates with combinations that will never get unruly or out of hand and with weights which will always remain safely between 0.0 and 1.0. For this intuitive nonlinear weighted regression, Regulatory and client preapproval will be easy to obtain.

21.3.2 Estimating Forward 12Mo Total Return

A close examination of the Efficient Market Hypothesis (*EMH*) shows that the expectation of rational investing decisions plays a significant role in the EMH conclusions in favor of passive index investing. Therefore, in addition to attempting to purchase cheap stocks (*via some estimate of future 12Mo earnings*), we would also like to purchase stocks from sellers whose decisions may not be as rational as the EMH might hope.

Normally each stock trades within its own average trading volume over the course of weeks and months. This trading volume can be expressed as a percent ***WeeksVolume*** **= (total number of shares traded today)/(total shares outstanding)**. For any given stock there will be periods of calm when weekly trading percent (*WeeksVolume*) is light compared to the its historical average, and periods of frenzy when the weekly trading percent (*WeeksVolume*) is very high compared to its historical average. Our assertion is that *when a trading frenzy is underway the buyer and seller are less rational than on normal trading days.*

The following nine input factors (*each of which combines some measure of trading frenzy or intrinsic value or both*) will be converted to z Scores (Anderson et al. 2002) and are defined as follows:

First we see that **z Panic Level** is computed from the nonlinear regression future 12Mo EPS estimate divided by the week's closing price (*i.e. the estimated future EPS yield*) times the percent of outstanding shares traded that week (*Weeks Volume*) times the current week's trading percent as in comparison with the prior 52 weeks trading percent (*Volume52WeekRange*). This input will be high when the estimated future earnings yield is high (*the stock is cheap*), when a high percent of outstanding shares traded this week (*Weeks Volume*), and when this week's trading volume is on the high side compared to the previous 52 week trading history for this stock. This

1.	**zPanicLevel** ($x0$)	((NLRFuture12MoEPS/WeeksClose) * WeeksVolume * Volume52WeekRange))	t12.1
2.	**zPriceMomentum**($x1$)	(Past52WeekReturn * WeeksVolume)	t12.2
3.	**zDollarVolume**($x2$)	(WeeksVolume * Shares * WeeksClose)	t12.3
4.	**zFutureEPSYield**($x3$)	(NLRFuture12MoEPS/WeeksClose)	t12.4
5.	**zSalesAttractiveness**($x4$)	(Current12MoSPS/WeeksClose) * WeeksVolume	t12.5
6.	**zCurrentValuation**($x5$)	(CurrentVPS/WeeksClose)	t12.6
7.	**zValuationAttractiveness**($x6$)	(CurrentVPS/WeeksClose) * WeeksVolume	t12.7
8.	**zWallStreetRank**($x7$)	Current Wall Street analysts ranking as a z Score	t12.8
9.	**zFinancialRank**($x8$)	Current Wall Street financial ranking as a z Score	t12.9

is a stock selling on much higher volume than normal with a very cheap future earnings yield. We use this input as a measure of panic on the seller's side. Since each of these inputs are z Scores, a high value for this input indicates that this stock is in a greater trading frenzy relative to other stocks this week.

Second we see that **z Price Momentum** is computed from the stock's past 52 week total return (*Past52WeekReturn*) times the week's trading volume (*Weeks Volume*). This input will be high for stocks with strong momentum selling on high trading volume. This is a popular stock, and we use this input as a measure of price momentum on the buyer's side. Since each of these inputs are z Scores, a high value for this input indicates that this stock enjoys greater momentum relative to other stocks this week.

Third we see that **z Dollar Volume** is an estimate of the total dollar value of the shares traded this week. This is a popular stock, and we use this input as a measure of relative dollar flow through this stock as opposed to other stocks this week. Since each of these inputs are z Scores, a high value for this input indicates that more dollars are flowing through this stock than other stocks this week.

Fourth we see that **z Future EPS Yield** is a measure of how cheap the future 12Mo EPS estimate divided by the week's closing price (*i.e. the estimated future EPS yield*) is compared to other stocks this week. This input will be high when the estimated future earnings yield is high (*the stock is cheap*). Since each of these inputs are z Scores, a high value for this input indicates that this stock is cheaper relative to other stocks this week.

Fifth we see that **z Sales Attractiveness** is computed from the current 12Mo SPS divided by the week's closing price (*i.e. the current sales yield*) times the percent of outstanding shares traded that week (*Weeks Volume*). This input will be high when the current sales yield is high (*the stock is cheap*), and when a high percent of outstanding shares traded this week (*Weeks Volume*. This is a stock selling on high volume with a very cheap current sales yield. We use this input as a measure of attraction on the buyer's side.

Sixth we see that **z Current Valuation** is a measure of the current enterprise value divided by the week's closing price (*i.e. the current VPS yield*). This input will be high when the current VPS yield is high (*the stock is cheap*). Since each of these inputs are z Scores, a high value for this input indicates that this stock is cheaper relative to other stocks this week.

Seventh we see that **z Valuation Attractiveness** is a measure of the current enterprise value divided by the week's closing price (*i.e. the current VPS yield*) times this week's trading volume (*Weeks Volume*). This input will be high when the current VPS yield is high (*the stock is cheap*), and trading volume is high. Since each of these inputs are z Scores, a high value for this input indicates that this stock is more attractive to buyers relative to other stocks this week.

Eighth we see that **z Wall Street Rank** is a measure of the current Wall Street analysts' rank for this stock. This input will be high when the Wall Street analysts' rank for this stock is high. Since each of these inputs are z Scores, a high value for this input indicates that this stock enjoys a higher Wall Street analyst ranking relative to other stocks this week.

Ninth we see that **z Financial Rank** is a measure of the current Wall Street analysts' financial rank for this stock. This input will be high when the Wall Street analysts' financial rank for this stock is high. Since each of these inputs are z Scores, a high value for this input indicates that this stock enjoys a higher Wall Street analyst financial ranking relative to other stocks this week.

The following single output factor (*what we train on*) will be converted to sigmoid score (Kleinbaum et al. 2010; Anderson et al. 2002) and is defined as follows:

1.	**sFuture12MoReturn** (*y*)	The actual Future 12Mo Total Return—as a sigmoid-score

Obviously we are not trying to predict actual future 12 month total return so much as we are trying to predict relative future 12 month total return. We don't really need to know actual future total 12 month returns. We only need to select the 100 Value Line stocks with the highest *relative* estimated future total 12 month return. This allows us the luxury of converting the output variable (**sFuture12MoReturn**) to a sigmoid factor, which allows us to perform a nonlinear logistic regression (Kleinbaum et al. 2010) of the following form.

$$\mathbf{logit}\big(\mathbf{f0}\,(\mathbf{x0},\mathbf{v0})\,,\mathbf{f1}\,(\mathbf{x1},\mathbf{v1})\,,\mathbf{f2}\,(\mathbf{x2},\mathbf{v2})\,,\mathbf{f3}\,(\mathbf{x3},\mathbf{v3})\,,\mathbf{f4}\,(\mathbf{x4},\mathbf{v4})\,,\mathbf{f5}\,(\mathbf{x5},\mathbf{v5})\,,\\ \mathbf{f6}\,(\mathbf{x6},\mathbf{v6})\,,\mathbf{f7}\,(\mathbf{x7},\mathbf{v7})\,,\mathbf{f8}\,(\mathbf{x8},\mathbf{v8})\big)\ \mathbf{where}\ \{\mathbf{op}\,(\aleph,+,-,\mathbf{min},\mathbf{max})\}$$

This simple and extremely intuitive nonlinear logistic regression will easily win regulatory and client preapproval. First of all this nonlinear regression will never produce unexpected or wild output. It will produce an orderly estimate for (**sFuture12MoReturn**) which will always lie between 0.0 and 1.0 for each stock. In essence, this nonlinear regression model will automatically rank each stock between 0.0 and 1.0 in terms of estimated future 12 month total return (*with 1.0 being the most desirable and 0.0 being the least desirable*). Let us start with the simplest case where the functions (**f0** thru **f8**) are all noops = $\aleph$, then the final result will always be like the following example:

$$\begin{aligned}\mathbf{sFuture12MoReturn}(y) = \mathbf{sigmoid}\big(&.34 * \mathbf{x0} + .16 * \mathbf{x1} + .81 * \mathbf{x2}\\ &+ .54 * \mathbf{x3} + .26 * \mathbf{x4} + .72 * \mathbf{x5} + .59 * \mathbf{x6}\\ &+ .21 * \mathbf{x7} + .91 * \mathbf{x8}\big)\end{aligned}$$

We have nine inputs in the form of z Scores for factors combining some measure of relative value and/or trading frenzy. Our model simply projects each weighted factor onto a relative ranking between 0.0 and 1.0—with the added benefit that, in the past 520 weeks, these weights have been the most successful in predicting next year's relative future 12Mo total return for the Value Line stocks.

Now moving on to the case where one of more of the functions (**f0** thru **f8**) are other than noops, then the final result will always be something like the following example:

$$\begin{aligned}\mathbf{sFuture12MoReturn}(y) = \mathbf{sigmoid}\big(&.34 * \mathbf{x0} + .16 * \mathbf{x1} + .81 * \mathbf{max}\,(\mathbf{x2}, \mathbf{x0})\\ &+ .54 * \mathbf{x3} + .26 * \mathbf{x4} + .72 * \mathbf{x5} + .59 * \mathbf{x6}\\ &+ .21 * \mathbf{x7} + .91 * \mathbf{x8}\big)\end{aligned}$$

Again we have nine inputs in the form of z Scores for factors combining some measure of relative value and/or trading frenzy. Our model projects each weighted factor or simple combination of factors onto a relative ranking between 0.0 and 1.0—with the added benefit that, in the past 520 weeks, these weights and these combinations have been the most successful in predicting next year's relative future 12Mo total return for the Value Line stocks.

21.3.3 Historical Returns

Applying all of these tools, techniques, and factors to the task of creating our semi-passive VEP100 index fund, we perform our 1502 regression runs for the first week in each year from 1990 thru 2009. We select the 100 Value Line stocks with the highest **sFuture12MoReturn** values. And hold them for 1 year. We then compare the results to the SP100 passive index, buying the 100 stocks with the highest ttmEP, and buying the 100 stocks with the highest ftmEP and present the results in Table 21.2.

Our VEP100 semi-passive index produced a much higher compound annual growth rate (*CAGR%*) than the SP100 index and the 100 ttmEP method. However, it cannot compete with the ideal ftmEP method (*where one can see into the future*). Nevertheless the total return of our VEP100 semi-passive index is impressive and will definitely appeal to a wide range of high net worth clients.

So have we beaten the Efficient Market Hypothesis? With a little bit of humor I can answer with a definite **Yes** and **No**.

Table 21.2 Returns VEP 100

Year	SP100 stocks	100 ttmEP stocks	VEP100 index fund	100 ftmEP stocks
1990	(6 %)	(17 %)	(22 %)	3 %
1991	24 %	40 %	47 %	111 %
1992	3 %	22 %	33 %	56 %
1993	8 %	9 %	23 %	46 %
1994	0 %	6 %	0 %	18 %
1995	36 %	22 %	30 %	49 %
1996	23 %	28 %	24 %	38 %
1997	28 %	27 %	31 %	51 %
1998	32 %	12 %	0 %	12 %
1999	31 %	38 %	30 %	22 %
2000	(13 %)	14 %	10 %	45 %
2001	(15 %)	11 %	38 %	56 %
2002	(24 %)	(15 %)	(6 %)	8 %
2003	24 %	52 %	62 %	67 %
2004	4 %	13 %	30 %	45 %
2005	(1 %)	17 %	29 %	43 %
2006	16 %	7 %	8 %	19 %
2007	3 %	(5 %)	13 %	20 %
2008	(37 %)	(28 %)	(42 %)	(17 %)
2009	19 %	43 %	131 %	120 %
CAGR%	6 %	14 %	17 %	37 %
Volatility	20 %	20 %	30 %	30 %
CAGR% 1990s	17 %	18 %	18 %	38 %
CAGR% 2000s	(4 %)	8 %	20 %	37 %

Note: Per annum total returns for each year

Yes, because the VEP100 CAGR% of 17 % is a whopping 9 % per annum greater than the SP100! This is a significant amount which will be of interest to a large class of serious investors. Furthermore, the performance of the VEP100 is more consistent across bull and bear decades with a CAGR % of 18 % in the bullish 1990s and a CAGR% of 20 % in the bearish 2000s. Coupled with the transparent and intuitive methodology of the VEP100, there is definite added value here.

No, because the EMH does not actually claim that one cannot make higher profits than the indices. The EMH claims that one cannot increase returns without also increasing volatility, and this is exactly what happens with the VEP100 semi-passive index. Volatility increases from 20 % with the SP100 to 30 % with the VEP100. So in an important way, the VEP100 is a classic confirmation of the Efficient Market Hypothesis.

21.4 Summary

Advances in both the industrial strength and accuracy of Symbolic Regression packages can help overcome the resistance to SR in the investment finance industry. Management trust, regulatory approval, and client acceptance, are no longer the severe hurdles that they were in the past. Improvements in SR robustness, result invariance, demonstrable accuracy, and regression constraint languages, such as Regression Query Language **RQL** (Korns 2010, 2013, 2014), now support regulatory and client preapproval of important component SR processes.

In this research work, as series of cascade linear and nonlinear SR regressions are used to create a transparent semi-passive index fund with significantly higher returns, over the 1990–2009 two decade period, than its Standard &Poors 100 index benchmark. Because of its transparent and algorithmic nature, the new VEP100 semi-passive index fund could enjoy much lower costs than a standard active fund and yet enjoy attractive returns—costs similar in nature to the SP100 passive index fund.

Future research will focus on other semi-passive indices with performance tailored to various diverse client needs and requirements, and regulatory approval issues.

AQ4

References

Graham, Benjamin, and David Dodd. 2008. Securities Analysis. New York, New York, USA. McGraw-Hill.

Kennedy, J.; Eberhart, R. 1995. Particle Swarm Optimization. *Proceedings of IEEE International Conference on Neural Networks*. **IV**. pp. 1942–1948.

Korns, Michael F. 2007. Large-Scale, Time-Constrained Symbolic Regression-Classification. In Riolo, Rick, L, Soule, Terrance, and Wortzel, Bill, editors, Genetic Programming Theory and Practice V, New York, New York, USA. Springer, pp. 299–314.

Korns, Michael F., and Nunez, Loryfel, 2008. Profiling Symbolic Regression-Classification. In Riolo, Rick, L, Soule, Terrance, and Wortzel, Bill, editors, Genetic Programming Theory and Practice VI, New York, New York, USA. Springer, pp. 215–228.

Korns, Michael F., 2009. Symbolic Regression of Conditional Target Expressions. In Riolo, Rick, L, Soule, Terrance, and Wortzel, Bill, editors, Genetic Programming Theory and Practice VII, New York, New York, USA. Springer, pp. 211–228.

Korns, Michael F., 2010. Abstract Expression Grammar Symbolic Regression. In Riolo, Rick, L, Soule, Terrance, and Wortzel, Bill, editors, Genetic Programming Theory and Practice VIII, New York, New York, USA. Springer, pp. 109–128.

Price, Kenneth, Storn, Rainer, Lampinen, Jouni 2009. Differential Evolution: A Practical Approach to Global Optimization. New York, New York, USA. Springer.

Guido Smits, Ekaterina Vladislavleva, and Mark Kotanchek 2010, Scalable Symbolic Regression by Continuous Evolution with Very Small Populations, in Riolo, Rick, L, Soule, Terrance, and Wortzel, Bill, editors, *Genetic Programming Theory and Practice VIII*, New York, New York, USA. Springer, pp. 147–160.

Flor Castillo, Arthur Kordon, and Carlos Villa 2010, Genetic Programming Transforms in Linear Regression Situations, in Riolo, Rick, L, Soule, Terrance, and Wortzel, Bill, editors, *Genetic Programming Theory and Practice VIII*, New York, New York, USA. Springer, pp. 175–194.

Trent McConaghy, Pieter Palmers, Gao Peng, Michiel Steyaert, Goerges Gielen 2009, Variation-Aware Analog Structural Synthesis: A Computational Intelligence Approach. New York, New York, USA. Springer.

J.A., Nelder, and R. W. Wedderburn, 1972, *Journal of the Royal Statistical Society, Series A, General*, 135:370–384.

John R Koza 1992, Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge Massachusetts, The MIT Press.

Korns, Michael F., 2011a. Accuracy in Symbolic Regression. In Riolo, Rick, L, Soule, Terrance, and Wortzel, Bill, editors, Genetic Programming Theory and Practice IX, New York, New York, USA. Springer (*to be published in winter 2011*).

Pham, D., Ghanbarzadeh, A., Koc, E., Otri, S., Rahim, S., and Zaidi, M. 2005. "The Bees Algorithm". Technical Report Cardiff University.

Parpinelli, R. S., and Lopes, H. S., 2011. New inspirations in swarm intelligence: a survey. *Int Journal of Bio-inspired Computation*. **Vol 3**. Number 1.

Bernstein, J., 2001. Momentum Stock Selection: Using The Momentum Method for Maximum Profits. New York, New York, McGraw Hill

Nicholas, J., 2000. Market-Neutral Investing: Long/Short Hedge Fund Strategies. New York, New York, Bloomberg Press.

Poli, Riccardo, McPhee, Nicholas, Vanneshi, Leonardo, 2009. Analysis of the Effects of Elitism on Bloat in Linear and Tree-based Genetic Programming. In Riolo, Rick, L, Soule, Terrance, and Wortzel, Bill, editors, Genetic Programming Theory and Practice VI, New York, New York, USA. Springer, pp. 91–110.

Korns, Michael F, 2011b. Accuracy in Symbolic Regression. In Riolo, Rick, L, Soule, Terrance, and Wortzel, Bill, editors, Genetic Programming Theory and Practice IX, New York, New York, USA. Springer.

Korns, Michael F., 2012a. A Baseline Symbolic Regression Algorithm. In Soule, Terrance, and Wortzel, Bill, editors, Genetic Programming Theory and Practice X, New York, New York, USA. Springer.

Korns, Michael F., 2013. Extreme Accuracy in Symbolic Regression. In Soule, Terrance, and Wortzel, Bill, editors, Genetic Programming Theory and Practice XI, New York, New York, USA. Springer.

Korns, Michael F., 2014. Extremely Accurate Symbolic Regression for Large Feature Problems. In Soule, Terrance, and Wortzel, Bill, editors, Genetic Programming Theory and Practice XII, New York, New York, USA. Springer.

Korns, Michael F., 2012b. Predicting Corporate Forward 12 Month Earnings, 2012. Theory and New Applications of Swarm Intelligence, ISBN 978-953-51-0364-6, edited by Rafael Parpinelli and Heitor S. Lopes, InTech Academic Publishers.

Kleinbaum, David G., and Klein, Michael, 2010. Logistic Regression: A Self-Learning Text (Statistics for Biology and Health), ISBN 978–1441917416, New York, New York, USA. Springer.

Anderson, David R., Sweeney, Dennis J., and Williams, Thomas A, 2002. Essentials of Statistics for Business and Economics, ISBN 978–0324145809, Southwestern College Publishers.

AUTHOR QUERIES

AQ1. Please note that the reference style has been changed from a Numbered style to a Name–Date style as per the style.

AQ2. Please check the change made in the sentence "In addition to our own ARC system . . . " and correct if necessary.

AQ3. Please provide department and institute/university details in the affiliation.

AQ4. Refs. "Kennedy and Eberhart (1995), Korns (2007, 2009, 2011a, b), Korns et al. (2008), Price et al. (2009), Pham et al. (2005), Parpinelli and Lopes (2011), and Poli et al. (2009)" are not cited in the text. Please provide the citation or delete them from the list.