Introductory Analysis of Successful Stock Portfolio Annual Returns Using Basic Regression Methods

Matthew Sears[1] and Semere Habtemicael[2]

Abstract— **The canonical forms of both linear and quadratic multiple regression models are used to determine if the combination of six specific stock-picking concepts interpreted as components in a mixture model prove to be a good indicator of stock portfolio annual returns. In this paper, we show if these regression methods can serve as a fairly accurate model to predict long-term annual return based on such stock-picking concepts. We also discuss the differences of each model and their individual performances. Additionally, Cook's Distance is used to make our models more robust, increasing the accuracy of each model by removing potential outliers.**

I. INTRODUCTION

During the Fall 2017 semester, we came across a data set that would serve as the perfect introduction for an undergraduate in Applied Mathematics to learn and apply multiple regression methods for data analysis in R. In addition, the exercise of typesetting these findings into LaTeX will provide such an undergraduate the proper experience for future endeavors. For this research, annual return is the only response we chose to analyze, out of the six available, in accordance with the six observations available. For a detailed explanation of how this data set was generated, how the simplex coordinate system is used, and how a more advanced methods of stock selection decision support systems are used, see the reference research paper [1].

II. OVERVIEW

A. Motivation

The motivation behind this research is to see if regression methods can be used as an accurate predictor of long-term annual returns given a set of stock-picking concepts that are weighted as components a mixture model. To verify this, we divided our data into training periods and testing periods. Models with a low mean square error (MSE) with our training data and strong correlation with our testing data will allow us to infer that such a model is useful for determining what mix of stock-picking concepts can achieve long-term high annual returns.

B. Stock-Picking Concepts

The first 4 stock-picking concepts are used to select stocks with the characteristics of high return rates. The remaining 2 are used to select stocks with characteristics of low risk. All 6 will be considered as observations, or components in a mixture model.

- **Large B/P** – book value-to-price ratio is used to compare a stock's book value to its market value. A large B/P ratio could indicate that a stock is undervalued.
- **Large S/P** – sales-to-price ratio is used to compare a stock's revenues to its stock price. A large S/P ratio could also indicate that a stock is undervalued.
- **Large ROE** – return on equity is used as a measure of a stock's profitability. A large return on equity could indicate that the company is efficient and profitable.
- **Large return rate in the last quarter** – a momentum factor that indicates if the return rate of a stock is high, it will continue to go higher in a midterm
- **Large market capitalization** – used to determine a company's size. A large market capitalization could indicate low risk and high liquidity.
- **Small systematic risk** – measures the fluctuation in stock returns relative to benchmarks in the market. A small systematic risk means a higher likelihood of maintaining a positive return.

C. Data

The data set used is comprised of 4 periods of stock market data, each consisting of 20 quarters within 5 years. See [1] for how the observations were scored and weighted as components in a mixture. For our training data, we use the first period, September 1990 – June 1995, and the second period, September 1995 – June 2000. For our first training set, we will test it on the second and third periods (up to June 2005). Our second training set will be used on only the third period. The fourth period is considered an anomaly and thus ignored as testing data since the extreme volatility caused by the 2008 stock market crash is unable to be accurately captured by our regression models.

D. Regression Methods

- **Multiple Linear Regression** – Multiple Linear Regression attempts to model the relationship of multiple observations and a response variable with a linear fit to the observed data. Since this data is derived from a mixture experiment, the standard form of a cartesian-based multiple linear regression model:

$$y = \beta_0 + \sum_{i=1}^{q} \beta_i x_i \tag{1}$$

[1] M. Sears is an undergraduate in the Department of Applied Mathematics, Wentworth Institute of Technology, Boston, MA 02115 `searsm1@wit.edu`

[2] S. Habtemicael is M. Sears' mentor, academic advisor, and Assistant Professor in the Department of Applied Mathematics, Wentworth Institute of Technology, Boston, MA 02115 `habtemicaels@wit.edu`

will need to be slightly modified. This is due to the fact that there are several specifications to be noted when modeling a mixture experiment [2]:

$$\beta_0 = 0$$

$$\sum_{i=1}^{q} x_i = 1$$

$$x_i \geq 0,\ i = 1, 2, \ldots, q$$

$$x_i = 0, \frac{1}{m}, \frac{2}{m}, \ldots, 1$$

where q denotes the number of independent factors and m denotes the number of mixture components (in this case, $q = m = 6$). Subsequently, a simplex coordinate system is substituted for the standard cartesian coordinate system. Additionally, the y-intercept, β_0, will be ignored because we do not expect to receive a return if we do not use any stock-picking concepts to participate in the market. This transforms Eq. (1) into the canonical form:

$$y = \sum_{i=1}^{q} \beta_i x_i + \epsilon \tag{2}$$

where β_i denotes the model's regression parameter coefficients and ϵ denotes the error/residual.

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

- **Quadratic Regression** – expanding on Eq. (2), the canonical form of the quadratic mixture model is:

$$y = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j}^{q} \beta_{ij} x_i x_j + \epsilon \tag{3}$$

where the parameters are defined similarly to Eq. (2). This will theoretically provide a more accurate fit to the data, since there are six levels of each of the six mixture components.

III. ANALYSIS

A. *Verifying Assumptions*

We use a quantile-quantile plot on our data to make sure that is normally distributed. Using this plot, shown in Fig. 1 and Fig. 2, allows us to infer that we're able to properly use our regression models for this data and also determine our errors based on the normal distribution.

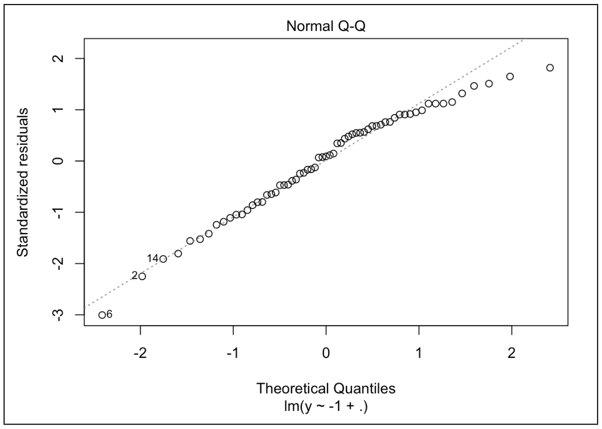


Fig. 1. Quantile-quantile plot of our linear regression that encourages us that our data is normally distributed, as it fits a straight line

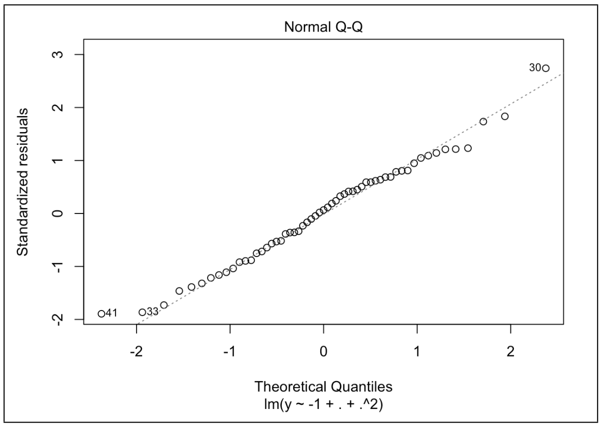


Fig. 2. Quantile-quantile plot of our quadratic regression that encourages us that our data is normally distributed, as it fits a straight line

B. *Regression Performance*

Table 1 shows the correlation between the predicted $\hat{Y}$ values generated by our regression models using the training data vs. the Y values of the testing periods. Here, we see that Linear 1, which is the linear regression from the first training set, outperforms the other models. It is strongly correlated with our third testing period, which indicates that our model trained from September 1990 – June 1995 data was able to predict annual returns one full decade later. This means that we can use this as a fairly accurate model to predict long-term annual return based on such stock-picking concepts. Comparing Fig. 3 and Fig. 4, we can see how the MSE of every model can be reduced by 6% by removing the outliers seen in Fig. 5.

- **Linear Regression** – In Table 1, Linear 2, which is the linear regression from the second training set, does not perform as well as Linear 1 to predict the third period, which suggests that our linear regression models are better suited for long-term prediction instead of short-term prediction.
- **Quadratic Regression** – Despite having a very low MSE, as shown in Fig. 3 and Fig. 4, our quadratic regression models were unable to perform well in its predictions, as shown in Table 1. This suggest that our

model was overfitting the data, which in turn limits its predictive capability.

TABLE I

CORRELATION OF Y AND $\hat{Y}$

	Test Period 2	Test Period 3
Linear 1	-0.4728	0.7702
Quadratic 1	0.0561	0.4571
Linear 2	**N/A**	-0.5596
Quadratic 2	**N/A**	-0.2126

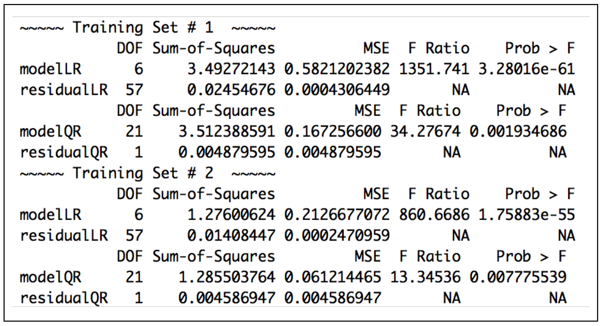
C. *Analysis of Variance (ANOVA)*

ANOVA, as shown in Fig. 3 and Fig. 4, consists of calculations that provide information about levels of variability within a regression model and form a basis for tests of significance. In other words, it is a powerful statistical technique that involves partitioning the observed variance into different components to conduct various significance tests.

- **Degrees of Freedom (DOF)** – for our models, this is defined as the number of observations/explanatory variables p. Regarding the residuals, or the error degrees of freedom, this is defined as $p - n - 1$ observations where n is the number of samples.
- **Sum-of-Squares** – This is defined as the sum, over all observations, of the squared differences of each observation from their mean. It relates to the total variance of the observations.
- **MSE** – This sample variance is estimated by the mean square, and is obtained by dividing the sum of squares by the respective degrees of freedom. The MSE is a measure of the quality of our predictive model, with values closer to 0 indicating higher accuracy.
- **F Ratio** – This statistic is a ratio of the model mean square and the residual mean square. Large values of this test statistic provide evidence against the null hypothesis that allows us to safely reject it. The null hypothesis, H_0, states that $\beta_1, \beta_2, \ldots, \beta_q = 0$, and the alternative hypothesis, H_1, simply states that at least one of the parameters $\beta_i \neq 0$ for $i = 1, 2, \ldots, q$.
- **Prob $>$ F** – This checks if the model as a whole is appropriate in predicting our Y values. We expect that with very small probability we can safely reject the null hypothesis H_0. We got a very small number for all our models which is consistent with our findings.

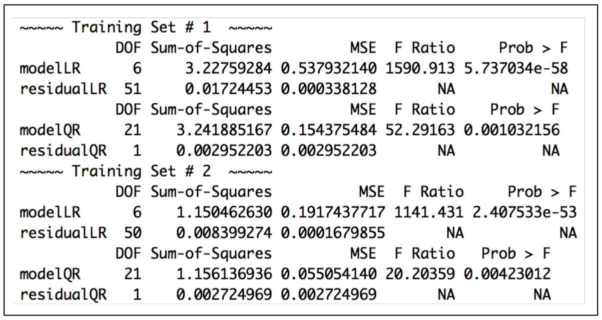
IV. CONCLUSIONS

One strong assumption that we made was that the error is normally distributed. To improve our results in the future, we could use an exponential or poisson distibrution for the error in our models to add more randomness to avoid overfitting and also give a better predictive results. What we understand from our results is that these stock-picking concepts were a good indicator for a long-term annual return prediction (10 years in the future) using linear regression, and not for the short-term.



```
~~~~~ Training Set # 1  ~~~~~
           DOF Sum-of-Squares          MSE  F Ratio    Prob > F
modelLR      6     3.49272143 0.5821202382 1351.741 3.28016e-61
residualLR  57     0.02454676 0.0004306449       NA          NA
           DOF Sum-of-Squares         MSE  F Ratio    Prob > F
modelQR     21    3.512388591 0.167256600 34.27674 0.001934686
residualQR   1    0.004879595 0.004879595       NA          NA
~~~~~ Training Set # 2  ~~~~~
           DOF Sum-of-Squares          MSE  F Ratio    Prob > F
modelLR      6     1.27600624 0.2126677072 860.6686 1.75883e-55
residualLR  57     0.01408447 0.0002470959       NA          NA
           DOF Sum-of-Squares         MSE  F Ratio    Prob > F
modelQR     21    1.285503764 0.061214465 13.34536 0.007775539
residualQR   1    0.004586947 0.004586947       NA          NA
```

Fig. 3. ANOVA table results for linear and quadratic regressions performed on training data before removal of outliers



```
~~~~~ Training Set # 1  ~~~~~
           DOF Sum-of-Squares         MSE  F Ratio     Prob > F
modelLR      6     3.22759284 0.537932140 1590.913 5.737034e-58
residualLR  51     0.01724453 0.000338128       NA           NA
           DOF Sum-of-Squares         MSE  F Ratio    Prob > F
modelQR     21    3.241885167 0.154375484 52.29163 0.001032156
residualQR   1    0.002952203 0.002952203       NA          NA
~~~~~ Training Set # 2  ~~~~~
           DOF Sum-of-Squares          MSE  F Ratio     Prob > F
modelLR      6    1.150462630 0.1917437717 1141.431 2.407533e-53
residualLR  50    0.008399274 0.0001679855       NA           NA
           DOF Sum-of-Squares         MSE  F Ratio   Prob > F
modelQR     21    1.156136936 0.055054140 20.20359 0.00423012
residualQR   1    0.002724969 0.002724969       NA         NA
```

Fig. 4. ANOVA table results for linear and quadratic regressions performed on training data after removal of outliers

ACKNOWLEDGMENT

The data set used in this research is sourced from the UCI Machine Learning Repository; namely, the ”Stock portfolio performance Data Set” by I-Cheng Yeh, where the data set of performances of weighted scoring stock portfolios are obtained with mixture design from the US stock market historical database [1].

REFERENCES

[1] Liu, Y. C., Yeh, I. C. Using mixture design and neural networks to build stock selection decision support systems. Neural Computing and Applications, 1-15. (Print ISSN 0941-0643, Online ISSN 1433-3058, First online: 16 November 2015, DOI 10.1007/s00521-015-2090-x)

[2] Cornell J. A., Experiments with Mixtures: Designs, Models and the Analysis of Mixture Data. 3rd edition, 2002, John Wiley & Sons, New York.

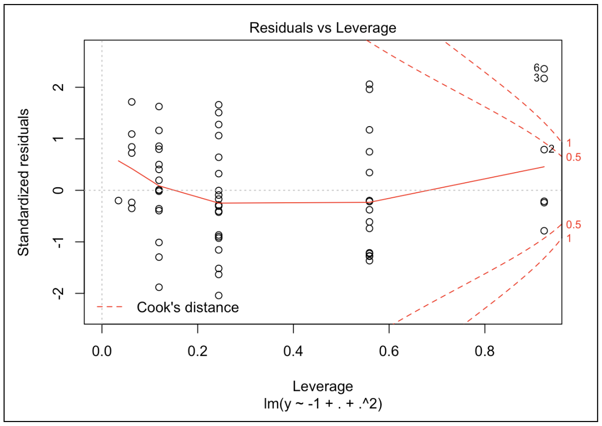


Fig. 5. Residuals vs leverage plot showing us which data points to consider as outliers via Cook’s Distance