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

Matthew Sears[1], Semere Habtemicael[2]

[1]Sophomore, Department of Applied Mathematics, Wentworth Institute of Technology, Boston, MA 02115

[2]Assistant Professor, Department of Applied Mathematics, Wentworth Institute of Technology, Boston, MA 02115

WENTWORTH INSTITUTE OF TECHNOLOGY

Regression Methods

Multiple Linear Regression

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

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

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

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

$$x_i \geq 0, \ i = 1, 2, \ldots, 6$$

Quadratic Regression

$$y = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j}^{q} \beta_{ij} x_i x_j + \epsilon \tag{2}$$

with the same parameters as Equation (1)

Data Set

Sourced from the UCI Machine Learning Repository – "Stock portfolio performance Data Set" by I-Cheng Yeh [1].

Data & Motivation

Annual returns: the percentage of the yearly gains or losses of the stock portfolio
Stock-picking concepts: methods used to pick stocks that show signs of high return rates and low risk

The data set used has 4 periods of stock market data, each of length 5 years.

Time frame	The beginning time of the 1st holding period	The beginning time of the 20th holding period
The 1st period	September 1990	June 1995
The 2nd period	September 1995	June 2000
The 3rd period	September 2000	June 2005
The 4th period	September 2005	June 2010

We use the first 2 periods as training data.

Can regression methods be used to predict long-term annual returns?

R Code Snippets

Extracting the Data

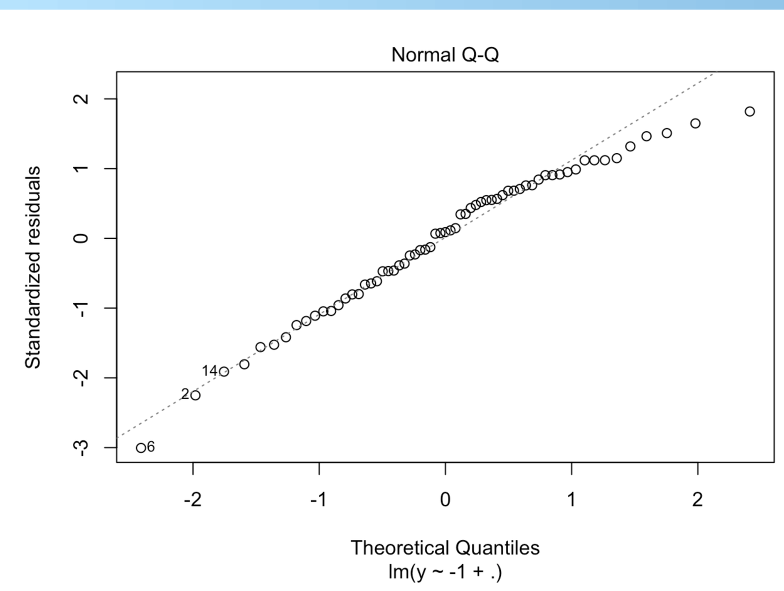
```
# Read in excel data
rawData <- data.frame(read_excel("stock portfolio performance data set.xlsx",
                      sheet = sheetList[i]))
rawData <- rawData[-1,]


# Extract x1 - x6 (weighted concepts) and y (annual return) as num
data <- data.frame(x1 = as.numeric(rawData[,c(2)]),
                   x2 = as.numeric(rawData[,c(3)]),
                   x3 = as.numeric(rawData[,c(4)]),
                   x4 = as.numeric(rawData[,c(5)]),
                   x5 = as.numeric(rawData[,c(6)]),
                   x6 = as.numeric(rawData[,c(7)]),
                   y  = as.numeric(rawData[,c(8)]))
```

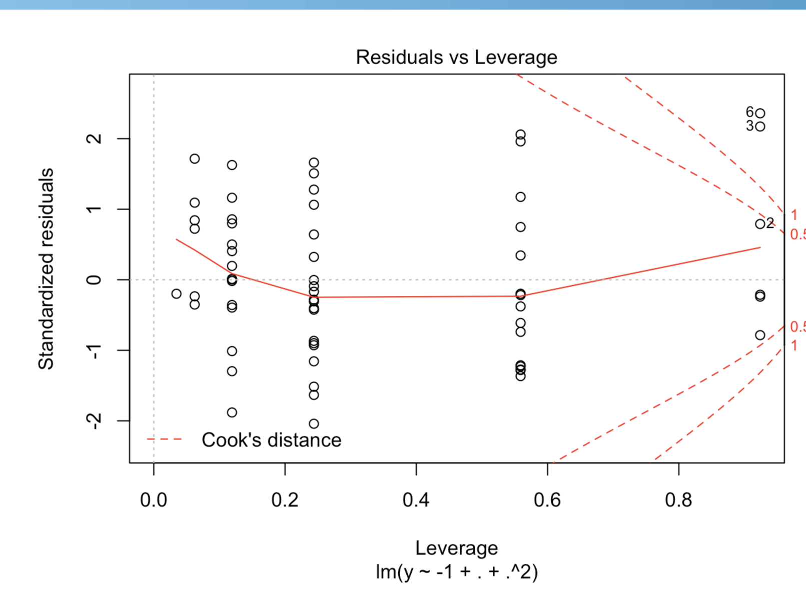
Generating the Models

```
# For training data (1, 2) remove rows containing outliers determined by Cook's Dist
# then perform regression. Only need to use predict() for one set of observations
# due to the mixture-model nature of the data
switch(i,
      {data      <- data[c(-1,-3:-6,-19),]
       linReg_1  <- lm(y ~ -1 + ., data)
       yHat_LR1  <- predict(linReg_1, sheet1[,c(1:6)])
       quadReg_1 <- lm(y ~ -1 + . + .^2, data)
       yHat_QR1  <- predict(quadReg_1, sheet1[,c(1:6)])},
      {data      <- data[c(-1:-6,-8),]
       linReg_2  <- lm(y ~ -1 + ., data)
       yHat_LR2  <- predict(linReg_2, sheet2[,c(1:6)])
       quadReg_2 <- lm(y ~ -1 + . + .^2, data)
       yHat_QR2  <- predict(quadReg_2, sheet2[,c(1:6)])})
```

Data Analysis



Plot of linear regression model supporting that our data and error are normally distributed



Plot showing which data points to consider as outliers via Cook's Distance

```
~~~~~ 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
```

Analysis of Variance (ANOVA): Table of various calculations that provides useful information about our regression models

Successful Model

$$\begin{aligned} \beta_1 &= 0.2879967 \\ \beta_2 &= 0.2832200 \\ \beta_3 &= 0.2779605 \\ \beta_4 &= 0.1754495 \\ \beta_5 &= 0.1410966 \\ \beta_6 &= 0.2517268 \end{aligned}$$

Linear Regression 1 (model trained with the 1st period data)

Results and Conclusion

- Quadratic Regression overfit the data, performed poorly on testing data
- Linear Regression 1 has the highest correlation with testing data from the 3rd period, 77%

$\Longrightarrow$ We can use a linear model as a fairly accurate model to predict long-term high annual returns

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.

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