ASSET SELECTION USING FACTOR MODEL AND DATA ENVELOPE ANALYSIS-A QUANTILE REGRESSION APPROACH

By

Abhay Kumar Singh and David E Allen

School of Accounting Finance & Economics, Edith Cowan University Australia

School of Accounting, Finance and Economics & FEMARC Working Paper Series
Edith Cowan University
October 2010
Working Paper 1005

Correspondence author:
David E. Allen
School of Accounting, Finance and Economics
Faculty of Business and Law
Edith Cowan University
Joondalup, WA 6027
Australia
Phone: +61 8 6304 5471
Fax: +61 8 6304 5271
Email: d.allen@ecu.edu.au
Abstract

With the growing number of stocks and other financial instruments in the investment market, there is always a need for profitable methods of asset selection. The Fama-French three factor model, makes the problem of asset selection easy, by narrowing down the number of parameters, but the usual technique of Ordinary Least Square (OLS), used for estimation of the coefficients of the three factors suffers from the problem of modelling using the conditional mean of the distribution, as is the case with OLS. In this paper, we use the technique of Data Envelopment Analysis (DEA) applied to the Fama-French Three Factor Model, to choose stocks from Dow Jones Industrial Index. We use a more robust technique called as Quantile Regression to estimate the coefficients for the factor model and show that the assets selected using this regression method form a higher return equally weighted portfolio.

Keywords: Asset Selection, Factor Model, DEA, Quantile Regression

JEL Codes: G11, G12, C21
1 INTRODUCTION

The main objective of asset selection for any portfolio manager is to select stocks, which can form a portfolio having higher expected return for given risk levels. The level of risk associated with the assets, stocks in our case, becomes a major deciding factor when choosing stocks from the ever increasing number of stocks in the financial market. Markowitz (1952), used stock return variance as a measure of risk and the prime deciding factor in stock selection in his pioneering portfolio theory. Jack Treynor (1961, 1962), William Sharpe (1964), John Lintner (1965) and Jan Mossin (1966) independently, proposed Capital Asset Pricing Theory, (CAPM), to quantify the relationship between beta of an asset and its corresponding return, which in potential applications, given appropriate assumptions, simplified the Markowitz portfolio theory by reducing the number of parameters required for asset selection.

The use of a single factor risk metric as in the CAPM oversimplifies a complex market. Eugene Fama and Kenneth French developed the Fama-French three factor model, which described “value” and “size” to be the most significant factors, outside of market risk, for explaining the realized returns of publicly traded stocks. The three factors, beta, SMB (for size effect), HML (for value), as proposed by Fama-French gives the projected return of a stock as a combination of these three factors. The natural approach to quantify the model is to apply OLS regression, which assumes a linear relationship across the mean of the distribution, and thus doesn’t quantify or assess the lower and upper tails of the return distribution which may play a major part when it comes to the efficient quantification of risk. A new and more robust alternative to OLS is Quantile Regression developed by Koenker and Basset (1978), which gives the capability of modelling the conditional quantiles across the distribution. Modelling the whole distribution becomes important when the return distribution becomes skewed due to adverse market conditions like the recent Global Financial Crisis, and the incapability of OLS to quantify the lower tails of the distribution can lead to wrong asset selection which could lead to greater loss.

Data Envelopment Analysis (DEA), Charnes et.al. (1978) and Banker et.al. (1984) is a powerful technique adopted from the operational research area. DEA is used for evaluating and comparing performances of organizational units in multi-attribute and multidimensional environment by determining the relative efficiency of a productive unit by considering its closeness to an efficiency frontier.

In this paper we use the Fama-French factor model coefficients as calculated from Quantile Regression as an input to DEA for asset selection process. We show by a comparative analysis of asset selected by means of OLS and assets selected by application of Quantile Regression that the assets selected by the latter give better returns when combined in a equally weighted portfolio. We also show that the assets selected by application of Quantile Regression not only give better returns in normal market conditions but also in conditions of extreme financial distress.

After the introduction in section 1, the rest of the paper is organized as follows: Section 2 provides the background with more insight into Fama-French Three Factor Model, Quantile Regression and Data
Envelopment Analysis. Section 3 discusses the research method employed in this work. Next, Section 4 presents and discusses the major results and finally Section 5 draws the conclusions.

2 BACKGROUND

2.1 THE FAMA-FRENCH THREE FACTOR MODEL

Jack Treynor (1961, 1962), William Sharpe (1964), John Lintner (1965) and Jan Mossin (1966) independently, proposed Capital Asset Pricing Theory, (CAPM), to quantify the relationship between the beta of an asset and its corresponding return. CAPM stands on the broad assumption that, that only one risk factor is common to a broad-based market portfolio, which is beta. (As derived from sweeping assumptions about common expectations, frictionless markets, etc). Modelling of the CAPM using OLS assumes that the relationship between return and beta is linear, as given in equation (1).

\[ r_A = r_f + \beta_A(r_m - r_f) + \alpha + \epsilon \]  

where

- \( r_A \) is the return of the asset
- \( r_m \) is the return of the market
- \( r_f \) is the risk free rate of return
- \( \alpha \) is the intercept of regression
- \( \epsilon \) is the standard error of regression

Fama and French (1992, 1993) extended the basic CAPM to include size and book-to-market effects as explanatory factors in explaining the cross-section of stock returns. SMB (Small minus Big) gives the size premium which is the additional return received by investors from investing in companies having a low market capitalization. HML (High minus Low), gives the value premium which is the return provided to investors for investing in companies having high book-to-market values.

SMB is a factor measuring "size risk", which comes from the view that, small companies (companies with low market capitalization), are expected to be relatively more sensitive to various risk factors, which is a result of their undiversified nature and their inability to absorb negative financial events. HML, on the other hand is a factor which proposes association of higher risk with “value” stocks (high B/M values) as compared to “growth” stocks (low B/M values). This is intuitively justified as firms or companies ought to attain a minimum size in order to enter an Initial Public Offering (IPO).

The three factor Fama-French model is written as:

\[ r_A = r_f + \beta_A(r_m - r_f) + s_A SMB + h_A HML + \alpha + \epsilon \]  

\[ (2) \]
Where \( s_A \) and \( h_A \) capture the security’s sensitivity to these two additional factors. Portfolio formation using this model requires the historical analysis of returns based on the three factors using regression measures, which quantifies estimates of the three risk variables involved in the model, i.e. \( \beta_A \), \( s_A \), \( h_A \), and the usual regression analysis using OLS gives us the estimates around the means of the distributions of the historical returns and hence doesn’t efficiently quantify the behaviour around the tails. Modelling the behaviour of factor models using quantile regression gives us the added advantage of capturing the tail values as well as efficiently analysing the median values. The coefficients obtained from lower quantiles (5% or lower) represent the lower tail risk in the return distribution of every stock, which is of interest when it comes to efficient asset selection.

2.2 QUANTILE REGRESSION

Linear regression represents the dependent variable, as a linear function of one or more independent variable, subject to a random ‘disturbance’ or ‘error’ term. It estimates the mean value of the dependent variable for given levels of the independent variables. For this type of regression, where we want to understand the central tendency in a dataset, OLS is an effective method. OLS loses its effectiveness when we try to go beyond the median value or towards the extremes of a data set.

Koenker and Bassett (1978), introduced Quantile Regression as an extension of classical ordinary least squares (OLS) estimation of conditional mean models to the estimation of an ensemble of models for conditional quantile functions for a data distribution. The central special case is the median regression estimator that minimizes a sum of absolute errors. The remaining conditional quantile functions are predicted by minimizing an asymmetrically weighted sum of absolute errors, weights being the function of quantile of interest. This makes quantile regression a robust technique even in presence of outliers. Taken together the ensemble of estimated conditional quantile functions offers a much more complete view of the effect of covariates on the location, scale and shape of the distribution of the response variable.

Quantiles refer to the generalized case of dividing a conditional distribution into parts. The technique of quantile regression extends this idea to build models which express the quantile of conditional distribution of the response variable as function of observed covariates.

Linear regression coefficient represents the change in the response variable produced by a one unit change in the predictor variable associated with that coefficient. Quantile regression coefficients gives the change in a specified quantile of the response variable produced by a one unit change in the predictor variable.

Quantiles as proposed by Koenkar and Bassett (1978) can be defined through an optimization problem. Similar to the problem of defining sample mean as the solution of the problem of minimizing the sum of squared residuals (as done in OLS regression), the median quantile (0.5) is defined through the minimization of sum of absolute residuals. The symmetrical piecewise linear absolute value function assures same
number of observations above and below the median of the distribution.

The other quantile values can be obtained by minimizing a sum of asymmetrically weighted absolute residuals, (giving different weights to positive and negative residuals). Solving

\[
\min_{\xi \in \mathbb{R}} \sum \rho_{\tau}(y_i - \xi)
\]  

(3)

Where \( \rho_{\tau}(\cdot) \) is the tilted absolute value function as shown in Figure 1, this gives the \( \tau \)th sample quantile with its solution. Taking the directional derivatives of the objective function with respect to \( \xi \) (from left to right) shows that this problem yields the sample quantile as its solution.

![Figure 1: Quantile Regression \( \rho \) Function](image)

After defining the unconditional quantiles as an optimization problem, it is easy to define conditional quantiles similarly. Taking least squares regression model as a base to proceed, for a random sample \( y_1, y_2, \ldots, y_n \), we solve

\[
\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2
\]  

(4)

which gives the sample mean, an estimate of the unconditional population mean, \( EY \). Replacing the scalar \( \mu \) by a parametric function \( \mu(x, \beta) \) and then solving

\[
\min_{\mu \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mu(x_i, \beta))^2
\]  

(5)

gives an estimate of the conditional expectation function \( E(Y|x) \).

Proceeding the same way for quantile regression, To obtain an estimate of the conditional median function, the scalar \( \xi \) in the first equation is replaced by the parametric function \( \xi(x_i, \beta) \) and \( \tau \) is set to
The estimates of the other conditional quantile functions are obtained by replacing absolute values by $\rho_\tau(\cdot)$ and solving

$$\min_{\mu \in \mathbb{R}} \sum \rho_\tau(y_i - \xi(x_i, \beta))$$ \hspace{1cm} (6)

The resulting minimization problem, when $\xi(x, \beta)$ is formulated as a linear function of parameters, can be solved very efficiently by linear programming methods. Further insight into this robust regression technique can be obtained from Koenkar and Bassett Quantile Regression monograph.

Quantile regression coefficients have the advantages that by including them in the analysis we can combine them by certain weighting schemes to yield more robust measurements of effect of the factors across the quantiles, in contrast to OLS estimates around the mean. Chan and Lakonishok (1992), originally proposed this approach in their work, which proved by means of simulations the applicability of quantile regression in beta estimation. We will use a symmetric weighting scheme to combine the coefficients obtained from each of the quantile levels (5%, 25%, 50%, 75% and 95%), to get a single coefficients for each of the three factors, which will be set as a input in the DEA model. The resulting estimators have weights which are the linear combination of quantile regression coefficients.

$$\beta_t = 0.05\beta_{(0.05,t)} + 0.2\beta_{(0.25,t)} + 0.5\beta_{(0.5,t)} + 0.2\beta_{(0.75,t)} + 0.05\beta_{(0.95,t)}$$ \hspace{1cm} (7)

$$s_t = 0.05s_{(0.05,t)} + 0.2s_{(0.25,t)} + 0.5s_{(0.5,t)} + 0.2s_{(0.75,t)} + 0.05s_{(0.95,t)}$$ \hspace{1cm} (8)

$$h_t = 0.05h_{(0.05,t)} + 0.2h_{(0.25,t)} + 0.5h_{(0.5,t)} + 0.2h_{(0.75,t)} + 0.05h_{(0.95,t)}$$ \hspace{1cm} (9)

### 2.3 DATA ENVELOPMENT ANALYSIS (DEA)

Data Envelope Analysis was originally introduced by Charnes et al. (1978) as a non-parametric linear programming approach, capable of handling multiple inputs as well as multiple outputs, Charnes et al. (1994) and Cooper et al. (2000). DEA measures the efficiency of the decision making unit (DMU) by the comparison with best producer in the sample to derive compared efficiency.

DEA submits subjective measures of operational efficiency to the number of homogenous entities compared with each other, through a number of sample’s units which form together a performance frontier curve that envelopes all observations, and hence this approach is called Data Envelopment Analysis. Consequently, DMUs which lie on the curve are efficient in distributing their inputs and producing their outputs, while DMUs which do not lie on the curve are considered inefficient.

Performance measurement using DEA method consists of determining the relative efficiency of a pro-
ductive unit by considering its closeness to an efficiency frontier. DEA efficiency is not the same as mean-variance efficiency in the Markowitz model where mean and variance are the only two parameters for optimization. In the DEA approach, efficiency is the objective function value of a multi-criteria linear programming model. The objective of the DEA is to determine relative performance indicators among productive units, considering specific groups of inputs and outputs. It is a multi-factor productivity analysis model for measuring the relative efficiencies of a homogenous set of decision-making units (DMUs). The efficiency score in the presence of multiple input and output factors is defined as:

\[
Efficiency = \frac{\text{Weighted Sum of Outputs}}{\text{Weighted Sum of Inputs}}
\]  

(10)

In the current analysis, the Fama-French three factor coefficients for the securities are used as an input to the DEA application and expected returns are the output. Assuming that there are \( n \) DMUs, each with \( m \) inputs and \( s \) outputs, the relative efficiency score of a test DMU \( p \) is obtained by solving the following model proposed by Charnes et al. (1978):

\[
\begin{align*}
\max & \quad \sum_{k=1}^{s} \frac{\vartheta_k y_{kp}}{\sum_{j=1}^{m} u_j x_{jp}} \\
\text{st} & \quad \sum_{k=1}^{s} \frac{\vartheta_k y_{kp}}{\sum_{j=1}^{m} u_j x_{jp}} \leq \forall i \\
& \quad u_k, u_j \geq 0 \forall k, j
\end{align*}
\]  

(11)

Where,
\( k=1 \) to \( s, j=1 \) to \( m, i=1 \) to \( n \),
\( y_{ki} \) = amount of output \( k \) produced by DMU \( i \),
\( x_{ji} \) = amount of input \( j \) utilized by DMU \( i \),
\( \vartheta_k \) = weight given to output \( k \),
\( u_j \) = weight given to input \( j \)

The fractional program shown in the above equation can be converted to a linear program as shown in equation below.

\[
\begin{align*}
\max & \quad \sum_{k=1}^{s} \vartheta_k y_{kp} \\
\text{st} & \quad \sum_{j=1}^{m} u_j x_{jp} = 1
\end{align*}
\]  

(12)
\[ \sum_{k=1}^{s} \vartheta_k - \sum_{k=1}^{s} \vartheta_k y_{ki} \leq 0 \forall i \]
\[ u_k, u_j \geq 0 \forall k, j \]

To identify the relative efficiency scores of all the DMUs the above problem is run n times. Each DMU selects input and output weights that maximize its efficiency score. DEA, has demonstrated some very satisfactory applications in finance; such as, in the evaluation (ex-post) of investment funds (e.g., Morey and Morey 1999, Gregorious 2003, Haslem and Scheraga 2003), although its initial applications had been predominantly to public organizations (e.g., Shen et al. 2005, Zhu 2003, Avkiran 2001, Calhoun 2003). Chen (2008), used DEA for stock selection to form portfolios and compare them against the benchmark market portfolio, by using the size effect in the model. Lopes, Ana, Edgar Lanzer, Marcus Lima, and Newton da Costa, Jr., (2008) applied DEA on the Brazilian stock market applying risk measures like variance and beta as inputs and quarterly returns as the output. Singh, Sahu, Bharadwaj (2009), used DEA to form efficient portfolio in their comparative analysis. With all these wide applications, DEA has never been tested in contrasting market situations, for example in the period before the financial crisis and during the financial crisis (2007-2008).

3 DATA AND METHODOLOGY

The study uses daily prices of the 30 Dow Jones Industrial Average Stocks, for a period from January 2005 to December 2008, along with the Fama-French factors for the same period, obtained from French’s website to calculate the Fama-French coefficients. Table 1, gives the 30 stocks traded in Dow Jones Industrial Average and used in this study.

Table-1: Dow Jones Industrial 30 Stocks used in the study.

<table>
<thead>
<tr>
<th>3M</th>
<th>EI DU PONT DE NEMOURS</th>
<th>KRAFT FOODS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALCOA</td>
<td>EXXON MOBILE</td>
<td>MCDONALDS</td>
</tr>
<tr>
<td>AMERICAN EXPRESS</td>
<td>GENERAL ELECTRIC</td>
<td>MERCK &amp; CO.</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>GENERAL MOTORS</td>
<td>MICROSOFT</td>
</tr>
<tr>
<td>BANK OF AMERICA</td>
<td>HEWLETT-PACKARD</td>
<td>PFIZER</td>
</tr>
<tr>
<td>BOEING</td>
<td>HOME DEPOT</td>
<td>PROCTER&amp;GAMBLE</td>
</tr>
<tr>
<td>CATERPILLAR</td>
<td>INTEL</td>
<td>UNITED TECHNOLOGIES</td>
</tr>
<tr>
<td>CHEVRON</td>
<td>IBM</td>
<td>VERIZON COMMUNICATIONS</td>
</tr>
<tr>
<td>CITIGROUP</td>
<td>JOHNSON &amp; JOHNSON</td>
<td>WAL MART STORES</td>
</tr>
<tr>
<td>COCA COLA</td>
<td>JP MORGAN CHASE &amp; CO.</td>
<td>WALT DISNEY</td>
</tr>
</tbody>
</table>

The methodology adopted in this research can be summarized in the following steps:
1. Calculate Fama-French Coefficients using Quantile Regression (5%, 25%, 50%, 75% and 95% quantiles) and OLS, using daily log returns for each stock along with the daily factors.

2. Use the symmetric weighting scheme (equation 7, equation 8 and equation 9), to combine the coefficients calculated from quantile regression to get a single measure for each of the three coefficients.

3. Standardize the coefficients and the annualized returns to make them positive to be used as deciding input and output parameters in the DEA model, using the following relation

\[ RZ_{ij} = \text{Abs}(\text{min}(Z_j)) + Z_i \]  

where

\[ RZ_{ij} \] is the rescaling for each j attributes for stock i.

\( \text{min}(Z_j) \) is the minimum of all the values of jth attribute for all stocks.

\( Z_i \) is the value of attribute j for stock i.

Finally all the attributes for all the stocks are divided by respective maximums, as shown in equation-14.

\[ MRZ_{ij} = \frac{RZ_{ij}}{\text{Column Maximum}} \]  

Where, \( MRZ_{ij} \) is the final rescaled value for each attribute for a stock.

4. We drop the stock having the minimum rescaled annualized return, which was zero in every case. The asset selection using the three factors for the rest of the 29 stocks is done for both OLS and quantile regression measures.

5. The coefficients are chosen as input to the variable return to scale input oriented DEA model and the annualized returns as the output, which minimizes the inputs (three risk factors) for a given level of output (return). Stocks having an efficiency score of 100% are chosen to form an equally weighted portfolio.

6. The resulting portfolios obtained for both the regression techniques are evaluated based on their return after a hold out period of one year.

7. All the steps are repeated for three year data period (2005-2007), resulting in three different asset sets selected during different time periods representing different market conditions, which included the period of Global Financial Crisis.
4 DISCUSSION OF RESULTS

Table 2 gives the stocks selected using the factor coefficients calculated from OLS, it also gives the return which the set of selected stocks gives when combined into an equally weighted portfolio after a hold out period of one year. Table 3 gives the stocks selected using factor coefficients calculated for different quantiles using quantile regression along with the hold out period equally weighted portfolio return. As evident from the results in these tables, the stocks selected by using quantile regression estimates gives better hold out period returns than the ones selected using OLS. The method not only gives better results in normal market conditions but also gives better performance at the time of the recent financial crisis (years 2007-2008).

Table-2: Assets Selected from DEA using OLS estimates

<table>
<thead>
<tr>
<th>YEAR</th>
<th>STOCKS</th>
<th>ONE YEAR HOLD OUT RETURN</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>Boeing, HP, INTEL, Johnson &amp; Johnson, Kraft Foods, Pfizer, Procter &amp; Gamble</td>
<td>12.03333</td>
</tr>
<tr>
<td>2006</td>
<td>AT&amp;T, General Motors, HP, IBM, Johnson &amp; Johnson, Kraft Foods, Merk &amp; Co., Walt Disney</td>
<td>4.181735</td>
</tr>
<tr>
<td>2007</td>
<td>Boeing, Chevron, Coca Cola, HP, INTEL, IBM, Johnson &amp; Johnson, Mac Donalds, Merk &amp; Co., Procter &amp; Gamble</td>
<td>-43.0281</td>
</tr>
</tbody>
</table>

Table-3: Asset Selected from DEA using Quantile Regression estimates.

<table>
<thead>
<tr>
<th>YEAR</th>
<th>STOCKS</th>
<th>ONE YEAR HOLD OUT PERIOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>AT&amp;T, General Motors, HP, IBM, Johnsons &amp; Johnsons, Kraft Foods, Merk &amp; Co., Microsoft</td>
<td>5.250369</td>
</tr>
<tr>
<td>2007</td>
<td>AT&amp;T, Chevron, Coca Cola, HP, INTEL, IBM, Johnsons &amp; Johnsons, Mac Donalds, Microsoft, Procter &amp; Gamble</td>
<td>-39.1764</td>
</tr>
</tbody>
</table>

The stocks selected based on applications of quantile regression estimates using DEA perform better, as the quantile measures provide an efficient quantification of the tail risk involved with the return distribution, which is not possible with OLS. Figure 2 gives the distribution of fitted values obtained for one of the sample stocks from the factor model using OLS and 5% and 95% quantile regression, the fitted values show that OLS is incapable of capturing the tail values and hence less efficient for the quantification of the overall risk.
5 CONCLUSION

The Fama-French Three factor model models the dependence of asset returns on three factors, the market return, size factor and value factor, but the traditional technique used for the estimation of model, i.e. OLS is incapable of describing the extremes of the distribution. In this research work we used a more robust technique of Quantile Regression which efficiently describes the required quantiles and hence gives better estimates of the risk described by the factors. The research provides a new dimension to the use of factors in DEA analysis for selecting efficient assets using Quantile Regression. The empirical results not only show that the assets selected by DEA perform better when quantile estimates are used, but they also show that the quantile based estimates also prevent the investor from some degree of loss to an extent at times of unpredictable financial distress situations. Quantile Regressions are now widely used in quantitative finance and also in specialized financial fields like econometrics. Further work can be done on applications of this robust tool with other optimization algorithms to test the capabilities and applicability of the technique.

Figure 2: Fitted Values obtained from OLS and Quantile Regression (5% and 95%).
6 REFERENCES


Barnes, Michelle and Hughes, Anthony (Tony) W., A Quantile Regression Analysis of the Cross Section of Stock Market Returns (November 2002). Available at SSRN: http://ssrn.com/abstract=458522


Sharpe, W., 1963, A simplified model for portfolio analysis, Management Science 9, 277-293.
