The Standard and Poor's Europe 350: A closer look

Wilco Lablans
9839224

12-04-2008
Master Thesis

First supervisor: Dr. P. Sengmüller
Second Supervisor:

University of Amsterdam
Faculty Economics and Business Economics
Business Economics Finance

Contents

Word of thanks

First, I would like to thank Dr. Paul Sengmüller for his useful comments.
Next I would like to thank my father for providing programming tutorials which were essential for data preparation. For filtering the dataset on errors and for preparing them for the three factor analysis, the data had to be transformed and sorted in the appropriate format. SPSS and Excel appeared too limited to do this job. Therefore programming language VBA was used. At last, I would like to thank Susan Fagg from Standard and Poor's for providing data for constructing the portfolio.

1. Introduction

A few databases exist researchers can use for empirical research. The Center for Research in Securities Prices is a commonly used database for researchers in the United States. Thomson DataStream is used by many European researchers for their empirical analysis. A database containing accurate information is important for analysis. Unfortunately, several studies have shown these databases are not without errors. Conducting analysis with data which may contain errors can lead to different outcomes, depending on the degree and magnitude of the errors. In this thesis an overview of literature will be provided included advises for screening datasets.

Indices are portfolio's consisting of multiple stocks. Indices are used for measuring markets. Standard and Poor's has introduced many indices, including an index which measures the overall European market factor. It is interesting to examine the compilation and returns of indices. In this thesis, an European index will be examined.

The Capital Asset Pricing Model (CAPM) has dominated financial analysis for a long time. The model explains the returns of stocks in relation to the return of an overall market factor. Because of discrepancies a better model was introduced by Fama and French: the 3 factor model. The three factor model appeared better in explaining stock returns, although some researchers remained unconvinced. It is also interesting from U.S. perspective to test results obtained from analysis performed in the United States market on non U.S. equity markets. The three factor model needs testing and the companies included in the Standard and Poor's Europe 350 provide an excellent sample which can be used for testing the model. The model will thus be tested with data for companies included in the index (Ince, Porter, 2006).

The thesis is organized as follows: section 2 describes the importance of accurate databases and illustrates with an example. Next, the need for testing the three factor model is explained. Section 3 describes the most commonly used database in the United States: The Center for Research in Securities Prices. Next, Thomson DataStream will be described. Section 4 provides information of the Standard and Poor's 350 Europe Index. Section 5 provides an overview of literature concerning the accuracy of databases, describes methods how errors can be detected and provides advises for handling datasets. Also, return data for all companies included in the index are screened for possible errors. Returns of an exchange traded fund which mimics the index are compared to returns calculated with data from DataStream. Section 6 explains the CAPM model and the three factor model. Section 7 describes applying the three factor model to the data of the companies included in the Standard and Poor's 350 Europe Index.

2. The importance of accurate databases and the three factor model

In section 2.1, the importance of accurate databases will be described and illustrated with an example. Section 2.2 will describe testing the three factor model.

2.1 The importance of an accurate database

According to Ince and Porter, international asset pricing is an important topic in financing literature. In order to conduct such research, the availability of high-quality equity return data is considered a necessary condition. Many studies use Thomson DataStream as a source for worldwide equity return data. Thomson is the only independent source for international equity data in Europe and Thomson DataStream covers many markets and many securities in every market. Ince and Porter examine the suitability of Thomson DataStream as an academic research database. The data covered by Thomson DataStream is evaluated by filtering the data from Thomson and comparing the results with data gathered from the Center for Research in Securities Prices (CRSP).

Ince and Porter compare market wide average returns calculated from the screened and raw data. One conclusion is that most problems apply to the smaller size deciles in the data. Next, sample statistics for size deciles portfolios and profits from momentum strategies are conducted to illustrate the consequences of the wrong data. Ince and Porter downloaded daily data for the period January 1, 1975 till 31 December 2002. Daily data is gathered for securities traded on U.S. exchanges and securities which were no longer traded. Monthly returns are calculated from end-of-month daily data. The same data for the same period is gathered from CRSP. The purpose of Ince and Porter is to illustrate the differences between CRSP and Thomson by calculating equally weighted returns by exchange, equally weighted market returns, the returns of two momentum strategies trading strategies and size decile returns. The Thomson dataset contained all equities traded on NASDAQ-NMS, NASDAQ-nonNMS, NYSE or AMEX. The CRSP dataset contained all common equity of U.S.-based companies traded on the NASDAQ, NYSE or AMEX. First, without screening the data, average monthly returns and standard deviations of both datasets are calculated. The statistics were calculated for an equally weighted average market return, value weighted market return, for each exchange, and for each decile formed from the common stock (Ince, Porter, 2006).

It appeared that the equally weighted returns calculated from the Thomson database were much higher than the returns calculated with the data obtained from CRSP. Thomson DataStream (TDS) reported an equally weighted market return of 2.40% and standard deviation of 7.53% per month, while CRSP reported an equally weighted market return of 1.41% and standard deviation of 5.69% per month. The correlation between the two market return series was 0.66. The value weighted market returns appeared much closer: from TDS data, the calculated value weighted market return was 1.14%, and the standard deviation was 4.40% per month. While from CRSP data the calculated value weighted market return was 1.13%, and the standard deviation was 4.57% per month. The correlation between returns was 0.998. This difference showed that the errors must be in smaller issues. Ince and Porter also formed size deciles and compared returns and standard deviations calculated for every decile. The largest differences were in the smaller deciles. Also, two momentum strategies were used to illustrate the effect of the different values in the databases: A 1090 strategy – this is a strategy which sorts all stocks by their average returns over a 11-month period t-2 through t-12 and forms an equally weighted portfolio which hold long the top 10% and holds short the bottom 10%, and holds this for one month before rebalancing – and a 3070 strategy which works congruent, but then for the top 30% and the bottom 30%. The result from the 1090 strategy was for TDS data an average return of 0.26% per month, and with CRSP data it resulted in a average return of 1.13% per month. The result from the 3070 strategy was for TDS data an average return of 1.02% per month and calculated from CRSP data it resulted in an average return of 0,95% per month.

Ince and Porter also isolated the effect of coverage on reported differences by matching the databases security by security. The matching in coverage rate rose from about approximately 20% in 1975 to almost 90% in 2002. The authors use matching codes for matching securities and also use key words or phrases to find corresponding securities. The results appeared similar to the sample with raw data. This implies that large differences in size decile returns, market returns and momentum returns are not only caused by the inclusion of other securities than common equity by Thomson.

Dividend information for both sources was compared and 93% had identical observations while 7% had positive dividend payments on CRSP and zero dividend information on Thomson. Market wide dividend yields were calculated for both databases, and the correlation coefficient was 0.982. Ince and Porter also found errors when they recalculated RI. RI is calculated using price and dividend data and represents return with dividends reinvested. Ince and Porter compared the results with the percentage change in the return index. When different results were found, Ince and Porter used the return they calculated using the price and dividend data. All observations in both databases with end-of-previous-month price less than $ 1.00 were dropped, because Thomson rounds prices to the nearest penny, and this practice can cause nontrivial differences in the calculated returns when prices are small. Many instances of data errors were found.

Some type of errors were screened according this method: when R_t or R_t-1 is higher then 300% and $(1+R_t)(1+ R_{t-1})$ is smaller than 50%, R_t and R_t-1 are set to missing values. This is the way abnormally high values are screened by Ince and Porter. Ince and Porter chose the value 300% arbitrary. Several values were used and records which failed the test were examined, and the 300% level appeared to work good. The result of the screened and corrected dataset were notable: for an equally weighted market return of all stocks the average return and standard deviation of the screened and corrected dataset of Thomson declined from 2.40% to 1.51% and from 7.53 % to 5.16% respectively. The average return and standard deviation calculated from data from CRSP was 1.29% and 5.46%. Correlation between the returns rose from 0.66 to 1.00. For a calculated value-weighted return the results appeared less eminent: average return and standard deviation of raw and screened and corrected data from Thomson were 1.14% and 1.13% for average return, and 4,40% and 4.47% for standard deviation. Average return and standard deviation for data from CRSP was 1.16% and 4.49% respectively. The correlation between the datasets was 1.00.

Individual equity data for Greece, Ireland, Germany and the United kingdom was also tested. The authors do this to show that the problems described are not unique to data for the United States. For the period January 1975 till October 2003 returns for value-weighted and equally weighted market portfolios for each country are calculated. Next, correlations for the data and total market indexes for each country are calculated. The correlations range from 0.80 for both Greece and Ireland and 0.27 for Germany and 0.21 for the united Kingdom. Also for this four markets the returns calculated from the raw and screened data differed substantially. The correlations between returns of raw and screened data appeared often low, and the screened data corresponded more closely to market indexes.

These implications of the investigation of Ince and Porter illustrate the need for an accurate database. The results for studies using this database for input can change by using more accurate databases. In section 5.6, advices are provided for using correct datasets. In section 5.4 the dataset for all companies included in the Standard & Poor's Europe 350 index is used for screening the data.

2.2 The three factor model needs testing

Next, the three factor model introduced by Fama and French is tested for robustness. A significant relation between firm size, book-to-market ratios and returns on securities for non-financial firms has been documented by Fama and French. Regressions were performed to see if relationships existed between variables and the cross-section of stock returns. Size and book-to-market ratio seemed the strongest predictors in their analysis. Also even the performance of managers can be tested using the premiums provided by size and book-to-market ratios. Because some researchers remained unconvinced about the robustness of this relationship, other researchers used different countries (Chan, Hamao and Lakonishok 1991) or different time periods (Davis 1994) or just a hold out sample to test the model for robustness. This was done to prove the robustness of the relationship and to see if the relationship would hold and was not the result of data-snooping. Fama and French also excluded financial firms from their analysis. This was because of the high leverage of these firms and this high leverage should have another meaning for these firms than other firms (Davis, 1994).

Barber and Lyon also analyse the relationship for a hold-out sample. Barber and Lyon tested whether this hold-out sample could alter conclusions and wanted to test the robustness of the findings of Fama and French. The result was that both financial and non-financial firms exhibited significant premiums. For example, Barber and Lyon test the three-factor model by dividing the period for analysis in two time periods. Barber and Lyon wanted to test if the premiums lasted for both time-periods. The factor model was tested over different classes of industries. The goal was again to test robustness of the model. The model will also be tested in this thesis for the sample of firms included in the Standard and Poor's Europe 350 during the period 01-01-1988 till 01-07-2006 (Barber, Lyon, 1997).

3. Databases

Section 3.1 will provide a description of About the Center for Research in Securities Prices. Section 3.2 will provide a description of Thomson DataStream.

3.1 The Center for Research in Securities Prices

According to the website of the Center for Research in Securities Prices - abbreviated as CRSP – is located in the Chicago financial district. The Center for Research in Securities Prices is also part of the University of Chicago’s Graduate School of Business. Leaders of the industry often partnered with CRSP to explore new ideas. In 1959 the initiative was taken by Louis Engel and Professor James H. Lorie to gather and clean prices and dividends of all stocks traded on the NYSE since 1926. The Center for Research in Securities Prices was established in 1960 and the purpose would be gathering and analysing information about all stocks listed on the NYSE since 1926. Since then, the database had been expanded. For researchers, it would now be easy to conduct analyses with an accurate and complete database. For researchers, a clean dataset for conducting analysis on stocks had been created. The Center for Research in Securities Prices provides an accurate database for financial analyses and stock market research to academic institutions and investment banks.

The Center for Research in Securities Prices was established for accurately measuring returns of common stocks listed on the NYSE. The data has since been extensively used for academical research. The coverage of the large database includes:

For US stocks, CRSP provides monthly and annual updates. Updates contain return, price and volume data on AMEX, NYSE and NASDAQ common stocks with basic market indices. The monthly data starts December 1925, daily data starts July 1962 for NYSE. For the AMEX both monthly and daily data start July 1962 and for NASDAQ both monthly and daily data start per December 1972. The Center for Research in Securities Prices covers about 26.500 stocks. For further illustration, the CRSP provides six databases:

- CRSP US Stock Database
- CRSP US Indices Database
- CRSP US Treasuries Database
- CRSP US Survivor-Bias-Free Mutual Funds Database
- CRSP/COMPUSTAT Merged Database

- CRSP/Ziman Real Estate Data Series

All data contained in the database is tested for its accuracy. The Center for Research in Securities Prices also uses multiple databases to check values and improving the quality of the database. Databases are updated monthly. About eighty percent of all academic research in stock analysis uses the database for its research.

Although CRSP is believed to contain all accurate information, some researchers believe CRSP has shortcomings. For example, Shumway finds that CRSP data is missing thousands of delisting returns. (Shumway, 1997). Shumway and Warther next investigate the delisting bias in CSRP's NASDAQ data, and find that missing returns for delisted stocks are large. A significant bias existed in the data, because most missing delisting returns are associated with negative events (Shumway, Wharter, 1997). Another study performed by Chakrabarty and Trcinka, found that discrepancies existed between returns on momentum portfolio's constructed with data from CRSP and data collected from Trade and Quote databases. This also provides evidence databases can contain errors and handle and record information differently (Chakrabarty, Trcinka, 2006).

3.2 Thomson DataStream

Thomson provides a large database for researchers and financial institutions. Researchers use financial data for a lot of empirical research. The main database which provides information on the European market is Thomson DataStream. DataStream is often viewed as an invaluable research tool for finance departments and university economics (Summers, P.M., 2003). According to Ince and Porter, who state that it is interesting from U.S. perspective to test results obtained from analysis performed in the United States market on non U.S. equity markets. They compared data from TDS for U.S. equities with data from CRSP. An investigation for European stocks thus is not performed. The largest dataset is held by Thomson DataStream. DataStream covers most the largest number of markets and the highest ratio of total securities in each market. However, when the databases were compared, differences in the data were found. This provides evidence databases are not 100% accurate. When Ince and Porter screened non-U.S. data obtained from DataStream it was shown DataStream was not 100% accurate (Ince, Porter, 2006).

The database from Thomson, DataStream has time series for financial instruments, indicators and securities for over 175 countries, 60 markets worldwide. For some series the data goes back 50 years. For equities, DataStream provides data from over 100 equity exchanges from 70 countries. Thomson collects the data direct from stock exchanges and from local market data organizations. Daily price history is kept for over 40 years of history prices. For equity indexes, Thomson offers about 100.000 indices. The data is collected since for example 1950 for the Nikkei225 and since 1978 for the FTSE 100. A rough overview of the coverage of DataStream, according to the website:

- Indices: DataStream covers over 100.000 indices, with market indices for 65 countries. Indices from for example FTSE, S&P, MSCI and Dow Jones Stoxx are included.
- Equities and Fundamentals: DataStream provides data of 75.000 active equities which cover over 64 developed and emerging markets. The data is provided to Thomson by international and local suppliers, exchanges and published reports.
- Commodities: DataStream provides over 3000 spot prices
- Bonds: DataStream covers over 80.000 fixed income instruments and over 60.000 warrants.
- International macroeconomics: DataStream covers over 400.000 time series including indicators for about 175 countries.
- Interest and exchange rates: DataStream covers Forec rates for over 45 currencies. More than 2500 interest rates are available of over 50 markets.

For equities, DataStream provides data from over 100 equity exchanges from 70 countries. DataStream collects the data direct from stock exchanges and from local market data organizations. Daily price history is kept for over 40 years of history prices. For equity indexes, Thomson offers about 100.000 indices. The data is collected since for example 1950 for the Nikkei225 and since 1978 for the FTSE 100.

Ulbricht and Weiner also address the question whether the choice of the financial databases can have impact on research results. COMPUSTAT North America by Standard & Poor's and Worldscope by Thomson Financial are compared and differences in coverage in the databases were found.

4. The Standard and Poor's Europe 350 Index

Section 4.1 begins with an introduction. Section 4.2 describes the eligibility criteria. Section 4.3 will describe additions an deletions of companies in the index. Section 4.4 will describe index construction. Section 4.5 explains the currency use for companies. Section 4.6 describes the investable weight factor. Section 4.7 describes index data. Section 4.8 describes index precision, and finally section 4.9 explains index mathematics.

4.1 Introduction

The S&P Europe 350 is developed to create an ideal benchmark for Europe. The S&P Europe 350 consists of shares of the 350 largest European companies and covers approximately 70 % of the region's market cap and spans 17 exchanges. The S&P Europe 350 is a benchmark for investors who like to invest in the European region. The S&P Europe 350 is supported by the S&P Index Committee and a team of economists and index specialists to ensure the index remains an accurate measure which reflects the risk and return characteristics of the European companies. The S&P Europe 350 is also one of the seven indices which make up the S&P Global 1200. Standard and Poor's constructs 7 indexes which span different regions. The seven indices are the S&P 500, S&P/Topix 150, S&P/TSX 60, S&P/ASX 50, S&P Asia 50, S&P Latin America 40 and the S&P Europe 350. All seven indices span different global regions and together make up the S&P Global 1200. The S&P 350 can be split up in S&P Euro, S&P Euro plus and S&P U.K. The indices are all market capitalization-weighted indices and real-time. The S&P Europe 350 index is drawn from 17 European markets, a specification of the 17 European markets is listed in table 1. The size of each market included in the S&P Europe 350 corresponds to its relative size in the index, based on its float-adjusted market value. The S&P Europe 350 is also constructed to provide investors an exposure to the most liquid and largest stocks from 17 major European markets. The weight of each country in the index reflects also the weight of the particular sector in the universe of stocks from the 17 major countries. By investing in the S&P Europe 350, an investor can invest in an index which reflects Europe.

Country Weights and number of Companies included in S&P Europe 350: (Table 1)
(base date June 30, 2006):

Table 1 Country Weights and number of Companies included in S&P Europe 350

Country	*Country Weights*	*Number of Companies*
Great Britain	36.73%	125
France	14.90%	46
Switzerland	10.55%	22
Germany	10.28%	32
Spain	5.96%	18
Italy	5.53%	26
Netherlands	5.12%	18
Sweden	3.16%	20
Belgium	1.78%	10
Finland	1.67%	5
Ireland	1.11%	6
Norway	0.90%	6
Denmark	0.61%	4
Portugal	0.58%	6
Greece	0.43%	3
Luxembourg	0.42%	1
Austria	0.28%	2

Source: Standard an Poor's, 2007

For illustration, some characteristics of the S&P Europe 350 are displayed below: (base date June 30, 2006)

Table 2 Index performance:

Returns	
1 Month	0.87%
3 Month	-2.47%
YTD	5.37%
Returns (% pa)	
1 Year	19.10%
3 Year	19.32%
5 Year	1.88%
7 Year	2.77%
Risk (% pa)	
3 Years Std Dev	8.61%
5 Years Std Dev	16.48%
Sharpe Ratio	
3 Year	0.5237
5 Year	0.0185

Source: Standard an Poor's, 2007

4.2 Eligibility criteria

All 350 stocks included in the S&P Europe 350 are selected based on their liquidity, size, country representation, and sector representation. The aim of Standard & Poor's is to design tradable indices that are highly liquid. The total market capitalization must approximate the market segment it wants to replicate while keeping the number of stocks at a minimum to create an easily replicable and highly cost-effective trading instrument. All eligibility factors are considered important and are taken into account by constructing the index. The eligibility factors for the stocks included in the index are as follows:

- *Market Capitalization*. Stocks will be included in the S&P Europe 350 when the stocks are among the largest stocks in the markets in terms of market capitalization. The weight of a stock in the index is based on the free-float-adjusted market capital of the stock. Strategic holdings are excluded from the float-adjusted market capital. Strategic holdings include: holdings of shares by private, corporate and governments and shares owned by trusts, venture capitalists, founders and directors of the company. After these holdings of stock are excluded from the float-adjusted market capital, a free-float-adjusted capital remains.

- *Liquidity*. Constituents included in the index are ranked on liquidity. Liquidity is measured by dollar value traded. The higher the 12-month value traded or float turnover of the stock is, the more likely a stock will be included in the index.

- *Domicile*. The Domicile of a stock is determined on the basis of a lot of criteria. Some criteria are listing of stock, place of operations, residence of senior officers, headquarters of the company and registration.

- *Eligible Securities*. Eligible for conclusion in the index are all common and preferred shares. Bonds, warrants, convertible stock, rights and preferred stock which provide a guaranteed fixed return are not eligible to be included in the index.

- *Sector Classification*. The index provides geographic and economic balance across 10 GICS sectors. GICS stands for Global Industry Classification Standard. The 10 GICS sectors are: consumer discretionary, Energy, Consumer Staples, Health care, Financials, Information technology, Industrials, Telecommunication services, Materials and utilities. A sector breakdown of the S&P Europe 350 is listed below.

Table 3 Sector Breakdown:

Sector Breakdown	
Financials	31.10%
Energy	11.20%
Health Care	9.40%
Cons Staples	9.00%
Cons Disc	9.00%
Materials	7.30%
Industrials	7.20%
Utilities	6.30%
Telecom Svc	6.10%
Info Tech	3.40%

Source: Standard an Poor's, 2007

4.3 Additions an deletions of Companies

Additions of companies in the index are made according to market size with liquidity, regional, country and sector representation preserved in the index. For initial public offerings holds that the company has to prove liquidity for at least six months before the company can be added to the index. When extraordinary large global offerings occur where the expected trading volume justifies inclusion, an exception can be made.

A company can be deleted from the index due to acquisitions, mergers and spin-offs or due to suspension or bankruptcies. When a company is deleted from the index, an other company is included for replacement.

4.4 Index construction

The level of the index reflects total market value of the companies included in the index relative to a particular base period. Total market value is calculated as the product of price of the stock times the number of shares available after float adjustment. How the market value is float adjusted is explained in section 4.6. The result of this calculation is represented by an indexed number to make the value easier to work with.

The value of the index is the quotient of total available market capitalization of all companies included in the index and the divisor of the index. The divisor is adjusted for changes in the constituents' share capital after the base date to assure continuity in index values. This includes right issues, share buybacks and issuances, additions and deletions of companies to the index, spin-offs, and adjustments in availability.

4.5 Currency of calculation

Before introduction of euro, all prices were converted to the European Currency Unit (ECU). From 01-01-1999 on, the index is calculated in euros. The prices are collected in local currencies via Reuters, and Reuter's real-time spot exchange rates are used to convert local prices to euros. The closing value of the index is calculated using the real-time exchange rates at 05:00 PM Greenwich Mean Time.

4.6 Investable weight factor (IWF)

Every company in the index is assigned an investable weight factor. The investable weight factor is defined as 1 – the sum of the percentage held by major shareholders. The investable weight factor thus ranges between 0 and 1, and is a factor which is used to account for the publicly available shares of a company. Three types of shareholders are considered as control blocks:

1. Other publicly traded corporations, private equity firms, strategic partners, venture capital firms or leveraged buyout groups.
2. Holdings in the company by government entities. All levels of the government are taken into account.
3. Holdings of founders of the company, former officers and directors of the company, directors or family trusts of officers. Holdings of trusts, pension funds, foundations, employee stock ownership plans and other investment factors associated with and which are controlled by the company.

Holdings of a group which exceed 10% of the total outstanding number of shares of a company, are excluded. Holdings in a group form a cumulative total, and holdings under 5% are only taken into account when they belong to clearly related shareholders such as board members or family members. Also, statutory limits on foreign ownership are recognized in calculating the IWF. The IWF thus is one minus all strategic holdings or the limit which is imposed by the statutory foreign ownership.

An exception of the rule is applied to Italian constituents. A so-called *patti di Sindicatto* is an agreement between some shareholders to act in unison on all voting matters. Also, they can give first right of refusal for their holdings to other members of the patti. These are considered as strategic blocs and when the aggregate holding is more then 5% it is also removed from IWF.

4.7 Index data

Daily return series are calculated and net return reinvested is calculated for the index. Net return is reflective to the return were dividends are reinvested after the deduction of withholding tax. Standard and Poor's uses the tax rate which is applied to non-resident institutions which not benefit from double taxation treaties. The tax-rates are maintained and updated, and are sourced from Ernst and Young's Worldwide Corporate Tax Guide and the Economic Intelligence unit. Net return reinvested is approximately the returns available to an investor based in Luxembourg. An investor in Luxembourg faces the highest tax-rates of Europe residents.

4.8 Index precision

For all calculations, the following decimal places are used:

- Index values are rounded to two decimal places.
- Exchange rates are stated to fourteen decimal places.
- Market capitalization is stated to four decimal places.
- Index values are calculated to fourteen decimal places.
- Outstanding shares are expressed in units.
- For all index calculations, the following level of precision is used:
- The Investable Weight Factor is rounded to decimal places.
- Share prices are rounded to six decimal places.

4.9 Index Mathematics

Most equity indices from Standard and Poor's are float-adjusted and market cap weighted. This means that each stock's weight in the index is directly proportional to its float-adjusted market value. Some indices are equal-weighted, each stock is then weighted equally in the index. Even other groups of indices are weighted to other factors. One concept - the index divisor – is crucial to all index calculations. The index divisor will be elucidated:

The index divisor: when an index would be expressed in dollar terms, it would be an unwieldy number. Because this number is uneasy to handle, the market value is divided by the index divisor to scale this number. Several factors cause the index value to change – for example the additions and deletions of companies - and the divisor is changed to offset the change in market value of the index. The divisor must provide a continuous measure of market valuation.

For capitalization weighted indexes, the formula to calculate the S&P Europe 350 is as follows:

$$\text{Index Level} = (\Sigma P_i \times Q_i) / \text{Divisor}$$

Where P_i is the price of stock i and Q_i represents total used number of shares outstanding for stock i. This is summed for all shares in the index, and the figure is divided by the divisor to obtain an easy to handle number.

The S&P Europe 350 index is float adjusted. When an index is float adjusted the total number of shares outstanding is reduced by closely held shares because these shares are not available to investors. Standard and Poor's calculates for each stock an Investable Weight Factor (IWF). The variable Q_i in the first equation is then replaced by the product of the IWF and the outstanding Shares:

$$Q_i = IWF_i \times \text{Total Shares}_i$$

The IWF is replaced by another multiplier at times other adjustments are made. In order to avoid unwanted double counting, the next formula is used:

$$\text{If } FA > FR \text{ then } IS = 1 - FA$$

$$\text{If } FA < FR \text{ then } IS = 1 - FR$$

With FA as the fraction of shares eliminated due to float adjustment, FR as the fraction of shares excluded for foreign ownership restrictions and IS as the fraction of total shares which is excluded on the basis of the combination between FA and FR. In addition, Q_i can next be written as:

$$Q_i = IS_i \times \text{Total Shares}_i$$

Every time the IWF or the share count changes, the index divisor must be adjusted to keep the index level unchanged.

The divisor is adjusted any time the number of stocks in the index is changed which alter total market value of the index while stock prices are hold constant. The next formula will show how the divisor is adjusted when one stock is deleted and one stock is added to the index:

Let at time t-1 stock r be deleted and at time t stock s be added to the index:

$$\text{Index Level}_{t-1} = ((\Sigma P_i \times Q_i) + P_rQ_r) / \text{Divisor}_{t-1}$$

$$\text{Index Level}_t = ((\Sigma P_i \times Q_i) + P_sQ_s) / \text{Divisor}_t$$

For simplicity, adjustments to share counts and IWF are ignored. To let the index level remain the same:

$$((\Sigma P_i \times Q_i) + P_rQ_r) / \text{Divisor}_{t-1} = \text{Index level} = ((\Sigma P_i \times Q_i) + P_sQ_s) / \text{Divisor}_t$$

So

$$\text{Divisor}_t = (\text{Divisor}_{t-1}) \times ((\Sigma P_i \times Q_i) + P_sQ_s) / ((\Sigma P_i \times Q_i) + P_rQ_r)$$

In this way, the new divisor is being calculated to hold the level of the index constant.

Adjustments to the divisor are done after the close of trading, and closing prices are used for calculations. Next, a description of possible changes in total market value is summed up. A distinction has been made for changes caused by the management of the index and changes due to corporate actions of the constituent companies.

Index management related changes: these changes occur when companies are deleted and added into the index or when the Investable Weight Factor changes. These changes usually occur only a few times per year. The divisor is adjusted as shown earlier.

Corporate action related changes: In table 4 the various corporate actions which alter the divisor are summed up:

Table 4 Corporate action related changes

Corporate Action	**Divisor Adjusted (Yes or No)**
Company is added or deleted	Yes
Change in shares outstanding	Yes
Stock split	No
Spin-off. Spun-off company not added to the index	Yes
Spin-off. Spun-off company is added to the index	No

Spin-off. Spun-off company is added to the index and another company is removed to keep number of names	Yes
Change in IWF causes market value to change	Yes
Special Dividend. Share price is assumed to drop.	Yes
Rights Offering where shareholders receive the right to by a proportional number of additional shares at a set price	Yes

Source: Standard an Poor's, 2007

Standard and Poor's also calculates the total return of the index. The total return index reflects both movements in stock prices and reinvestment of dividend income in the overall index. Every day, total dividend paid per day is calculated and converted into points of the price index:

$$\text{Total Daily Dividend} = \Sigma \text{ Dividend}_i \text{ x Shares}_i$$

Where Shares_i represents shares and Dividend_i represents dividend per share paid for stock i. For most stocks dividend is zero except for four times a year, and for some stocks Dividend is always zero. Total Daily Dividend is then converted to index points. The Divisor is divided for the underlying price index:

Index Dividend = Total Daily Dividend / Divisor

Total Return is usual calculated as:

Total Return = $(P_t + D_t) / P_{t-1}) - 1$

Next,

$$\text{DTR}_t = ((\text{ Index level}_t + \text{Index Dividend}_t) / \text{Index level}_{t-1}) - 1$$

DTR is defined as daily total return for the index. DTR is next used for calculation of total return index from one day to the next:

$$\text{Total Return Index}_t = (\text{ Total Return Index}_{t-1}) \text{ x } (1 + \text{DTR}_t)$$

Calculations of price earnings ratios, earnings per share and dividend yields are calculated by using the divisor as if it represents shares for a company. For example, earnings per share for the index is calculated as:

$$\text{Index EPS} = (\Sigma\, eps_i \times Shares_i) / \text{Divisor}$$

Where Index EPS represents earnings per share for the overall index, eps_i represents earnings per share for stock i, and $Shares_i$ represents the shares used to calculate the index.
The price-earnings ratio is calculated as the ratio between the index level and the index EPS:

$$\text{IndexPE} = ((\Sigma\, P_i \times Shares_i) / \text{Divisor}) / ((\Sigma\, eps_i \times Shares_i) / \text{Divisor})$$

The same approach can be applied to other index fundamentals, like price-to-book and book value per share.

5 Errors in databases

Section 5.1 provides a review of other studies and literature. Section 5.2 describes about comparisons between different data sources. Section 5.3 describes methods for detection of errors. Section 5.4 describes different types of errors and section 5.5 describes their implications. Finally, section 5.6 concludes with solutions and advices.

5.1 Introduction and review other studies and literature

The research in mutual fund studies has increased enormously since large databases are kept which hold fund characteristics and fund returns. Among the most widely used databases are CRSP and Morningstar in the United States. Thomson is among the most widely used database in Europe. Although the data contained in the databases has been gathered with lot effort, the databases are not without errors. Courtenay and Keller examined the CRSP shares-outstanding data. Courtenay and Keller stated that recently researchers rely on machine-readable data, and do not use original documents. The quality of the research performed by researchers is thus dependent on this machine-readable data (Courtenay, Keller, 1994).

Several empirical studies have been dedicated in investigating the quality of databases. For example, Elton Gruber and Blake compare data in two different databases. They compared the

CRSP Survivor Bias Free U.S. Mutual Fund Database and the Database from Morningstar. Return characteristics kept by both databases are compared to each other. The database of CRSP contains a form of survivorship bias – which they refer to as omission bias – and is significant. The return data for the CRSP files is for some companies recorded monthly, for some files annually and for some companies even no return data was available. Thus if researchers use monthly return data, they use a sample which understates the proportion of mergers and liquidations. The result is that performance is overstated. Elton Gruber and Blake use the survivorship-bias-free sample constructed earlier by Elton et al and compare the sample to the CRSP database. The analysis was conducted for the period 1977 to December 1993. Elton Gruber and Blake concluded that of the 154 mutual funds – with year-end 1976 total net assets less than 15 million dollar – for 93 funds complete monthly returns were available, 10 funds had some monthly returns and 51 funds had no monthly returns recorded by CRSP. The bias due to omitting returns for funds was estimated 35.9 basis points, for an estimated alpha. The Fama-French three factor model was used to estimate alphas. The bias found was 44 basis points (Elton, Gruber, Blake, 2001).

Also other sorts of data has been analyzed by researchers. For example, Guenther and Rosman examined differences between SIC-codes assigned to companies by CRSP and COMPUSTAT. Sic codes are used to classify stocks to the right industry. SIC codes are also extensively used by researchers, so it is important to examine whether these codes produce homogenous industries. The conclusion of the study was that large differences exist between the SIC codes reported. A study including this codes was re-examined and Guenther and Rosman also concluded that using a different database can affect the results of empirical research. Guenther and Rosman do not state which source is better, but caution researchers because it may influence findings in empirical research. Guenther and Rosman advise that researchers should investigate whether the empirical results of analyses are sensitive to the method of industry classification used by the database the researchers use (Guenther, Rosman, 1994).

San Miguel reports in his study the findings of a study which compared the Research and Development expense data in Form 10-K reports to the Securities and Exchange Commission to the information reported in COMPUSTAT. In this study also several differences were found between the two resource databases. The study analysed a sample of 256 companies from 100 different COMPUSTAT industries in 1972. The 10-K reports are a reliable source of data, so COMPUSTAT was tested for its credibility. In this study, information for 78 companies, or 30% appeared different (San Miguel, 1977).

Rosenberg and Houglet also argued that erroneous data can destroy research effort and can damage management decisions based on the data when the data contains too much errors. In a study the researchers compared monthly price relatives of CRSP and COMPUSTAT for NYSE stocks. Price relatives for 97 utilities from March 1962 to June 1968 and for 844 industrials from January 1963 to June 1968 were compared. Price relatives were defined as $P_t / P_{t-1.}$ In total, 5,939 utility price relatives and 35,357 industrial price relatives were matched. Large errors were infrequent, but some appeared. When the databases were compared, 2.91 per cent differed by more than 100 basis points return over the month, 0.712 per cent differed by more than 500 basis points return over the month and 0.303 per cent differed by more than 2000 basis points in return over the month. According to the researchers, returns over 100 basis points but below 500 basis points could be regarded as rounding errors, but returns over 500 basis points should be taken seriously. Errors were corrected by finding the original published resources for the utility sample (Rosenberg, Houglet, 1974).

Later, Bennin updated the study performed by Rosenberg and Houglet. Returns including all distributions were compared where Rosenberg and Houglet only compared returns from prices. For this study, the authors used the new tape introduced by COMPUSTAT, the Price Earnings Dividend Tape. For this tape, COMPUSTAT claimed the price information of primary importance. Accuracy was tested by comparing the database with the same data obtained from CRSP. Monthly returns were matched of each NYSE company. Returns were compared for the period January 1962 through July 1978. Two types of mistakes were possible: errors in price, and errors in dividends. The error rate in the study of Rosenberg and Houglet was 0.0071 (294/41296) and the error rate in the study of Bennin was 0.0025 (471/187460). It appeared that four companies had an unusually high number of dissimilar returns. Bennin assumed that recently published information would contain more errors because users had less time to uncover errors and because of the pressure to produce a timely tape, but this appeared not true. Since 1975 they concluded that the two tapes are equally reliable, and the cross-checking technique shows that CRSP had an error rate of 1/10,000 and COMPUSTAT had an error-rate of 1/1,000 during the period 1962-1978 (Bennin, 1980).

5.2 Comparison different data sources

Some studies which investigate errors in databases make use of different databases from different sources. Elton, Gruber and Blake recognized in their study that both databases had different information resources. The two databases were the *CRSP Survivor Bias Free U.S. Mutual Fund Database* and the Database from Morningstar. The returns were both examined whether the funds had the same return data in both databases. The return data was also corrected for known biases. Common stock funds were selected from the Growth and Income group that had over 15 million dollar total net assets in 1998. Both data sources contained completely information of monthly returns from January 1979 through December 1998. Survivorship and omission bias were eliminated. Five-year samples were constructed, and the three-factor model of Fama and French added with a momentum factor was applied to obtain information about alphas. The following differences between the data in the two databases were of frequent occurrence:

First, in earlier years the differences appeared larger then in later years. Second, differences were larger for small funds rather then for large funds. For example, for small funds the difference was 61 basis points per year in the period 1979-1983, and 16 basis points in difference for alpha for large funds. Elton, Gruber and Blake did not investigate which database contained the correct return data, they only showed that different conclusions can be drawn by researchers depending on which database is used (Elton, Gruber, Blake, 2001).

In another study, Courtenay and Keller compared the reported distributions of CRSP for the period January 1 through December 1989 to a secondary data source, Moody's dividend record. They compared record dates, distribution codes, ex-dividend dates and distribution dates. Differences found between the databases were reconciled by reviews of annual reports. To illustrate, the following differences between the databases were found: 20 ex-date differences, 91 coding differences, 5 arithmetic errors, 8 instances of late updates and 17 reporting differences (Courtenay, Keller, 1994).

Comparing databases using different information resources is a useful tool to detect errors. The method is fairly easy to apply. Researchers having access to different databases can compare the databases to obtain a reliable source of data. In this study a comparable dataset is not available, so the method applied by the researchers in these studies will not be applicable. Instead, other methods will be used to detect errors. The methods used to detect errors will be explained in section 5.3.

5.3 Methods for detection of errors

If a comparable database is not available, other methods have to be used to detect errors. Rosenberg and Houglet concluded in their study that the nature of the errors is such that when an error is made, the wrong number is completely unrelated to the other prices. Also, in time series, the mistake is usually followed by the next number which has a value in the opposite direction. The return data provided by Thomson DataStream for all stocks included in the Standard and Poor's 350 for the period 01-01-1988 till 01-08-2007 will be screened for the next errors:

1. Return data will be inspected for jumps up or down by more than three standard deviations in time-series. Suspicious outliers are defined as numbers which are more then three standard deviations from the mean and are followed by a return number in the opposite direction, which are not more then 50% difference in value of the number preceding the outlier. Standard deviation is calculated for each stock individually using daily return data.
2. The dataset will be screened if return data remains the same for several days. Table 6 summarizes how many of these sequences of zero-return series appeared in the dataset.

The next section will show the results of screening the dataset for errors.

5.4 Different types of errors and errors in the data

Several causes for errors in databases can occur. For example, Rosenberg and Houglet found the following errors in databases. For stock splits: for some companies, errors resulted because the stock split was applied one month too early, because prices between two months were reversed, the price of another company was entered by mistake, a reversal in digits was keypunched, and for some errors the source was obscure. In summary, two kinds of errors were found: errors due to treating splits and errors due to recording prices wrong. COMPUSTTAT contained an error rate of 1/600 and CRSP contained no major errors for the sample investigated by Rosenberg and Houglet. In treating stock splits, CRSP contained an error rate of 1/32 and COMPUSTTAT contained an error rate of 3/32. Rosenberg and Houglet defended COMPUSTAT because monthly prices are viewed as least important in the database of COMPUSTAT, while prices are the primary content of CRSP (Rosenberg, Houglet, 1974).

Possible causes for the errors detected in DataStream could occur due to prices between days are reversed, prices of other companies are entered by mistake, or reversals in digits are keypunched. Other missing data occurs since no return data is available for a particular firm or particular time period, or returns remain the same for a certain period of time. The sources for errors in the dataset are not further investigated. All return data for all companies included in the Standard and Poor's Europe 350 index is used for screening for errors. Thus when a company was included in the index from 01-01-1993 till 01-01-1997, and return data for the particular company for the period 01-01-1988 till 01-08-2007 was available, all returns in the period 01-01-1988 till 01-08-2007 are screened for possible errors. Table 5 and 6 summarize how many errors were discovered using the techniques.

Table 5 Missing return data

Total number of months missing return data	Total number of companies
1 month	2
3 months	4
9 months	1
23 months	1
32 months	1

The same return data was checked for suspicious outliers. Of all observations, 2.04 % were detected as outliers. Of these outliers, only 9.7 % appeared suspicious. To summarize for all observations, 0.19% appeared suspicious.

Table 6 summarizes how many times the same return data a certain number of days in sequence occurred. Daily returns for a total of 2.385.890 observations were screened:

Table 6 Sequences of zero returns

Number of days of zero return in sequence	Absolute number of observations	Relative Number of observations
1	134765	0.056484163
2	32318	0.013545469
3	7962	0.003337119
4	2340	0.000980766
5	1018	0.000426675
6	552	0.00023136
7	306	0.000128254
8	217	9.09514E-05
9	144	6.03548E-05
10	95	3.98174E-05
11	84	3.5207E-05
12	51	2.13757E-05
13	50	2.09565E-05
14	37	1.55078E-05
15	42	1.76035E-05
16	19	7.96349E-06
17	19	7.96349E-06
18	22	9.22088E-06
19	9	3.77218E-06
20	11	4.61044E-06
21	5	2.09565E-06
22	7	2.93392E-06
23	5	2.09565E-06
24	4	1.67652E-06
25	5	2.09565E-06

As can be concluded from the table, sequences of zero returns do not account for a large proportion in the dataset. For example, it occurred 552 times in the whole dataset a sequence of 6 days of no return appeared, which accounted for 0.0231 % of total observations. When one would delete all sequences of 3 days no return data or larger that would account for a deletion of 2.2 % of total observations. When one does this for respectively sequences of 5 and 10 zero-return, this would account for deletion of respectively 0.83 % and 0.26% of total returns.

5.5 Implications errors

Rosenberg and Houglet calculated the mean, skewness, variance and kurtosis of the samples to see how large the differences were. The samples were the original samples, and samples where errors were deleted. The result was that the mean was affected by a factor of less than 1/1000, variances were affected by a factor of 5/3, skewness by 11 and kurtosis by 18. The errors also caused an upward bias in any arithmetic index of monthly returns. The errors in the CRSP database caused a bias of 0.06 per cent per annum, while only one price relative in 1500 was erroneous. Also, beta regressions were performed to the samples. The result was that the estimated betas were much widely dispersed and had larger standard errors. Estimated variance of specific risk increased by a factor of 10 or more and R^2 decreased to almost zero.

In this analysis errors are detected and counted to show the occurrence of these errors. Because of the lack of a comparable database, alternate resources were not available. To provide a description of possibilities of errors and the rate at which they occur, tables 5 and 6 provided a description of occurrences of strange values. It would also be interesting to compare returns and correlation of returns with the reconstructed portfolio and the portfolio of Standard and Poor's . Unfortunately, Standard and Poor's only provides this data for a large fee, so no comparison could be made. The next section will provide solutions and advises for using databases.

5.6 Solutions and advices

Other possibilities for researchers to use datasets which are free of errors are discussed. One easy way is just to restrict the sample with funds were returns were available. Other possibilities are proposed by researchers:

In cases more than one database exists, databases can be compared. Courtenay and Keller also used different secondary data information sources. Discrepancies in data were solved by review of annual reports. Also, *The Wall Street Journal index* and *Standard & Poor's Quarterly Dividend Record* were reviewed to find the right pieces of information and errors were corrected (Courtenay, Keller, 1994).

San Miguel in his research for errors in R & D data proposes several recommendations for researchers: When several computerized databases exist, researchers can compare the databases using little effort and cost. When other computerized databases do not exist, statistical sampling techniques could be utilized by researchers to test the reliability. When small samples are analyzed, the authors recommend researchers to use the original information rather then information extracted from other databases (San Miguel, 1977).

Rosenberg and Houglet conclude that CRSP has a database with a lower error-rate than COMPUSTAT, but they also recognize that COMPUSTAT has a greater complexity, and a higher degree of effort is needed to provide high accuracy. They state that outliers are always suspect because they will be unlikely be generated by an economic consequence and the manner errors arise cause errors to be large as stated earlier. Because outliers also have a dominant effect on traditional statistical analysis, it is important to inspect them. The authors state for conclusion three principles:

1) They state that using a statistical method which minimizes the effect of outliers will be a good tool. A researcher can do this by
   - replacement of outliers with truncated values which are at least three standard deviations away from the mean.
   - discard the outliers.
   - use estimation procedures which are more “robust” than the quadratic methods earlier discussed.
2) Knowledge of the underlying economic process and the sources of errors also helps in detecting the errors. For this cause, price relatives were used instead of prices alone.
3) Using comparative databases is highly efficient. It is least expensive. Because the probability that an error remains undetected is the product of the error rates in the databases and the probability that given the databases make the same error in the same item, the errors will be a close enough to be indistinguishable.

In this analysis the errors will not be replaced by other numbers. Tables are provided for illustration how many data types appeared suspect.

5.7 Correlation in value-weighted monthly returns

It is interesting to investigate whether one can replicate the index using data collected from DataStream. Because return data for the Standard and Poor’s Europe 350 index could only be obtained from Standard and Poor’s for a large fee, returns of an exchange traded fund that mimics the index were collected. Return data were collected from the website of Yahoo, for the period 01-09-2002 to 01-05-2006. The 44 monthly value-weighted returns of the exchange traded fund were compared to the 44 monthly value-weighted returns calculated from the data from DataStream. The correlation between the returns was 0,80. We also verified whether these returns differed, using a t-test. Table 7 summarizes the statistics.

Assuming $v_1, v_2 \ldots v_{44} \sim N(\mu ; \sigma^2)$.

H_0: $\mu = 0$ $\alpha = 0.05$

H_1: $\mu \neq 0$

Where v_i is defined as the difference in monthly-weighted return for the i-th month.

Test-statistic T = $\bar{v}$ / (s / √ n) = 0.00476 / (0.0279 / √ 44) = 1.129

Under H_0: $T \sim t_{43}$

Because the P-value (=0,265) is larger then α there is no evidence to reject the H_0.

Table 7 Statistics comparing value-weighted monthly returns

t-Test: Paired Two Sample for Means

	Exchange Traded Fund	DataStream
Mean	0,016299923	0,01154
Variance	0,001810952	0,002049
Observations	44	44
Pearson Correlation	0,799077427	
Hypothesized Mean Difference	0	
df	43	
t Stat	1,129516173	
P(T<=t) one-tail	0,132470965	
t Critical one-tail	1,681071353	
P(T<=t) two-tail	0,264941931	
t Critical two-tail	2,016690814	

6 The three factor model

Section 6.1 describes the CAPM-model. Section 6.2 describes the methodology of Fama and French and their empirical findings. Section 6.3 provides a conclusion.

6.1 The CAPM-model and the cross-section of average returns on stocks explained

In general, when investors are exposed to higher risk, a higher return is expected. Two models have been developed to show the relation between these concepts. When returns are inspected, a relationship becomes clear that the higher the risk an investor faces, the higher the returns an investor earns. Investors are risk-averse, and when risk increases, investors need higher expected returns to be offset for the risk they are exposed to. Risk is measured by volatility, and returns with higher risk are more volatile than returns with lower risk. The more a portfolio is diversified, the less risky a portfolio is.

The systematic risk of an asset is measured as beta. Beta is calculated using the historical returns of a security and the return of the market, where the market is represented by a broad index like the Standard and Poor's 500. The beta of a portfolio is just the average of the betas included in the portfolio, weighted by the market capitalization of each security.

The CAPM model quantifies the relationship between the beta of an asset and its corresponding return. A few assumptions are made for the CAPM model:

- Investors only care about volatility and the expected returns, so expected returns given a certain volatility are always maximized.
- The only risk factor is the market portfolio.
- Investors have homogenous beliefs about the trade off between risk and reward in the market.

The CAPM-model is defined as follows:

$$E(R_a) = R_f + \beta\,(R_m - R_f).$$

E(Ra) is defined as the expected return on an asset, R_f is defined as the risk-free rate, β is defined as the beta of an asset and R_m represents the rate of return of the market as a whole.

Active fund managers try to outperform the market. The CAPM model provides an estimate of the given return investors need to be compensated for bearing a certain risk. The difference between the real return and the return computed from the CAPM is defined as alpha. The beta of a certain stock is usually calculated from the monthly returns of a security, with 3 to 5 years of data. The beta is then calculated from a time series regression from R_f, R_m and Ra.

The next regression can be run:

$$R_a - R_f = \alpha + \beta\,(R_m - R_f)$$

Although the CAPM model is widely excepted, academic research has found anomalies in the CAPM model. The CAPM-model is often incorrect in explaining the expected returns on securities. The expected returns can often be explained by more than only one risk-factor. Multi factor models include more factors to improve the predictive power of a model.

A lot of academic research thus has been dedicated to explain the behaviour of stock returns. Fama and French analyse the cross-section of stock returns because the average returns seem to show little relation to the market β of the CAPM asset-pricing model. For decades, the asset-pricing model predicted that the market portfolio is mean-variance efficient. When a market portfolio is efficient, the asset pricing model has two implications:

1. a positive linear function exists between the expected returns of securities and the betas of the securities, where beta is defined as the slope in the regression of the return's of a security and the return of the market.

2. The beta of the market is seen as being sufficient to describe the cross-section of all expected returns.

But when researchers analysed securities over time, several empirical studies contradicted the CAPM asset-pricing model. Researchers have investigated whether size (measured as stock price times total number of shares), earnings/price, leverage, book-to-market equity and book equity show relation to the average returns on stocks. Several researchers found evidence that not alone the market's beta had explanatory power for expected returns on stock. Examples of these studies are the size effect of Banz (1981), the empirical evidence of Bhandari (1988) and the findings of Basu (1983). Banz finds that market equity also has some explanatory power in explaining the cross-section of returns of securities. Bhandari (1988) finds in his empirical study a positive relation between expected return and leverage of firms. Finally, Basu concludes from tests which take into account size and market beta, that earnings-price ratios add to the explanatory power of the cross-section of returns.

Fama and French investigated whether several characteristics of firms appeared some correlations with the returns on stocks. Fama and French analysed the cross-section of stock returns because the average returns seem to show little relation to the market beta of the Sharpe-Lintner asset-pricing model. Fama and French investigated whether other factors could relate to the cross-section of average stock returns. In earlier analysis, Fama and French showed in 1992 that market beta has little information about average returns. The two variables size and book-to-market equity seemed to explain the cross-section of average stock returns well. They evaluated the explanatory power of the market β, earning-price ratios, size and book-to-market equity in explaining the cross-section of average returns on stocks. The stocks they investigated are all listed on AMEX, NYSE and NASDAQ. Fama and French find that only a weak relation exist between the beta of a security and average return for the period 1963-1990. The main results of their empirical work during the period 1963-1990 are:

1. the cross-section of average stock returns is not explained by using beta alone.
2. book-to-market equity and size absorb the roles of earnings-price ratio and leverage.

Size measured by market equity and book-to-market equity seemed powerful in explaining the cross-section of stock returns. Firm size and book-to-market ratios emerged as the strongest predictors of the security returns in their analysis. A high book-to-market ratio, which means a low stock price relative to book value, means low earnings on assets. A low book-to-market ratio is associated with low earnings. Also small firms tend to have lower earnings on assets than larger firms (Fama, French, 1992)

Later, Fama and French expanded the analysis to government and corporate bonds. In this analysis Fama and French regress monthly returns to the market portfolio, size and book-to-market ratio's. Also in this analysis size and book-to-market equity added to the explanation of the cross-section of average returns on stocks.

6.2 Methodology of Fama and French and empirical findings

Fama and French use the return files of all non-financial firms of the NYSE, NASDAQ and Amex and the merged COMPUSTAT industrial files of balance-sheet data and income statements from the Center for Research in Security Prices. Financial firms were excluded from analysis because these firms have high leverage, while high leverage doesn't have to mean financial distress for financial firms. Although later, financial firms were included in the analysis by Barber and Lyon, and the same conclusions remained (Barber, Lyon, 1997)

When stocks are sorted into portfolio's, the betas are better measured because a portfolio has lower variance when the stocks are isolated. Also, because individual stocks are volatile, the hypothesis that all average returns are the same cannot be rejected. Thus the stocks are grouped in a portfolio, portfolio variance is reduced and average returns can be inspected more thoroughly (Cochrane, 2000).

Six portfolio's were formed by Fama and French as follows:
Every year, all stocks (NYSE, AMEX, NASDAQ) are ranked on size in June. Size is estimated as market equity. Market equity is defined as price times shares outstanding. Size rankings are based on the market value of equity in June of year *t*. The median size is used to split all stocks in two groups: small and big (S and B). The two groups do not have to contain an equal number of firms. Next, all stocks are divided in three groups based on their book-to-market ratio: 30% in Low, 40% in Medium and 30 % of all stocks is divided in High. This splits are arbitrary, and Fama and French have not searched for alternate splits.

The groups are cross-sected in six groups: S/L, S/M, S/H, B/L, B/M and B/H). To illustrate, the group S/L contains stocks in the small market equity group which also are in the Low book-to-market equity group. Group B/H contains stocks in the Big market equity group which also are allocated to the High book-to-market equity group.

Next, monthly value-weighted returns are calculated for each portfolio from July in year t till June in year t+1. Every year the portfolios are reformed again in June.

To be included a stock must have data for the prices for December of year t-1 and June of year t and book equity for year t-1. Fama and French proceed as follows:

For size: the factor SMB is each month the difference between simple average returns on the 3 portfolios S/L, S/M and S/H, and the simple average of the returns on the portfolio's B/L, B/M and B/H. SMB is thus defined as the difference between the returns on small- and big-stock portfolios which have about the same weighted average book-to-market equity.

For book-to-market equity: The portfolio HML is each month the difference between simple average returns on the two portfolios S/H and B/H, and the simple average of the returns on S/L and B/L. HML is thus defined as the difference between the returns on low- and high-BE/ME portfolios which have about the same size.

Their proxy for the market factor is defined as the return on a value-weighted portfolio of all stocks included in the 6 portfolios minus the return on the one-month bill rate. The one-month bill rate is defined as the risk-free rate.

For all stocks, 25 portfolio's are formed on size and book-to-equity. All stocks are then allocated to the right size quintile. Portfolio's are reformed every year. Later, portfolio's were formed also on the basis of E/P (earnings/price) and D/P (dividend/price) to check robustness of the model. The stocks are sorted by size and by book-to-market equity in June. For book-to-market equity, book-to-market ratio's are calculated as the ratio between the market value of the equity in December of year *t* – 1 and the book value of equity which is reported on the firms balance sheet in year *t*. When companies were ranked according to their book-to-market ratio, the rankings were used to construct book-to-market portfolio's from July of year t through June of year *t* + 1. Based on these rankings, firms are placed in the appropriate size quintile portfolio. The portfolios are reformed in June of each year. All stocks are accordingly allocated to the quintiles, and 25 portfolios are constructed. Value-weighted monthly returns are calculated for each portfolio from July to June t+1. For returns on stocks, excess returns were used in the analysis. Excess return are defined as the returns on stocks minus the return on the one month bill rate.

The average return per portfolio tend to decrease from the small- to the big-size portfolios. Average return also tended to increase with book-to-market equity. Average risk factors are just average values of a explanatory variable.

Fama and French performed three types of regressions:

1. $R_m – R_f$ is used to explain stock returns
2. SMB and HML are used to explain stock returns
3. SMB, HML and $R_m – R_f$ are used as explanatory variables.

The results of the regressions showed that $R_m - R_f$ captures much common variation. But some variation still remained unexplained. SMB and HML capture substantial time-series variation in the stock returns. But the best results came from the last regression: much of the variation seemed to be explained by these three factors.

Also, the intercepts for each regression for each of the three regressions-formulas were compared. When $R_m - R_f$ was used as the only explanatory variable the intercepts showed a size effect. Intercepts for smaller size portfolios exceed those in higher portfolios. They proved that market beta leaves cross-sectional variation in average stock returns.

For SMB and HML a time-series regression of excess stock returns on SMB and HML produced large intercepts and often the intercepts closed to more than two standard deviations from zero. The large intercepts showed that SMB and HML do not explain average premiums on stock returns, where premiums are defined as average stock return over the one-month bill rate.

For the regression of $R_m - R_f$, HML and SMB, the intercepts resulting from the regressions appeared near zero. The regressions absorbed the common time series variation in the returns on the stocks. From the regressions appeared that the SMB and HML factors explained the differences in average returns on stocks, and $R_m - R_f$ explained why the stock returns are larger than the one-month bill rate (Fama, French, 1993)

6.3 Conclusion Fama and French

The main result of the analysis performed by Fama and French was that size and book-to-market equity both seem to describe the cross-section of average stock returns. Fama and French don’t think that the analysis will only hold for their sample because there is no evidence that the explanatory power of book-to-market equity will deteriorate through time and for both sub periods in the analysis the relation holds strong. Also, preliminary work performed by Fama and French suggests that low-book-to-market equity firms tended to be persistently good earners relative to high-book-to-market equity firms. If the analysis of Fama and French will be persistent in the future it will influence portfolio formation and performance evaluation of investors.

7 Methodology

Section 7.1. describes the data. Section 7.2. describes the process of searching and matching companies included in the Standard and Poor’s Europe 350. Section 7.3 describes missing data. Section 7.4 describes currencies. Section 7.5. describes the model. Section 7.6 shows regression results. Finally, section 7.6 ends with the conclusion.

7.1. Data

The 3 factor analysis extends for the period 01-01-1988 till 01-07-2006. Standard and Poor’s provides on their website the list of companies included in the Standard and Poor’s Europe 350 index for the past seven years. Because a longer period for conducting the 3 factor model was feasible, information was obtained from Standard and Poor’s which included all the companies added and deleted in the Standard and Poor’s Europe 350 index for the period 01-01-1988 till 01-01-2007. The list obtained included exact dates for inclusion and deletion of companies in the Standard and Poor’s Europe 350. Whenever a company is included in the index there is also a company deleted from the index at the same date. For all companies the name of each company was provided, and for most but not all companies RIC codes and Sedol codes were provided. The latter codes – Sedol codes – simplified searching the corresponding companies in Thomson navigator. The RIC code of a security is another code used by Reuters and stands for Reuters Instrument Code.

A Sedol code stands for a Stock Exchange Daily Official List code and exists for over 35 years. Sedol codes have been allocated to companies for administration. Sedol codes consist of a 7 digit number and each company is assigned an unique Sedol code. A unique global identifier increases efficiency and leads to reduction in trading errors. One disadvantage of sedol codes is that the codes do not recognize the market the security trades.

As stated earlier, unfortunately some companies could not be found in DataStream. When data for an included company could not be found in Thomson DataStream, this company is excluded from analysis.

The Thomson navigator contains information for all security types. The information needed for the securities was obtained from DataStream of Thomson. All information – the names of the companies, RIC and Sedol codes – was used for finding the corresponding companies in DataStream from Thomson. The next section will outline how all information was used to find the corresponding companies in DataStream.

7.2 Searching and matching companies included in the Standard and Poor's Europe 350

With the information obtained from Standard and Poor's which included the exact dates and constituent companies in the Standard and Poor's Europe 350 index, the corresponding companies were searched for in DataStream. DataStream contains the current constituent list of most indices, so the constituent list of the Standard and Poor's Europe 350 index in DataStream was downloaded and corresponding companies were matched. Companies were matched manually to avoid mistakes. The reason this method was used was that names of companies included in DataStream sometimes differed from the names included in the list obtained from Standard and Poor's. Examples are the companies TF1-TV Francaise which is called TF1 and Deutsche luft-rg which is called Deutsch Lufthansa in DataStream. Because the names of some companies were not exactly corresponding, all companies were linked manually to ensure faultless combinations. Because DataStream only contains the current constituent list, former included companies were matched in alternate ways.

All companies included in the list obtained from Standard and Poor's were entered in the search-module of DataStream. First, Sedol codes were used to find the corresponding companies in DataStream. Because Sedol codes sometimes change over time, names of corresponding companies were compared to match the corresponding companies in DataStream. Most companies were reasonably easy to find using this method. When entering a Sedol code in the search engine of Thomson showed no result, other methods were used to find the corresponding company. Every company has an unique RIC code. Unfortunately, the search engine of DataStream does not allow users to search for companies using RIC codes. Names of companies were used to find corresponding companies. Also, base dates of companies and markets where securities traded were used to find the corresponding companies. The markets where securities traded were distracted from RIC-codes. A RIC-code is a code of 1 to 4 characters, followed by 1 or 2 characters which indicate the exchange the stock is traded. For example, NRK.L is the RIC code for the company Northern Rock. The letter L after the dot states that the security is traded on the London Stock Exchange. This information was also used for matching the companies included in the index. Using these methods, most companies in the navigator were matched.

7.3 Missing data

With the constituent list of the Standard and Poor's Europe 350 obtained from Standard and Poor's, almost all constituent companies over the period 01-01-1988 till 01-07-2006 were matched with the corresponding companies in DataStream. For all companies, returns, book values and market equity were collected for the period each company was included in the index. In total, 688 companies needed to matched in DataStream. Of these 688 companies, 79 were not matched using the methods explained above. Thomson was contacted to find the corresponding companies. With help of several employees of Thomson, 66 companies were matched. Even with help from several employees of Thomson, 13 companies could not be matched. The analysis is performed with omission of these companies, because information of these companies was not available in DataStream. Table 5 summarizes the missing return data for the companies included in the index. The firms lacking data are excluded from analysis.

7.4 Currencies

Because the Standard & Poor's Europe 350 represents the European market, several European countries are represented in the index. Currency's are important because the market equity of the companies included in the index must be converted to one currency. Before 01-01-1999, Standard and Poor's calculates market equity of the Standard and Poor's Europe 350 using the European Currency Unit (ECU). From 01-01-1999 on, all prices are converted to euro using the real-time spot exchange rate from Reuters. Prices are collected in local currencies via Reuters. The same method is conducted in this analysis. For market-equity based calculations prices are converted to European Currency Unit (ECU) till 01-01-1999. From 01-01-1999 on, market equity is converted to euro.

7.5 The three factor model applied

Three stock market factors are used for explaining the cross-section of stock returns: the overall market factor, a factor for size and a factor for book-to-market equity. These factors will be used to explain the cross-section of the average returns of the companies included in the index. Excess returns are used for analyses, the UK interbank 1-month offered rate is used as proxy for the risk free rate. If the model appears good, intercepts of regressions will be near zero. The returns of 16 portfolio's formed on size and book-to-market equity are explained. Programming for data-manipulation was necessary to handle the large datasets.

In June of every year t from 01-06-1988 till 01-07-2006 all stocks of the companies included in the Standard and Poor's Europe 350 are ranked on size. Total market equity is used for analysis. The group of companies formed is next every year split in half by the median size firm. Two groups are thus formed: small and big (S and B).

Next, In June of every year *t* from 06-1988 till 07-2006 all stocks of the companies included in the Standard and Poor's Europe 350 are ranked on book-to-market equity. Book-to-market equity is defined as the ratio between book common equity of year *t*-1 and market equity at the end of December *t*-1. The group is then split every year into three groups: the breakpoints are the 30% for the bottom, 40% of the middle and top 30% of all ranked stocks. Three groups are formed, respectively L, M and H.

Next, six portfolio's are formed from the cross-sections of each group of stocks: S/L, S/M, S/H, B/L, B/M and B/H. Monthly value-weighted returns are calculated from July till June *t* +1. The portfolio's are reformed in June *t* +1 and monthly value-weighted returns are calculated again.

Three factors are calculated for the period 06-1988 till 07-2006: the market factor, HML and SMB. The three factors are defined as follows:

- SMB (Small minus Big) is defined as the simple average of the returns on S/L, S/M, S/H minus the simple average on the returns on B/L, B/M and B/H, every month.
- HML (High minus Low) is defined as the simple average on the returns on S/H, B/H minus the simple average on the returns on S/L, B/L, every month.
- $R_m - R_f$, or the market factor, is defined as the difference between the value-weighted return on all stocks included in the index UK interbank 1-month offered rate, measured at the start of each month.

Next, 16 portfolio's are formed on size and book-to-market equity. Fama and French use in the factor analysis an average of 3174 firms every year. Because in this analysis less companies are used, each fourth and fifth quintile are combined to fill the extreme quintiles with companies. Excess returns are calculated and are used as dependent variables in time-series regressions. In June each year *t* all stocks included in the index are sorted on size and the ratio of book-to-market equity. Market equity in June is used for analysis. For book-to-market ratio, the book value of the company for year *t*-1 is divided by the market equity of the company at the end of December of year *t*-1. All stocks are next divided into 5 quintiles for both criteria and 25 portfolios are formed by their intersections. Each fourth and fifth quintile are combined to fill the extreme quintiles, thus 16 portfolio's remain. Equal-weighted monthly returns are calculated for each portfolio from July of year *t* till June year *t*+1. Table 8 shows descriptive statistics and information for the 16 portfolio's:

Table 8 Descriptive statistics

	Average of annual number of firms in each portfolio			
	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	78.33	37.50	28.17	42.33
2	25.94	14.17	9.50	11.22
3	12.94	7.06	6.50	5.06
4 and 5(Big)	14.56	7.33	5.11	3.72

	Average of the annual averages of firm size			
	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	3211.86	3433.28	3501.23	3169.03
2	10429.05	10328.29	10636.10	9928.71
3	20150.95	21402.13	20709.01	20558.65
4 and 5(Big)	47382.85	43158.82	37370.95	46434.83

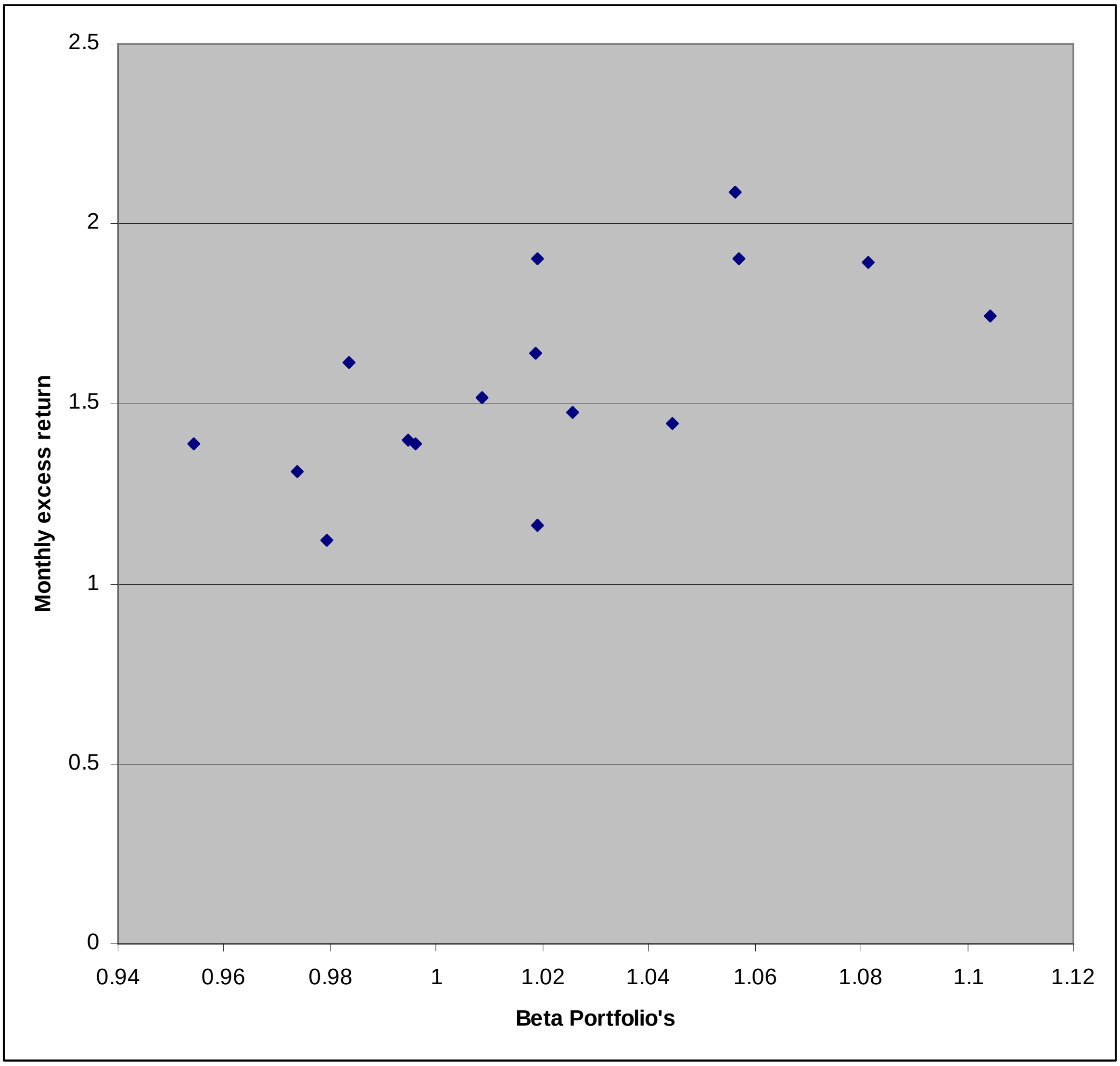
	Average of the annual book-to-market ratios for each portfolio			
	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	0.26	0.52	0.69	1.13
2	0.27	0.52	0.69	1.06
3	0.26	0.51	0.70	1.07
4 and 5(Big)	0.26	0.51	0.70	1.03

The tables show characteristics of the 16 portfolio's formed by size and book-to-market-ratio. First, the average of annual number of firms in each portfolio is shown. Most stocks are allocated to the lower quintiles. Next, the averages of annual averages of firm size and the average of the annual book-to-market ratios for every portfolio are shown.

7.6 Regression results

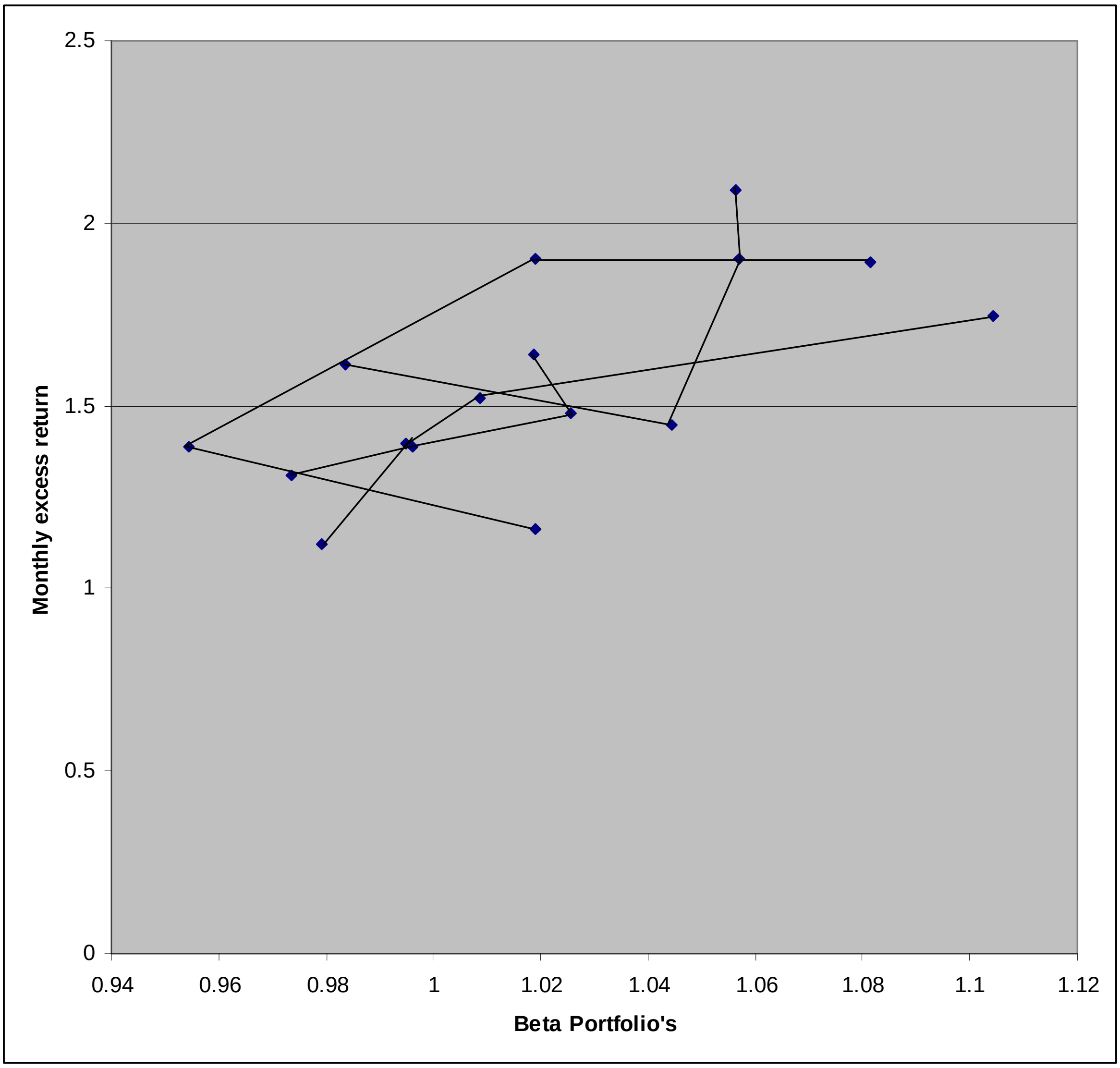
When the CAPM was tested, the model appeared not explaining the excess returns very well. When average excess returns were plotted versus the betas of stocks, the slope of the regression appeared too flat and the slope did not crossed any plausible risk-free rate. Figure 1 shows the monthly excess returns of the 16 portfolio's formed on size and book-to-market ratio and the betas of the portfolio's. As can be seen from figure 1 a lot of dispersion exist in average excess return for portfolio's which have nearly identical beta's.

Figure 1 Monthly excess returns and betas of 16 portfolio's



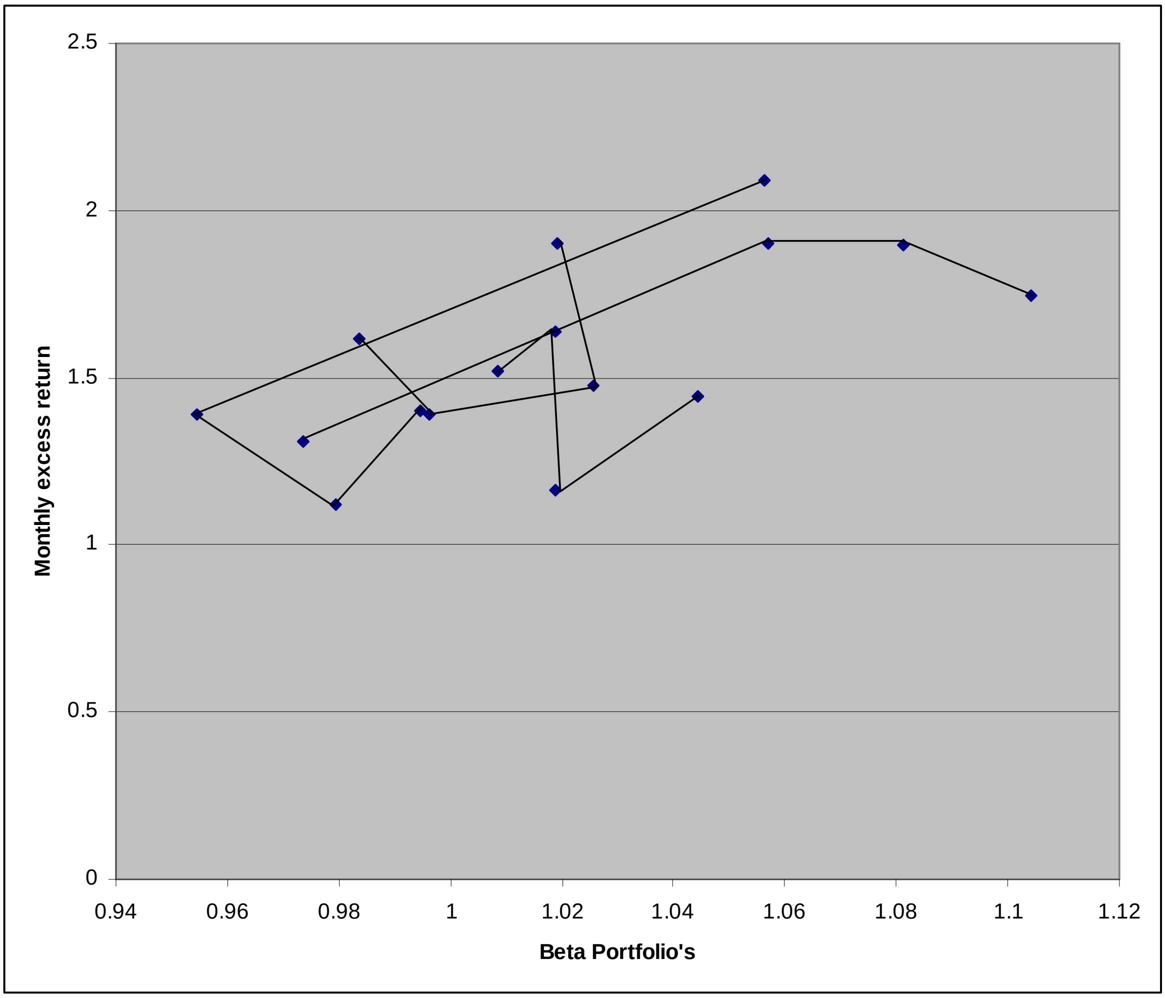
In Figure 2 the dots are connected for the portfolio’s that have the same book-to-market ratio, but different sizes. The variation in size produces a variation of average excess returns which is positively related to beta.

Figure 2 Average monthly excess return and betas. Portfolios for same book-to-market ratio but different size are connected.



In figure 3 the dots are connected for the portfolio’s that have the same size but different book-to-market ratios. The variation in excess returns is not positive related to beta.

Figure 3 Average monthly excess return and betas. Portfolios for same size but different book-to-market ratio are connected.



When the 3 factor model is constructed, figures 4 and 5 show the excess monthly returns predicted by the 3 factor model and the actual excess monthly returns of each portfolio. When the model is well constructed, all points should lie on a 45^{o} line. In figure 4 and 5 the points lie closer to this prediction than in figures 2 and 3. Figure 4 connects the portfolio's with the same book-to-market ratio and figure 5 connects the portfolio's which have the same size.

Figure 4 Actual mean average return and returns predicted by 3 factor model. Portfolios with same book-to-market ratios but different size are connected.

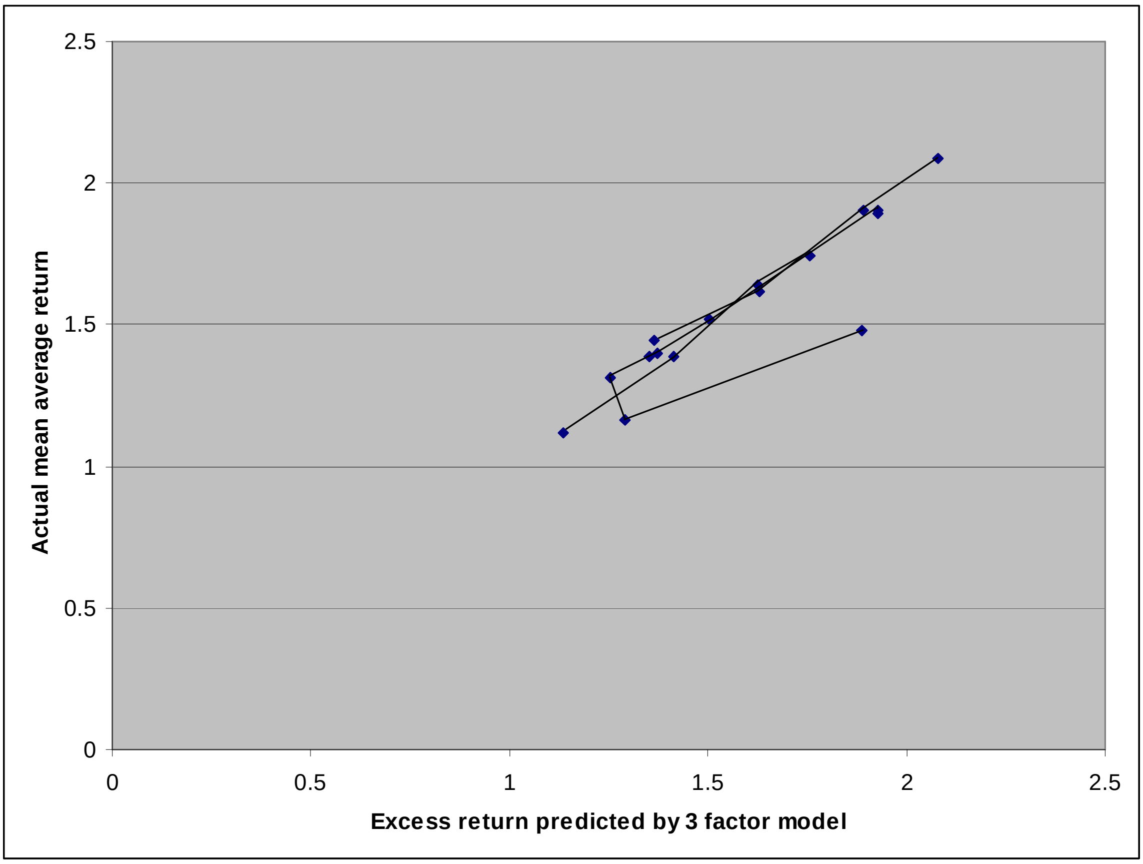
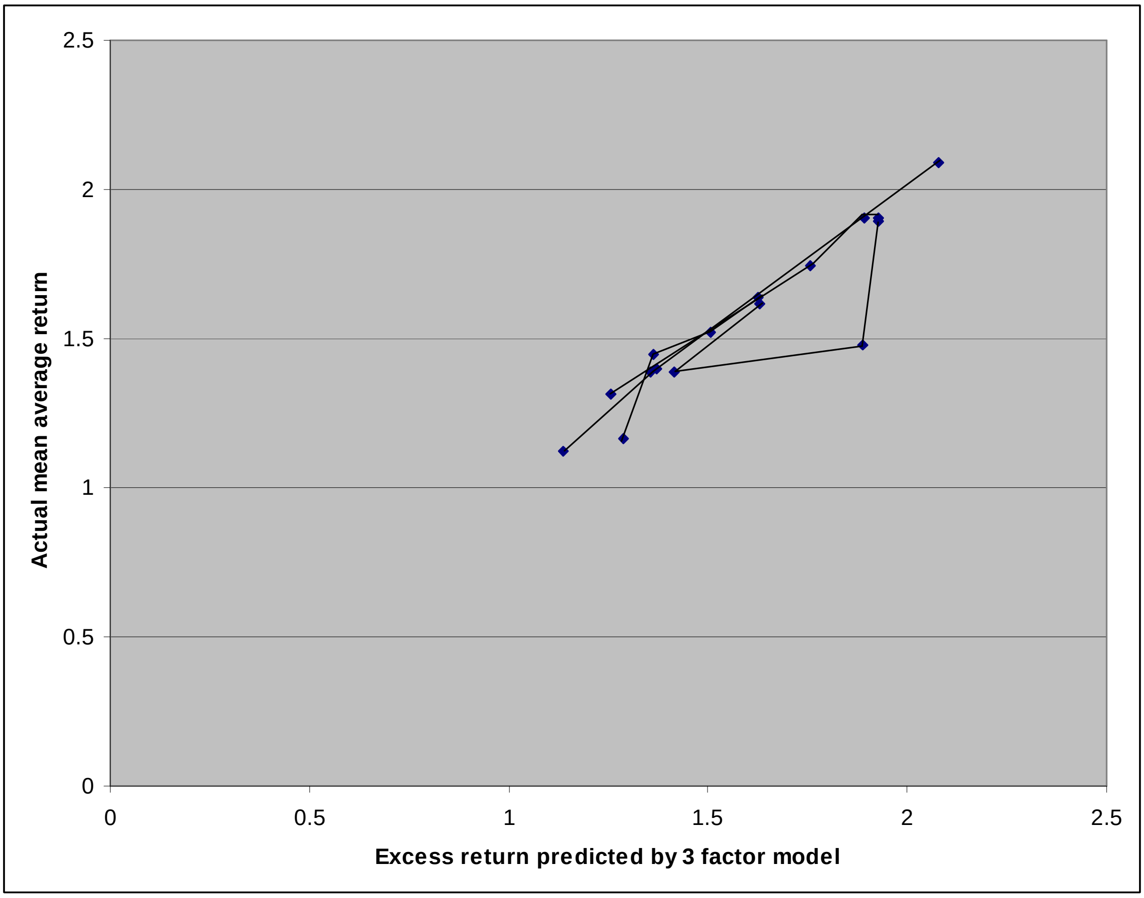


Figure 5 Actual mean average return and returns predicted by 3 factor model. Portfolios with same size but different size book-to-market ratio are connected.



The results of the regressions will now be discussed in more detail. The tables are summed in the appendix. Table 9 displays the mean and standard deviation for the factors $R_m – R_f$, SMB and HML. For the factor $R_m – R_f$, the average return for the period 06-1988 till 07-2006 yielded 1.97 % per month (t = 4.57). The average premiums for SMB and HML yielded -0.58 % (t = -2.42) and 0.46 % (t = 2.59) per month respectively. Correlations between the explanatory factors are shown in table 10. Correlations between the variables are of low value, with correlation between $R_m – R_f$ and SMB of 0.14 being highest.

Next, table 11 displays the means, standard deviations and t-statistics for the excess returns on the portfolio's on stocks. The means on the 16 portfolio's vary from 1.30 % to 2.20 % per month. All t-statistics for means are significant at $\alpha = 0.05$. The table shows evidence that average returns tends to increase with book-to-market ratio, but increases of average returns with size do not seem evident.

Slopes and values for R^2 are used to provide evidence whether the risk factors capture the variation in the returns of the stocks. The role of the stock-market factors are examined in three regressions. In the first regression, excess return of all stocks in the index – used as proxy for the market portfolio - is used to explain the cross-section of excess returns. Next, the factors SMB and HML are regressed with excess returns of the stocks. Last, excess returns of all stocks in the index and SMB and HML are used to explain the cross-section of excess returns. Thus three regressions are performed:

- $R(t) - R_f(t) = a + b*[R_m(t) - R_f(t)] + e(t)$
- $R(t) - R_f(t) = a + s*SMB(t) + h*HML(t) + e(t)$
- $R(t) - R_f(t) = a + b*[Rm(t) - R_f(t)] + s*SMB(t) + h*HML(t) + e(t)$

Where $R(t) - R_f(t)$ is defined as excess return of the stocks and $R_m(t) - R_f(t)$ is defined as excess return of the entire portfolio. The results of performing the regressions and statistics are summarized in tables 12, 13 and 14. R^2 values are the residual standard errors and s(e) are adjusted for degrees of freedom.

The values of R^2 in the first regression indicate that some variation in the excess returns might be explained by other factors than the overall market factor. R^2 values range from 0.70 to 0.91, and only 4 out of 16 portfolios have R^2 values of 0.90 or higher.

For the regression of excess returns with SMB and HML, it appears from table 13 that the factors SMB and HML capture not that much variation in returns.

Next, table 14 shows the results of the regression of excess returns with $R_m - R_f$, SMB and HML. The three factors used in regressions capture more variation in returns when the factor $R_m - R_f$ is used alone. The values of R^2 all appear above 0.90 or near 0.90 except for the lower-right quintiles. For the values of beta holds that every value appears more than 20 standard deviations from zero. For the slopes of SMB holds that the t-statistics appear largely from zero, for the upper left portfolios. For those portfolios, SMB surely captures some shared variation in stock returns which is missed by $R_m - R_f$. The slopes of the factor SMB do relate to size. In every book-to-market equity quintile the values of the slopes on SMB decrease from smaller- to the bigger-size quintiles.

For the slopes on HML holds that there appears no real explanatory power. Fama and French concluded in their analysis that the slopes of HML increased monotonically from lower book-to-market equity quintiles to higher book-to-market equity quintiles. The values of the slopes all appeared more than five standard errors from zero. In this analysis this does not appear. In this analysis, HML captures no real shared variation of stock returns.

Surprisingly, Fama and French in their analysis also find lower values of R^2 in the smallest-size quintiles and highest-book-to-market equity ratio quintiles. Here, adding the factors SMB and HML to the market factor for explaining the excess returns, resulted in increases of R^2 in 13 of the 16 baskets, 2 baskets remained the same value of R^2, and in only one basket decrease of R^2 appeared. It thus seems clear that adding the two factors in the regression does seem to add to the explanatory power of the regression-model. The values of R^2 and the slopes in the regressions show that R_m - R_f and SMB proxy for risk factors. The factor HML does not seem to capture much variation in the excess stock returns.

We now examine the intercepts of the regressions. In tables 15, 16 and 17 the intercepts of the three regressions are shown. Fama and French concluded that in the regression where only the market factor was used as explanatory variable, the value of the intercept appeared higher for smaller-size quintiles then for bigger-size quintiles. Also, higher book-to-market equity ratio quintiles appeared to have higher intercepts then lower book-to-market equity ratio quintiles. This results showed that the market beta leaved cross-sectional variation in the average return on stocks which was related to book-to-market ratio and size. It does however not appear from table 15 that intercepts of the smallest-size quintiles exceed those for the biggest size quintiles. The intercepts do increase within each size quintile when one moves from lower to higher book-to-market quintiles.

In the regression where the factors SMB and HML are used as explanatory variables the intercepts appeared relatively large. The results are shown in table 16.The intercepts are similar in size supports the prediction that the size and book-to-market factors can help explain the differences in average returns of stocks.

In table 17 the intercepts of the regression where all explanatory variables are included are shown. The intercepts appear lower than the intercepts shown in table 16. This means that the regressions which use all explanatory variables explain the cross-section of returns well.

7.7 Interpretation and conclusion

Fama and French found evidence that book-to-market equity and size are related to firm profits. High book-to-market equity firms tended to show poor earnings and low book-to-market equity firms tended to show high earnings. The factor HML captured variation in the returns in their analysis. The common risk factors in stock returns have been examined and tested. Table 10 shows there appears low correlation between the factors R_m - R_f, SMB and HML. This made it possible to examine the separate roles of these factors in explaining the cross-section of average returns. It has been evaluated if adding the factors SMB and HML captures variation in returns. Regressions have been performed and several statistic values have added to the proof that the factors offer more explanatory power for stock returns. From the analysis appeared that the factors R_m - R_f and SMB had explanatory power for the excess returns on stocks. The factor HML in this sample showed no real explanatory power, the factor SMB does. It has though been interesting to test the 3 factor analysis to this sample of European companies.

It is always interesting to find factors which add to the explanatory power of a model which tries to explain the cross-section of stock returns. Performance of managers can be evaluated by regressing their portfolio performance by these factors. Because the three factor model explains the returns of stocks better then the CAPM firm specific events - for event studies - can be better analyzed for what their specific influence is on stock returns.

Bibliography

Banz, R.W. (1981). The relationship between return and market value of common stocks. *Journal of financial economic.* Vol. 9, 3-18.

Barber, B.M., Lyon, J.D. (1997). Firm size, book-to-market ratio, and security returns: a holdout sample of financial firms. *The Journal of finance.* VOL LII, 2, pp. 875-883.

Basu, S. (1983). The relationship between earnings yield, market value, and return for NYSE common stocks: further evidence. *Journal of financial economics.*12, pp. 129-156.

Bhandari, L.C. (1988). Debt equity ratio and expected common stock returns: empirical evidence. *Journal of finance*.43,pp. 507-528.

Bennin, R. (1980). Error rates in CRSP and COMPUSTAT: a second look. *The journal of finance*. Vol. 35, 5, pp. 1267-1271.

Chakrabarty, B., Trzcinka, C. (2006). Momentum: Does the database make a difference? *The Journal of financial research.* Vol. XXIX, 4, pp. 441-462.

Chan, L., Hamao, Y. and Lakonishok, J. (1991). Fundamentals and stock returns in Japan. *Journal of Finance*, 46, 1739–1764.

Cochrane, J.H. (2000). *Asset pricing*. Princeton : Princeton University Press.

Courtenay, S.M., Keller, S.B. (1994). Errors in database revisited: An examination of the CRSP shares-outstanding data. *The Accounting Review*. Vol. 69, 1, pp. 285-291.

Davis, J. L., (1994). The cross-section of realized stock returns: the pre-COMPUSTAT evidence. *The journal of finance.*49, pp. 1579-1593.

Elton, E.J., Gruber, M.J. and Blake, C.R. (2001). A first look at the accuracy of the CRSP Mutual Fund Database and a comparison of the CRSP and Morningstar Mutual Fund Databases. *The Journal of finance.* Vol. LVI, 6, pp. 2415-2430.

Fama, E.F., French, K.R. (1992). The cross-section of expected stock returns. *The Journal of finance.* Vol. XLVII, 2, pp. 427-465.

Fama, E.F., French, K.R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics.* 33, pp. 3-56.

Guenther, D.A., Rosman, A.J. (1994). Differences between COMPUSTAT and CRSP SIC codes and related effects on research. *Journal of Accounting and Economics.*18. pp. 115-128.

Ince, O.S., Porter, R.B. (2006). Individual equity return data from Thomson DataStream: Handle with care! *The Journal of Financial Research.* Vol. XXIX, 4, pp. 463-479.

Rosenberg, B., Houglet, M. (1974). Error rates in CRSP and COMPUSTAT databases and their implications. *The journal of finance*. 29, 07, pp. 1303-1310.

San Miguel, J.G. (1977). The reliability of R&D data in COMPUSTAT and 10-K reports. *The Accounting Review*. Vol. LII, 3, pp. 638-641.

Shumway, T. (1997). The delisting bias in CRSP data. *The* Journal *of finance.* Vol. LIL, 1, pp. 327-340.

Shumway, T., Warther, V.A. (1999). The delisting bias in CRSP's Nasdaq data and its implications for the size effect. *The Journal of finance.* Vol. LIV, 6, pp. 2361-2379.

Summers, P.M. (2003). Data surveys. *The Australian Economic Review*. (36), 1, pp. 128-133.
Cochrane, J.H. (2000). *Asset pricing*. Princeton, NJ: Princeton University Press.

Ulbricht, N., Weiner, C. (2005). *Worldscope meets Compustat: a comparison of financial databases*. Proefschrift, Humboldt University.

Datastream

Websites:

Center for research in Security Prices (2008). Homepage (http://www.crsp.com), 5 february.

Harvard Business School (2007). http://www.library.hbs.edu, 10 october.

Standard and Poor's (2007). Homepage (www.standardandpoors.com), 9 september.

Thomson (2007). Homepage (www.thomson.com), 15 september

Yahoo (2008). Homepage (http://finance.yahoo.com), 10 April

Appendix

Table 9 Statistics explanatory variables.

Name	Mean	St.dev	t-statistic
$R_m - R_f$	1.97	6.16	4.57
SMB	-0.58	3.43	-2.42
HML	0.46	2.57	2.59

Table 10 Correlations explanatory variables.

Name	$R_m - R_f$	SMB	HML
$R_m - R_f$	1.00	0.14	-0.02
SMB	0.14	1.00	-0.12
HML	-0.02	-0.12	1.00

Table 11 Statistics excess monthly returns for each portfolio.

	Means			
	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	1.57	1.47	1.71	2.00
2	1.38	1.73	1.48	1.60
3	1.38	1.83	1.97	2.02
4 and 5(Big)	1.48	1.25	2.18	1.47
	Standard Deviations			
	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	6.70	6.57	6.62	6.86
2	6.36	6.98	6.98	6.78
3	6.71	7.36	7.36	7.85
4 and 5(Big)	6.50	6.53	7.65	7.46

t-statistics for means

Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	3.36	3.2	3.69	4.18
2	3.1	3.54	3.02	3.36
3	2.93	3.56	3.83	3.68
4 and 5(Big)	3.25	2.74	4.07	2.81

Table 12 Statistics for regression $R(t) - R_f(t) = a + b^*[R_m(t) - R_f(t)] + e(t)$

R_m is defined as the value-weighted monthly return on all stocks included in the index. R_f is defined as the UK interbank 1-month offered rate. SMB is defined as the difference between return on portfolio's consisting small and big-stocks. HML is defined as the difference in return between high and low book-to-market equity portfolios. The 16 size-BE/ME portfolio's are formed as follows: In June of every year t from 1988 to 2006 all stocks sorted by size and book-to-market equity. Size (measured as price times share outstanding)is measured at the end of June to allocate all stocks into 5 quintiles. Similarly, all stocks are allocated to 5 quintiles by BE/ME, where BE is measured at the end of December in year t-1 and ME is market equity at the end of December t-1. The 25 portfolio's are formed by the interceptions of the quintiles. For both criteria, quintile 4 and 5 are grouped to fill the groups with stocks. As a result, 16 portfolios remain. Value weighted monthly returns are calculated for the portfolios from July of year t to June of t + 1. The period of investigation extends from 1988 to 2006**.**

R^2

Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	0.91	0.86	0.84	0.85
2	0.91	0.88	0.86	0.83
3	0.86	0.87	0.81	0.75
4 and 5(Big)	0.91	0.90	0.68	0.70

S(e)

Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	2.05	2.46	2.65	2.66
2	1.86	2.46	2.62	2.84
3	2.54	2.69	3.2	3.90
4 and 5(Big)	1.91	2.03	4.33	4.11

b Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	1.04	0.99	0.99	1.03
2	0.99	1.06	1.05	1.00
3	1.00	1.11	1.08	1.11
4 and 5(Big)	1.01	1.01	1.03	1.01

t(b) Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	44.28	35.25	32.60	33.88
2	46.52	37.93	35.22	30.98
3	34.91	36.33	29.54	24.89
4 and 5(Big)	46.40	43.61	20.82	21.62

F Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	1960.68	1242.81	1063.05	1148.12
2	2164.25	1438.99	1240.39	960.03
3	1218.92	1320.16	872.67	619.59
4 and 5(Big)	2153.16	1901.74	433.54	467.60

Table 13 Statistics for regression: $R(t) - R_f(t) = a + s*SMB(t) + h*HML(t) + e(t)$

R^2 Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	0.075	0.15	0.12	0.12
2	0.035	0.085	0.073	0.088
3	0.046	0.062	0.07	0.04
4 and 5(Big)	-0.005	0.013	-0.003	0.012

s(e)	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	6.44	6.07	0.062	6.43
2	6.25	6.68	6.72	6.48
3	6.56	7.13	7.09	7.69
4 and 5(Big)	6.52	6.49	7.66	7.41

s	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	0.57	0.76	0.71	0.72
2	0.39	0.62	0.57	0.62
3	0.45	0.57	0.59	0.51
4 and 5(Big)	0.13	0.29	-0.023	0.32

t(s)	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	4.28	6.1	5.52	5.45
2	3.01	4.54	4.11	4.63
3	3.35	3.88	4.02	3.23
4 and 5(Big)	0.99	2.14	-0.15	2.1

h	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	0.088	0.093	0.05	0.061
2	0.037	0.017	-0.1	0.043
3	-0.074	0.2	0.3	0.15
4 and 5(Big)	0.035	-0.0037	0.24	-0.016

t(h)	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	0.5	0.55	0.29	0.35
2	0.22	0.0993	-0.56	0.24
3	-0.41	0.99	1.52	0.7
4 and 5(Big)	0.2	-0.021	1.16	-0.078

F Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	9.18	18.64	15.3	14.91
2	4.69	10.39	8.97	10.77
3	5.92	7.67	8.65	5.27
4 and 5(Big)	0.48	2.33	0.71	2.24

Table 14 Statistics for regression: $R(t) - R_f(t) = a + b*[Rm(t) – R_f(t)] + s*SMB(t) + h*HML(t) + e(t)$

R^2 Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	0.93	0.93	0.89	0.90
2	0.92	0.91	0.89	0.86
3	0.87	0.88	0.83	0.76
4 and 5(Big)	0.92	0.90	0.70	0.70

s(e) Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	1.77	1.75	2.16	2.15
2	1.81	2.14	2.36	2.26
3	2.43	2.51	2.99	3.84
4 and 5(Big)	1.86	2.03	4.19	4.11

b Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	1.01	0.95	0.95	0.99
2	0.98	1.03	1.03	0.97
3	0.99	1.1	1.05	1.09
4 and 5(Big)	1.02	1	1.05	1.01

t(b)	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	49.28	47.14	38.22	39.98
2	46.91	41.99	37.77	32.98
3	35.58	37.67	30.47	24.6
4 and 5(Big)	47.55	43.07	21.75	21.33

s	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	0.3	0.52	0.46	0.46
2	0.14	0.35	0.3	0.36
3	0.19	0.29	0.31	0.23
4 and 5(Big)	-0.14	0.0024	-0.3	0.052

t(s)	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	8.22	14.16	10.19	10.37
2	3.7	7.97	6.12	6.84
3	3.81	5.45	5.02	2.84
4 and 5(Big)	-3.48	0.56	-3.4	0.61

h	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	0.03	0.034	-0.0079	0.00048
2	-0.023	-0.046	-0.17	-0.017
3	-0.14	0.13	0.23	0.081
4 and 5(Big)	-0.027	-0.065	0.18	-0.13

t(h)	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	0.61	0.71	-0.13	0.008
2	-0.46	-0.79	-2.55	-0.24
3	-2.02	1.86	2.83	0.76
4 and 5(Big)	-0.53	-1.17	1.56	-1.17

F	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	889.82	890.47	571.09	621.77
2	770.71	654.98	523.99	408.59
3	450.79	514.04	341.84	215.7
4 and 5(Big)	757.65	634.06	159.26	156.47

Table 15 Statistics of intercepts regression $R(t) - R_f(t) = a + b*[R_m(t) - R_f(t)] + e(t)$

a	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	-0.0047	-0.0048	-0.0023	-0.0016
2	-0.0056	-0.0036	-0.0059	-0.0038
3	-0.0061	-0.0036	-0.0015	-0.0016
4 and 5(Big)	-0.0051	-0.0073	-0.0016	-0.0053

t(a)	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	-3.08	-2.63	-1.19	-0.082
2	-4.13	-2	-3.08	-1.8
3	-3.28	-1.81	-0.64	-0.55
4 and 5(Big)	-3.62	-4.91	0.51	-1.74

Table 16 Statistics of intercepts regression $R(t) - R_f(t) = a + s*SMB(t) + h*HML(t) + e(t)$

a	Book-to-Market Equity Quintile			
Size Quintile	1(Low)	2	3	4 and 5 (High)
1(Small)	0.019	0.019	0.021	0.024
2	0.016	0.021	0.019	0.019
3	0.017	0.021	0.022	0.015
4 and 5(Big)	0.015	0.014	0.021	0.017

t(a) Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	4.11	4.29	4.7	5.18
2	3.54	4.34	3.84	4.15
3	3.55	4.04	4.26	4.01
4 and 5(Big)	3.29	3.05	3.73	3.12

Table 17 Statistics of intercepts regression $R(t) - R_f(t) = a + b*[Rm(t) - R_f(t)] + s*SMB(t) + h*HML(t) + e(t)$

a Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	0.0015	-0.0011	0.0011	0.0033
2	-0.0045	-0.00079	-0.003	-0.001
3	-0.0041	-0.0021	-0.00026	-0.00026
4 and 5(Big)	-0.0059	-0.0069	-0.0014	-0.0043

t(a) Size Quintile	Book-to-Market Equity Quintile 1(Low)	2	3	4 and 5 (High)
1(Small)	-1.88	-0.84	0.67	2
2	-3.3	-0.49	-1.65	-0.51
3	-2.22	-1.08	-0.11	-0.09
4 and 5(Big)	-4.21	-4.45	-0.43	-1.38