Does size matter?
An empirical study modifying Fama & French's three factor model to detect size-effect based on turnover in the Swedish markets

Bachelor thesis
Economics
Spring 2014

Daniel Boros
Claes Eriksson

Spelar storleken roll?
En empirisk studie med Fama & French:s tre-faktor modell modifierad för att undersöka storlekseffekt baserat på omsättning på den svenska marknaden
ABSTRACT

This thesis investigates whether the estimation of the cost of equity (or the expected return) in the Swedish market should incorporate an adjustment for a company's size. This is what is commonly known as the size-effect, first presented by Banz (1980) and has later been a part of models for estimating cost of equity, such as Fama & French’s three factor model (1992). The Fama & French model was developed based on empirical research. Since the model was developed, the research on the size-effect has been divided and today there are empirical studies contradicting its existence. Arguments against the size-effect are to some extent supported by the fact that there is no solid theoretical explanation for it. It seems however that market participants in the Swedish markets do adjust for the size.

A limitation of the Fama & French model is that market data is required for the estimation. Our starting point is to investigate if there is a presence of the size-effect in the Swedish markets using a modified version Fama & French model. In our modified model a proxy for the market value of the firm has been introduced, namely the firms turnover. This is motivated by the fact data regarding a company's turnover is available for private firms as well. In the case that size-effect is observable using the turnover as a proxy this would allow to extend the model to estimate the cost of equity for private firms. In the case where a consistent estimated marginal effect of the turnover is observed, our model could be used to estimate cost of equity with reasonable precision.

Historical data on Swedish companies from each of the OMX Large, Mid & Small cap lists is used in a regression setting to investigate if any statistical significant results can be observed on whether the logarithm of the turnover affects the expected return.

Our results indicate that the marginal effect of the turnover is positive, contradicting previous research and economic intuition that size of a company should be negatively correlated (or uncorrelated) with the expected return. By investigating the internal and external validity of the results, comparison to previous research and assessing data quality, we conclude that errors originating from these factors are not plausible to cause the unintuitive results. We therefore conclude that the use of turnover as a proxy for market value is not viable, which may be attributed to the fundamental relationship between the turnover and cost of equity in valuation formulas.
Conclusively we cannot draw any further conclusions regarding presence of size-effect in the Swedish equity markets and discard the possibility of using our modified model for estimating cost of equity for private firms.
PREFACE

We would like to thank Malin Gröndahl Maass for her help with making this thesis possible. Further, we would like to direct a special thank you to our supervisor, Göran Hägg, for continuous support and constructive criticism as well as our opponents, Ellinor and Pontus, for their feedback. Lastly, we would like to thank Niklas for sparking our interest in modelling cost of equity with company data.

Thank you!

__________________________________________
Daniel Boros

__________________________________________
Claes Eriksson

Linköping, 2014-06-09
# TABLE OF CONTENTS

1. Introduction .................................................................................................................. 6
   1.1 Background ................................................................................................................ 6
   1.2 Discussion of the problem at hand ............................................................................ 7
   1.3 Purpose and questions at issue ................................................................................ 8
   1.4 Scope .......................................................................................................................... 8
   1.5 Disposition of the thesis ........................................................................................ 9

2. Cost of equity .................................................................................................................. 10
   2.1 Cost of equity and the equity risk premium ............................................................ 10
      2.1.1 CAPM ............................................................................................................... 10
      2.1.2 The Fama-French Three Factor Model ............................................................... 11
      2.1.3 Arbitrage pricing theory .................................................................................. 11
      2.1.4 Additional theory on the cost of equity .............................................................. 12
   2.2 The size effect ........................................................................................................... 13
   2.3 Turnover as a proxy ................................................................................................. 15

3. Econometrics ................................................................................................................. 16
   3.1 OLS-regression ......................................................................................................... 16
   3.2 The R^2 measure ....................................................................................................... 17
   3.3 Functional form analysis ......................................................................................... 17
      3.3.1 The Logarithm transform ................................................................................. 17
   3.4 Residual analysis ..................................................................................................... 18
      3.4.1 Tukey-anscombe plot ..................................................................................... 18
      3.4.2 Quantile-quantile plot ................................................................................... 18
      3.4.3 Asymptotic properties ...................................................................................... 20
      3.4.4 Autocorrelation ............................................................................................... 20
   3.5 Students t-test .......................................................................................................... 21
   3.6 Robust regression .................................................................................................... 21

4. Methodolgy .................................................................................................................... 22
   4.1 Model ......................................................................................................................... 22
   4.2 Data ............................................................................................................................ 23
   4.3 Model estimation ...................................................................................................... 24
      4.3.1 Transform ......................................................................................................... 24
      4.3.2 Regression exercise ......................................................................................... 25
      4.3.3 Residual analysis ............................................................................................. 26
      4.3.4 Robustness analysis ......................................................................................... 26
   4.4 Methodological critique ......................................................................................... 27
1. INTRODUCTION

1.1 BACKGROUND

The cost of equity is one of the fundamental parts in many fields of finance, such as corporate valuation and capital allocation. It is however unobservable, which has led to decades of research in providing viable models to estimate the cost of equity. The Capital Asset Pricing Model (CAPM), attributed to Treynor, Sharpe and Leitner, is perhaps the most famous of these models. It is one of the cornerstones in modern portfolio theory and has been widely popular, partly due to its simplicity as it assumes that no other parameters than the company’s co-movement affect the estimate (and therefore only needs this estimate).

This fundamental result of CAPM was however questioned in 1980 when Rolf W. Banz presented his research on the effect of a company’s size on the observed returns of the stock. He suggested that smaller firms had higher return than larger firms than could be explained by CAPM. Eugene Fama and Kenneth R. French incorporated the effect of a company’s size in their proposed Three Factor model to estimate the risk premium of a corporation in 1992, which is presented below.

\[
E[r_{i}] - r_{f} = \alpha_{i} + \beta_{i} \left[ E[r_{m}] - r_{f} \right] + s_{i} \cdot SMB_{i} + h_{i} \cdot HML_{i} \quad (1)
\]

\( r_{i} \) = cost of equity, \( r_{f} \) = risk-free rate, \( r_{m} - r_{f} \) = market risk premium, \( SMB_{i} \) = difference in return between diversified portfolios of small and large firms, \( HML_{i} \) = difference in return for high/low book-to-market value

Although the model did not have the solid theoretical foundation CAPM did, instead based on empirical results, it did seem to provide more useful results. This spawned a significant amount of research on the subject, as well as investment firms eager to take advantage of the possible excess returns. Today the presence of the size-effect is questioned with several empirical studies (e.g. Horowitz et al., 2000) to support the notion that either the effect was never present or has disappeared since it was first discovered.

The studies that have been presented on the subject all entirely rely on market data (often from the US) and rightly so, since the models all are entirely based on market data. Although this is intuitive it proposes problems when estimating the cost of equity for private firms where no market data is available.
1.2 DISCUSSION OF THE PROBLEM AT HAND

The several models used to estimate the cost of equity differ quite extensively in their fundamental assumptions. To some extent, motivation for choosing between CAPM and Fama & French should be possible to observe in the markets. If the parameters used in Fama & French are observed to give increased explanation for the observed market data that would to some extent (given the choice between the two models) speak in favour of using this model. However, the divide in the research on the size-effect internationally, combined with the limited research in the Swedish markets, do not give any clear justification for the use of size-premiums in the Swedish markets. Whether the size-effect exists or not; Swedish practitioners seem to adjust their required return on investment for the size. The global auditing firm PricewaterhouseCoopers conducts an annual survey of the equity risk premium used by different actors in the Swedish markets, where the premium in use for the largest companies compared to the smallest differed on average by 2.2% in 2012. This provides a need for scientific research in whether there is actually any motivation for Swedish investors to use this premium.

Although they are two of the most commonly used models, CAPM and Fama & French, do not provide any solution of how to estimate the expected return on investment (i.e. the cost of equity) for private firms, there has to our knowledge been limited studies in how more fundamental data can be used to model the expected return. It seems to us that the turnover is intuitively highly correlated with the size of a company, which is our starting point for using the turnover to adjust for the size of the firm. Finding a functional relationship between the cost of equity and the size, described by turnover, would allow estimation for the cost of equity for private firms where market data is not available. This requires us to do the analysis at the specific company level, something that is not possible to do with the original model, as it adjusts for portfolios of companies with different sizes.

By analysing the Swedish stock market with an adjusted version of the Fama & French model using turnover as a proxy for the market value, we hope to give some insight in both whether there is a size-effect present in the Swedish markets as well as it is possible to use the turnover instead to estimate the size-effect for private firms. Furthermore, to be able to use the principle of the Fama & French model for estimating the equity risk premium (and by this the cost of equity) for private companies we would like to investigate whether there are consistent estimates for the marginal effect of the turnover at the equity risk premium.
1.3 PURPOSE AND QUESTIONS AT ISSUE

The purpose of this thesis is to assess whether the size-effect is present in the Swedish equity markets and to investigate if any robust conclusion can be drawn regarding the magnitude of the size-effect based on turnover.

In order to reach conclusions connected to the purpose of this thesis, we propose the following questions.

- Does a company's turnover seem to provide a viable proxy for estimating the size-effect in the Swedish markets?
- Can any conclusions regarding the presence of the size-effect in the Swedish equity markets be drawn?
- Can external validity and a functional relationship of the turnover be demonstrated through repeated estimation of the model's parameters for varied subsets of the available market data?

1.4 SCOPE

Firstly, as our main focus is the marginal effect of the turnover, other possible relevant fundamental factors from the companies' income statement will not be incorporated in our study.

Secondly, this study aims at taking a closer look at the Swedish markets and will therefore only use data from the NASDAQ OMX Stockholm large, medium and small cap lists. NASDAQ OMX Stockholm has been chosen, as it is the largest regulated stock market in Sweden, which gives us big data sample.

We will not consider additional costs of investing, i.e. transaction-costs, or taxes when studying the expected return on investments. Due to that not all companies have reported the results for 2013 at the time of this study (done in the spring of 2014), the study will truncate the time series after the 2012 observations.
1.5 DISPOSITION OF THE THESIS

The thesis will have eight major chapters and an overview can be seen in figure 1.

Chapter two will present commonly used models for the cost of equity, current research on the size-effect and discuss the use of turnover as a proxy for the size-effect.

Chapter three presents some additional theory for the econometrical tools we will use in this study we deem to be outside the scope of ordinary econometrics.

Chapter four explains the methodology used in our study. This includes a discussion of the model, the data and the regression model that will be used. The critique of the method will also be included in this chapter.

Chapter five presents our results, i.e. the transformation of the turnover parameter, the results of the regression and the internal and external validity of our model. However, no conclusions will be drawn regarding the results in this chapter.

Chapter six discusses our results. We will approach this in a way that analyses the results in a systematic manor where the discussion regarding the most fundamental parts is done last.

Lastly, we will present a summarised conclusion, which answers the questions at issue in our study in chapter seven as well as give suggestions for further research in chapter eight.
2. COST OF EQUITY

This chapter aims at giving an overview of the economic theory that this study relies upon.

2.1 COST OF EQUITY AND THE EQUITY RISK PREMIUM

The main focus of our study lies in modelling the expected return (or cost of equity, from the company’s perspective). The cost of equity (COE) for a company is the return shareholders require on their investment. This number is one of the fundamental parameters for any investment decision involving an equity investment. In contrast to lenders, who receive a predetermined interest rate on the capital lent (and are therefore only exposed to credit risk), equity investors are not guaranteed any level of return on their investment. Due to the importance of this parameter there has been a lot of research in modelling the expected return on an equity investment. In this section, we will discuss the most well-known of these. The excess return requirement of investors for investing in equity (and not in risk-free assets) is called the equity risk premium.

2.1.1 CAPM

The Capital Asset Pricing Model (CAPM) (Treynor, 1962; Sharpe, 1964; Leitner, 1965) is perhaps the most widely used model for describing the relation between risk and return of a company and could be considered one of the cornerstone for modern portfolio theory. It is a one factor model where beta often is estimated by linear regression. The core result is a set of formulas for the behaviour of excess return of a company \( r_i \), presented below.

\[
E[r_i] - r_f = \beta_i * (E[r_m] - r_f) \tag{2}
\]

where,

\[
\beta_i = \frac{cov(r_i, r_m)}{var(r_m)} \cdot \tag{3}
\]

Here \( r_m \) is the market return and \( r_f \) is the risk free rate. Note that \( E[r_m] - r_f \) is the expected market premium.\(^1\) Although being a realistic and comprehensible model with strong and intuitive predictions, it has been empirically criticized (e.g. Black, Jensen & Scholes, 1972; Fama & French, 2004) to the extent that some researchers has declared the model invalid. Furthermore, in 1978 Blume & Friend showed that a large portion of individuals are highly undiversified in their investments. It followed that variance is, in these cases, a better measure

\(^1\) The reader is presumed to be familiar with this model, and for further description we refer to any textbook covering modern portfolio theory.
of risk, whereas both CAPM and Arbitrage Pricing Theory both rely on covariance as the relevant risk measure.

Despite the criticism it is still widely used and since it is practically impossible to test the exact model (as this would require finding the true market portfolio, testing mean-variance efficiency (Roll, 1976)), some claim that the criticism is invalid. Levy (2010) also suggest that the model is valid but that empirical studies fail due to that investors rationality can be questioned, referring to Kahneman & Tversky’s studies which discredit the notion of “the rational investor”.

2.1.2 THE FAMA-FRENCH THREE FACTOR MODEL

Eugene Fama and Kenneth R. French developed their Three Factor Model to describe the relation between risk and return in 1992 (Fama & French, 1992). The main argument for this model was that several studies had shown that other factors could add to the explanation of the risk-return relationship. Fama & French saw that several of these factors were different ways of scaling a firms’ stock price and ended up with a model which had three factors, described in formula 1 and repeated below.

\[
E[r_i] - r_f = \alpha_i + \beta_i \left( E[r_m] - r_f \right) + s * SMB_i + h * HML_i
\]

\(r_i = \text{cost of equity}, \ r_f = \text{risk-free rate}, \ r_m - r_f = \text{market risk premium}, \ SMB_i = \text{difference in return between diversified portfolios of small and large firms}, \ HML_i = \text{difference in return for high/low book-to-market}\)

Here, SMB is the difference between return on diversified portfolios of small and big stocks and HML is the difference between stocks with high contra low book-to-market value. Note that the beta in this model is not the same as beta in the CAPM. These two parameters help in taking the size effect first presented by Banz (1980) (see separate section) and the explanatory power of high (or low) book-to-market ratio as presented by e.g. Stattman (1980).

2.1.3 ARBITRAGE PRICING THEORY

Another family of models are the Arbitrage Pricing Theory models. The fundamental argument for the arbitrage pricing theory (Roll & Ross, 1984) is that all expected risks will be incorporated in market prices. The risk premium that needs to be estimated is the exposure of a security to unforeseen events, systemic risk. These cannot be foreseen but the sensitivity to them can be measured for the specific security. Furthermore, there are only a finite set of these systemic risks. It is possible to create portfolios that have similar exposure to these factors, and two
assets with the same risk exposure must have the same expected return. This is expressed below (Roll & Ross, 1984) as

\[ r_i = a_i + b_{i,1} * f_1 + \cdots + b_{i,n} * f_n + \epsilon_i. \]  (4)

Here \( r_i \) is the return of asset \( i \), \( a_i \) is a constant for asset \( i \), \( f_n \) is the \( n \)th systemic factor, \( b_{i,n} \) is asset \( i \)'s sensitivity to the \( n \)th systemic factor and \( \epsilon_i \) is the diversifiable risk. The expected return of asset \( i \) can be expressed as

\[ E[r_i] = r_f + b_{i,1}(f_1 - r_f) + \cdots + b_{i,n}(f_n - r_f). \]  (5)

where \( r_f \) denotes the risk-free rate.

The arbitrage pricing theory has been criticized. For example Shanken (1982), argues that in Ross’ model there is no exact linear risk-return relation. Dhrymes et al. (1982) notes that without attributing economic meaning to the factors in the model, it would be difficult to gain any usefulness of the theory for either explanatory or predictive purposes.

---

### 2.1.4 ADDITIONAL THEORY ON THE COST OF EQUITY

Several other aspects have been identified that provides empirical explanatory power for the cost of equity. For example, the equity risk premium seems to have time-variation dynamics, expected to be cyclical in sync with the business-cycle. Ameer et al. (2013) confirmed this and found that the risk premium reached record levels after the 2008 crisis. The momentum in stock returns is also well-documented, as discussed by e.g. Lewellen (2002), i.e. that stock returns are positively autocorrelated.

Other parameters that have been found to explain cross-sectional variation are dividend yield and leverage (Berk, 1995). Xu (2007) examines the dividend yield and repurchasing yield as predictors of expected returns and finds that dividend yields seem to affect the expected return indirectly and repurchasing yield affect the expected return directly.

Some research on other, more qualitative, factors has also been made. E.g Ziegler et al. (2002) investigate different aspects of corporate social responsibility’s (CSR) effect on share prices between 1996 to 2001. They look at environmental and social performance. In the study they conclude that stocks in industries with high environmental performance has a higher average monthly return compared to companies in other industries with similar financial performance. On the other hand, social performance had a negative correlation with the stock returns (although this conclusion was less robust). Alter & Oppenheimer (2006) find that for short term movements, stocks with easily pronounced names consistently outperformed stocks with names.
that were hard to pronounce. This was valid for both a laboratory setting and for real world market data and they mention that: "Whereas financial analysts delve into the differential performance of industries and market sectors, a straightforward psychological principle cuts across these categories and predicts, quite simply and robustly, that companies with names like Barnings Incorporated will initially outperform companies with names like Aegeadux Incorporated".

Finally, we would like to mention the possible effect of illiquidity. Amihud (2002) investigates the illiquidity premium for individual stocks. A positive return-illiquidity relationship is observed. The effect is seen to be stronger for small stocks and proposes that the size-effect to at least some extent could be explained by the illiquidity premium.

### 2.2 THE SIZE EFFECT

One of the main goals with this study is to find out whether the size-effect is present in the Swedish markets. The size-effect was first discussed by Banz in 1980, but had been observed up to 40 years earlier. The study found that there was an empirical relationship between return on NYSE common stocks and the market capitalization of these stocks. The relationship was that smaller firms had higher risk adjusted returns than larger firms, but this effect was not observed to be linear in market value. The main effect was seen for very small firms. Banz concludes that there is a pronounced effect for small companies, but mentions that this effect is not stable through time (different sub-periods showed different results) and the study could not conclude whether market capitalization was the true cause of the effect or if it was a proxy for some other factor. The majority of research started in the early 90's after Fama & French presented their three factor model.

Banz does not offer any theoretical foundation for the size-effect, but mentions that it is possible that the lack of information for smaller companies introduces an "estimation risk" (a concept introduced by Klein & Bawa in 1977). Investors with asymmetric information will limit their diversification. It is shown by Banz, in his 1980 paper, that securities held only by a smaller fraction of investors have a higher risk-adjusted return. Chan et al. (1982) argues that the size-effect is perfectly compatible with a multifactor pricing equation based on the Arbitrage Pricing Theory as the higher average returns of smaller firms are compensation for the additional risk borne in an arbitrage free market.

Different studies observe different results when it comes to the existence of the size effect. For example, Fama & French (2011) examines four major geographical regions in search of empirical evidence of the size-effect. They conclude that in North America, Europe and Asia
Pacific there is no pronounced size-effect, apart from that the value premiums for all regions which is negatively correlated with the size. In Japan however, the same observations were not found.

Van Dijk (2011) reviews the validity and the debate of the size-effect. He quotes a multitude of empirical studies (we refer the reader to Van Dijk’s paper for further details), both in the US markets and internationally. He concludes that there is contradictory research where empirical studies suggest that the size-effect is “dead” as well as other research suggesting models for the size-effect. He also points out that in recent years the size-effect in the US markets has been pronounced and that more robust empirical research is needed in the European markets. Furthermore, he compiles the main theoretical arguments for and against the existence of the size effect to be:

- The size is a proxy for the exposure to several of the risk factors that contribute to the systemic risk
- The size-effect stems from being a compensation from trading costs or liquidity risk
- Investors are not fully rational; either due to the fact that investors prefer growth stocks, are overconfident about their abilities or the incomplete information on small firms. The observation that average household-investors tend to tilt their investment toward small stocks with higher risk is also observed by Barber & Odean (2000)
- The size-effect is a statistical fluke. Some research suggests that data mining is the foundation for the effect and that only the surprising results have been published. The time-variation of the effect (ranging from negative premiums to positive ones depending on the time studied) seem to be in line with the notion that the sample chosen will affect the outcome

Horowitz et al. (2000) use the conventional approach for investigating the size effect, by sorting the NYSE stocks into deciles based on market value. They analyse the effect with three methodologies; annual compound returns, monthly cross-sectional regressions and linear spline regression between 1980 and 1996. All three methodologies concluded that there was no observable size-effect during the time-period.

The discovery of the size-effect quickly gave rise to small-cap mutual funds, eager to take advantage of this opportunity. Horowitz et al. mentions that the size-effect might have disappeared since it was first discovered, or perhaps even gone into reverse.
2.3 Turnover as a Proxy

To be able to adjust for the size-effect for a private company, a model using fundamental data from the balance sheet and income statement is needed. Furthermore, using company specific data allow us to study the relationship between the parameter and the size-effect for the specific company and not looking at a portfolio level. By intuition, using the company’s turnover as a proxy for size seems to be a reasonable starting point. This intuition is further motivated by the way that the value of a company is estimated theoretically.

One of the simplest ways of valuing a company is Gordon’s formula (Gordon & Shapiro, 1956), which only requires three input-parameters for determining the value of a company.

\[
P = \frac{D}{r_E - g} \tag{6}
\]

Here \( P \) is the price of a company, \( D \) is the dividends (assumed to be constant perpetually), \( r_E \) is the cost of equity and \( g \) the assumed perpetual growth. The parameter for including the adjustment for size in the Fama & French model is the market value (i.e. \( P \) in formula 2). Assuming the net earnings ratio and the dividend payout ratio to be constant, higher turnover should increase the value since

\[
D = Turnover \times Net\ earnings\ ratio \times Dividend\ payout\ ratio \tag{7}
\]

The turnover, as we propose as a proxy for the value of the firm, would therefore directly influence the firm value and be a viable proxy for the market value.
3. ECONOMETRICS

The methodology (further discussed in the next chapter) for obtaining our results are econometrical methods. To be able to draw statistically significant results, we will need to assess the internal and external validity of the assumptions made and the data used. We therefore deem it necessary to start this section with our definition of the concepts internal and external validity. We have chosen to define these in the same manner as (Stock & Watson, 2011), namely;

"A statistical analysis is said to have internal validity if the statistical inferences about causal effects are valid for the population being studied. The analysis is said to have external validity if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings”.

The methodology section will thoroughly describe in which way the modelling approach aims at being valid in both the above senses and the discussion sections the indicators of this being fulfilled or not.

Further, to support the practical part of our study, we provide a short review of basic econometric theories essential in our study. This is complemented with the somewhat more involved tools, methods and concepts used in our validity assessments.

3.1 OLS-REgression

In the thesis, the methodology specified to generate results used with which our result questions may be answered is much based on OLS-regression. We recall that, for OLS-estimates to be unbiased and consistent, several assumptions have to be fulfilled (Stock & Watson, 2011). As control of these assumptions is essential in the assessment of the quality of our selected method, they are listed below;

1. Correct specification of the functional form, namely the linear relationships used in the regression are true.
2. Exogeniety, meaning that the expected value of the error is to be zero when conditioned on the included dependent variables. This means that the residuals may not contain any systematic deviations from zero and the errors and independent variables must be uncorrelated.
3. In the case of cross sectional data, e.g. data for the same individual at different time points, the assumption that each observation is drawn independently from the same distribution is added. This is commonly referred to as the observations being independent and identically distributed.
4. Large outliers for the dependent and independent variables are unlikely.

Further, the following two assumptions are be added when one aims at performing student’s t-tests for the estimates;

5. In order for the t-test to be valid, the assumption that the errors are normally distributed has to be made.

6. The errors are homoscedastic, meaning that the variance of the errors is constant across all observations. A violation of this assumption may be resolved by using robust regression, which is discussed below.

3.2 THE R$^2$ MEASURE

One way of assessing the explanatory power of our model is the $R^2$ measure. The $R^2$ is a measure commonly used to assess how much of the variability in the sample of the dependent variable, namely $\{y_i\}_{i=1}^n$, is explained by the sample of independent variables, $\{x_1, x_2, ..., x_j\}_{i=1}^n$ (Stock & Watson, 2011).

It is worth noting that $R^2$ increases whenever a new independent variable is added and is thus not optimal when deciding whether to include a variable or not into the model. In this paper, it will instead be used to assess the change in the models descriptive power when varying the data sets used to fit the same model with the same independent variables.

3.3 FUNCTIONAL FORM ANALYSIS

For the assessment on whether the decided upon functional form is correctly specified (or if e.g. a logarithm transform is needed), scatter plots between the dependent variable and each of the independent variables are useful tools (Stock & Watson, 2011). If the plots indicate a linear relationship, the currently used transforms of the dependent and independent variables should suffice.

3.3.1 THE LOGARITHM TRANSFORM

To achieve the wished upon linearity, a common transform used when the data is distributed with substantially heavier tails than the normal distribution is the logarithm transform, namely

$$ y = \log_b x \leftrightarrow x = b^y. \ (8) $$

The log-transform, as found in any introductory literature on calculus, has several useful characteristics, namely:
• It is a function mapping $\mathbb{R}^+ \rightarrow \mathbb{R}$
• The logarithm function is strictly increasing
• It is a monotone transform, i.e. $\forall \, i, j: \, x_i > x_j \rightarrow log_b(x_i) > log_b(x_j)$
• Used on a regression data-set, since it is the exponent of the base $b$, it will compress the data sample to a smaller range and distribute the data points more evenly over the new range.

The listed benefits suggest that the logarithm-transform is a reasonable candidate transform in our methodology.

### 3.4 RESIDUAL ANALYSIS

When the regressions have been run, among other measures, estimates of the coefficients and residuals are received. Since the residuals are themselves estimates of the error terms in the observations, they may advantageously be used to investigate whether the assumptions regarding the errors of the model, as stated in the OLS section, seem to be fulfilled. Performing thorough residual analysis in the regression results allows us to draw conclusions as to the internal validity of the OLS-estimates. There exist several graphical tools commonly used in residual analyses.

#### 3.4.1 TUKEY-ANSCOMBE PLOT

One of the qualitative graphical tools is the Tukey-Anscombe plot, where the estimated residuals are plotted on the $y$-axis against the fitted values of the model on the $x$-axis (Anscombe, 1973). In this way, one may investigate if the residuals are scattered randomly around zero, with constant variance, for the entire interval of fitted values. Thus, the Tukey-Anscombe plot enables assessment as to whether the errors are exogenous and homoscedastic.

#### 3.4.2 QUANTILE-QUANTILE PLOT

Another tool is the Quantile-Quantile plot. Page 236 in (Hult, Lindskog, Hammarlid, & Rehn, 2012) describes the Quantile-Quantile-plot (QQ-plot) in the generic case. However in our specific context, it will refer to the plot of the theoretical quantiles of the normal distributions against the empirical quantiles of a data set. It will be used to study the normality assumption of the errors. If the data indeed comes from a normal distribution, the empirical quantiles should indicate a linear function. Deviations from this line indicate thicker or thinner right and/or left tails than the normal distribution. In Figure 2 below, three common probability density
functions (pdfs) have been plotted, the standard normal, the student’s t with one degree of freedom and the standard log normal.


By the description of the QQ-plot, a sample drawn from the standard normal should follow a straight line in the QQ-plot. The QQ-plot for a sample from the student’s t distribution should indicate thicker right and left tails, namely, the lower quantiles should be for lower values and the higher quantiles for higher values than the standard normal reference. So, the QQ-plot should look like an inverted S. The log-normal distribution has a thinner left tail and a thicker right tail than the normal. Thus, the QQ-plot for a sample drawn from the log-normal should indicate higher values for all residuals, giving the plot a U-shape. This is confirmed by the QQ-plots in Figure 3.

FIGURE 3: QQ-PLOTS FOR A SAMPLE FROM (A) THE STANDARD NORMAL (LEFT), (B) THE STUDENT’S T WITH ONE DEGREE OF FREEDOM (MIDDLE) AND (C) THE STANDARD LOG-NORMAL (RIGHT).
### 3.4.3 Asymptotic Properties

In the investigation of the normality assumption regarding the OLS-errors, assessment of the distribution’s asymptotical behaviour of the errors is valuable. By investigating if the errors seem to converge against a normal distribution when increasing the dataset for the regression, a conclusion of the validity of this assumption may be drawn. This is done on the theoretical basis that the asymptotic of the empirical cumulative distribution is

$$
\hat{F}_n(x) \rightarrow a.s. \rightarrow F(x). \quad (9)
$$

Namely that it tends towards the true underlying distribution of the errors. This is done on the theoretical basis that the asymptotic of the empirical cumulative distribution function (van der Vaart, 1998).

### 3.4.4 Autocorrelation

Another important concept when studying data over time-period (called time-series) is how the datasets depend on previous observed data in the same time-series. In a time series setting, we have the set of $T$ random variables, $\{X_t\}_{t=0}^T$ and the set of corresponding observations $\{x_t\}_{t=0}^T$. These represent the variables and observations of the same quantity at the times $t = 0, 1, \ldots, T$. When trying to assess how this quantity evolves, it makes sense to investigate how it depends on the past values of the same quantity, namely how $X_k$ is correlated to $X_l, l < k$. This relationsship between a quantity and the quantity shifted time steps backwards is called the autocorrelation of shift $k - l$. Given a sample of $T$ observations, the sample autocorrelation of shift $h$ may be calculated as

$$
\hat{\rho} = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \quad (10)
$$

where $\hat{\gamma}(h)$ is the sample autocovariance of shift $h$, and is calculated as

$$
\hat{\gamma}(h) = \frac{1}{T} \sum_{t=1}^{T-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}) \quad for \quad -T < h < T \quad (11)
$$

where $\bar{x}$ denotes the arithmetic mean. The presence of autocorrelation is naturally plausible in economic time series, which is why plausibility and the severity of this assumption possibly being breached in our thesis will be discussed.
3.5 STUDENTS T-TEST

To investigate the internal validity of our proposed model we will use the students’s t-test to test the significance of the turnover variable.

When testing the null hypothesis; \( H_0: \beta_k = \beta_{k0} \), against the alternative hypothesis; \( H_1: \beta_k \neq \beta_{k0} \), the two sided student’s t-test will be applied. The null hypothesis is the hypothesis that there is no actual relationship between the dependent variable and the explanatory variable. The p-value obtained in the test states the probability of falsely rejecting the null hypothesis (Stock & Watson, 2011). In the context of assessing if a variable, e.g. the turnover is significant, we thus wish \( p \) to be small. The specifics and background of the t-test may be found in the appendix.

3.6 ROBUST REGRESSION

As mentioned earlier, using a robust regression method may mitigate a possible violation of the homoscedasticity assumption for the OLS. The method used in this paper is the one presented in (Holland & Welsch, 1977), namely the iteratively reweighted least-squares (IRLS) method, which, in short, solves an optimization problem, using an iterative method where each step consists of solving a weighted least squares problem (WLS). In the weighted least squares, an adjustment to the OLS-approach is made. Instead of minimizing the sum of squares of the normal residuals, namely \( \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \), we minimize the weighted residuals \( \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \), where each observation gets a weight. This allows us to decrease the weight of outliers and thus the models sensitivity to these. In IRLS, the WLS is initially run with some generic weight set \( \{w_i^0\}_{i=1}^{n} \). The algorithm then adjusts and standardizes the residuals of the WLS and use these and a robust weighting function to define a new set of weights, \( \{w_i^1\}_{i=1}^{n} \). The WLS is then run again, using \( \{w_i^1\}_{i=1}^{n} \) and the IRLS then produces \( \{w_i^2\}_{i=1}^{n} \) and so on, until a set of weights such that \( \sum_{i=1}^{n} w_i^M (y_i - \hat{y}_i)^2 \) is minimized is found. The method is implemented in the Statistical Toolbox of MATLAB and in the analyses of this thesis; this implementation is used for the robust regressions, to remedy the existence of heteroscedasticity in our sample.
4. METHODOLOGY

This chapter will give the reader an understanding of the methodology used in this essay, discussing both what data we used and how we used it. Lastly we will discuss the possible shortcomings of the methodology and how these are handled.

4.1 MODEL

As presented in the theoretical framework, there are several models aiming at estimating the expected return. In our study however, we need a model that is able to indicate the effect of a company’s size to describe the expected return.

We consider the Fama-French three factor model (equation 1). The model adjusts for company size, which is consistent with the aim of our study. It has also been previously used for studies of the size-effect, which allows for comparing our results with previous studies to some extent. Further, it adjusts for previous valuation of a company by having a parameter comparing market value to the book value. The model is also used in the study made by Horowitz et al. (2000) where they question the existence of the size-effect. Although we will not be using the exact same approach as Horowitz’ study, using the same model gives some comparability with their study.

The CAPM does not adjust for this as it assumes full diversification and that all company-specific data (except for the co-movement with the market) is irrelevant as one can be diversified. Therefore we will not use CAPM in our study.

Secondly, arbitrage pricing theory also proposes some obstacles for being viable to use in our study. The assumption that all expected risks are incorporated in the current prices directly contradicts that any size-effect should be observed. Furthermore, the model’s parameters are all systematic risks and unspecified. Consequently, the model is based on assumptions that do not fit with our study and ensuring a realistic implementation would be difficult.

The problem with using the Fama-French model is that we also look for a functional relationship between the size in terms of turnover and the cost of equity, to be able to extend the model for estimating cost of equity for private firms. As discussed in previous chapters, we propose a small change to the original Fama & French model and account for the size by introducing the turnover as a proxy for size. Our proposed model is stated in the equation below.

$$E[R_i] - R_f = \alpha + \gamma * f(TO_i) + \beta * (E[R_m] - R_f) + \theta * PTB_i , (12)$$
resulting in the following model for the regression analysis;

\[ R_i - R_f = \alpha_i + \gamma \cdot f(TO_i) + \beta \cdot (R_m - R_f) + \theta \cdot PTB_i + \varepsilon_i \cdot i = 1, ..., n \] (13)

In (12) and (13), \( R_i \) is the return for company \( i \), \( R_f \) is the risk free rate, \( TO_i \) is the turnover for company \( i \), \( R_m - R_f \) is the market premium and \( PTB_i \) is the price-to-book ratio for company \( i \). Note that we are using the price-to-book ratio instead of the book-to-market value which is proposed in the original model. However, the price-to-book is simply the inverse of the original model so the only effect we expect from this is a negative estimate of the associated parameter than what is usually observed.

4.2 DATA

We move on to present the data used for the regression of the model presented. The data collected is from NASDAQ OMX Stockholm using Reuters Eikon. The companies that data has been collected for are companies on NASDAQ OMX with turnover registered for two ends of a ten year period, between 2002 and 2012. Note that for each year, the date used is the 31st of December, so referring to 2012 means referring to the last day of 2012. Using data from these companies we believe that we to some extent avoid collecting misrepresentative data from e.g. companies with outlying returns due to unrealistic expectations and similar distorting factors. This will however introduce a survival-bias\(^2\), which we are aware of and will further discuss in our analysis.

Furthermore, the data is collected annually since the companies’ annual turnover is used, as well as the book value, and is from the same point in time. Data for each company each year is treated as an individual observation, i.e. the data is not treated as time-series. We expect that the price-to-book ratio will handle some of the dependence of previous performance (the value premium in the Fama-French model). In the case where data for any of the parameters is missing for a specific year and specific company, the data has been discarded and no assumption has been made. The result became a total of 1940 data points for 197 different companies over the above mentioned period.

The turnover (TO) is taken as Revenue for each company for each year in Reuters. From the over 1100 tickers the turnover was the first filtering parameter for which companies to use. To

\(^2\) Survivorship bias is a logical mistake easily made in scientific studies. When choosing a sample to use for a study, the process for data selection introduces a bias as the process selects a sample that has survived the requirements of the selection process. Carhart et al. (2002) demonstrated that for increased sample length, a higher bias is introduced for the average annual performance for US mutual funds.
do this, it was required that the companies we chose data for must have had a registered turnover in the year 2002 and 2012.

The market return \((R_m)\) is based on the annual relative difference between the OMX All Share index for each year. The data is the annual level at the end of each year between 2002 and 2013 (as the 2012 return is calculated as the difference between 2012 and 2013), divided by the former year’s level.

The risk free rate \((R_f)\) is simply taken as the yield to maturity of each corresponding year for the 10 year Swedish government bond. The maturity of 10 years has been chosen with the motivation that 10 years is a plausible investment horizon in the context of company valuation.

The price-to-book ratio (PTB) is taken from Reuters Eikon directly, where it is calculated as the market price at the specific time, for each company, taken in comparison to the book value of the equity at the same time.

The data will be used for annual observations for each company, as mentioned. However, as the different time points will be treated as independent data points the only regard that will be taken to time is to synchronize the data. This means that for company \(i\) at time \(t\), the rate, turnover and price-to-book ratio will be for time \(t\) while the return for company \(i\) as well as the return for the market will be based on the return between time \(t\) and time \(t+1\).

### 4.3 Model Estimation

The model specified will be estimated by using regression, for which the foundation was presented in the theoretical framework. The details of our approach will be described below.

---

#### 4.3.1 Transform

As discussed in the theoretical framework, the function \(f\) in formula 12 and 13 aims at finding a transform of \(TO\) that introduces a significant linear relationship with the untransformed variable \((R_t - R_f)\). The first transform investigated is \(f_1(x) = x\), which would indicate a linear relationship directly between \(TO\) and \((R_t - R_f)\). A scatter plot is used to qualitatively investigate if the assumption of a linear relationship is plausible.

Another plausible relationship between the turnover and the risk premium could be of the form \((R_t - R_f) \propto k\), where \(10^k = TO\). This would mean that the risk premium is proportional to the exponent to which 10 has to be raised to equal \(TO\), namely the logarithm of \(TO\) with base 10, the resulting transform if this is the case being \(f_2(x) = \log_{10}(x)\).
### 4.3.2 Regression Exercise

The regression exercise has been conducted over several increasing subsets of the data, where robust regressions have been run for each individual set using MATLAB. The decision process for which data points each set should contain may be seen below.

Primarily, the data points are sorted increasingly with regard to turnover, meaning that \((TO)_i < (TO)_j < (TO)_k\) for \(i < j < k\). The intervals are then defined as the sets \(I_m = \{(TO)_i : i = 1, \ldots, \frac{n}{K}\}\), meaning that the first interval contains the first \(\frac{1}{K}\)-th of the observations, the second \(\frac{2}{K}\)-th and so on, ending with the \(K\)-th interval containing all the observations. Note that \(I_n \subset I_{n+1}\), namely the intervals are increasing. Figure 4 gives a visualization of how the intervals are built.

![Figure 4: Visualization of how the intervals are defined.](image)

The reason for introducing these increasing sets of data points is to assess the external validity of the model. If it is the case that the estimate for \(\hat{\beta}\) in interval 1, denoted as \(\hat{\beta}_1\), is close to that of interval 2, \(\hat{\beta}_2\), the size effect estimated with interval 1 is also valid for interval 2. In the event that it is additionally observed that \(\hat{\beta}_m \approx \hat{\beta}_{m+1}\) and \(\hat{\beta}_{m+1} \approx \hat{\beta}_{m+2}\), inductive reasoning stipulates that it is plausible that the estimate \(\hat{\beta}_1 \approx \hat{\beta}_2 \approx \cdots \approx \hat{\beta}_{m+2} = \hat{\beta}_K\), and thus also the model, is valid for all possible turnovers, namely for \(TO \in \mathbb{R}^+\). This in turn would indicate that \(\hat{\beta}_m\) is not only valid in the population and setting it was estimated from, but also for all other settings and populations, i.e. it is externally valid.

In each regression, the p-value for \(\hat{\beta}_m\) is also calculated and investigated. As stated in the theoretical framework, the lower the p-value is, the lower is the probability of the rejection of \(H_0: \beta = 0\) being faulty.

In our setting, we have 1940 observations, namely \(n = 1940\). Prime factorization of 1940 yields \(1940 = 2 \times 2 \times 5 \times 97\). By this, in order for \(n\) to be divisible with \(K\), it is decided that \(K \in \{2, 4, 5, 10, 20\}\). For the main analysis, \(K\) is set to 5, yielding the interval sets seen in table 1.
### Table 1: Specification of the Five Intervals Used.

<table>
<thead>
<tr>
<th>Interval</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>First observation</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Last observation</td>
<td>388</td>
<td>776</td>
<td>1,164</td>
<td>1,552</td>
<td>1,940</td>
</tr>
<tr>
<td>Turnover, first [mSEK]</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Turnover, last [mSEK]</td>
<td>125</td>
<td>533</td>
<td>1,488</td>
<td>7,054</td>
<td>310,367</td>
</tr>
<tr>
<td>Number of observations</td>
<td>388</td>
<td>776</td>
<td>1,164</td>
<td>1,552</td>
<td>1,940</td>
</tr>
</tbody>
</table>

---

**4.3.3 Residual Analysis**

The notion of external validity only makes sense if the model also experiences internal validity. When we lack internal validity, the external validity just indicates that the model is invalid not only for the population and setting of the estimation, but for all settings and populations.

To investigate if the model is internally valid, residual analysis for all regressions is performed. As discussed in the theoretical framework, the biggest threats to internal validity in our setting are:

- **Endogeniety**, namely that the error is independent when conditioned on the dependent variables in the model; occurring here if $E[\varepsilon \mid T(O), (R_m - R_f), PTB] \neq 0$, and
- **Non-normal errors**, namely $\varepsilon \notin N(0, \sigma^2 I_n)$.

Note that the possible heteroscedasticity of the errors is accounted for when using robust regression. The normality assumption is controlled for through inspection of the QQ-plots and the histograms of the residuals received in each regression. By investigating that the histograms are fairly well centred around zero, the assumption of exogeniety is also controlled for.

---

**4.3.4 Robustness Analysis**

In order to investigate the results’ sensitivity to the data, the regression series are run for interval divisions corresponding to $K = 4$ and $K = 10$ as well. Also here, the behaviour of the estimates for the (TO)-parameter and the p-values are recorded as the regression is run on the increasing intervals. By investigating these values when $K = 4$ and $K = 10$ to those obtained...
when $K = 5$, the robustness of the model towards data may be assessed. This is as it seems unlikely that statistical flukes which appear as a result of our choice for the intervals are present for other choices of intervals as well.

### 4.4 Methodological Critique

Our study is mainly a quantitative study, which makes it highly dependent on the data. We believe that our dataset is large enough to give statistically significant outcomes. However, the data is collected from a third party, which gives room for errors in the data which we have not been able to detect. This includes that adjustments in the data have been done by the third party. Also, due to the large amount of data, a more rigorous quality check of the data has not been possible from a time-perspective due to the limited scope of this study.

The transformation of the turnover will be made by an inspection of the plotted graphs of turnover against the corresponding return. This introduces a qualitative aspect to our study as the assessment of the transformation to use will be qualitative.

To the extent it is possible, we will chose intervals for the regression exercise that will not distort the results. Predefining these to be simply an increasing subset of our entire dataset should give some objectivity, but the choice could still affect our outcome. We hope to give more robustness to our study by verifying our result with regression over different interval sets.

As we will treat observations for the same company as separate data points, no consideration will be taken for either which year or which company the observation is made. This introduces one main problem. Possible autocorrelation for a company's returns between years will not be observable in the study. Autocorrelation could possibly be observed either for one singular company's performance (i.e. a company that performs well tend to continue to do so) or due to macro factors over the years. Since the data will be sorted based on turnover, any autocorrelation in the residuals based on either of these will not be observable. To some extent, this will be mitigated by our model accounting for past performance of a company with the price-to-book variable. Also, the risk free rate should to some extent account for macro-factors.

The data is collected for the large, mid and small cap lists on NASDAQ OMX Stockholm. Since these are different lists, there could be a need for accounting that the risk premium might be affected simply by which list the specific stock is listed on. This will be out of scope for the study to investigate further, but could to some extent distort the results.

The t-test used to investigate the significance of the parameters in each regression analysis assumes that the regression errors are normally distributed. This is an assumption that could
possibly be wrong, which would distort the results of the performed t-tests. The risk of this is however considered minor as the normality of the residuals is controlled for in all of the regressions.

Some companies included in our study could present abnormal returns between years due to high expectations from the market or situations giving similar effects. This could be, for example, for pharmaceutical companies with drug close to approval or other companies where the market only reacts to potential future earnings. To avoid these outliers in our sample we have used companies that are at least 10 years old. This introduces the previously discussed survival bias, but allows for reducing the number of outliers.

Theories that are the foundation for the study could clearly be a source of possible errors. To avoid this, literature and papers used are from reliable sources; well-known published literature and peer-reviewed journals. Internet-based sources are only used where the source is well known and considered reliable. This should give a solid theoretical foundation for the study.
5. RESULTS

There have been several aspects to the exercise that gave us the results needed for our study. We will walk through these in the order they were done in the study.

5.1 TRANSFORM

As mentioned, we use visual inspections of the plotted values of the turnover to the risk premium to investigate the linearity of the relationship. We present the resulting graphs used below, as well as a short explanation of their interpretation.

In upper plot of Figure 5, the scatter plot with risk premium, $R_p$, on the y-axis and turnover, TO, on the x-axis may be seen. For larger values of turnover, the assumption of a linear relationship between the two seems plausible. However, if a linear relationship exists, it does not seem constant, especially not when approaching lower bounds of the turnover spectrum. Below, the histogram for (TO) may be found. The histogram indicates a very high occurrence of observations of low to very low turnover, as opposed to those in the turnover class of hundreds of billions of SEK.

FIGURE 5: SCATTER PLOT WITH TO VS. RP (UPPER PLOT) AND HISTOGRAM OF TO (LOWER).
The plots in Figure 6 contain the same plots as seen in Figure 5, but with TO transformed to $\log_{10}(TO)$. In the scatter plot, a much clearer linear relationship is identifiable in the entire $\log_{10}(TO)$ scope. Further, the scatter plot, together with the histogram, indicates that by transforming the turnover, a sample with similar data point density over the entire scope and less extreme observations is obtained. On the basis of this, $f_2(x) = \log_{10}(x)$ is the transform that has been used in the regression exercises.

**FIGURE 6: SCATTER PLOT WITH LOG(TO) VS. RP (UPPER PLOT) AND HISTOGRAM OF LOG(TO) (LOWER).**

### 5.2 REGRESSION EXERCISE

The results of the regressions on the initial five increasing turnover intervals $I_i$, $i = 1, \ldots, 5$, as specified in the methodology section, may be seen in Table 2. For each interval, the estimated values of the parameters, $\alpha, \gamma, \beta$ and $\theta$ are presented, together with the p-value when performing a t-test for each of the four $H_0: \alpha_i/\gamma_i/\beta_i /\theta_i = 0$. The column containing $\beta$, namely the estimate of the effect of $\log_{10}(TO)$, is highlighted.
TABLE 2: RESULTS OF THE REGRESSION PARAMETER ESTIMATIONS FOR THE FIVE MAIN INCREASING INTERVALS SPECIFIED IN THE METHODOLOGY SECTION.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Intercept</th>
<th>Log(TO)</th>
<th>Market risk prem.</th>
<th>Price to book</th>
<th>R^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.20256</td>
<td>0.094233</td>
<td>1.0238</td>
<td>-0.0071808</td>
<td>0.317</td>
</tr>
<tr>
<td></td>
<td>0.010246</td>
<td>0.053907</td>
<td>1.13E-15</td>
<td>1.06E-07</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.20651</td>
<td>0.088358</td>
<td>0.98156</td>
<td>-0.0070752</td>
<td>0.315</td>
</tr>
<tr>
<td></td>
<td>0.00021844</td>
<td>0.00085205</td>
<td>5.09E-37</td>
<td>3.44E-10</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.2039</td>
<td>0.076297</td>
<td>0.99963</td>
<td>-0.0064919</td>
<td>0.326</td>
</tr>
<tr>
<td></td>
<td>2.75E-06</td>
<td>1.66E-05</td>
<td>4.58E-69</td>
<td>6.40E-11</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.18665</td>
<td>0.062633</td>
<td>1.0604</td>
<td>-0.006184</td>
<td>0.353</td>
</tr>
<tr>
<td></td>
<td>1.47E-07</td>
<td>1.07E-06</td>
<td>1.35E-115</td>
<td>1.68E-11</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.13554</td>
<td>0.038622</td>
<td>1.0708</td>
<td>-0.0054116</td>
<td>0.381</td>
</tr>
<tr>
<td></td>
<td>1.80E-07</td>
<td>1.56E-06</td>
<td>2.16E-168</td>
<td>7.91E-11</td>
<td></td>
</tr>
</tbody>
</table>

We observe from the log_{10}(TO) column of Table 2 and Figure 7, visualizing the observations:

- The p-value decreases, meaning that the significance level at which $H_0$ may be rejected increases, when including companies with larger and larger turnover.

- When increasing the sample to include larger companies, $\hat{\gamma}$ becomes smaller and smaller, indicating that the influence of the logarithm of the turnover decreases with the size of the companies. Furthermore, the estimate of the turnover parameter implies that the risk premium increases with larger companies.
It is also noteworthy that $R^2$ increases when extending the sample, indicating that the more data points are used, the better the variability in the model is captured.

When using the entire data sample, the linear relationship obtained between $\log_{10}(TO)$ and the risk premium is visualized in Figure 8. Note that the most extreme outliers have been truncated from the smaller sub plot, in order to properly visualize that the linear relationship is not zero.

![Figure 8: The fitted linear relationship between log(Turnover) and the risk premium.](image)

5.3 Residual Analysis

In Figure 9, the QQ-plot and the histogram of the residuals for the regression when run on interval 1 may be seen. From the QQ-plot we see that the residuals deviate from normality mainly in the right tail. The histogram however indicates that this right-skewness is mainly caused by a few extreme outliers and that apart from these, the distribution seems fairly normal.
FIGURE 9: QQ-PLot (Left) and Histogram (Right) for the Residuals from Regression on Interval 1.

Figures 10 to 12 show the QQ-plots and histograms for the residuals obtained in the intervals 2, 3 and 4. The QQ-plots consistently show that the right tail becomes thinner and thinner when the sample is extended. This is confirmed by the histograms, which indicate a convergence towards normality.

FIGURE 10: QQ-Plot (Left) and Histogram (Right) for the Residuals from Regression on Interval 2.

FIGURE 11: QQ-Plot (Left) and Histogram (Right) for the Residuals from Regression on Interval 3.
Finally, the QQ-plot and histogram for the residuals from the regression where the entire data sample is included may be seen in Figure 13. The trends observed above may be extended to the residuals in the entire interval. Both plots show that the assumption of residual normality is plausible.

### 5.4 Robustness Analysis

In Table 3, the regression results when splitting the log$_{10}$(turnover) interval into four increasing intervals, rather than five, are presented. The highlighted column for log$_{10}$(TO) and the visualization of the same in Figure 14 show the same results as in the case with five intervals, namely that

- The p-value decreases, meaning that the significance level at which $H_0$ may be rejected increase, when including companies with larger and larger turnover.
- When increasing the sample to include larger companies, ŷ becomes smaller and smaller, indicating that the influence of the logarithm of the turnover decreases with the size of the companies.
- The parameter estimate shows that increasing turnover will have a positive marginal effect on the risk premium.
Table 3 presents the regression results when the data is split into four increasing intervals. As in the case when investigating the results for two intervals, the parameter estimate and p-values for \( \log_{10}(TO) \) both decrease with increased intervals. This may also be seen in Figure 15.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Intercept</th>
<th>Log(TO)</th>
<th>Market risk</th>
<th>Price to book</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.19127</td>
<td>0.084285</td>
<td>1.023</td>
<td>-0.0074732</td>
<td>0.315</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0087078</td>
<td>0.04222</td>
<td>1.02E-19</td>
<td>1.03E-08</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.20572</td>
<td>0.083884</td>
<td>0.97318</td>
<td>-0.0069057</td>
<td>0.311</td>
</tr>
<tr>
<td>p-value</td>
<td>0.000027859</td>
<td>0.000093496</td>
<td>9.32E-49</td>
<td>6.55E-11</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.19803</td>
<td>0.068939</td>
<td>1.0443</td>
<td>-0.006056</td>
<td>0.342</td>
</tr>
<tr>
<td>p-value</td>
<td>1.47E-07</td>
<td>9.41E-07</td>
<td>8.74E-102</td>
<td>1.06E-10</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.13554</td>
<td>0.038622</td>
<td>1.0708</td>
<td>-0.0054116</td>
<td>0.381</td>
</tr>
<tr>
<td>p-value</td>
<td>1.80E-07</td>
<td>1.56E-06</td>
<td>2.16E-168</td>
<td>7.91E-11</td>
<td></td>
</tr>
</tbody>
</table>

Figure 14: \( \log(TO) \) parameter estimate when increasing \( \log(T0) \) scope (blue line and circles) and the corresponding p-value (green line and circles), four intervals.
### Table 4: Regression Results When the Dataset Is Divided Into 10 Increasing Intervals

<table>
<thead>
<tr>
<th>Interval</th>
<th>Intercept</th>
<th>Log(TO)</th>
<th>Market risk prem</th>
<th>Price to book</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.21953</td>
<td>0.1949</td>
<td>0.84158</td>
<td>-0.010204</td>
<td>0.345</td>
</tr>
<tr>
<td>p - value</td>
<td>0.05167</td>
<td>0.00068248</td>
<td>6.82E-04</td>
<td>5.33E-08</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.20256</td>
<td>0.094233</td>
<td>1.0238</td>
<td>-0.0071808</td>
<td>0.317</td>
</tr>
<tr>
<td>p - value</td>
<td>0.010246</td>
<td>1.1306E-15</td>
<td>1.13E-15</td>
<td>1.06E-07</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.18556</td>
<td>0.077987</td>
<td>0.96449</td>
<td>-0.0070016</td>
<td>0.302</td>
</tr>
<tr>
<td>p - value</td>
<td>5.42E-03</td>
<td>2.04E-22</td>
<td>2.04E-22</td>
<td>1.82E-08</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.20651</td>
<td>0.088358</td>
<td>0.98156</td>
<td>-0.0070752</td>
<td>0.315</td>
</tr>
<tr>
<td>p - value</td>
<td>2.18E-04</td>
<td>5.09E-37</td>
<td>5.09E-37</td>
<td>3.44E-10</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.20572</td>
<td>0.083884</td>
<td>0.97318</td>
<td>-0.0069057</td>
<td>0.311</td>
</tr>
<tr>
<td>p - value</td>
<td>2.79E-05</td>
<td>9.32E-49</td>
<td>9.32E-49</td>
<td>6.55E-11</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-0.2039</td>
<td>0.076297</td>
<td>0.99963</td>
<td>-0.0064919</td>
<td>0.326</td>
</tr>
<tr>
<td>p - value</td>
<td>2.7507E-06</td>
<td>4.5786E-69</td>
<td>4.58E-69</td>
<td>6.40E-11</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.19471</td>
<td>0.068697</td>
<td>1.0255</td>
<td>-0.0063809</td>
<td>0.338</td>
</tr>
<tr>
<td>p - value</td>
<td>6.92E-07</td>
<td>1.28E-90</td>
<td>1.28E-90</td>
<td>1.73E-11</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.18665</td>
<td>0.062633</td>
<td>1.0604</td>
<td>-0.006184</td>
<td>0.353</td>
</tr>
<tr>
<td>p - value</td>
<td>1.4745E-07</td>
<td>1.3511E-115</td>
<td>1.35E-115</td>
<td>1.68E-11</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-0.18576</td>
<td>0.059656</td>
<td>1.0706</td>
<td>-0.0053222</td>
<td>0.367</td>
</tr>
<tr>
<td>p - value</td>
<td>1.63E-09</td>
<td>3.92E-140</td>
<td>3.92E-140</td>
<td>1.01E-09</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-0.13554</td>
<td>0.038622</td>
<td>1.0708</td>
<td>-0.0054116</td>
<td>0.381</td>
</tr>
<tr>
<td>p - value</td>
<td>1.801E-07</td>
<td>0</td>
<td>2.16E-168</td>
<td>7.91E-11</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 15:** Log(TO) Parameter Estimate When Increasing Log(T0) Scope (Blue Line and Circles) and the Corresponding P-Value (Green Line and Circles), Ten Intervals.
6. ANALYSIS AND DISCUSSION

Our main result is that the parameter for the logarithm of the turnover seems to be significant in all subsets of the whole dataset studied. The parameter estimate does however indicate a positive relationship between the turnover and the risk premium, i.e. that there is a marginally higher expected return for companies with higher turnover. The significance is increasing when including larger companies in the larger subsets and from subset three to five, the probability of Type I errors are essentially zero. However, it should be noted that the increasing amount of data points used affects the precision of the t-tests and therefore the method will inherently be likely to show increased significance by just expanding the interval studied. This means that the increased significance can stem from both observed size-effect and the size of the dataset.

The other observation made is that the estimation of the turnover parameter is decreasing when including more data points with companies having a higher turnover. This indicates that the marginal effect of the turnover on the expected return becomes less for companies with higher turnover.

Since our results significantly differ from what we expect to observe based on previous research and intuition we will analyse possible sources of errors. We begin with analysing the output of the study (i.e. results) and move back through the thesis to end with analysing our input and model.

6.1 THE TRANSFORMATION

Using the transformation of the logarithm instead of the turnover makes sense both from a mathematical and economical point of view. The heavily right-skewed data sample for turnover is, as expected (as seen in the Results chapter), transformed into a more homogenous dataset. Since the transform is monotone, the internal order is maintained and thus it is still the marginal effect of the turnover that is investigated in the regression exercise.

From the economic perspective our argument is that using the logarithm with base 10 will distribute the companies according to the magnitude of the turnover. I.e. when not using the logarithm, an increased turnover from between 100 MSEK to 200 MSEK (100% increase) should affect the expected return the same amount as an increase from 1000 MSEK to 1100 MSEK (10% increase). Instead, when using the logarithm, one assumes that an increase from 10 to 100 MSEK should affect the expected return in a similar fashion as an increase from 100 to 1000 MSEK.

---

3 Type 1 errors is the incorrect rejection of the null hypothesis
The use of the logarithm transform therefore does not seem likely to have distorted our results since it is simply used to achieve a more linear relationship between the turnover and the expected return, retaining all desired properties of our data.

6.2 ANALYSIS OF INTERNAL VALIDITY

Another possible source of errors is if the internal validity of our results implies that the results obtained do not have any statistical significance. When analysing the internal validity, one main observation that was seen in the results was that with increasing number of data points, i.e. including a larger subset containing data for larger companies, the normality assumption of the residual errors becomes increasingly valid. As explained, this increases the validity of the p-values, i.e. the measures of the significance of the parameters, obtained from the t-tests. It is also noteworthy that the residual errors are centred around zero which implies exogeneity. In other words, the errors do not seem to be correlated with the independent variables.

However, our method does not allow observing any time-dependency for specific companies’ expected return over the timespan in the analysis. This is a known possible weakness in our methodology since positive autocorrelation in stock returns is an observable phenomenon. Due to the fact that survivorship-bias also may have been introduced during the data selection process, it is possible that the autocorrelation may have been introduced or increased. If this is the case in our cross-sectional data set, possible dependence across the sorted data set would be the effect. This would suggest that it is possible that the assumption of independent identically distributed (i.i.d) data points may be false. If the observations really are i.i.d, then the asymptotic behaviour of the empirical cumulative distribution function implies that for increasing data sets the error distribution should converge to the true distribution (here, the normal distribution). Since this is what we observe for the larger datasets, we deem the t-tests in our study, and thus the internal validity, as reliable.

6.3 ROBUSTNESS

Our method for studying the robustness of our model is a very hands-on approach. The aim is to investigate possible bias arising from the choice of our subsets of the entire dataset. Therefore the only thing that is being changed in these regression exercises is the choice of intervals. When applying the technique, the results obtained indicated the same results as for five intervals when choosing ten and four intervals instead. When using four intervals, the results look essentially the same as in the original regression exercise. However, using ten intervals we observe some fluctuations in the p-value and estimated parameter for the logarithm of the turnover. The
strictly decreasing values, which were previously observed, now fluctuates some in the first few intervals. We do however attribute this to the fact that the first few intervals now become substantially smaller than the ones previously used. It thus becomes likely that for these the assumptions for the t-tests are violated and that outliers might get large enough significance to affect the estimation of the parameters. Since these fluctuations are small, we still consider the ten interval regression exercise to support our results.

6.4 THE OTHER PARAMETERS

Since the remaining variables of the model we used were the same ones used in the Fama & French three factor model, we did not transform these. We observe that both the price-to-book parameter, as well as the market risk premium, has high significance for all intervals, implying that the variables are useful in the model. The market risk premium parameter estimate is close to 1 for every interval. This makes intuitive sense since the market risk premium and expected return variables are both partly built up by the risk free rate.

The price-to-book parameter estimates goes closer to zero for increasing intervals, similar to the estimates seen for the turnover parameter estimate. We observe a negative parameter estimation for the price-to-book variable. This is consistent with what we expect to see, since the usual variable is the inverse (book-to-market) with a positive parameter estimation (e.g. Fama & French, 2011). A high price-to-book ratio implies a high valuation compared to the asset value of the company. Although this might indicate a company with high returns, it is likely that several of the companies are overvalued and therefore will at some point not achieve the returns to defend this valuation. Similarly, undervalued companies can achieve returns, which would drive comparably higher return on equity.

Observing parameter estimates that are in line with what we would expect from previous studies for the unmodified variables, but not for the turnover (which we did introduce), does to some extent imply that the turnover parameter are not a viable proxy for the size. We will return to this discussion.

Lastly, it is worth mentioning that with increasing interval, the explanatory power, measured by the $R^2$ factor, is increasing. This should imply that the variation in the observed dataset is increasingly well-described by the model with larger intervals. However, as with the p-values, this could be explained to some extent by increasing the number of data points used.
Our study suggests that there is a positive size-effect in the Swedish equity markets, assuming that turnover is a valid proxy for the size of a company. As it has two major differences towards the previous research (the model alternation and being specifically for the Swedish markets) the comparisons will be with similar subject but for other geographical regions and using other models. The results contradict most studies done at the subject proposing either no observable size-effect or a negative relationship between size and expected return. Van Dijk (2011) mentions a study by Dimson and Marsh (1999) observed the size-effect in the U.K. as it was 5.9% per year between the years 1955-1988 while -5.6% per year between 1989-1997.

As is common for studies done in this area, our model does lack some fundamental theoretical foundation and is based on the models presented by previous empirical studies. Earlier we have presented that there are several different theories for the size-effect, and due to the lack of theoretical foundation, our results should be assessed from each of these perspectives in the manner it is possible. One of the theories is that the size-effect is a compensation for illiquidity and transaction costs. Our results are clearly not in line with the illiquidity theory. Regarding the transaction cost premium proposed, this theory assumes that household investments are tilted toward small stocks, which had previously been observed. With the financial crisis as well as the Euro-crisis being recent, it is possible that the risk-appetite of these investments has been mitigated and that more investments go into the safer big stocks. If this could be the case the argument could be reversed. We are however not able to investigate this further due to the limited scope of our study.

One explanation Banz gives in his original paper on the size effect (Banz, 1980) is that investors will demand a compensation for the estimation risk as a consequence of limited available information. Investors would not be keen on investing in smaller stocks (since these generally have less information available), as this would be more risky, and therefore the premium is needed to compensate for this. This theory is not in line with our results. With the advancements in information technology it is likely that the information flow for smaller companies has increased significantly and thus mitigating the effect this theory proposes.

Another theory of the size-effect is that size is simply a proxy for some of the factors of systemic risk. We are unable to draw any conclusions on whether this is the case from our results, but they should imply that there is some systemic risk factors for which increased size should increase the exposure. This does not seem very intuitive, but possible explanations would be e.g. that smaller companies are not as often internationally active and that larger companies therefore would have more exposure to economies currently under pressure.
Finally, the possibility of the size-effect being a statistical fluke should be brought up. The divide in research, as well as the lack of a solid theoretical foundation, makes the suggestion that this could be the case. The results of our study could be argued to support this theory, as well as contradict it. The results contradict most previous studies, supporting the notion of the size-effect being a statistical fluke. However, the rigorous testing of the internal validity and external validity suggests that for the dataset it is unlikely that the observations are statistical flukes.

**6.6 THE MODEL AND DATA**

From the above discussion we conclude that the transform is unlikely to distort our results and the significance of the estimated parameters suggests that we are not simply getting meaningless results. However, the results obtained are not consistent over different data sets, based on our robustness study, which implies instability in the positive size-effect we are observing. Lastly, we are not able to reconcile our results with any of the theories presented for the existence of the size-effect. This leads us to assess the input to our study; the model and the data.

The data used in the study comes from Reuters Eikon. Although we have limited quality control of our data (as mentioned), we believe that, since Reuters is a recognised provider of financial data, the errors from the source are limited and should not able to distort the results to the observed extent. Our use of the data on the other hand does introduce sources of errors. While using the data at a company specific level (i.e. turnover and price-to-book), we have taken little consideration to other company specific factors. This includes industry, which in turn includes different profit margins, growth rates etc. Since both profits and growth will clearly affect the expected return of a company, not adjusting for these will introduce errors. A company having a turnover of 1 BSEK and profits of 1 MSEK will not be valued to the same price as a company with the same turnover but 100 MSEK in profits. In our study however, we do only account for the effect of the turnover. Possible differing expectations of the returns for different time-periods are not adjusted for which will in turn introduce variation in our input that could distort our results. More specified studies where only certain industries are considered might get different results as the data should have less variation.

The most fundamental source of error, which we chose to discuss lastly after trying to eliminate all other sources as the cause, is the model itself. As discussed above, the variables used according to the Fama & French model does seem to yield the results we are expecting to see, but the variable we have specified did not. The motivation for choosing the turnover was based both on its characteristics (i.e. being available for private firms, allowing study at a company
specific levels) as well as the intuition that the turnover should be a good measure of size. The theoretical motivation we gave was based on the fundamental relation between a company's size and its dividends, according to Gordon's formula;

\[ P = \frac{D}{r_E - g}. \]

This in turn was related to the turnover by using the relationship

\[ \text{Dividends} = \text{Turnover} \times \text{Net earnings ratio} \times \text{Dividend payout ratio}. \]

Applied in practice, Gordon's formula does not yield precise results and can thus not be used to estimate precise functional relationships between the parameters. However, it should give some indication of how these relate. Our argument was that the dividends, and therefore also the size, relate to the size of the turnover. This should be a valid conclusion. What we failed to assess how the cost of equity could also be related to the turnover, as well as size, by this formula. Combining the relationships above and solving for the cost of equity we get the following result:

\[ r_E = \frac{\text{Turnover} \times \text{Net earnings ratio} \times \text{Dividend payout ratio}}{P} + g \]

This relation shines some light on the results of our study. Firstly we can observe the relation between the cost of equity and the price, where we have an inverse relationship. An inverse relationship between the price and the cost of equity is in line with what is commonly observed, that increased price reduces the cost of equity. Secondly, and most importantly, we observe a positive relationship between the turnover and the cost of equity; i.e. a marginal increase of the turnover increases the cost of equity. This is in line with what has been observed in our study. Since we do not expect these relationships to be precise, we do not know which relationship is stronger. We assumed that the turnover affects the price, which is likely to be true, but we now also observe that it is likely to affect the cost of equity as well. If in fact the turnover affects the cost of equity more than it affects the price, which our study indicates, then our original assumption of the turnover as the proxy for size is not valid. We deem this as the main source of the unintuitive results we have obtained.
The purpose of this thesis was to study whether there is a size-effect in the Swedish markets and whether a functional form for estimating this based on turnover could be observed. To do this, we have used data collected from NASDAQ OMX Stockholm and an alternate version of the Fama & French three factor model. We observed that transforming the turnover with the logarithm with base 10 gives a somewhat linear relationship between the company return and applied econometric tools to estimate this relationship. Using a large data set for companies with over 10 year history we observe that for this dataset there exists a positive size-effect in the Swedish equity markets, but that there is no consistency in exactly how this depends on the logarithm of the turnover. We also observe that the marginal effect of the logarithm of the turnover on the expected return is decreasing when using datasets that includes larger companies. We obtain similar results for each choice of subset definition.

After ruling out the transform and insignificant parameter estimations as the cause of error, deeming the results robust and being unable to reconcile our results with previous studies we deem the input to be the reason for our unintuitive relationship between turnover and expected returns.

Using a company's turnover as a proxy for size does not seem to be viable. We conclude that from the results we have observed it is likely that the turnover has a stronger positive relationship with the cost of equity itself than it has with the size of the company, which in turn then would affect the cost of equity. Neither could we find any consistent estimation of how strong the effect of a company's turnover is on the cost of equity, suggesting that no functional relationship could be observed.

Due to the misspecification of the model used in our study, we cannot draw any conclusions regarding the presence of a size-effect in the Swedish markets. Our sole observation regarding this is that turnover does seem to increase the expected return.

Based on our results, we do not recommend using the turnover as a proxy for size when estimating the cost of equity for private firms. If a model based on fundamental company data are to be used for estimating the cost of equity for private firms, careful consideration regarding the relationships between the variables need to be taken. Although a relationship between the turnover and cost of equity is found, the inconsistencies in the parameter estimations suggest that the positive relationship is not useful for practical application. Lastly, we are unable to answer whether today's use of size-premiums in the Swedish markets is motivated or not.
We suggest the following areas for further studies to build on our results.

1. Studying the size-effect in the Swedish equity markets using the original Fama & French model to assess whether today’s use of size-premiums are motivated.

2. Are consistent parameter estimation observable when studying specific industries and adjusting for data of different time-periods?

3. Could other fundamental data from a company’s balance sheet and income statement be used as more successful variables when estimating cost of equity?

4. Could other models (e.g. Arbitrage Pricing Theory) yield other results?
BIBLIOGRAPHY


%% Data processing script
% Daniel Boros, Claes Eriksson, Bachelor's thesis, May 13th, 2014
% Used for cleaning of data.
% Clear the workspace, close all figures and clear the command window.
clear all; close all;

%% Cell 1; Read data from source;
% Read data for TO from thesis_data_2014_05_09.xlsx.
turnover = xlsread('thesis_data_2014_05_09.xlsx',1,'C2:M228');
% for PTB
price_to_book = xlsread('thesis_data_2014_05_09.xlsx',2,'C2:M228');
% for the market price of companies
marketprice = xlsread('thesis_data_2014_05_09.xlsx',3,'C2:N228');
% For the risk free interest rate
risk_free_rate = xlsread('thesis_data_2014_05_09.xlsx',4,'C2:M228');
% For the market return
omx_pi = xlsread('thesis_data_2014_05_09.xlsx',5,'C2:N228');

%% Cell 2; Missing points
% Set all missing points to 0, in the data sets there exists such
% observations.
turnover(isnan(turnover)) = 0;
price_to_book(isnan(price_to_book)) = 0;
marketprice(isnan(marketprice)) = 0;

%% Cell 3; Calculate the yearly returns
% Calculate return for each company but only if there are no data points
% missing in the time series for the company.
% Specify storage matrix.
yearly_return = zeros(227,11);
% For all 227 companies
for i = 1:227
    % For all 11 years.
    for j = 1:11
        % If the market price at year t or t+1 is zero, set to 0
        if marketprice(i,j) == 0 | marketprice(i,j+1) == 0
            yearly_return(i,j) = 0;
        % Else, set to the relative return.
        else
            yearly_return(i,j) = (marketprice(i,j+1)-
                                    marketprice(i,j))/marketprice(i,j);
        end
    end
end

%% Cell 4; Calculate the yearly market return
% Calculate return for the market but only if there are no data points
% missing in the time series for the company.
% Specify storage matrix.
yearly_market_return = zeros(227,11);
% For all eleven years.
for j = 1:11
    yearly_market_return(:,j) = (omx_pi(1,j+1) - omx_pi(1,j))/omx_pi(1,j);
end

%% Cell 5; Clean negative turnovers
% If the turnover for any comopany for any year is less than 0, set it to zero.
for i = 1:227
    for j = 1:11
        if turnover(i,j) < 0
            turnover(i,j) = 0;
        end
    end
end

%% Cell 5; Gather processed data
% The turnover is transformed to be expressed in millions of sek [msek].
AllData(:,:,1) = turnover/10^6;
AllData(:,:,2) = yearly_return;
AllData(:,:,3) = price_to_book;
AllData(:,:,4) = risk_free_rate/100;
AllData(:,:,5) = yearly_market_return;
% Specify ID for companies
AllData(:,:,6) = repmat([1:227]',1,11);
% Specify observation year
AllData(:,:,7) = repmat([2002:2012], 227,1);

%% Cell 6; Clear missing variables
% Clear observations where a value for one of the variables is missing.
for i = 1:227
    for j = 1:11
        for k = 1:5
            if (AllData(i,j,1) == 0 | AllData(i,j,2) == 0 | AllData(i,j,3) == 0 | AllData(i,j,4) == 0 | AllData(i,j,5) == 0)
                for k = 1:6
                    AllData(i,j,k) = 0;
                end
            end
        end
    end
end

%% Cell 7; Investigate final data
% Count number of data points left
numnonzeros = zeros(5,1);
for i = 1:227
    for j = 1:11
        for k = 1:5
            if AllData(i,j,k) ~= 0
                numnonzeros(k,1) = numnonzeros(k,1) +1;
            end
        end
    end
end
end
end
% Count number of companies left.
numcompanies = 0;
for i = 1:227
    if any(AllData(i,:))
        numcompanies = numcompanies +1;
    end
end

%% Cell 8; Format matrix to fit regression format;
% Reshape each layer of the 3D matrix to a long vector.
AD_1 = reshape(AllData(:,:,1),227*11,1);
AD_2 = reshape(AllData(:,:,2),227*11,1);
AD_3 = reshape(AllData(:,:,3),227*11,1);
AD_4 = reshape(AllData(:,:,4),227*11,1);
AD_5 = reshape(AllData(:,:,5),227*11,1);
AD_6 = reshape(AllData(:,:,6),227*11,1);
AD_7 = reshape(AllData(:,:,7),227*11,1);

% Specify matrix containing these vectors as columns
AD = [AD_1 AD_2 AD_3 AD_4 AD_5 AD_6 AD_7];

%% Cell 9; Sort data
% Sort the data on the basis of TO.
[Y I] = sort(AD(:,1));
AD = AD(I,:);
numzeros = length(AD_1)-numnonzeros;
% Obtain final matrix that will be used as input in MainScript.m
Data_sorted = AD(numzeros+1:end,:);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  

REGRESSION ANALYSIS
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  

%% Main script
% Daniel Boros, Claes Eriksson, Bachelor's thesis, May 13th, 2014
% Used for the robust regression exercises and for all figure generation.

% Clear the workspace, close all figures and clear the command window.
clc; clear all; close all;

%% Cell 1; Import data.
% Load the data set for the variables presented in the Data section.
% These are, in the .mat file sorted as
% [TO R PTB R_f R_M Company_ID Year] and are internally sorted after the
% value of TO.
% The matrix in the .mat file is named: Data_sorted
load('Data_TO_sorted.mat');

%% Cell 2; Collect meta info
% Specify the number of observations.
n_obs = length(Data_sorted(:,1));

%% Cell 3; Specify the variables
% Turnover in million SEK
to = log10(Data_sorted(:,1));
% Market risk premium
mrp = Data_sorted(:,5) - Data_sorted(:,4);  
% Risk premium
rp = Data_sorted(:,2) - Data_sorted(:,4);  
% Price to book ratio
ptb = Data_sorted(:,3);  
% Company ID
co_id = Data_sorted(:,6);  
% Year of observation
o_y = Data_sorted(:,7);

% Specify the indata matrix of the independent variables.
indata = [rp to mrp ptb co_id o_y];

%% Cell 4; Specify methodology settings
% Specify how the number of intervals; this has been altered between 4, 5 
% and 10, as described in the thesis.
n_iv = 5;  
%[4,5, 10]
% Define the upper bounds of the data set intervals.
iv_ub = [1:n_iv].*n_obs/n_iv;
% Extract the turnover values of the interval upper bounds.
to_iv = 10.^[to([1 iv_ub])];

%% Cell 5; Regressions
% Introduce variable to store regression residuals in.
r = zeros(1940,3,n_iv);
% Introduce a loop counter.
count = 1;
% Run the regression for all the intervals.
for N = 1:length(iv_ub);
% Extract the data for the current interval.
reg_data = indata(1:iv_ub(N),:);
% Fit a linear model using robust regression.
mdl = LinearModel.fit(reg_data(:,2:4), reg_data(:,1), 'RobustOpts',
'on');
% Store the residuals of each model.
r(1:iv_ub(N),:,N) = [double(mdl.Residuals(:,1)) reg_data(:,5:6)];
% Store the parameters and p-values of each variable of each model.
coeffs(:,:,N) = [double(mdl.Coefficients)];
% Store the parameters and p-values for log(TO) in a separate matrix.
to_cp(:,:,N) = [coeffs(2,1,N); coeffs(2,4,N)];
% Store the degrees of freedom of the t-distribution for the t-tests.
tao_dof(N) = iv_ub(N)-2;
% Store the R2 och each model.
R2(N) = mdl.Rsquared.Ordinary;
% ------------------Plots for residual analysis----------------------
subplot(n_iv,2,count)
% QQ-plot of residuals
qqplot(double(mdl.Residuals(:,1))); grid off
subplot(n_iv,2,count+1)
% Histogram for residuals.
plotResiduals(mdl)
set(gca,'fontSize',10)
grid off
count = count +2;
end
%% Cell 6; Plot p-value and parameter estimates
figure()
% Plot
plot([1:N], to_cp','o-')
xlabel('Interval number')
ylabel('Value')
legend('Turnover parameter estimate', 'Corresponding p-value')
set(gca,'FontSize',14)
gridd on

%% Cell 7; Scatter plot, TO vs. risk premium, and histogram of TO
figure(2)
subplot(2,1,1)
plot(Data_sorted(:,1), rp, 'o')
xlabel('Turnover [mSEK]')
ylabel('Risk premium')
grid off
set(gca,'fontSize',14)
subplot(2,1,2)
hist(Data_sorted(:,1),50)
grid off

%% Cell 7; Scatter plot, log(TO) vs. risk premium, and histogram of log(TO)
figure(3)
subplot(2,1,1)
plot(to, rp, 'o')
xlabel('log(Turnover)')
ylabel('Risk premium')
grid off
set(gca,'fontSize',14)

%% Cell 8; Scatter plot with fitted line for log(TO)
figure(5)
C = coeffs(1,2,5);
plot(to, rp,'o', [min(to) max(to)], C*[min(to) max(to)],'r')
xlabel('log(Turnover)')
ylabel('Risk premium')
grid off
figureHandle = gcf;
set(findall(figureHandle,'type','text'),'FontSize',14)

set(findall(figureHandle,'type','text'),'FontSize',14)
RESIDUAL PLOTS FOR FIRST 5 INTERVALS, 10 INTERVALS
The base approach for detecting inference used in this paper is multiple linear regressions, where the aim is to fit the linear model

\[ y_i = \sum_{j=0}^{k} x_{ij} \beta_j + \epsilon_i, \quad i = 0, 1, 2, \ldots, n - 1, n, \ x_{i0} = 1 \]

to data (Stock & Watson, 2011). In the above setting \( y_i \) is the dependent variable, \( x_{ij} \) the \( i \)th observation of the \( j \)th independent variable, \( \epsilon_i \) the error term for observation \( i \) and \( \beta_j \) the coefficient for independent variable \( j \). As seen above, the total number of observations is \( n \) and the total number of independent variables used is \( k \).
For ease of notation, the linear model may be written in the following matrix form:

\[
\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n0} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}
\]

or even more compactly

\[ Y = X\beta + \varepsilon. \]

There exists several approaches for the estimation of \(\beta\), but they all have in common that they aim at explicitly specifying the functional form, namely

\[ Y = X\hat{\beta} + \hat{\varepsilon}. \]

Here, \(\hat{\beta}\) are the estimates and \(\hat{\varepsilon}\) are called the residuals. The method for estimation used in this study is Ordinary Least Squares (OLS) estimation, which aims at estimating the coefficients in a way that minimizes the sum of the squared residuals. It can be shown that this estimate becomes

\[ \hat{\beta} = (X^TX)^{-1}X^TY, \]

where \(X^T\) is the transpose of the matrix \(X\) defined above.

---

**STUDENT’S T-TEST**

In the test, we use the fact that the variable

\[ t_k = \frac{\hat{\beta}_k - \beta_{k0}}{SE(\hat{\beta}_k)} \]

is, under the assumption that the null hypothesis is true, an observation from the student’s t distribution with \(n - 2\) degrees of freedom, denoted \(T_{n-2}\). The variable \(SE(\hat{\beta}_k)\) is the standard error of the estimate \(\hat{\beta}_k\), namely an estimator of \(\sigma_{\hat{\beta}_k}\), the standard deviation of the sampling distribution of \(\hat{\beta}_k\). The standard error is estimated as

\[ SE(\hat{\beta}_k) = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \frac{\varepsilon_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}. \]

The p-value of the t-statistic states the probability of observing an estimate \(\hat{\beta}_k\) deviating from its true value \(\beta_{k0}\) as much as the one observed if the null hypothesis is true, namely
\[ p = P_{H_0}(|\hat{\beta}_k - \beta_{k0}| \geq |\hat{\beta}_k^{obs} - \beta_{k0}|). \]

The probability statement above may be extended to
\[ p = P_{H_0}\left(\frac{|\hat{\beta}_k - \beta_{k0}|}{SE(\hat{\beta}_k)} \geq \frac{|\hat{\beta}_k^{obs} - \beta_{k0}|}{SE(\hat{\beta}_k)}\right) = P_{H_0}(|t| \geq |t_{obs}|), \]
and we know that \( t \) is \( t_{n-2} \) distributed, meaning that
\[
p = P_{H_0}(|t| \geq |t_{obs}|) = P_{H_0}\left(t \geq t_{obs} \cup t \leq -t_{obs}\right) = \\
= P_{H_0}(t \geq t_{obs}) + P_{H_0}(t \leq -t_{obs}) = 2P_{H_0}(t \geq t_{obs}) = \\
= 2\left(1 - P_{H_0}(t \leq t_{obs})\right) = 2\left(1 - F_{t_{n-2}}(t_{obs})\right) = 2F_{t_{n-2}}(-t_{obs}).
\]

The third equality follows from the events \( t \geq t_{obs} \) and \( t \leq t_{obs} \) being disjoint, the fourth and seventh from the fact that the student's t-distribution is symmetrical around zero. \( F_{t_{n-2}}(x) \) denotes the cumulative distribution function of a \( t_{n-2} \) distributed variable.

Thus, given an observed \( p \), the probability of observing the \( \hat{\beta}_k^{obs} \), when \( H_0 \) actually is true is exactly \( p \). By this reasoning, the probability of a type I error, namely rejecting \( H_0 \) even though it is actually true, is \( p \). This is called a rejection of the hypothesis with a significance level of \( p \).