OTDK-dolgozat

2019
RELATIVE EQUITY VALUATION: AN ALGORITHMIC MULTIPLE REGRESSION APPROACH IN PYTHON

Illés Ferenc

Kézirat lezárásának dátuma: 2018.03.19.
1. Introduction ........................................................................................................................................... 1

2. Conventional Relative Valuation ........................................................................................................... 2
   2.1. The theoretical background of relative valuation ............................................................................. 3
   2.2. The multiples approach from a statistical viewpoint ....................................................................... 6
   2.3. Empirical validation of the different methods .................................................................................... 9
      2.3.1. Introducing the test concept ..................................................................................................... 10
      2.3.2. Testing as a DCF approximation ............................................................................................... 10
      2.3.3. Assumptions ............................................................................................................................. 10
      2.3.4. Selection criteria ...................................................................................................................... 12
      2.3.5. Evaluation .................................................................................................................................. 14
      2.3.6. Algorithm .................................................................................................................................... 15
      2.3.7. Empirical test ............................................................................................................................ 16

3. Relative valuation using statistical methods ............................................................................................ 19
   3.1. Model selection .................................................................................................................................. 20
      3.1.1. Variables selection, exploratory data analysis and relationships in the data ......................... 20
      3.1.2 Estimator selection ..................................................................................................................... 28
         3.1.2.1. Loss functions ....................................................................................................................... 29
            3.1.2.1.1. Squared error ............................................................................................................... 30
            3.1.2.1.2. Absolute error ............................................................................................................... 30
            3.1.2.1.3. Percentage squared error ............................................................................................ 31
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.2.2. Coefficient estimators for quadratic loss</td>
<td>31</td>
</tr>
<tr>
<td>3.1.2.2.1. Selection criteria</td>
<td>32</td>
</tr>
<tr>
<td>3.1.2.2.2. Ordinary least squares</td>
<td>33</td>
</tr>
<tr>
<td>3.1.2.2.3. Ridge regression</td>
<td>35</td>
</tr>
<tr>
<td>3.1.2.2.4. Percentage least squares</td>
<td>36</td>
</tr>
<tr>
<td>3.2. Model assessment</td>
<td>37</td>
</tr>
<tr>
<td>3.2.1. Random subsampling</td>
<td>38</td>
</tr>
<tr>
<td>3.2.2. Out-of-sample</td>
<td>39</td>
</tr>
<tr>
<td>3.3. Dataset</td>
<td>40</td>
</tr>
<tr>
<td>3.4. Variable selection</td>
<td>42</td>
</tr>
<tr>
<td>3.4.1 Principal component analysis</td>
<td>43</td>
</tr>
<tr>
<td>3.5 Model building</td>
<td>44</td>
</tr>
<tr>
<td>4. Conclusions</td>
<td>50</td>
</tr>
<tr>
<td>5. References</td>
<td>51</td>
</tr>
</tbody>
</table>
1. Introduction

The main motivation behind the paper is to present multiple linear regression models as an alternative for market multiples, which appear to be the only way of doing relative equity valuation. To be more precise, the goal is to find a method for valuing the equity of private companies based on their accounting metrics by relying on comparison with the valuation of public companies – bypassing a lengthy bottom-up calculation. Despite being seemingly more complicated than traditional techniques, the outlined regressions are far easier to use once understood, as they do not require manual peer selection and are more or less applicable in Excel.

According to Damodaran (2006), one of the most famous equity appraisal theorists, “Most equity research reports and many acquisition valuations are based upon a multiple such as a price to sales ratio or the value to EBITDA multiple and a group of comparable firms.” Even when a discounted cash flow model is used to calculate the value of equity, multiples are often used as a sanity check for the results. Despite their popularity, however, the literature on multiples is already very sparse in general and there is barely anything available on other relative valuation methods. Damodaran himself advocates the use of multiple linear regression as an alternative approach (Damodaran, 2006), but his model lacks technical statistical considerations, and is more of a side note than a true demonstration. This deficiency is especially striking as nowadays big data, predictive analytics and machine learning seem to be everywhere. For equity valuation this means hundreds of unique features for every public company, of which there are tens of thousands globally. All of that easily accessed through data vendors like Bloomberg or Capital IQ, to one of which most business schools, corporate finance advisors and investment firms are subscribed to. Yet most of that incredible wealth of information is wasted when relative valuations consists of simple multiples.

Now neither the ease of using multiples, nor the business educational background of potential users is lost on the author, therefore the paper will try to remain as non-technical as possible and try to extensively use intuitive reasoning that is easy to understand even with limited statistical knowledge. It is also worth mentioning, that even though the paper relies entirely on Python algorithms, the results are easy to implement is any statistical software or programming language and are also more or less applicable in Excel, which is the golden standard in corporate
finance. Python itself is completely open source –and therefore free– which can offset the cost of time invested in learning its use.

Since multiples are the “industry” golden standard, any suggested alternative is expected to provide some proof of its usefulness. In this case, that proof will be a demonstration of superior estimation accuracy and generality. The first section of the paper will be dedicated to introducing the multiples approach, its assumptions and the empirical evaluation of its performance. Afterwards the second section will present statistical concepts that can help build a better model, demonstrate multiple example models and measure their accuracy using the multiples as benchmarks. It should be emphasized that the paper does not claim to have found the best possible model or even anything remotely close. In fact, the final section will identify multiple conspicuous problems that need to be solved and should be subject to further investigation. Even so, these models are easy to use and can very well outperform even the best multiple estimations. With the apparent lack of similar research the findings can also be useful from a theoretical point of view.

2. Conventional relative valuation

There are four main avenues of equity valuation:

*The income approach* indicates the market value of a business enterprise based on the present value of the cash flows that the business can be expected to generate in the future. Such cash flows are discounted at a discount rate (the cost of capital) that reflects the time value of money and the risks associated with the cash flows.

*The relative or market approach* indicates the market value of equity or business enterprise based on a comparison of the target to comparable publicly traded companies or transactions in the relevant industry.

*The net asset approach* is based on the summation of the individual piecemeal values of the underlying assets and liabilities.

*The real option approach* values a company based on future volatility and fixed maximal loss, treating it as an investment right, as opposed to an obligation.
The paper will be obviously focusing on relative valuation, and to a certain extent on the income based approach, also known as the discounted cash flow (DCF) method. The differences in value between discounted cash flow valuation and relative valuation come from different views of market efficiency. In discounted cash flow valuation, the assumption is that markets make mistakes, that they correct these mistakes over time, and that these mistakes can often occur across entire sectors or even the entire market. The DCF value is often called “fundamental value” and has little reason conform to short term market sentiment. In relative valuation, the assumption is that while markets make mistakes on individual stocks, they are correct on average. The goal is not necessarily finding the fundamental value, but following the market sentiment. (Damodaran, 2006)

The current study does not attempt to estimate a better value for public companies than their official market price, but rather tries to apply that price (regardless their fundamental correctness) to private companies. Finding the fundamental value is probably impossible with a completely relative valuation, therefore also no claim of predicting stock prices by identifying under or overpriced public companies. Some might object by stating that private companies cannot be valued based on public companies, as that would be comparing apples and oranges. In their opinion only M&A transactions could be appropriate benchmarks. However, there is significant literature suggesting otherwise by examining the liquidity discounts on private firms’ equity (Officer, 2007; Cañadas and Rojo Ramirez, 2011) –demonstrating that public and private company equity is strongly connected. In the authors’ personal experience these discounts are widely used in practice, as even discounted cash flow approaches are usually applied with public company based WACCs, which inevitably results in public equity values that need further liquidity discounts. Therefore, it can be safely assumed that any valuation scheme using public companies has practical use outside the stock markets.

2.1. The theoretical background of relative valuation

There are two main approaches to multiples valuation. One estimates the equity value (E) of a company and the other estimates the enterprise value (EV) of a company. To understand how these approaches are different the concept of equity value, enterprise value and their respective cash flows needs to be understood. Equity value denotes the value of owning a company’s equity, the source of the value being the cash flows the shareholders are entitled to. On the other hand,
enterprise value is a more general concept, representing the value of all claims by all claimants, and the value is based on all free cash flows the company generates. The difference between the two values is the net debt, which is the total value of a company's liabilities and debts less the total value of its cash, cash equivalents and other liquid assets. Differentiating between these two types of values is important not just because the financial performance metrics used to measure them are distinct, but also because net debt is handled very different from equity during appraisals.

It seems intuitive that multiples based on earnings that are due to both shareholders and other claimants—usually creditors—as well should not be used to estimate the value of the equity, which is exclusively owned by the shareholders. Conversely, earnings that represent cash flows owed to the shareholders should not be good estimators of the value of debt. However, things are not always that simple in practice. For instance, since creditors are paid before the owners of equity, a company with plenty of cash flow available for its shareholders have already paid the owners of its debt, driving up both enterprise and equity value. Since one of the aims of the study is to evaluate multiples valuation, the validity of having separate multiples for equity and enterprise value will have to be verified as well. In fact, it will be demonstrated in the following, that both the E and EV approaches estimate E, and unless all multiples perform with the exact same accuracy, a single best estimator always exists.

The relative valuation approach is essentially a simplification of the discounted cash flow (DCF) method. Even though these two models are very different, the financial theory behind them builds on the same principles. While every cash flow based multiple is somewhat different, they all follow the same line of thought, so presenting a detailed breakdown for only one of them should suffice for the study’s purposes. The general workings of the earnings multiples can be shown with the following, somewhat simplified, derivation of a common multiple: EV/EBIT.

The enterprise value (EV) of a company is the sum of the present value of all future free cash flows to firm (FCFF). This is exactly what the DCF model uses when valuing a company. Thus

\[ EV = \sum_{t=1}^{\infty} \frac{FCFF_t}{(1 + r)^t} \]

4
FCFF is commonly calculated by taking the EBIT, multiplying it with \((1 - \text{tax rate})\), adding depreciation and amortization, and subtracting changes in working capital and capital expenditure. In some cases a number of further refinements and adjustments may also be made to try to eliminate distortions. It is clear, that if we assume no change in working capital and that the depreciation is more or less equal to the necessary capital expenditure, FCFF is fairly close to EBIT. By approximating FCFF with EBIT (earnings before interests and taxes) it reduces to

\[
EV = \sum_{t=1}^{\infty} \frac{EBIT_t}{(1 + r)^t}
\]

Additionally, an appropriate discount factor has to be used. Since EBIT eliminates for among other things different capital structures, we need a discount factor that takes the capital structure into consideration. Moreover, the ‘tax shield’ effect of interest has not been accounted for in FCFF, therefore it needs to be incorporated in the cost of debt part of the discount rate. A commonly used discount rate is the Weighted Average Cost of Capital or WACC. The previous equation then becomes

\[
EV = \sum_{t=1}^{\infty} \frac{EBIT_t}{(1 + \text{WACC})^t}
\]

This is where the so called comparables come into the picture, which are companies that are similar enough according to some criteria to be considered analogous in their DCF valuation. In other words, the value of these companies can be easily compared, hence the name *comparables*. For a group of so called peer or comparable companies, if their WACCs are uniform and the ratio of their EBITs stays constant in time, it can be easily seen that their EVs are going to be directly proportionate to their EBITs. As a consequence, if the EBIT of all the peer companies is known and the EV of at least one of them, all the unknown EVs can be estimated by using the known EV/EBIT ratio. Naturally, no two companies have the exact same risk (and cost of capital), calculating with EBIT is hardly ever the same as using FCFF, and the companies’ earnings fluctuate dynamically in time so EBIT ratios between peer companies will also change even under ideal conditions, and more so under realistic ones. Accordingly, the multiples approximation of the DCF
method is inherently inaccurate to an extent, as is the case with all simplifications. For this reason it is not enough to find a single comparable company in practical applications, instead the mean EV/EBITDA of multiple peers is used to decrease the estimation errors.

Similarly, equity value is the sum of the present value of all future free cash flows to equity (FCFE). FCFE is FCFF minus net borrowing (principals raised less principals paid) and interest expense multiplied by \((1 - \text{tax rate})\). In this case FCFE is approximated with net income (NI) and the discount rate with the cost of equity \(r_E\), but everything else is the same as in the previous case. The question of whether E can be estimated with revenues (REV), EBITDA or EBIT seems similar to asking whether the FCFE can be approximated with them. Theoretically it all depends on the peers, as companies with the exact same income, cost and debt structure will have directly proportional REV, EBITDA, EBIT and NI as well, so it does not matter which one is used. However, that is a purely theoretical scenario, as there are no twin companies. If that was the case, it would not matter what multiple is used at all, and there would not be a multitude of them.

2.2. The multiples approach from a statistical viewpoint

It needs to be highlighted, that the comparables approach depends on the availability of E and EV values for the peer companies. The whole point of using this method is that it allows users to bypass the resource (time, information and knowledge) intensive process of the DCF model, by comparing the company to similar companies that have been previously valued by others. Practically, that means publicly traded companies with public share prices and outstanding share amounts, which can be used to easily determine the true market value of their E. Unfortunately, no true EV is publicly available, because the market value of net debt is seldom known at all, much less publicly. What is available instead is the book value of net debt which can be automatically calculated from the balance sheet of any company and added to the equity value to approximate the enterprise value \((EV=E+\text{net debt})\) for the reference companies. In order to determine the true market enterprise value, the market value of net debt would have to be estimated by hand, but that would defeat the purpose of the comparables approach and is therefore taken at book value with minor adjustments at best. What logically follows, is that both EV and E multiples estimate E, as only E has a public market value. The only difference between them is that EV multiples use the net debt as well in the process, while E multiples do not. This can be more clearly seen by looking
at multiples valuation from a statistical/mathematical viewpoint. The relative valuation approach can be introduced in the following way:

1. Multiples valuation is a form linear regression. A regression is any model used to estimate quantitative variables using other quantitative and qualitative variables. It can also be viewed as a function approximation. A linear regression is any regression model, where the mapping between the explanatory (independent, exogenous, regressor), variables or their transformed value and the dependent (endogenous, regressand) variable is affine. That is, given a population \((Y_i, X_{i1}, \ldots, X_{ij}), i=1,\ldots,n\) the true underlying relation between the dependent variable and the explanatory/independent variables can be formulated as:

\[
Y_i = \beta_0 + \beta_1 \phi_1(X_{i1}) + \cdots + \beta_j \phi_j(X_{ij}) + \varepsilon_i \quad i = 1, \ldots, n
\]

Where \(\phi_1, \ldots, \phi_j\) are arbitrary functions and \(\varepsilon_i\) are the (random) errors unexplainable by the model. The linear part of the designation only relates the appearance of the regression coefficients \((\beta_j)\), which is linear in the above relationship. Since only the coefficients have to be linear, a linear model can be effectively used to model nonlinear relationships, especially if numerous independent variables are included. However, multiples valuation is a two-variable –also known as simple or univariate– linear regression, where no transformation or basis expansion is applied to the independent variable. The true coefficients of regressions are almost never known, so they have to be estimated as well from an available sample of dependent variables and the corresponding independent variables. The coefficients are usually estimated using the ordinary least squares (OLS) method, but many other methods are also possible. Multiples estimation is one such method, where the intercept is always zero and the slope coefficient is equal to the mean of the chosen sample multiples (i.e. the sample coefficients). The explanatory variables not being transformed or expanded upon and the intercept being zero means that the modeled relationship in a multiples regression is truly linear –as opposed to being simply affine.

2. The sample of \(Y_i (y_i)\) used to estimate the coefficients is the market capitalization of public companies, while \(Y_i\) is the market value of all companies’ equity.
3. Equity value is the only dependent variable. Enterprise value might appear to be another potential dependent variable at first glance, however, it is but the sum of equity value and net debt, and the latter is taken at book value for two main reasons. Firstly, depending on the accounting principles used, its book value is more or less kept close to market value. Secondly, its true market value –unlike the equity value of public companies’– is not publicly available, if available at all, and calculating the market value manually is unfeasible for a multiples valuation. Since the book value of net debt is always available –not just for public companies– it is an independent variable. This means, that even during enterprise valuation, only the equity value is estimated.

4. The other independent variables usually used are the revenues, EBITDA, EBIT, net income and the equity’s book value. Intuitively, the former three are for estimating enterprise value and the latter 2 equity value. It should be noted, that the multiple using the equity book value is not a direct approximation of the DCF method.

Since all multiples estimate the same quantity –the market value of equity– if one of them outperforms the others, there seems to be no point in using anything but the optimal multiple. However, some multiples are more widely applicable than others. The previous section did not explicitly mention it, but multiples valuation generally only works for companies with positive incomes, as companies in the red have non-negative value despite operating at a loss, not because of it. Such a company only has a positive value if at some point in the future it is expected to turn profitable –potentially by being valued as a real option– or based on the value of its assets, none of which can be grasped by a negative earnings ratio. Consequently, E/NI (also called P/E –price/earnings) ratios can only be utilized for the set of profitable companies, while EV ratios can be used more generally, especially the revenue multiples, as revenues cannot be negative. This means that EV multiples can still be useful even if they underperform E multiples, by being applicable where the latter are not.

It should also be highlighted, that even if EV multiples only value E, net debt is far from being irrelevant. Estimating E with EV multiples (e.g. EV/EBITDA - net debt) does not equal to valuing E with the EV variables directly (eg. E/EBITDA), even if the latter has gained some popularity
with analysts (Damodaran, 2006). The importance of net debt in the statistical framework of the comparables model can be stated as:

1. Using the EV variables and net debt individually is suboptimal because of the presence of negative confounding. Confounding arises because companies with higher REV, EBITDA or EBIT do not necessarily have higher market capitalizations, as in many cases they also have more net debt as well. Naturally, if we anchor the level of net debt, higher cash flows imply higher market cap. Similarly, having higher net debt decreases the market value of equity ceteris paribus, but the negative relationship is masked by the fact that companies with more net debt are often larger, have more earnings and therefore higher market caps. So what happens is that the cash flow proxy variable of choice and the net debt negatively confound each other, decreasing each other’s linear correlation with the market capitalization.

2. Equity value is more linearly correlated with the EV variables than net debt. The latter is only useful as a predictor, if used with EV variables.

3. Since multiples valuation is a simple linear regression, it only works with a single explanatory variable –which is optimally a cash flow variable rather than net debt due to the aforementioned stronger relationship– and net debt has to be included in some other way in the regression. Adding net debt to equity value, and using the EV variable to estimate this sum achieves this goal. It is equivalent to adding net debt as a second input variable in the equation with a fixed coefficient of minus one.

4. Adding net debt in this particular way improves the estimation and makes it potentially more accurate than net income or equity book value based models.

2.3. Empirical validation of the different methods

The main objective of the paper is to introduce a new approach to relative valuation, but in order to prove its usefulness, the efficiency of the orthodox approach has to be measured somehow, so that it can later be used as a benchmark. The empirical performance test of different multiples actually achieves multiple goals at the same time, as the same measurements used to benchmark the more advanced regression models in the second part of the paper can also be used to decide
which multiples are the most accurate and whether there is a point in separating E and EV valuations.

2.3.1. Introducing the test concept

Unfortunately, evaluating the performance of the multiples is not a trivial task. Using the database of Standard and Poor’s Capital IQ database there were around 42 thousand operating public companies in 2017 with market data going back decades. On the other hand, multiples valuation is meant to be used with small peer groups which are supposed to match the valued company according to some criteria. Testing the accuracy on all the possible peer groups or even a random sample of them is unfeasible with desktop scale computing capacities as well as theoretically unsound, since performance is only relevant on similar companies, not all possible subsets of given size. To make things worse, building peer groups requires human oversight, involves forecasting, non-numeric data (like business descriptions), and is highly subjective in general. Fortunately, there are ways to get around these obstacles.

2.3.2. Testing as a DCF approximation

Companies with similar earnings trends, growth rates and costs of capital should—according to theory—have similar valuation and therefore could be considered good peer groups. The most subjective part is the growth rate which has to be forecasted manually, however, by looking at historical data it is possible to find companies that experienced uniform income growth trends for extended periods. Using hindsight, it is actually possible to select better past peers algorithmically than what would be possible by hand in the present. Naturally, this approach is only valid if relative valuation is indeed applied in practice with the ultimate goal of approximating the DCF model.

2.3.3. Assumptions

The main assumption is that multiples users try to build peer groups that are comparable in a DCF framework and actually manage to do so with reasonable success. Most of the time analysts define comparable firms to be other firms in the firm’s business or businesses (Damodaran, 2006). This is so, because it is very hard to find companies that are actually similar based on rigorous theory (as it will be demonstrated later), and one does not know which companies are similar based purely on financial performance until having forecasted the incomes and estimated the discount.
rates for them manually. Forecasting for dozens or hundreds of companies is actually much more difficult than a mere DCF calculation, so it is not surprising that no one performs it. If there is a critical mass of people investing based on multiples obtained by theoretically flawed peer selection or inaccurate forecasts, the market price will reflect it and the actual peers will give a bad estimation.

The next assumption is, that the income level (revenues, EBITDA, EBIT, net income) used to find the peers should be the earnings level used in the multiple that is being tested (i.e. peer groups for EV/EBITDA should not be based on net income growth), while the cost of capital should be either the cost of equity or the WACC depending on the multiple. It might seem unnecessary to use the WACC, as only equity is being valued (as was shown above), but the discounted value of the FCFF of a company is not independent of the financing structure, even if the net debt is deducted at the end.

According to the capital asset pricing model (CAPM), the cost of equity is an affine function of the company’s beta, which also means that their linear correlation is 1. While properly introducing the model and its assumptions is out of the present study’s scope, it is suffice to say that should CAPM hold, for the purpose of building peer groups betas can be used instead of the proper costs of equity. The beta is the slope coefficient in the OLS regression between the independent variable of a stock index’s yield and the dependent variable of a stock’s yield, while the intercept of the regression should be zero according to the model (Jensen, Black and Scholes, 1972). From a known beta the cost of equity is acquired by multiplying it with the market risk premium rate (in this case it will be 5.5%) and adding the country’s risk free rate (0.5%), which is a government bond’s yield denominated in the same currency as the stock’s dividend yield. Therefore public companies on the same stock market have identical costs of equity if they have identical betas. It should be noted, that even though CAPM is widely accepted as an acceptable approximation, it has been shown time and time again to fail empirical tests (Reinganum, 1981; Fama and French, 1992; Basu and Chawla, 2010). Additionally, public companies with small and sparse daily trade volumes often have small or even negative empirical betas, far below their true market risk. The resulting false betas can result in biased peer groups. Therefore it is not
inconceivable that including betas in the empirical test will fail to improve the estimations’ quality or even worsen it by being biased and/or adding random noise.

As for estimating the WACC, the previously introduced cost of equity, the cost of debt and the corporate tax rate is necessary. The cost of debt also incorporates the risk free rate, but a credit rate spread needs to be estimated. In this case the interest coverage ratio and a credit rate spread table based on these ratios will be used, that can be accessed at Damodaran’s website (Pages.stern.nyu.edu, 2018). The used corporate tax rate will be 21%, but the effective tax rate is different for every company. The test will not account for these differences.

To summarize, the different multiples’ performance can be tested by finding public companies from the same stock market with similar betas/WACCs and similar income growths, as the company’s value should be directly proportional to its income all else equal.

2.3.4. Selection criteria

Naturally, there is no objective threshold beyond which companies’ betas/WACCs and earnings can be considered similar and finding exact matches is not only highly unlikely, but it is also pointless, as no one would be able to manually select a mathematically perfect peer group anyway. In the subsequent section the peer companies will be selected on the following criteria:

1. The growth rates will be compared by their CAGR (compounded annual growth rate), which is the geometric mean of the period growths. Naturally, uniform growth for 5 years does not guarantee long term uniform growth –which is a more important value driver– but it should be an acceptable proxy. The examined earnings (EBITDA, EBIT and net income) must be positive both in the valuation year (t-5) and at the end of the comparison period (t-0). This is necessary for two reasons. Firstly, multiples valuation breaks down if the earnings are negative, as mentioned earlier. Secondly, calculating CAGR involves taking the n\(^{th}\) (in this case the 5\(^{th}\)) root of the ratio of change. Even though an odd root of a negative number yields a real solution, the result is meaningless (it implies that the sign of the earnings is alternating annually). For the test’s purpose the maximal allowed CAGR difference of the examined earnings is 10%-50% depending on the scenario (in percentage of the CAGR, not absolute arithmetic percentage point difference).
2. The maximal allowed difference in beta/WACC is 10%-50% (again, in percentage of the beta/WACC).

3. As growth rates are compared by CAGR, the cash flows between the first and the last year of the comparison period are not taken into account. To adjust for this, the linear correlation coefficient ($r$) is also examined. It might seem intuitive to only compare with correlation – as perfect correlation means equal CAGR. However, linear correlation is insensitive to differing CAGR rates, and sensitive to fluctuating cash flows, even though the long term growth rate is a far more important value driver than the short-term ebb and flow. That means that any feasibly lax correlation condition will fail to ensure similar CAGR rates on companies with non-linear earnings trends. For the test purposes the correlation coefficient between the relevant cash flows during the comparison period must be at least 0.9-0.5. Where the allowed difference is 10%, the $r$ will be 0.9, where it is 20%, it will be 0.8, etc. Of course, correlation is not interpretable in a linear or percentage fashion, but relaxing the correlation criteria in tandem with the other two allows for more comprehensible outputs.

Also, for the purpose of the empirical test there must be at least 5 peers for every company to be valued with multiples. There is no fundamental reason for the cutoff being 5 peers, just as the filtering criteria could be different. Although this choice is unlikely to heavily influence the outcome, in order to reduce the room for any biases, multiple scenarios will be examined in a fashion not unlike a sensitivity test. The peer group selection process can be done differently by:

1. Using only 2 of the previous 3 criteria at a time. It could also be done on single variables, however, those peers groups would be too broad to conform to the original theoretical basis.
2. Changing the allowed difference between the peers.
3. Changing the number of peers.
4. Changing the date.

Introducing all these dimensions to the sensibility test would actually be counter-productive, as even a few dimensions can produce clear results –as will be demonstrated shortly– and more dimensions would only make the evaluation much more difficult to comprehend. Additionally, there is no apparent reason to presume the presence of confounding or multicollinearity, which would necessitate examining all the dimensions together. On these
grounds, the new examination aspects will be introduced one at a time. First the best peer selection criteria will be chosen, afterwards the effect of the peer group sizes will be evaluated and finally the results will be inspected at different dates.

2.3.5. Evaluation

In the later parts (chapter 3.1.2.) there will be a lengthy discussion of loss functions and multiple evaluation criteria. However, in the case of multiples, that would overcomplicate things. For now the measure of loss will be the standard squared error, as it can be considered the only type of error considered in financial theory and literature. Therefore, the accuracy will be measured with $R^2$ (coefficient of determination) values. $R^2$ has multiple interpretations, the most widespread being the ratio of variance explained by the independent variables. However, for our purposes another interpretation is more convenient: proportional reduction in error, the error being the squared difference between the estimated and true values and the reduction is measured against estimating with the dependent variables’ sample mean. Essentially this means estimating the market capitalization for every company that has enough peers using the multiples approach and also with the mean market capitalization of US public companies. Next, the difference between both estimations and the true market capitalizations is squared and summed up across the companies, and their ratio (the numerator being the total error of the multiples approach) is deducted from one.

This leads to the convenient property of being interpretable as a percentage improvement in estimation accuracy achieved by using the independent variables as opposed to not knowing them. For a standard OLS linear regression $R^2$ is always between 0% and 100%, unless the data used for fitting the regression is different than the data used for calculating the errors (more on that later in chapter 3.2.). However, multiples have no guarantee of being at least as good predictors as the mean E of US public companies, so the score can easily be negative.

Another important factor is the number of companies that can be valued with a given multiple at a given accuracy ($n$). In chapter 2.3.3., during the discussion of the assumptions the difficulty of finding peers was highlighted, it is therefore expected, that even at 50% allowed difference a large percentage of companies will be left without comparables.
2.3.6. Algorithm

In the previously outlined fashion the testing procedure can only be carried out by writing a custom algorithm, which can be found in appendix1.py. The code uses the following open source packages: Pandas (McKinney, 2010), NumPy (Oliphant, 2007) and NetworkX (Hagberg, Schult and Swart, 2008). Pandas is necessary to handle the data in the dataframes with textual labels. The alternative would the use of pure matrices, which would make the inspection and debugging of results much more complicated during the programming phase. NumPy is necessary for some matrix procedures that cannot be done with Pandas alone and also for general convenience where labels are unnecessary. NetworkX allows easy manipulation of data as graphs. That is convenient, as every company can be considered a node in a graph where peer companies are the connected nodes. The algorithm goes through the following process:

1. It imports the accounting metrics from an Excel spreadsheet;

2. calculates the WACC for every company;

3. creates four dataframes with only a specific variable type (REV, EBITDA, EBIT, NI);

4. removes companies that would have negative CAGR and/or multiples from all the dataframes;

5. depending on the scenario calculates the correlation,

6. iterates over every possible pair of nodes and adds an edge between them if they fit the filtering criteria;

7. creates a dictionary out of every company with enough peers (key) and its peers (value);

8. estimates the E or EV for every “key” company based on the corresponding “value” companies from the dictionary;

9. calculates the $R^2$ and places it in a pandas dataframe along with the number of values companies (number of keys in the dictionary).
Unfortunately, the code itself is very ad-hoc, since it did not need to be flexible or reusable, which might make it difficult to follow and understand for potential readers, even with the included comments. The author apologizes in advance for any inconvenience that might cause.

2.3.7. Empirical test

The test results are presented in the table below (Figure 1.). The columns show the maximal allowed deviation between peers, and the metrics displayed for every multiple are the $R^2$ measuring accuracy and $n$ measuring generality. There is no scenario without CAGR as it is assumed to be the most important peer selection criteria.

The first thing that can be deducted from the results is that EV/Revenues estimations have very large negative $R^2$ values. That means using revenue multiples is a far less accurate method than estimating with the average US public company market capitalization. In other words EV/REV is completely useless for general predictions. It might perform well in special cases, but not market wide. In the further scenarios it will be excluded for simplicity’s sake.

The table also reveals, that the number of companies valued ($n$) is overall negatively correlated with accuracy ($r=-0.515$). Therefore it is safe to say that accuracy comes at the price of generality in most cases.

Furthermore, the results also seem to provide strong evidence against using betas/WACCs for peer selection. Comparing the “CAGR, r, risk” and “CAGR, r” results, the number of valued companies dropped by ~10-18% on average (depending on the multiple), while the accuracy also decreased by ~0.5-13%. When comparing “CAGR, r” with “CAGR, risk”, the results are similar.

Figure 1. Performance of different multiples on 31.12.2012.

<table>
<thead>
<tr>
<th>Filtering</th>
<th>Multiple</th>
<th>Metric</th>
<th>50%</th>
<th>40%</th>
<th>30%</th>
<th>20%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAGR, risk</td>
<td>NI</td>
<td>R2</td>
<td>75.9%</td>
<td>79.5%</td>
<td>77.0%</td>
<td>72.1%</td>
<td>68.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>795</td>
<td>781</td>
<td>768</td>
<td>737</td>
<td>541</td>
</tr>
<tr>
<td></td>
<td>EBITDA</td>
<td>R2</td>
<td>79.9%</td>
<td>78.9%</td>
<td>78.2%</td>
<td>63.9%</td>
<td>-37.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>1108</td>
<td>1093</td>
<td>1079</td>
<td>1052</td>
<td>915</td>
</tr>
<tr>
<td></td>
<td>EBIT</td>
<td>R2</td>
<td>76.2%</td>
<td>75.6%</td>
<td>70.5%</td>
<td>76.7%</td>
<td>66.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>969</td>
<td>961</td>
<td>948</td>
<td>923</td>
<td>776</td>
</tr>
<tr>
<td></td>
<td>REV</td>
<td>R2</td>
<td>-355.9%</td>
<td>-389.8%</td>
<td>-566.7%</td>
<td>-616.9%</td>
<td>-175.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>1445</td>
<td>1434</td>
<td>1414</td>
<td>1363</td>
<td>1180</td>
</tr>
<tr>
<td></td>
<td>NI</td>
<td>EBITDA</td>
<td>EBIT</td>
<td>REV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-----</td>
<td>--------</td>
<td>-------</td>
<td>------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAGR, r</td>
<td>R2</td>
<td>R2</td>
<td>R2</td>
<td>R2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>83.1%</td>
<td>86.2%</td>
<td>78.2%</td>
<td>-635.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>798</td>
<td>1125</td>
<td>976</td>
<td>1503</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>87.9%</td>
<td>85.6%</td>
<td>72.3%</td>
<td>-2338.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>789</td>
<td>1113</td>
<td>960</td>
<td>1487</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>81.4%</td>
<td>86.5%</td>
<td>79.3%</td>
<td>-2294.7%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>716</td>
<td>1059</td>
<td>909</td>
<td>1447</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>91.3%</td>
<td>89.8%</td>
<td>78.4%</td>
<td>-58.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>539</td>
<td>918</td>
<td>775</td>
<td>1337</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>95.9%</td>
<td>86.0%</td>
<td>86.1%</td>
<td>13.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>237</td>
<td>645</td>
<td>503</td>
<td>1048</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>83.1%</td>
<td>84.3%</td>
<td>85.7%</td>
<td>-198.5%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>769</td>
<td>1093</td>
<td>1062</td>
<td>1429</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>84.3%</td>
<td>83.9%</td>
<td>80.0%</td>
<td>-3647.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>732</td>
<td>989</td>
<td>923</td>
<td>1394</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>85.7%</td>
<td>81.3%</td>
<td>82.3%</td>
<td>-308.6%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>610</td>
<td>846</td>
<td>846</td>
<td>1327</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>98.6%</td>
<td>91.0%</td>
<td>86.7%</td>
<td>-139.6%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>361</td>
<td>661</td>
<td>661</td>
<td>1150</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>-748.5%</td>
<td>-2294.7%</td>
<td>-70.6%</td>
<td>19.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>1429</td>
<td>1394</td>
<td>1327</td>
<td>660</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Naturally, this is very remote evidence against CAPM, however the findings in Figure 1. are definitely in agreement with the studies cited in the assumptions chapter (2.3.3.). As for WACCs, the results can be blamed on the rudimentary and most likely inaccurate calculation technique, but the betas are not fairing any better. As a consequence, only the CAGR and correlation criteria will be used in the following.

Besides these general trends, EBITDA multiples generally appear to overperform the others. Looking at the “CAGR, r” rows, EBITDA multiples have greater accuracy in all but one of the cases, but even then, the difference is absolutely marginal. At the same time, by definition more companies can be valued based on EBITDA than EBIT, and the more general EV/REV multiples have abysmal accuracy. Net income multiples have higher $R^2$ scores in most scenarios, but the added accuracy comes at a heavy price of far lower $n$ values. Still, that is not reason enough to disregard E multiples yet.

Figure 2. Performance of different multiples on 31.12.2012.

<table>
<thead>
<tr>
<th>Min. peers</th>
<th>Multiple</th>
<th>Metric</th>
<th>50%</th>
<th>40%</th>
<th>30%</th>
<th>20%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>NI</td>
<td>R2</td>
<td>83.0%</td>
<td>87.8%</td>
<td>81.3%</td>
<td>82.3%</td>
<td>88.1%</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>810</td>
<td>803</td>
<td>790</td>
<td>691</td>
<td>397</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EBITDA</td>
<td>R2</td>
<td>86.2%</td>
<td>85.6%</td>
<td>86.7%</td>
<td>79.7%</td>
<td>89.2%</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>1129</td>
<td>1126</td>
<td>1111</td>
<td>1026</td>
<td>800</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>NI</td>
<td>R2</td>
<td>83.1%</td>
<td>87.9%</td>
<td>81.4%</td>
<td>91.3%</td>
<td>95.9%</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>798</td>
<td>789</td>
<td>716</td>
<td>539</td>
<td>237</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2. illustrates the effect of using different restrictions for the minimal size of peer groups. There is barely any change in n depending on the leniency of the size criteria, although with 10% allowed deviation the size of the peer groups is an important factor. Even so, the overall conclusions did not change much and there is still a very noticeable tradeoff between accuracy and generality. Perhaps the only important feature is, that while previously the superiority of EBITDA multiples was not clear cut, when different peer group sizes are allowed EV/EBITDA has greater accuracy and generality in almost all cases. Whenever E/NI has a higher $R^2$ score, it achieves it on substantially fewer companies.

Figure 3. Performance of different multiples on 31.12.2011.

<table>
<thead>
<tr>
<th>Min. peers</th>
<th>Multiple</th>
<th>Metric</th>
<th>50%</th>
<th>40%</th>
<th>30%</th>
<th>20%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>NI</td>
<td>R2</td>
<td>80.1%</td>
<td>81.0%</td>
<td>75.1%</td>
<td>69.7%</td>
<td>95.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>787</td>
<td>782</td>
<td>757</td>
<td>676</td>
<td>349</td>
</tr>
<tr>
<td></td>
<td>EBITDA</td>
<td>R2</td>
<td>86.4%</td>
<td>87.6%</td>
<td>86.8%</td>
<td>86.8%</td>
<td>75.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>1078</td>
<td>1068</td>
<td>1052</td>
<td>990</td>
<td>775</td>
</tr>
<tr>
<td>5</td>
<td>NI</td>
<td>R2</td>
<td>79.9%</td>
<td>80.9%</td>
<td>80.0%</td>
<td>74.5%</td>
<td>95.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>780</td>
<td>755</td>
<td>700</td>
<td>520</td>
<td>224</td>
</tr>
<tr>
<td></td>
<td>EBITDA</td>
<td>R2</td>
<td>86.5%</td>
<td>87.7%</td>
<td>86.9%</td>
<td>87.0%</td>
<td>74.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>1068</td>
<td>1057</td>
<td>1015</td>
<td>883</td>
<td>632</td>
</tr>
<tr>
<td>10</td>
<td>NI</td>
<td>R2</td>
<td>79.1%</td>
<td>79.5%</td>
<td>82.3%</td>
<td>85.1%</td>
<td>87.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>950</td>
<td>927</td>
<td>862</td>
<td>745</td>
<td>469</td>
</tr>
<tr>
<td></td>
<td>EBITDA</td>
<td>R2</td>
<td>79.1%</td>
<td>79.5%</td>
<td>82.3%</td>
<td>85.1%</td>
<td>87.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>950</td>
<td>927</td>
<td>862</td>
<td>745</td>
<td>469</td>
</tr>
</tbody>
</table>

A quick look at the new results immediately reveals that not much has changed. The accuracies have been shuffled a bit, but the overall picture is the same: the EBITDA/EV multiples are more accurate almost everywhere while being valid for more companies. The accuracy-generality tradeoff has not vanished either. Of course, the date of the test is only a single year earlier, so a huge difference would have been more than surprising. Checking more dates is hardly necessary though, as these two years already give a good idea about the maximal empirical efficiency of multiples and the whole setup is highly hypothetical to start with. In other words, the
exact performance of an entirely unrealistic setting does not have significant added value over its approximation. For the proposed linear models more rigorous evaluations will be applied however.

Overall, the balance between accuracy and universality clearly makes the CAGR and correlation based peer selection the best choice in the test, with the EBITDA multiple being the best estimator by far. In some limited cases the E/NI could potentially give a better estimate, however. This cases also happen to be the scenarios with the highest required forecast accuracy, making their usage an overoptimistic proposition.

On another note, the “victory” of EV/EBITDA should definitely not come as a surprise as EBITDA has the highest linear explaining power even without the introduction any peer groups or even net debt. The linear correlation coefficient between EBITDA and Market Capitalization for the whole stock market in 2012 was 0.94, which translates into an $R^2$ of 88%, while net income only has $r=0.92$ ($R^2=85\%$). If EBITDA is regressed on EV, the $R^2$ is 89.4%. Revenues also only had 51.7% and 38% respectively (yes, they were better predictors with E). In fact, the only surprise is the lackluster performance of EBIT, as its $R^2$ is also 88.1% on E and 88.2% on EV. Of course, before drawing any undue conclusions, it needs to emphasized that these $R^2$ values are from a model which has seen all the market capitalizations before fitting the best possible trend line on them, therefore it is overfitted and is not representative of how accurately a regression could appraise a private company it has never seen before. The previous tests have only seen the peers before estimating the equity value, therefore they can be considered as a decent approximations of the truth.

On the other hand, even though these results are probably not overfitted, they paint a very optimistic picture of multiples estimation. At best, they should be treated as an upper cap of possible efficiency. Achieving such accuracy required perfect, or at least extraordinarily precise knowledge of the future cash flows of every single US public company for five years ahead. The chance of finding these in practice peers is very slim to say the least.

3. Relative valuation using statistical methods

Now that a ballpark figure has been established for the accuracy of the E and EV multiples a more important question follows: Are multiples the optimal substitute of a DCF calculation?
Besides the previously highlighted practical problems (difficulty of finding peers with uniform attributes), it cannot be emphasized enough, that the empirical accuracy of the EBITDA multiple approach calculated in the previous test should be treated as an upper bound.

As was established before, the goal of the paper is to show that a better approach indeed exists, yielding more universal and yet more accurate estimations. The first hint at the existence of a better model is the fact, that the multiples approach is an unorthodox simple linear regression, with zero guarantees of the peers’ mean being an optimal or even good estimator of the coefficients. As it turns out, not even the completely non-conventional (from a statistical point of view) method of fitting the regression can change the fact that the relationship between the equity’s market value and the chosen estimator variable is assumed to be linear. Even more important is the reason why a quick look at the linear correlation coefficients successfully predicts the triumph of the EBITDA multiple. It implies that even if we partition the sample into smaller peer groups based on future information, the local estimations fail to produce linear trends contrasting global ones, which casts serious doubts on the raison d'être of doing it in the first place. In other words, it seems very likely, that a generic global multiple linear regression without selecting any peer groups at all could provide superior or at least equal performance.

3.1. Model selection

There are many different ways of building a regression model, the multiple linear approach being merely one of them. The regression method should be picked based on the nature of the relationships between the variables, which means that the first step should be deciding on the list of potential variables.

3.1.1. Variables selection, exploratory data analysis and relationships in the data

The second part of the paper still maintains the assumption that any relative valuation should ideally be the approximation of the DCF method. Consequently, the variables that are assumed to have explanatory power are those that represent the cash flows, the growth rate or the discount factor in the formula. The cash flows can be included through the revenues, EBITDA, EBIT and net income figures, while the growth rates can be estimated directly by simply including the historic values of these cash flow variables and also by calculating the historic growth. Finally,
the discount factor needs to be estimated somehow. It would seem natural to use the WACC or the beta, however their previous underwhelming performance was not an accident. According to the DCF model, the market value of equity is supposed to be a reciprocal function of the discount rate (either cost of equity or the WACC), so one might expect to find no linear relationship, however, a reciprocal function should still result in a strong negative linear correlation and an observable hyperbole on a scatterplot, neither of which can be observed in the sample data. As it will be demonstrated shortly, both the betas and WACCs are remarkably independent of the market capitalization of public companies, the previous independent variables and of each other as well. In a nutshell: they are useless for estimations, even if transformed nonlinearly –unless, of course, there is some confounding variable masking their effect. Unfortunately, there is no apparent confounding variable candidate, and besides, even if they could be used for predictions, only public companies have betas, which essentially makes them inapplicable for the paper’s purposes, as estimating discount rates for private companies involves a high degree of uncertainty and subjective factors. Fortunately, the primary business sector a company or company division is operating in can be easily –and arguably objectively– determined, and companies with similar economic activities are expected to have similar market risks. By nature, the industry sector is a categorical variable, which makes using it in a regression somewhat complicated, but not impossible. Even though it was stated previously that the DCF formula should determine the list of possible variables, the inclusion of other performance metrics should be considered as well. Basically, there is no guarantee that the DCF approach could give a perfect appraisal for every company as its goal is not even estimating the current market value, but the “fundamental value”. There could be other unknown factors –like overoptimistic investor sentiment or an asset bubble– affecting the temporary value. Plus, as was mentioned before, some companies need to be valued as real options or at asset value. Therefore the book value of equity could also be included for starters, as it could perhaps give an approximate lower bound for certain companies, should the DCF method break down. The list of variables will be expanded later, but in order to start looking for the ideal regression model and necessary data transformations the current set should suffice.

The following exploratory data analysis relies on two Python graphical packages, Matplotlib (Hunter, 2007) and Seaborn (Waskom et al., 2017). For clarity the units have been removed from the figure axes on the pair-plots. Data scales are unimportant for inspecting the
relationships, while also being very difficult to read at such small sizes. Figures 4, 5 and 6. present a grid of pair-plots from the whole dataset. EBIT and data earlier than t-1 (the year before the valuation year) is not included due to their redundancy. The diagonals contain the histograms of the given variables. Although these histograms contain very little useful information, they are still better than empty space and demonstrate the presence of power laws in the data’s distribution.

Figure 7. and 8. are scatter plots of net debt and the book value of equity (equity BV) respectively with market capitalization on the vertical axis, where different industry sectors are colored separately. The net debt plot had its the outlier values removed to give a better view of high density areas.

The scatter plots (Figure 4. and 5.) make it crystal clear that even though there is substantial noise in some cases, every explanatory variable exhibits a global linear relationship with the market capitalization. Net debt seems to fit a linear trend very loosely and the book value of equity suffers from similar problems in its raw form. The problem is not their lack of linearity, it is the impossibility of properly modeling them with a function at all. That is where the industry sectors come in handy: as it turns out the sector subsamples have far less noise (see Figure 7 and 8.). Grouped like that, the messy scatterplots suddenly start to make sense and it is clear that separate linear trends can be fitted for the different groups. Even though the improvement is most evident for the Equity BV, the other variables too can be improved in a similar fashion. As for the beta and WACC, Figure 6. demonstrates the aforementioned lack of any hyperbole or negative trend that would be expected if they were good estimates of the discount rates.

When every observable global trend is linear, the optimal model to use is obviously a linear regression. That is very fortunate, as linear models also tend to be relatively simple and hopefully implementable in Excel. Naturally, there are many ways to estimate a linear regression, so the next step of model selection is comparing the different ways of fitting a linear trend.
Figure 4. Pair plot of potential variables with histograms in the diagonal, 2017.12.31.
Figure 5. Pair plot of potential variables with histograms in the diagonal, 2017.12.31.
Figure 6. Pair plot of potential variables with histograms in the diagonal, 2017.12.31.
Figure 7. Scatter plot of net debt and market cap by sectors (data in million USD), 2017.12.31.
3.1.2 Estimator selection

The goal of the paper is to predict the market value of equity \( (Y_i) \) as accurately as possible conditional on publicly available information \( (X_i) \). It has been established, that the target model is globally linear. The expectation is that even with perfect knowledge of the underlying factors, some random errors would always remain \( (\epsilon_i) \). As it was previously stated, a linear model is only linear in its coefficients \( (\beta_j) \), as \( X_i \) can be transformed using arbitrary functions \( (\phi_j) \). This leads us to the previously introduced formula for \( E \).

\[
Y_i = \beta_0 + \beta_1 \phi_1(X_{i1}) + \cdots + \beta_j \phi_j(X_{ij}) + \epsilon_i \quad i = 1, \ldots, n
\]

However, the formula above with the true \( \beta_j \) and \( \phi_j \) is not available, so in order to predict \( Y_i \), we have to find the appropriate functions for the regressors and estimate the coefficients. Fortunately,
the variables do not appear to need any transformations, as they visibly exhibit a linear relationship already. All that needs to be done is to find a method for estimating the coefficients that will maximize the accuracy of estimating the dependent variable, which means minimizing the following function:

$$E[L(Y_i - \hat{Y}_i)]$$

where $\hat{Y}_i$ is the estimated value of $Y_i$ and $L$ is the loss function measuring the accuracy as some function of the distance between the true and estimated market cap. Now $L$ is a random quantity which is inconvenient to use, so its expected value, the risk function is minimized instead. The way to minimize the risk function is to choose an appropriate function estimating $\beta_j$ conditional on some available sample of $Y_i (y_i)$ and the corresponding elements of $X_i (x_i)$. Naturally, $\varepsilon_i$ is irreducible by definition, but the rest of the risk depends on the estimate of $\hat{\beta}_j$, because the exogenous variables are perfectly known and:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(x_{i1}) + \cdots + \hat{\beta}_j(x_{ij}) + r_i$$

In other words the accuracy of estimating the dependent variable is equivalent to the accuracy of the coefficient estimates of the linear regression.

3.1.2.1. Loss functions

In order to further proceed with the coefficient estimator selection a loss function and corresponding risk needs to be chosen first. The definition of errors and residuals is separated in some cases in statistics, errors ($\varepsilon_i$ in the previous section) usually denoting the random deviations that cannot be explained by the estimation procedure, while residuals are the ‘errors’ that result from the imperfect estimation ($r_i$). Since the first type (errors) cannot be eliminated by definition, they are not relevant for the study’s purpose in most cases and therefore the expression “error” will not be reserved for them by default. Instead, both of them will be colloquially called errors. In the special cases where making a distinction between errors and residuals is important, it will be explicitly stated.
3.1.2.1.1. Squared error

The most widespread loss function by far is the quadratic error, with the risk function being the mean squared error (MSE). It has multiple benefits, which are being differentiable continuously, penalizing both negative and positive deviations equally and separately – i.e. equally large errors with inverse signs are not the same as zero errors –, and penalizing larger deviations more heavily. The first of these is universally advantageous, as it makes using it mathematically convenient in optimization problems – e.g. coefficient estimations with least squares methods – while the latter two are situational. In the case of valuing companies, undervaluing a company is equally harmful as overvaluing it, and the negative and positive errors do not cancel out each other, so the gross and symmetrical penalty on errors seems like a good idea, even if a company can only be undervalued by 100% at most, while there is no limit on overestimations. On the other hand, the over penalization of larger errors can be dangerous, especially in the presence of heteroscedasticity and outliers. The topic of parameter estimation will be discussed in the next section, but in advance it is suffice to say, that under these conditions simply minimizing the MSE is inefficient in estimating the coefficients because it gives more weight to high variance observations than optimal. Intuitively, if the dataset happens to be noisy on the higher end, the noise will be amplified in the coefficient estimations. Looking back at the previous scatter plots (Figures 4. and 5.) or considering the underlying mechanism of equity valuation (the errors are not additive but multiplicative, i.e. estimating the market cap of multibillion companies is bound to have higher absolute errors than estimating ones worth only a few million dollars) should make it obvious that both of these phenomena are at large in the dataset.

3.1.2.1.2. Absolute error

The mean absolute error (MAE) is as the name suggests, the average of the absolute errors. The absolute value function handles errors the same way regardless of size and direction, but is not differentiable everywhere, which is huge drawback, as it makes finding the set of coefficients that minimizes the MAE much more difficult. Unlike a least squares regression, least absolute deviations (LAD) regression does not have an analytical solving method (Barrodale and Roberts, 1973). Therefore, an iterative approach is required, the most widespread being a modification of the simplex method, the Barrodale-Roberts algorithm. Although that is hardly an obstacle in
Python, there is no way to calculate the coefficients in Excel without VBA programming, which makes a LAD regression inconvenient for the present study’s purpose. Moreover, even though a LAD fitted regression will not be influenced by outliers and heteroscedasticity, robustness is a double-edged sword. Under typical statistical assumptions, the median is the statistic for estimating location that minimizes the expected loss experienced under the absolute-error loss function, i.e. MAE is the median response of the regressand conditioned on the regressors. That means that a LAD approach suffers from the same disadvantages the median does, primarily that it does not take into account the precise value of each observation and hence does not use all information available. Despite the advantages they offer, absolute error based loss functions will not be considered in the following.

3.1.2.1.3. Percentage squared error

The mean squared percentage error (MSPE) is the expected value of the squared relative loss. The error is measured in percentage of the observed data, not the estimate. The squared percentage deviation is without a doubt an uncommon loss metric, however it is almost as convenient mathematically as the simple quadratic loss due to a square being taken instead of an absolute value. Minimizing its risk function is robust against the heteroscedasticity arising from a multiplicative error model, which is the most likely culprit in the current case. When measured relative to the size of the dependent variable, the errors of estimating the E of large companies have the same weight as smaller ones. The problem of heteroscedasticity is especially acute considering the fact that the frequency of companies has an inverse relationship with their size. This more or less results in a paradoxical situation where most companies are being misvalued in order to properly appraise behemoths like Apple, even though the most probable targets of any relative valuation in corporate finance will be smaller firms.

3.1.2.2. Coefficient estimators for quadratic loss

Before examining the estimators, it should be noted that the exact algorithms used to fit (or later assess) the different regression will not be introduced and derived mathematically, as open source Python packages will be utilized instead of manually writing the code where possible. When adjustments are necessary to standard usage, the changes will be highlighted in detail. The details necessary for implementation in Excel will also be presented.
3.1.2.2.1. Selection criteria

For simple quadratic loss, the risk that needs to be minimized in this case is the MSE of $\hat{\beta}_j$. MSE can be decomposed into two parts: the bias and the variance. Bias is the squared difference between the expected value of the parameter estimate and its true value, while variance is the mean squared difference between estimates calculated from the samples and the expected value of the estimate. (Hastie, Tibshirani and Friedman, 2009)

$$E\left[ (\beta - \hat{\beta})^2 \right] = (\beta - E[\hat{\beta}])^2 + E\left[ (\hat{\beta} - E[\hat{\beta}])^2 \right]$$

Normally, either the minimum variance unbiased estimator or the minimum variance estimator of the coefficients is selected for a regression. Trading bias for variance is giving up on capturing the true relationships in the data in order to decrease the magnitude of the residuals in the individual cases. Is it worth estimating the wrong quantity precisely, or is it better to estimate the right quantity inaccurately? Since our goal is to minimize MSE, exchanging extra bias for lower variance is desirable, as long as the overall MSE decreases, so a minimum variance estimator is preferred.

Besides being unbiased and having low variance, consistency and efficiency are also desirable properties which are closely related to MSE. Consistency means that as the sample size increases, the resulting sequence of estimates converges in probability to its true value, while an unbiased estimator is called efficient if its variance equals the Cramer-Rao Lower Bound. (Hu, 2018) An unbiased estimator achieving this lower bound does not always exist, but if does, then it is the minimum variance unbiased estimator.

It is important to highlight, that the difference between the potential methods is moderate, especially compared to the multiples estimators and selecting any statistical best practice should provide a superior performance. The question of selecting a best estimator does not always have a clear-cut answer, and the topic itself is generally very technical and reaches far beyond the scope of the current study. Therefore, the estimator selection will not be an exhaustive overview of the statistical literature available, but rather a brief reasoning in support of using certain estimators, even if they may not be the best possible choice.
3.1.2.2. Ordinary least squares

Without doubt the most well-known regression fitting approach is the ordinary least squares (OLS) estimator. The estimator minimalizes the following residual sum of squares:

\[ \hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_i \beta_j \right)^2 \right\} \]

Differentiating with respect to \( \beta \) the widely known solution can be obtained:

\[ \hat{\beta} = (X^T X)^{-1} X^T y \]

Assuming there is no perfect multicollinearity or larger amount of variables than sample elements (neither which should be a problem), the above formula always leads to a unique solution.

What makes it special, is that for the linear regression model it is the best (i.e. minimum variance) unbiased linear estimator (BLUE) according to the Gauss-Markov theorem with the following assumptions (Hastie, Tibshirani and Friedman, 2009):

1. \( E[\varepsilon|X] = 0 \)
   The errors of the linear model (not the residuals) have zero expected value conditional on the exogenous variables (\( X \)).

2. \( E[\varepsilon|X] = \sigma^2 < \infty \)
   The errors of the linear model (not the residuals) are homoscedastic, i.e. they have the same finite variance conditional the exogenous variables (\( X \)).

3. \( E[\varepsilon_i \varepsilon_j|X] = 0 \) for \( i \neq j \)
   The errors of the linear model (not the residuals) are uncorrelated conditional the exogenous variables (\( X \)).

If the errors are also Gaussian (i.e. normally distributed), then it is the best unbiased estimator (BUE) of all unbiased estimators, not just the linear ones, because it achieves the Cramér–Rao lower bound (Hu, 2018). However, there are estimators willing to sacrifice unbiasededness for lower variance and overall lower MSE –most notably ridge regression and the James-Stein estimator (the
latter being non-linear and also not-included in following). Without requiring unbiasedness, finding the minimum variance estimator is not straightforward. (Eldar, 2004).

Moreover, it was already established that the model suffers from heteroscedasticity, so for equity valuation purposes OLS is neither the BLUE nor the BUE. Still, heteroscedasticity only means that the OLS estimator is inefficient, as it is still unbiased and consistent. Retaining consistency means, that applying it on a large enough sample can still yield acceptable variance, that is, with a large enough dataset it can be arbitrarily small. Moreover, there is no straightforward remedy for heteroscedasticity, as there is no single unbiased estimator that could unconditionally give a lower variance estimation. The commonly used weighted least squares (WLS) and generalized least squared (GLS) methods can only improve upon OLS if the structure of the error terms is known. Simply estimating the error covariance matrix introduces N degrees of freedom for N observations, which makes these methods unfeasible to use on their own. (Shalizi, 2018) Therefore OLS cannot be discarded just yet.

Looking back at the pair-plot of the variables (Figure 4. and 5.), the REV, EBITDA, and NI data from successive years are visibly highly correlated. Although multicollinearity does not affect unbiasedness, consistency or even efficiency, only individual coefficient estimates –it is still and important issue. In an OLS multiple regression adding too many variables leads to overfitting (without cross-validation the MSE is a strictly decreasing function of the number of regressors), therefore a list of optimal variables needs to be selected. Multicollinearity increases the variance of coefficient estimates, which makes it harder to reliably tell which variables need to be dropped. Not to mention, that dropping a correlated variable changes the performance of the other non-orthogonal variables, so every possible variable combination needs to be evaluated, as it is not enough to look at variables one by one. Although the estimator’s consistency means that the coefficient estimates’ standard errors can be made arbitrarily small –and therefore the subset of variables can be reliably selected–, in practice the sample size is not large enough to disregard multicollinearity. The problem of overfitting and cross-validation will be discussed in a later section (3.2.).

In conclusion the OLS method is very accessible and widely known (even among finance professionals), has very strong properties and is also easy to interpret its results. However, the
assumptions underpinning its desirability are severely violated in the present case, which is a very good reason to disregard it altogether. Still, even if suboptimal, most corporate finance analysts are using Excel for their statistical and valuation needs alike, and also favor easily interpretable results, therefore an OLS approach will be attempted.

3.1.2.2.3. Ridge regression

Ridge regression is very similar to the OLS linear regression both in formula and calculations. What is different, is that ridge shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares.

\[
\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_i \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}
\]

Where \( \lambda \) is the complexity parameter that controls how much the coefficients should be shrunk towards zero (and each other) through the introduced penalty. Evidently, if \( \lambda = 0 \), it is the previously introduced OLS estimator. The main advantage of ridge regression is that its MSE is always strictly less than that of OLS for some \( \lambda > 0 \), and the difference increases with the multicollinearity between the regressors (which is very prevalent in the present case). The catch is of course, that no single \( \lambda \) can beat \( \lambda = 0 \) (i.e. the OLS case) and there is no automatic way to find that value. (Hoerl and Kennard, 1970). Intuitively what happens, is that by shrinking the coefficients they can deviate from their expected value less and less, therefore estimation variance decreases, while squared error from bias increases. The proof of Hoerl and Kennard concerns the existence of a sweet spot, where the increase in bias is smaller than the decrease in variance. Since the coefficients’ estimation variance tends to explode for non-orthogonal variables, it is easy to see why ridge works best for correlated variables.

It can also be easily seen that the solution found by this formula is not equivariant under scaling of \( x_i \), so the data has to be standardized before use. In addition, the intercept \( \beta_0 \) has been left out of the penalty term, as its penalization would make the procedure depend on the origin chosen for \( y_i \) - adding a constant to each of \( y_i \) would not simply result in a shift of the predictions by the same amount \( c \). There are multiple solutions to the minimization problem. One is to estimate
the intercept as the mean of the sample regressands ($\beta_0 = \bar{y}$), then center every regressor and estimate without an intercept. The other approach is to simply set $\lambda = 0$ for $\beta_0$. Afterwards the formula for the coefficients is: $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$, where $X$ either has $p$ columns for $p$ variables, or $p+1$ with the constant, but in the latter case the corresponding diagonal of $I$ is 0. (Hastie, Tibshirani and Friedman, 2009).

The main challenge of using ridge regression is that the optimal value of $\lambda$ — a continuous quantity with no mathematical upper bound — is unknown and has to be found by cross-validation. Fortunately in practice — especially after standardization — finding the scale parameter is manageable as the optimal shrinkage usually stays pretty close to single digits. However, variable selection is necessary for ridge as well, even if the smaller effect of multicollinearity makes it easier. Another problem with ridge regression is, that it is vulnerable to heteroscedasticity the same way OLS is.

Implementing ridge regression is also possible in Excel without any VBA coding once the optimal value for $\lambda$ is found. While there is no built in ridge regression functionality like there is for the OLS method, there are built in functions for taking the dot product of matrices or transposing and inverting them, which allows calculating the coefficients manually.

3.1.2.2.4. Percentage least squares

It is clear, that both previous estimators suffer from heteroscedasticity. However, if the risk function is the expected value of the quadratic relative loss, heteroscedasticity only persists if the relative losses themselves are heteroscedastic. There is a very good reason to believe that it is not the case and controlling for company size eliminates or greatly mitigates the problem. The paper Least Squares Percentage Regression by Chris Tofallis demonstrates, that if the risk function is the MSPE, then:

$$\hat{\beta} = (X^T D^2 X)^{-1} X^T D^2 y$$

can be used to find the coefficient estimates, where $D$ is an $n$ by $n$ division matrix containing $1/y_i$ in the $i^{th}$ diagonal position. Naturally, that means that $y_i$ cannot be equal to zero, however the market value of equity is always positive. The above formula is the result of applying the OLS method after dividing the $i^{th}$ row of the sample $x_i$ matrix of exogenous variables (including the constants) by $y_i$ — i.e. differentiating the relative errors with respect to $\beta_i$. After the division the constant is no
longer a constant, instead it will ‘behave like other variables’ in the estimation. This is very important, as this will necessitate some adjustments to the Python packages typically used for fitting linear regression, because they handle constant terms separately. If the new model errors \((e_i/y_i)\) satisfy the Gauss-Markov assumptions, the resulting estimates of the parameters will be the maximum likelihood estimation, and the estimator is also shown to be unbiased. (Tofallis, 2009)

Unfortunately there is no proof available for the efficiency or consistency of the estimation. It should be emphasized that this method is not better at minimizing the MSE than OLS, only at minimizing the MSPE. On the other hand, measuring the accuracy with MSPE is potentially even better for the paper’s goals than the MSE, because the model performance is equally important on companies of all sizes.

Once again, the estimator is easy to implement is Excel, because it can be calculated the same way an OLS regression is after a dividing the variables (both the dependent and independent ones) by \(y_i\), as the “data analysis” regression tool has the option to fit a regression without an intercept. But even without it, the matrix operations described above could be carried out with the built in matrix functions. Using ridge regression on relative errors is a bit trickier but can be carried out manually in the same fashion as the vanilla ridge introduced before.

### 3.2. Model assessment

It is no secret that the any practical application of a regression model is at risk of overfitting. In fact, in the authors’ personal experience any time a potential statistical model is mentioned in a financial or economic setting, the first reaction is usually to dismiss it as “probably being overfitted”. Accordingly, the question of cross-validation is perhaps even more crucial than the previous model selection.

Overfitting is a general term for the phenomena that for a sample the accuracy of the model is strictly increasing with model complexity (in this case the number of variables), while it also becomes less and less applicable for different samples of the same population. In the extreme case, if there is a variable for every single sample element, the model can simply learn every datapoint, which is obviously harmful, because the goal is to find a general pattern that can be used with other data as well. The process of assessing how well a model will generalize to an independent dataset
is called cross-validation and selecting the best scheme depends on the underlying data and applied models. As a result, doing it can be just as complicated as selecting an estimator. In the following however a simplified approach will be taken, as the ultimate goal of the paper is not to exactly determine the accuracy of potential models. Of course, it is a very important information, but an approximate value is more than enough, since the improvements over the multiples approach are going to be substantial. Any assessment method that successfully prevents the models from increasing their complexity indefinitely can be considered valid.

3.2.1. Random subsampling

Perhaps the most common way of handling overcomplexity in models is splitting the dataset into a training and a test subsample before applying any supervised learning procedure. Afterwards the model coefficients are estimated on the training dataset and assessed on the test dataset. Since the resulting score will be random depending on the drawn subsamples, the process has to be repeated many times in order to get conclusive results, utilizing the fact that the empiric parameters tend converge by the law of large numbers to their true population value. Unfortunately this convergence can be very slow, easily requiring 1000-10,000 iterations just to be sure of the first two decimals of the test $R^2$. Although the random splits are not computationally intensive themselves, applying a model many times is. The other issue with train-test subsampling is that that by having a reduced sample size for train sets the estimations are going to have more variance as well, and therefore the accuracy will be lower than it would be for an actual estimation using all the data. This problem is very acute for industry sectors like utilities that are made up of a few dozen constituent companies. In fact, if every single year is separately cross-validated in the sample, then “all US public companies” might not be a large enough dataset on its own. Finally, multiyear random subsampling means testing models fitted on future data and testing it on the past, which does not seem right.

For model assessment purposes all the company data will be merged together from 2017 onwards all the way back to 2012. As it will be demonstrated later, even after the necessary pre-filtering that still means ~10,000 elements. Afterwards the training set will be 2/3 and the test set will be 1/3 of the starting sample. This is arbitrary of course, but an often used convention. Since the data is not expected to be entirely stationary year over year due to macroeconomic factors (e.g.
market risk premium or risk free rate), additional variables will have to be included that allow regressions to fit different slopes for the variables in different years. For example the market average EV/EBITDA for the S&P500 companies together with the EBITDA features can be a good proxy for asset inflation if included in the model as a constant and also as an interaction variable – the constant multiplied with the cash flow variables.

3.2.2. Out-of-sample

Because of the problem of testing future data on past data, non-stationarities and serial correlation, practitioners usually resort to out-of-sample (OOS) evaluation instead, where a section from the end of the series is withheld for evaluation. It can be said, that in the traditional forecasting literature OOS evaluation is the standard evaluation procedure. (Bergmeir, Hyndman and Koo, 2017). In the current case that means, that every year except the first will be used as a test set, and for every test set all the previous years’ data will be the training set.

Now OOS sampling is extremely convenient, as unlike the previous case that means five iterations for the six year data instead of thousands. As one might expect, there is no such thing as a free lunch. The trade-off with OOS sampling is that unless some assumptions are made, there is very little guarantee that five OOS samples give a better evaluation than 5 random subsamples. Therefore the following is assumed:

1. The composition of companies on the US stock market stays more or less the same year-over-year – which is obviously true. More or less means that only a marginal amount of public companies go private/merge/get acquired/default/etc. and similarly, few new ones enter. In a random subsample the composition is entirely random and almost surely very different. In other words the conditions vary much less in practice than in a random sample.

2. When the true valuation is carried out, the model will be fitted on historical data as well with the previously mentioned fixed composition. Consequently, using random subsamples is a poor simulation and is expected to underestimate accuracy because it assumes more difference between the train and test samples than reasonable.

Naturally, as the introduction clearly stated, the target companies for a valuation are going to be private companies. The market value of public companies is already known and no attempt
is made at trying to beat the market. The given price is accepted as optimal for better or worse. Dealing with a fixed composition training set that is similar to the test set might give a realistic picture of accuracy when valuing public companies, but there is no such guarantee for private ones. At this point, one has to decide whether to believe in private companies having a composition similar to the American stock market. This might be a serious leap of faith, but the same assumption has been implicitly made for the market multiples already. This can be most easily seen by the fact that even the widest filtering criteria only managed to find peers for half of the public companies. If the standard methods cannot even find comparables for public companies, how well could it find them for private ones? If the OOS sample results are at risk of overestimating the accuracy of the linear regressions, then the same could be said about the benchmark scores. The methods outlined in the following do not have to be perfect to be useful, they only have to be better than the standard approaches.

### 3.3. Dataset

For the model demonstration all US based companies traded on any US stock exchange (excluding OTC markets) have been included for the period between 31.12.2012 and 31.12.2017. Since every sample contains historical data going back 4 years (t-4), the earliest features are from 2008. For these companies the following features were downloaded in all of all the six years:

1. Market capitalization
2. Net debt
3. Book value of equity
4. Revenues (t-0, t-1, t-2, t-3 and t-4)
5. EBITDA (t-0, t-1, t-2, t-3 and t-4)
6. EBIT (t-0, t-1, t-2, t-3 and t-4)
7. EBIT (t-0, t-1, t-2, t-3 and t-4)
8. Net income (t-0, t-1, t-2, t-3 and t-4)
9. S&P 500 average EV/EBITDA multiple
10. Primary business sector (first digit)

In the next step financial companies having no EBITDA and EBIT metrics have been dropped from the dataset. They can be valued in the exact same fashion based on their EBT however. Although the model is completely compatible with them, for simplicity’s sake they are excluded. Companies having negative EBITDA in t-0 or t-1 have also been removed, as they can hardly be valued based on their income and are therefore out of scope. Finally, some companies had to be filtered out because their income metrics were identical or equal to zero in consecutive years indicating potentially erroneous/missing data or a default. The remaining dataset has ~10,000 elements –25 features each— for the six years.

The database of Capital IQ has data for both currently public companies, and companies that used to be public, but for various reasons are not anymore. Unfortunately, data can only be downloaded en masse in their Excel add-on for companies that are still being traded at the moment. Additionally, relatively fresh IPOs cannot be included either, as they do not have data going back all the way to 2008, which is necessary for the OOS cross-validation. These two factors can easily result in selection bias, so that should be kept in mind.

The industry sector could be included as a dummy variable with one-hot encoding and 11 sectors corresponding to 10 new binary variables. On top of that, for every variable that is expected to be more accurate when accounting for sectors, an additional interaction variable would have to be included for all of the 10 starting dummies to allow for a different slope beyond the different intercept. Besides net debt and Equity BV –for which the potential improvement was visible on Figures 7. and 8. – potentially every other variable could use interaction features, as the sectors are the proxies for the risk (i.e. the discount rate) on cash flows. It is easy to see how dummy features could get out of hand quickly. Instead, a different approach is proposed: fitting a different regression for every sector. The resulting sector specific custom hyperplanes do the same job as the otherwise necessary countless synthetic variables.

Finally, some synthetical variables are also included. These include the calculated growth rates, corresponding interaction variables and the interaction variables for the market average EV/EBITDA.
3.4. Variable selection

The main challenge of valuing equity with regressions is keeping model complexity in check by cross-validation. The starting 24 independent variables are bound to overfit the model even before adding synthetic variables, so most of them has to be eliminated as a first step. There are many ways to select variables. These range from simple univariate analyses to lasso regressions. None of these will be considered however, as none of them can easily deal with the very high levels of multicollinearity in the data. In some cases, the income variables are so self-similar across time, that they could be considered a single underlying feature with some random noise added every year (see seen on Figure 4.). While 24 features is too much, they can be easily grouped. On one hand there are the net debt, equity BV and average EV/EBITDA variables that are relatively independent of everything else –especially the latter which is a constant every year. Moving on to the income variables, they are mainly correlated with other income features of the same kind, and also with the other income features from the same year. If every cash flow variable could be somehow compressed together year-by-year (i.e. cross-sectionally) the resulting general annual income variables would suddenly reduce the number of variables to 10. If the same types of incomes could be merged together for the consecutive periods (time-serially), that would leave only 9 features, one for every earnings level. The potential opportunities for compressing features can be easily visualized with a cluster-heatmap of the correlations (Figure 9.).

The revenues are by far the most correlated, followed by EBITDA, EBIT and net income, which makes a lot of sense considering the fact, that the lower levels of an income statement have more added noise from the different potential expenses. The heatmap also happens to follow intuition in that features from direct consecutive years are less orthogonal than those more further apart.

Ridge regression is perhaps far less sensitive to the multicollinearity and could more reliably select among the features, but it can only do so with the right $\lambda$ parameter, which needs to be found with cross-validation. Combining the few hundred iteration of lambda with the multimillion potential combinations of the 24 variables, the resulting process is too much for a desktop computer.

### 3.4.1 Principal component analysis

Principal component analysis (PCA) is a statistical procedure using orthogonal transformations to convert a set of correlated variables into a set of values of linearly uncorrelated variables called principal components. Since there are infinitely many ways of transforming the variables into principal components, the transformation is defined in such a way that the first principal component has the largest possible variance (that is, it accounts for as much of the
variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. (Zaki and Meira, 2014) There are three serious issues however. Firstly, there is no underlying principle guaranteeing that the maximum variance directions of the data contain the most information. In fact, PCA is sensitive to the relative scaling of the original variables as multiplying a feature by a constant number increases its variance without adding any explanatory power, but the increased variance gives it more weight in the final components after the process. Secondly, principal components are far less interpretable than the original features. Finally, the algorithm for PCA involves calculating eigenvalues at one point, which cannot be done in Excel using the built-in functions and tools.

On the upside, since the variables that need PCA are the same types of incomes, their scale is also the same, so there is no need for standardization. Moreover, even though the regression cannot be run on principal components in Excel, PCA can still be used to reduce the number of variables to a manageable level, select the best subset of them and then simply use the original features of the optimal principal components to appraise equity, especially if a ridge regression is used, as it does not care much about multicollinearity.

By examining the heatmap (Figure 9.), the following will be carried out:

1. Merging every revenues feature into a single component. They are almost identical, so a single component is more than enough.
2. Merging the EBITDA into 2 components. Since the cash flows closer to the valuation are expected to have more explanatory power, the first principal component will be extracted from the t-0 and t-1 features and the second from the remaining three.
3. Extracting 2 primary components from EBIT the same way as before.
4. Leaving NI t-4 as is, since it is relatively uncorrelated.
5. Extracting 2 components from the remaining NI variables (t-0 and t-1, t-2 and t-3).

3.5 Model building

The packages used for fitting the models are Pandas, NumPy, Statsmodels (Statsmodels.org, 2018), Scikit-learn (Pedregosa et al., 2011) and itertools. As previously, Pandas
and NumPy are for moving data in dataframes and performing matrix operations respectively. Statsmodels is used for fitting the linear regression for the percentage least squares estimator. This package could have been substituted with Scikit-learn as well, but the idea of switching out the functions in the original code was discarded due to time constraints. Finally itertools is used to generate the combinations of a given size and it’s included in Python 3.6 by default. The code itself is contained in appendix2.py Finally, it should be noted that the actual algorithms deviate from the theoretical regression estimates, as a lower bound is placed manually on the end result, which is zero. Of course, the estimators will not optimize the coefficients accordingly, but that is actually a good thing. An estimator exploiting the bounded downside error would seriously undervalue the market on purpose, as its downside risk is smaller than the upside. Additionally, such an estimator would most likely have no analytical solution, only iterative ones, making it inapplicable in Excel.

The squared error will be measured in \( R^2 \) for easy interpretation and comparability with the multiples tests. Since \( R^2 \) is inversely proportional to MSE, maximizing it is the same thing as minimizing the MSE. The biggest problem with using the MSE directly is that its size depends on the scale of the data.

The process will be the following:

1. Importing the data from Excel spreadsheets;
2. removing companies with negative EBITDA t-0 and t-1;
3. removing companies that have a zero or “not a number/NaN” for any feature;
4. performing PCA by transforming variables in the previously outlined fashion;
5. fitting a regression for every variable combination up to size \( x \) for every OOS training sample;
6. calculating test scores on the OOS test datasets.

Beyond the previously outlined principal components, NI t-4, net debt, equity BV and the mean market EV/EBITDA ratio are also included in the first trial for a total of 12 starting features.

During the first run the maximal size of combinations is 10. Additionally, the inclusion of “Net Debt” and “Sector” features is set to mandatory. The OOS cross-validation results in a 6
variable optimal setup, which means that it is more than capable of capping model complexity (it would have picked 10 variables otherwise). The test $R^2$ is 86%, with “Equity BV”, “SPX EV/EBITDA”, “REV PC”, “EBITDA PC t-0, t-1”, “Net Debt” and “Sector” being the optimal setup. When testing this same setup with random training/test splits, the average test $R^2$ is 86,6%, which is already higher than what the different scenarios achieved during the multiples benchmarking (Figures 1, 2 and 3.). The average number of valued companies is $n=1703$, which is far in excess of the comparables approach. Forcing the model to use net debt and sectors can seem to be arbitrary, however previous experience with many different tests not included in the final paper and random subsampling always ended with net debt and the sectors being included in the final models. During continuous development of the code, the functionality of fitting regressions without sectors was not even updated after a certain point, as it seemed pointless to put effort into such approaches. The truth is, that at the end of the day, forcing these variables is not going to bias the results upwards, as cross-validation is performed on thousands of entirely random subsets of the 6 year data after OOS fitting, so it is not about fine tuning for the latter. If anything, this can only decrease potential performance by restricting the variable optimization.

In the next step the growth rates and the corresponding growth- and SPX EV/EBITDA interaction features for the EBITDA principal component are added. Running the previous OOS cross-validated variable selection process for a maximum of 10 combinations once again, the new optimal setup is “Equity BV”, “SPX EV/EBITDA”, “EBITDA SPX int”, “EBITDA CAGR int”, “REV PC”, “EBITDA PC t-0 and t-1”, “Net Debt” and “Sector”. In short, besides the previous six variables the two new synthetic variables are selected by the cross-validation, as was expected. The OOS method still manages to avoid overfitting though, as the maximal amount of allowed variables has still not been reached. The new test $R^2$ is 88%. The random subsampling with 1000 iterations yields the following test score: $R^2=88.6%$. The $n$ is still 1703 companies as before. Looking at the multiples (Figures 1, 2 and 3.), the regression model is clearly starting to come out ahead.

Since these results used PCA, the accuracy should be tested on the vanilla variables to assess what could be expected when using Excel. Therefore,
“Equity BV”, “SPX EV/EBITDA”, “EBITDA SPX int”, “EBITDA CAGR int”, “REV PC”, “EBITDA PC t-0 and t-1”, “Net Debt” and “Sector”

is replaced by:

“Equity BV”, “SPX EV/EBITDA”, “EBITDA SPX int”, “EBITDA CAGR int”, “REV t-0”, “EBITDA t-0”, ”EBITDA t-1”, “Net Debt” and “Sector”.

Where the interaction features are the product of the constants and EBITDA t-0. The combined REV principal component will simply be substituted with REV t-0, as all the revenues variables are so similar that it should not matter which one is used (see Figure 9.). Therefore, the latest one is picked. Without PCA, multicollinearity becomes an issue yet again, so a ridge regression will be applied on them. Ridge regression is included in scikit-learn, however that version is not compatible with a PLS estimation, so another open source code was used (Gist, 2018) which was easier to modify. Ironically, the ridge model eventually turned out to be heavily underperforming for percentage least squares, therefore it will not be introduced. The results of 100 random train/test splits can be observed on Figure 10. At \( \lambda = 0 \), the \( R^2 \) of the OLS approach is still above 88.5%, while even at the peak, with \( \lambda = 1,2 \) the \( R^2 \) only goes as high as 88.6%. That is hardly enough reason to substitute an OLS estimator with a ridge one, as the former is much more widely known and easier to use in Excel.
There is only one thing left that needs to be checked, the PLS estimators. Even though these results will not be directly comparable with market multiples, it was stated that optimizing for minimal relative errors could be desirable. By taking a look at Figure 11, it becomes very clear why it is not enough to target maximal $R^2$ (or equivalently minimal MSE). As expected, the relative errors are marginal on the higher end of market capitalization, but easily reach multiple hundreds (or thousands in %) by following an extreme power law. When the same variables are fit using the PLS approach, the relative errors look very tame in comparison (Figure 12.). However, optimal feature selection will not be performed on percentage squared errors, as the established theoretical background behind the PLS is very limited to say the least.
Looking back at multiples, more precisely the 500-1000 large set of companies that cannot be valued by multiples, their ranks are very likely made up from these smaller stocks that elude global regression models as well. By getting rid of every company under 5000 million USD, estimation accuracy would increase substantially.

4. Conclusions

It should go without saying, that the model demonstrated above is very far removed from optimal solutions. Still, it performs favorably against standard market multiples. It is also easily implemented by anyone frequently using multiples for equity valuation. All that needs to be done is downloading some key financial metrics for all the companies in the same first digit business sector for the past few years –which is a trivial task for anyone having access to either Bloomberg or Capital IQ–, generating the synthetic variables with a quick multiplication and running a regression in Excel on the aforementioned features (latest revenues, latest EBITDA and EBITDA
for the year before, the book value of equity, net debt, the mean EV/EBITDA ratio for the S&P500, the 5 year EBITDA CAGR, the product of the latest EBITDA and the CAGR, and finally the product of the latest EBITDA and the mean EV/EBITDA), and using the resulting coefficients for the appraisal. While seemingly complicated, this can be actually done on a completely automatic basis with VBA macros or in Python of course. It also gets rid of the manual peer group selection procedure, which can easily be more time intensive then the suggested regression process.

As for potential future research, the most critical deficiency of the current model is its inability to value small companies regardless of using percentage least squares or ordinary least squares. Perhaps the most important improvement would be a classifier algorithm, that could tell whether a company can be valued by the larger and easy to grasp public companies, or not.

5. References


