The Relationship Between Corporate Governance and Company Performance

New Factors, Models and Approaches to Causality

Conor Reid B.A, B.A.I, H.Dip

A Dissertation submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business, University College Dublin

August, 2018

Supervisors: Dr. James McDermott and Dr. Miguel Nicolau

Head of School: Professor Ciarán Ó hÓgartaigh

Dedication

To my...

Contents

List of figures **vii**

Acknowledgements **viii**

Abstract **ix**

1 Introduction **1**

1.1 Background 1
1.2 Business Motivation 3
1.3 Academic Contribution 4
1.4 Research Goals and Scope 4
1.5 Document Outline 6

2 Literature Review **7**

2.1 Introduction 7
2.2 Company Performance - Measures and Influencers 8
2.2.1 Introduction 8
2.2.2 Financial Ratios 9
2.2.3 Environmental Considerations 21
2.2.4 Corporate Social Responsibility 24
2.2.5 ESG Disclosure 27
2.2.6 Executive Compensation 29
2.3 Corporate Governance and Company Performance 30
2.3.1 Introduction 30
2.3.2 Existing research 31
2.4 Inferring Causation 34

2.4.1 Introduction . . . 34
2.4.2 Matching . . . 36
2.4.3 Minimal-Model Semantics . . . 42
2.5 Research Gap . . . 44

3 Methodology **46**
3.1 Introduction . . . 46
3.2 Data Acquisition . . . 47
3.2.1 Core Data . . . 47
3.2.2 New Factors - Independent . . . 48
3.2.3 New Factors - Dependent . . . 48
3.3 Data Pre-Processing and Reduction . . . 49
3.3.1 Missing Values . . . 49
3.3.2 Outliers . . . 51
3.3.3 Thresholding . . . 52
3.4 Data Mining, Algorithms and Software . . . 53
3.4.1 Classification . . . 53
3.4.2 Regression . . . 54
3.4.3 Causal Estimation . . . 56
3.5 Causal Estimation - Motivating Statements . . . 57
3.6 Interpretation . . . 58
3.6.1 Classification . . . 58
3.6.2 Regression . . . 59
3.6.3 Causal Estimation . . . 59

4 Results **60**
4.1 Introduction . . . 60
4.2 Classification . . . 60
4.3 Regression . . . 64
4.3.1 SPX . . . 64
4.3.2 SXXP . . . 67
4.3.3 EEBP . . . 69
4.4 Causal Estimation . . . 71

4.4.1 BOD.Age.Rng - SPX 71
4.4.2 Indep.Chrprsn.Feml.CEO.or.Equiv - SXXP 73
4.4.3 CEOPayOverMedian - SPX 75

5 Discussion **77**
5.1 Introduction 77

6 Conclusions and Future Research **78**
6.1 Introduction 78

Program code **79**

Appendices **79**

Bibliography **80**

List of Notation **85**

List of Tables

2.1 Enron Scandal - Mahama (2015) 17

4.1 Classification Results - SPX, Tobin's Q 61
4.2 Classification Results - SPX, Altman Z Score 61
4.3 Classification Results - SXXP, Tobin's Q 62
4.4 Classification Results - SXXP, Altman Z Score 62
4.5 Classification Results - EEBP, Tobin's Q 63
4.6 Classification Results - EEBP, Altman Z Score 63
4.7 Regression Results - SPX, Tobin's Q 64
4.8 Regression Results - SPX, Altman Z Score 65
4.9 Regression Results - SPX, MScore 66
4.10 Regression Results - SXXP, Tobin's Q 67
4.11 Regression Results - SXXP, Altman Z Score 68
4.12 Regression Results - EEBP, Tobin's Q 69
4.13 Regression Results - EEBP, Altman Z Score 70

List of Figures

2.1 Logit Regression for Beneish M-Score (Source: Herawati (2015)). 18
2.2 Beneish M-Score for Prosecuted Malaysian Companies (Source: Kamal *et al.* (2016)). 20
2.3 Correlation between corporate environmental protection spending and economic success (Source: Schaltegger and Synnestvedt (2002)). 22

4.1 BOD.Age.Rng / SPX / Tobins Q 71
4.2 BOD.Age.Rng / Altman Z 72
4.3 Indep_Chrprsn_Feml_CEO_or_Equiv / Tobins Q 73
4.4 Indep_Chrprsn_Feml_CEO_or_Equiv / Altman Z 74
4.5 CEOPayOverMedian / Tobins Q 75
4.6 CEOPayOverMedian / Altman Z 76

Acknowledgements

Thanks to...

Abstract

This project...

Chapter 1

Introduction

This chapter forms the introduction to the current study, laying out its motivations and aims. We begin with a background of the domain, followed by the business motivation for this study. We follow this with a summary of the academic contribution, the research goals and scope and end with an outline of the document structure as a whole.

1.1 Background

This study is primarily concerned with the relationship between corporate governance and company performance, particularly with how the former can be optimised to positively influence the latter. Corporate governance is a widely discussed, debated and researched topic that is as relevant today as it has ever been. The governance of a company dictates it's policies and motivations, ensuring that all stakeholders [1] have input to how the company is run and share a vision of where it is going. Governance policy also acts to mitigate financial and ethical pitfalls, by setting clear standards. Thus it is fair to say that corporate governance has a wide ranging influence within and also outside the company.

[1] A stakeholder is someone who has a stake, or a personal interest, in the company. Employees, the local community and the media are all stakeholders.

We frequently see instances of corporate governance failure, which can lead to disastrous consequences financially and reputationally. Reputation is often of high importance in both the public and private sectors, which can lead to a promotion of more ethical and fair behaviour in order to protect it. Instances where companies fail in this regard often make eye-catching headlines, for example McLaughlin (2017), McVeigh (2015) and Kirkpatrick (2009). In a hyperconnected world where news spreads quickly, the importance of a functioning governance structure is more important than ever.

The interests of different parties can conflict in any business, between entities such as the shareholders [2] or directors. There is much debate on how best to align these interests, with suggested initiatives like structuring executive compensation to be at least partly dependant on firm performance. Shareholder interests can also conflict with the interests of the wider public as a whole, or the stakeholders. This is especially true for companies that heavy rely on natural resources as a driver for business. In this case, sustainability not just of the company but of finite natural resources that the public depend on must be closely governed and managed. This is often the responsibility not just of those within the company, but those outside it too.

It is reasonable to argue that corporate governance influences all aspects of the company, not least its economic success. Moldovan and Mutu (2015) studied this relationship, collecting data on corporate governance and using it to predict corporate success as measured in various ways. They were able to learn models that did this successfully, resulting in a number of rules dictating governance of high performing companies. The current study uses this as a starting point, and looks to address some limitations within in a number of areas, outlined in section 1.4. Modern work around causation is also studied, with a view to applying it in this domain. This would act to strengthen previ-

[2] A shareholder is often an investor who has equity in the company. They often have no personal interest in the company, solely financial.

ously derived relationships, and presents an opportunity to study the deeper causal influencers of corporate economic success.

1.2 Business Motivation

Moldovan and Mutu (2015) state the conclusions they reached in their research, and point to the business significance of each. For example, they found that for US based companies the number of women on the board of directors was positively connected to company performance. They also found that in Western Europe, companies should employ larger audit teams that in turn lowers the risk of bankruptcy. In Eastern Europe, their main finding is that an independent chairman best influences economic success.

The business benefit of the above is obvious. By deriving a number of relationships between economic success and corporate governance, the authors first prove that a relationship does in fact exist in the first place. That is, high corporate governance performance is strongly linked with success. Secondly they are able to put forward recommendations for governance best practice and show what elements are most influential, with geographic context. A key element of management is identifying levers with which to effect outcomes in a positive way, which leads into the motivation for this research.

We look to first verify some of the above findings, but also look to find new influencers of economic success by expanding the research to include other predictors outside of governance features. This would in effect expand the array of tools available to corporations for effecting economic change. We also aim to strengthen these findings by seeking causal influencers, which would add a hierarchical element to the range of levers for change and direct efforts to spaces that are most likely to yield real success.

1.3 Academic Contribution

A key element of this study is the exploration of causality research and the application of these techniques in this domain. There is continual active research in this area, with interested parties offering new techniques and thought processes for making steps towards proving causation in a variety of domains. To our knowledge, causation research has not be applied in the area of corporate governance and its effect on outcomes, and thus would represent a novel endeavour that stands to contribute to the field in a meaningful way.

For example, the rules proposed by Moldovan and Mutu (2015) are backed by strong correlations drawn from highly accurate statistical model. They make no steps towards estimating a cause and effect element to those relationships, or any other type of deeper analysis. We propose that a significant academic contribution would be had by exploring how causality is reached and applying it here, to see if more can be said of the aforementioned rules.

As mentioned above, this study plans to expand the work of Moldovan and Mutu (2015) to include other types of company actions and activity. This would help gain a more wholistic view of how economic success can be promoted across all company functions.

1.4 Research Goals and Scope

There are a number of key goals that this study aims to achieve. They are presented below, along with a discussion of how success will be measured at each stage.

1. **Reproduce the findings of Moldovan and Mutu (2015).**
   As mentioned, Moldovan and Mutu (2015) made findings that point to interesting relationships between corporate governance and company performance. It would be useful to use similar data to reproduce some of these findings using the same techniques as the authors.

2. **Improve on these findings**

   Next, the aim is to improve on these results using three methods. The first involves considering other predictors of corporate success beyond governance, such as a company's social responsibility performance or their environmental impact. The second involves using alternative measures of corporate success itself, that may better reflect how successful a company is. These may take the form of more informative financial ratios etc. The third method involves using alternative statistical techniques with a view to improving model performance using the standard measures of model accuracy. There are a plethora of techniques and algorithms not considered in the original study that may prove useful. The way in which data is preprocessed may also be altered as part of this step. For example, the authors discretised corporate success and perform classification on the resulting data. It may be advantageous to perform regression analysis here to gain greater granularity.

3. **Apply modern work on causality.**

   A number of conclusions on the influence of corporate governance on company performance have been reached, using established statistical analysis and subsequently discovered correlation. In order to strengthen these findings and and gain deeper insight into the underlying mechanisms of the domain, modern work in causality will be applied. This will involve significant research into the ways in which this can be achieved, including data requirements and required pre-processing. The aim here is to gain a much deeper understanding of the causal influencers of corporate economic success, to drive best practice and contribute to knowledge base in this area.

It is equally important to discuss what is out of the scope of this study. A distinction is not made in this study between public and private companies, although regulations dictating how public companies must govern are often more stringent and strictly enforced than private companies. Regulations include preventative measures for avoiding bankruptcy etc. Privately held firms often how more freedom and flexibility here. This can be especially true when

dealing with audits and so on.

Further, regulatory differences from country to country are not considered. Some countries may introduce certain taxation and laws that influence the decisions made by local companies, like a carbon emissions tax that may make companies take their environmental footprint more seriously.

1.5 Document Outline

This document is laid out as follows. Chapter 2 contains a brief literature review of this topic including how corporate success can be measured, other predictive corporate features that may be included, a review of other similar studies and concludes with a summary of research in the area of causation. Chapter 3 contains details of this studies methodology, including a summary of the data used and its pre-processing, algorithms used and methodology around applying causal techniques. Chapter 4 contains the results of this study. Included in chapter 5 is a discussion of these results, with some concluding remarks and opportunities for future research outlined in chapter 6.

Chapter 3

Methodology

3.1 Introduction

This chapters outlines the methods followed during this analysis, with a view to explaining the steps taken in detail with a view to aiding replication. The Knowledge Discovery in Databases (KDD) process as outlined by Fayyad *et al.* (1996) was followed where possible. This is a well established end-to-end framework for deriving knowledge from raw data. To that end, data acquisition and the raw data characteristics are discussed first. This is followed by a discussion of the preprocessing and reduction required to make this data useable, including how missing data and outliers are handled. An outline of the data mining techniques employed is given, including analysis of the advantages and disadvantages of the various statistical methods and associated software packages available. Following this, the steps taken for each element of this study are outlined in detail. That is, the replication of and expansion on the work by Moldovan and Mutu (2015) and the application of causal research in this domain. An analysis of the methods for interpreting results is given, including what measures of algorithmic success should be used and related matters. Overall, this chapter represents the technical aspect of this study and aims to facilitate the replication and expansion of this analysis by future researchers.

Git and GitHub was used as a version control and task tracking tool. All code modules written in fulfilment of the present study, including all source files of this report, are available at `https://github.com/ReidConor/dissertation`. Using source control has various uses, not least acting as a cloud storage mechanism in case of local machine failure. In addition, changes to the project over time can be much more easily managed, including changes to the datasets used. Making this code available on GitHub also facilitates collaboration and the communication of ideas between collaborators.

3.2 Data Acquisition

3.2.1 Core Data

The primary source of data for the current analysis comes from the authors of the paper it extends. Darie Moldovan and Simona Mutu were kind enough to provide the data they used in their analysis, providing the complete dataset and granting permission to use it here. This is highly beneficial for a number of reasons. Firstly, the results from the current analysis can be placed in a much clearer context, since we can directly and numerically compare the findings of this study to the original and identify areas of achieved improvement. Secondly, being granted access to a purpose built dataset prior to undertaking this analysis represents a significant catalyst for progress, and expedites the process of gaining greater understanding in this domain. The statements made by the authors based on identified correlations can also be used in the causal inference stage of this study as the basis for the formulation of research questions.

Three datasets where provided by Darie Moldovan and Simona Mutu, each covering 52 features for three distinct stock indexes. They are; the S&P 500 based in the United States, the STOXX 600 based in Europe and the STOXX 300 based in Eastern Europe. Combined, these datasets total 1400 records of companies from the year 2014. The authors scrapped this data using the

Bloomberg financial data repository, which contains a vast amount of historical financial data on companies across the world. In their study, the authors decided to analyse each market in isolation rather than in combination, inferring that the relationship between corporate governance and performance is characteristically different between markets. For the present study, this same logic is adopted.

3.2.2 New Factors - Independent

Part of this study is the exploration of new factors that could be introduced into the analysis to better explain corporate success. In other words, new independent variables to append onto the core dataset that extend the original research. A Bloomberg terminal was used to acquire all data outlined in this section, due to the ease as which features could be found, extracted and integrated with the original dataset.

Discussed in section 2.2.3 is the importance of environmental performance and its influence on overall economic health. To the end an exploration of the data available in Bloomberg was conducted, however it was found that the amount of missing data for the majority of environment-related features in the year in question was prohibitive to their inclusion. Thus, it was decided to use a propriety score formulated by Bloomberg themselves as a proxy for performance in this area. The justification for this is outlined in section 2.2.5.

Another new independent feature introduced here is total CEO compensation for each company in this study, as discussed in section 2.2.6. CEO compensation is readily available in Bloomberg for the year under study, and so it is relatively easy to extract and append this measure onto the core dataset.

3.2.3 New Factors - Dependent

Another goal of this study is to explore other dependant variables, or in other words auxiliary indicators that characterise levels of success. A single new

dependent variable is included, namely the Beneish M Score as outlined in section 2.2.2. The M Score uses an aggregate of various company-specific financial ratios to calculate the probability of that company having intentionally manipulated it's reported earnings. There are two variations on this score, one using a combination of five financial ratios and the other adding an additional three. All variables were derived from Bloomberg for the appropriate years, and appended to the original dataset.

3.3 Data Pre-Processing and Reduction

3.3.1 Missing Values

Moldovan and Mutu (2015) decided to remove any observations in their data that had missing values in the dependant variable, or not enough information to calculate those values. It could be argued that these emissions are justified, since incomplete data could unfairly skew the properties of that observation and misrepresent it in the data. Any conclusions that were made using these observations could be fundamentally flawed. Below is table outlining the degree of missing dependant variables per dataset.

Dataset	Row Count	Missing Tobins Q Score	Missing Altman Z Score
SPX	500	4	81
SXXP	600	4	127
EEBP	300	3	65

For the classification stage of the current study, where an attempt is made to replicate the findings of the original authors, these rows will be removed in much the same way.

However, when it comes to the regression and causal inference stage, missing data represents a more complex issue. Horton and Kleinman (2007) state that it is critically important to address missing data, particularly in observational

analysis with may predictors (as in the current study), as it arises frequently in almost all investigations using real world data. There are a number of reasons for the presence of missing data, from randomly missing data points (i.e. the propensity to be missing is independent of the value itself) to non-randomly missing data points (where the true value influences the propensity for it be to missing). The table below outlines the degree of missing values in each dataset, and indicates the dimensions of each if complete-case analysis were carried out.

Dataset	Row Count	Complete Cases
SPX	500	56
SXXP	600	2
EEBP	300	0

It is clear that complete case analysis is infeasible here, and so some other method of handling incomplete data is required. Horton and Kleinman (2007) discuss a range of methods for addressing this issue, with the specific aim of enabling the fitting of a logistic regression model on a sample dataset. One such method is multiple imputation, which they describe as a multi-step approach to estimating incomplete data that relies on an assumption that values are missing at random. First, the missing entries are filled in m times. These new values are drawn for a distribution that is different for each entry and variable. There is then an analysis stage where the m completed data sets are studied in isolation. Finally, the m datasets are pooled into a final result. Rubin (2004) states that if the imputation method is correctly implemented, then the results are valid for statistical modelling.

The companies represented in these datasets are public, and so are responsible for accurately reporting along a number of dimensions like company directorship and board composition, as well as financial statements as audited by a third party. However this does not cover all features involved in this study, which may be optional for reporting purposes. Regulation in this regard also differ between markets involved in this analysis, making it difficult to deduce whether data is missing at random or not. Jakobsen *et al.* (2017) state that

in this case, multiple imputation may be suitable.

A popular implementation of this technique is the `MICE` package in `R`, as outlined by van Buuren and Groothuis-Oudshoorn (2011). `MICE` imputes using chained equations, which involves specifying the imputation model on a variable by variable basis and using the other variables as predictors. At each step in the algorithm, an imputed value is generated and used in the imputation of the next variable. This process is repeated for each iteration, until convergence is reached as specified by the Gibbs sampling procedure, outlined in more detail by Yildirim (2012). This method can handle both continuous and discrete variables, as is required with the present data. `MICE` will be used to prepare each dataset for the regression as well as causal inference stages of the current study.

3.3.2 Outliers

Moldovan and Mutu (2015) state that they remove outliers in the data, citing a desire to prevent *"data errors"*. They do not provide an explanation of what characterises an outlier or a data error, nor do they present any evidence that outliers are fair emissions. It is generally accepted that outliers must be proven to be mistakes at the data collection stage, or invalid in some other way to justify leaving them out of the analysis. Without such justification, outliers are valid data points and may prove crucial to the formulation of a faithful model. While they state that removal in total only discounts 122 observations, this amounts to roughly 9% of the original dataset.

Since outlier detection and omission can have a significant impact on model performance, as shown for example by Pollet and van der Meij (2017) and Zijlstra *et al.* (2011), this study will conduct identical analysis with and without the presence of outliers in order to assess the utility of their inclusion. As mentioned above, the original study neglects to detail how the authors characterised or identified outliers, and so some methodology for doing so must be chosen. Cousineau and Chartier (2010) reviews several different methods of

outlier detection, one of which is Cook's distance originally proposed by Cook (1977). Cook's distance considers the influence of a given case i on all n fitted values in a regression analysis, and is calculated as

$$D_i = \frac{e_i^2}{pMSE}\Big(\frac{h_i i}{(1-h_i i)^2}\Big) \tag{3.3.1}$$

where e_i is the i^{th} element of the residual vector, $h_i i$ is known as the leverage and MSE is the mean square error. As noted before, both original dependent variables (the Tobins Q score and Altman Z score) are continuous in nature. Moldovan and Mutu (2015) threshold on this value to transform the problem to a classification task, whereas the current study both replicates this and performs regression on the original values. For this reason, a method to identify outliers that relies on regression analysis is chosen, and is used before any thresholding takes place.

3.3.3 Thresholding

In their analysis, Moldovan and Mutu (2015) threshold on both the Tobins Q score and Altman Z score to create classes from the continuous measures. This frames the problem as a classification task rather than regression which might be a more natural framing. Similar classes are created here, to facilitate a comparison of results. The original continuous measures are retained to facilitate regression analysis.

Tobins Q is discretised using the median to split observations into two categories, as per the original authors who cite Creamer and Freund (2010) as suggesting such a split. This is suitable for both the classification stage and causal estimation stage. The discretisation of the Altman Z score is more involved. Moldovan and Mutu (2015) create three classes here, as suggested by Altman (1968). Those classes are listed as "distress', "grey" and "safe" referring to the company's risk of bankruptcy. The following values are used to create these classes.

Threshold	Class
AZS >2.99	Safe
2.99 >AZS >1.81	Gray
1.81 >AZS	Distress

This is suitable for the classification stage of the current study. However, the causal estimation stage requires a binary class as the dependant variable and so the "grey" and "distress" classes are merged into one. Causal results referring to the Altman Z score thus refer to estimating the effect of some treatment on a "safe" or "not safe" level of bankruptcy risk.

3.4 Data Mining, Algorithms and Software

As referenced numerous times above, there are three main stages to this study; classification, regression and causal inference. Each of these steps requires a distinct approach and choice of toolset and software. Here, a discussion on these choices is included with justification for each.

3.4.1 Classification

Moldovan and Mutu (2015) in the original study approached the research question as a classification problem, thresholding the continuous dependant variables and using appropriate algorithms to achieve that goal. For each dataset and measure of success, they implemented four distinct classification algorithms and compared performance between each using identical metrics. This study implements a subset of these algorithms to facilitate a limited verification and comparison. Since the main objectives of this study lie elsewhere, this study neglects to include each and every algorithm used originally.

The `Adaboost M1` algorithm proved to be one of the highest performing implementations in the original study, and so is included here also. The `adabag` package in `R` is used, which implements the algorithm as proposed by Freund and Schapire (1996). As per the name, this algorithm uses *boosting*, which

the authors state can be used to significantly reduce the error of weak learners by unifying them in a weighted sum that represented the final output of the boosted classifier. In this sense, the performance of each individual weak learner can be poor, however as long as each is better than a random guess the final result converges to what is known as a *strong learner*. `Adaboost` represents an improvement on bagging, attempting to build multiple models using randomly chosen training instances and eventually combining into a single model with improved accuracy. Adaboost adds an adaptive layer to this, by disproportionally weighting up poorly modelled instances (those with higher error) in subsequent models.

The next most performant algorithm used by Moldovan and Mutu (2015) was `J48`, which is a `Java` implementation of the `C4.5` decision tree algorithm. This algorithm builds decision trees from training data using information entropy, which represents the expected value of a random variable that describes the amount of information contained in a particular split or location in a decision tree. In this way the algorithm choses attributes of the data that most effectively and efficiently splits the samples into subsets enriched in one class. Since the current study uses `R`, an translated implementation is required. The `C5.0` algorithm was chosen, which is a next generation version of the `C4.5` algorithm, implementing various runtime efficiencies in terms of speed and memory usage.

3.4.2 Regression

For the regression stage of this analysis, both Tobins Q and the Altman Z score remain as continuous variables which replace the thresholded values used above as the dependant variables. A standard linear model is of the form

$$Y_i = X_i^T \beta + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2),\ i = 1, ...n \tag{3.4.1}$$

The optimal linear unbiased estimator for β is found by solving

$$(X^T X)\beta = X^T y \tag{3.4.2}$$

For equation 3.4.2 to be solvable, the matrix $X^T X$ must be invertible, which is not the case when the number of independent variables is much larger than the number of observations in the dataset or if there is collinearity between variables that are previously believed to be independent. In the current dataset, the first condition is not met although ratio of observations to features is as low as approximately 6:1 for the `eebp` dataset and so may still be a valid concern. The collinearity within this dataset is certainly a source of concern, due to the large number of features included. Analysis showed that of 1830 possible pairings of features in the `spx`, `sxxp` and `eebp` datasets there were 796, 944 and 392 pairings with a statistically significant collinearity coefficient respectively. A high degree of collinearity between independent variables can cause the regression coefficients to become very sensitive to small changes in the model. It can also reduce the precision of the estimate coefficients, reducing the statistical power of the model. It is clear then that a regression methodology capable of dealing with this is required, that uses an alternative estimator for β.

Hoerl and Kennard (1970) proposed a modification of equation 3.4.2 based on a perturbation, denoted by λ, to the matrix $(X^T X)$ making it invertible. The equation thus becomes

$$(X^T X + \lambda I)\beta = X^T y \tag{3.4.3}$$

where I is the identity matrix. The estimator of β thus becomes

$$\hat{\beta}(\lambda) = (X^T X + \lambda I)^{-1} X^T y \tag{3.4.4}$$

and is called the *ridge estimator*. Hoerl and Kennard (1970) state that there exists some constant λ that leads to a mean square error less than that achieved by ordinary least squares regression. This holds even when the original matrix in equation 3.4.2 is in fact invertible. The question now arises about how to chose the optimal value for λ. There is much research on this topic, however

for the current study the software will be computing values for λ itself taking into account the fact that we require the MSE to be as small as possible.

A shortcoming of ridge regression is that is does not inherently include any variable selection stage, in that it estimates coefficients but does not act to reduce any to exactly zero. This can be considered suboptimal when only a few predictors are likely to be influential or where model interpretability is important. Lasso regression (least absolute shrinkage and selection operator) aims to address these issues, by performing both the regularisation step carried out by ridge regression as well as a variable selection stage. Fonti and Belitser (2017) state that feature selection is a vital step in data analysis, and can act to simplify the final model by removing features that are not important as well as reducing the size of the problem to enable other algorithms to operate more quickly. Lasso works to minimise the sum of the squared errors with an upper bound on the total sum of the (absolute) values of the actual model parameters.

The current study implements both lasso and ridge regression under the general umbrella of regularised linear regression on all three dependent variables (Tobins Q and Altman Z scores, as well as the new Benish M-Score where available). The `glmnet` package in `R` is used to achieve this, which fits a generalised linear model using a regularisation parameter `alpha`. `Alpha` dictates where along the spectrum between lasso and ridge the penalty factor lies. A value of 1 leads the algorithm to use lasso, a value of 0 leads the algorithm to use ridge. Values inbetween lead to elastic net regression, described by Fonti and Belitser (2017) as a combination of the two. For each dataset and dependent variable, a series of models are built using a range of `alpha` values between 0 and 1 in 0.1 increments. Analysis is then carried out to see which implementation minimises the MSE.

3.4.3 Causal Estimation

For the causal estimation portion of this study, a propensity score matching algorithm is used as discussed in section 2.4. In order to achieve this, a

module named `causality` written in `python` and made available in GitHub is used. This project is available at `https://github.com/akelleh/causality`. Among other pieces of functionality, this module allows the direct formulation and execution of propensity score matching and provides an interface for graphically judging the quality of the matching process. `causality` was installed using `pip`, and interfaced with using custom written `R` and `python` scripts that prepared the data, executed the module, and collected the results.

The first step in this stage of the study is the formulation of research questions, many of which are derived from the correlations and *'if-this-then-that'* style conclusions of Moldovan and Mutu (2015). Many of those conclusions are market specific. However in this study each statement is tested across all markets in order to verify the applicability of each on a global scale. To those statements, this study adds new research questions motivated by existing research in this domain outlined in the literature review. A list and discussion of those questions can be found in section 3.5.

Each research question informs the choice of *treatment* and *effect* pairs. Treatments are variables contained in the data already and represent some condition that may cause a variation in the outcome. The effect then is another term for a dependant variable, as used in both the classification and regression stages of this study. We use both the Tobin Q score and Altman Z score, as well as the Benish M-Score where possible.

3.5 Causal Estimation - Motivating Statements

As mentioned above, Moldovan and Mutu (2015) make eight statements regarding the governance influencers on corporate success. These are listed below, each referring to the market that the finding was original attributed to.

1. For the American companies inside the S&P 500 index, we found a positive correlation between the percentage higher than 20pct of women in the board and the Tobins Q ratio.

2. For the American companies inside the S&P 500 index, we found...the presence of an independent lead director in the company along with a financial leverage higher than 2.5 incur a higher risk of bankruptcy.

3. When analysing the Eastern European companies data, we found that a smaller age range for the board members is positively related with the companies performance...

4. When analysing the Eastern European companies data, we found...that a financial leverage less than 4 is needed in order to be on the upper side of the Tobins Q ratio.

5. When analysing the Eastern European companies data, we found...to be on the safe zone of the Altman Z-score it is important to have an independent chairperson or even a woman as CEO.

6. For the Western European companies....the presence of an independent lead director or a former CEO in the board could be a sign of weaker performances, being negatively correlated with Tobins Q

7. For the Western European companies.... a large percentage of women in the board could also affect negatively the performance.

8. For the Western European companies....for the companies with large financial leverage in order to be in the "safe" zone of the Altman Z-score it could be a good idea to adopt an Auditing Committee with more than four people.

3.6 Interpretation

3.6.1 Classification

Moldovan and Mutu (2015) used a number of metrics for measuring performance. They first show the accuracy of each model, which is simply the number of correctly identified observations in the data. Next they show the precision

for each class, which is the number of correctly identified instances over the total predicted instances for that class. In the literature, these measures are also referred to the sensitivity and specificity.

Finally, they list the area under the receiver operator characteristic (ROC). A ROC curve is created by plotting the sensitivity against the fall-out (or 1 - specificity) at various discrimination thresholds. The discrimination threshold describes the cutoff imposed on the predicted probabilities required for assigning an observation to a given class. The area under the curve (AUC) then, as described by Flach (2007) is a single numerical measure of the model's performance, equivalent to the probability that a uniformly drawn random positive is ranked before a random negative.

3.6.2 Regression

As mentioned above, regularised linear regression is use in this study and varies with the penalty applied from ridge to lasso regression. The r^2 for each model is presented, as an indication of the quantity of the variance in that data that is described by the model. Alongside this, the root mean square error (RSME) is included as an measure of the difference between the values predicted by the model and the actual values observed.

3.6.3 Causal Estimation

Section 3.5 outlines the research questions proposed for this section of the current study, and will form the basis for a number of results sets. For each question (and thus treatment and effect pair), a number of performance related metrics are presented.

Chapter 5

Discussion

5.1 Introduction

In this chapter we examine ...

Chapter 6

Conclusions and Future Research

6.1 Introduction

Program code

Insert snippets of important code here
Point to github where code can be found

`https://github.com/ReidConor/dissertation`

Bibliography

Altman, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, **23**(4): 589–609.

Bebchuk, C. A., L. and A. Ferrell. 2005. What matters in corporate governance? *Harvard Law School John M. Olin Center Discussion*.

Bebchuk, L. and A. Cohen. 2005. The costs of entrenched boards. *Journal of Financial Economics*, pages 409–433.

Beneish, M. 1999. The detection of earnings manipulation. *Financial Analyst Journal*, pages 24–36.

Beneish, M. and D. Nichols. 2007. The predictable cost of earnings manipulation. *SSRN-id1006840*.

Bhagat, S. and B. Bolton. 2008. Corporate governance and firm performance. *Journal of corporate finance*, **14**(3): 257–273.

Bolton, P., H. Chen and N. Wang. 2011. A unified theory of tobin's q, corporate investment, financing, and risk management. *The journal of Finance*, **66**(5): 1545–1578.

Chung, K. H. and S. W. Pruitt. 1994. A simple approximation of tobin's q. *Financial management*, pages 70–74.

Cook, R. D. 1977. Detection of influential observation in linear regression. *Technometrics*, **19**(1): 15–18.
URL `http://www.jstor.org/stable/1268249`

Core, R. H., J. and D. Larcker. 1999. Corporate governance, chief executive officer compensation, and firm performance. *Journal of Financial Economics*, pages 371–406.

Cousineau, D. and S. Chartier. 2010. Outliers detection and treatment: a review. *International Journal of Psychological Research*, **3**(1): 58 – 67. ISSN 20112084.
URL `https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=61167387&site=ehost-live`

Creamer, G. and Y. Freund. 2010. Learning a board balanced scorecard to improve corporate performance. *Decision Support Systems*, **49**(4): 365–385.

Dehejia, R. H. and S. Wahba. 2002. Propensity score matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, **84**(1): 151–161.

Eidleman, G. J. 1995. Z scores-a guide to failure prediction. *The CPA Journal*, **65**(2): 52.

Esarey, J. 2015. Causal inference with observational data.

Fatemi, G. M., A. and S. Kaiser. 2017. Esg performance and firm value: The moderating role of disclosure. *Global Finance Journal*.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth. 1996. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, **39**(11): 27–34.

Flach, P. 2007. Putting things in order. on the fundamental role of ranking in classification and probability estimation.

Fonti, V. and E. Belitser. 2017. Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*.

Freund, Y. and R. Schapire. 1996. Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, page 148–156.

Guo, K. T., R. and T. Nohel. 2008. Undoing the powerful anti-takeover force of staggered boards. *Journal of Corporate Finance*, pages 274–288.

Hausman, J. A. 1978. Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, pages 1251–1271.

Herawati, N. 2015. Application of beneish m-score models and data mining to detect financial fraud. *Procedia - Social and Behavioral Sciences*, pages 924–930.

Hoerl, A. E. and R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1): 55–67. ISSN 00401706. URL `http://www.jstor.org/stable/1267351`

Horton, N. and K. Kleinman. 2007. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistician*, **61**: 79–90.

Jakobsen, J. C., C. Gluud, J. Wetterslev and P. Winkel. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, **17**(1): 162.

Kamal, M., M. Salleh and A. Ahmad. 2016. Detecting financial statement fraud by malaysian public listed companies: The reliability of the beneish m-score model. *Jurnal Pengurusan*, pages 23 – 32.

King, G., C. Lucas and R. A Nielsen. 2014. The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*.

Kirkpatrick, G. 2009. The corporate governance lessons from the financial crisis. *OECD Journal: Financial Market Trends*, **2009**(1): 61–87.

Mahama, M. 2015. Detecting corporate fraud and financial distress using the altman and beneish models. *International Journal of Economics, Commerce and Management*, **III**.

McLaughlin, E. C. 2017. Man dragged off united flight has concussion, will file suit, lawyer says.
URL http://edition.cnn.com/2017/04/13/travel/united-passenger-pulled-off-flight-lawsuit-family-attorney-speak/

McVeigh, I. 2015. Volkswagen scandal: Bad governance is often a sign of trouble ahead.
URL http://www.telegraph.co.uk/finance/comment/11893886/Volkswagen-scandal-Bad-governance-is-often-a-sign-of-trouble-ahead.html

Moldovan, D. and S. Mutu, 2015. Learning the relationship between corporate governance and company performance using data mining. In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 368–381. Springer.

Orlitzky, M., F. L. Schmidt and S. L. Rynes. 2003. Corporate social and financial performance: A meta-analysis. *Organization studies*, **24**(3): 403–441.

Pearl, J. and T. S. Verma. 1995. A theory of inferred causation. *Studies in Logic and the Foundations of Mathematics*, **134**: 789–811.

Pollet, T. V. and L. van der Meij. 2017. To remove or not to remove: the impact of outlier handling on significance testing in testosterone data. *Adaptive Human Behavior and Physiology*, **3**(1): 43–60. ISSN 2198-7335. doi:10.1007/s40750-016-0050-z.
URL https://doi.org/10.1007/s40750-016-0050-z

Rosenbaum, P. R. and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, pages 41–55.

Rubin, D. B. 2004. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Schaltegger, S. and T. Synnestvedt. 2002. The link between green and economic success: environmental management as the crucial trigger between

environmental and economic performance. *Journal of environmental management*, **65**(4): 339–346.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, **25**(1): 1.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of money, credit and banking*, **1**(1): 15–29.

van Buuren, S. and K. Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, **45**(3): 1–67.
URL `https://www.jstatsoft.org/v45/i03/`

Wahba, H. 2008. Does the market value corporate environmental responsibility? an empirical examination. *Corporate Social Responsibility and Environmental Management*, **15**(2): 89–99.

Yildirim, I. 2012. Bayesian inference: Gibbs sampling.

Zijlstra, W. P., L. A. van der Ark and K. Sijtsma. 2011. Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*, **36**(2): 186–212.

List of Notation

Entries are listed in the order of appearance. The "Ref" is the number of the section, definition, etc., in which the notation is explained.

Symbol	Description	Ref
$PREFST$	The liquidating value of a firm's preferred stock	