universität wien

MAGISTERARBEIT

Titel der Magisterarbeit

"Treatment effects and the causality between CSR and financial performance"

Verfasserin

Ida Margareta Nilsson

Angestrebter akademischer Grad

Magistra der Sozial- und Wirtschaftwissenschaften (Mag.rer.soc.oec)

Wien, im März 2010

Studienkennzahl lt. Studienblatt:	A 066 913
Studienrichtung:	Magisterstudium Volkswirtschaftslehre
Betreuer:	Univ.- Prof. Dr. B.Burcin Yurtoglu

Index

1. Introduction **5**

2. Treatment effects **6**

2.1 Differences 7

2.2 Difference-in-difference 9

2.3 Propensity score matching 11

2.3.1 Roy-Rubin Model and selection bias 14

2.3.2. Assumptions and implementation 15

2.3.3 Choice of model 17

2.3.4 Choice of variables 18

2.3.5 Choice of matching algorithm 20

2.3.6 Overlap and common support 23

2.3.7 Assessing the matching quality 24

2.3.8 Variance estimation 25

2.3.9 Sensitivity Analysis 26

2.4 Instrumental Variables 27

2.4.1 IV estimation of a multiple regression model 28

2.4.2 Two stage least squares 30

3. Company performance **33**

4. Corporate Social Responsibility **37**

4.1 Measurements of CSR 40

4.2 Global Report Initiative 40

4.3 Global Compact 41

5. The Data **43**

5.1 Financial performance data 46

6. Matching and results **52**

6.1 First method 52

6.2 Propensity score matching 53

7. Conclusions **59**

References **63**

Appendix **67**

Anhang 73

Abstract 75

Zusammenfassung 77

Curriculum Vitae 79

1. Introduction

A natural experiment is, unlike an experiment, the process of assessing a phenomenon that can not be pre-tested in a laboratory or pre-simulated by researchers. The outcome of a natural experiment can only be estimated by trying a large amount of different estimation methods or models.[1] These natural experiments also called "quasi-experiments" often tries to find the causality between two phenomena. It can be the causality between wage and education, between savings and income or assessing the use of a medicine on health. The number of possible causality relations to estimate is infinite. When measuring the treatment effect, i.e. a particular treatment's casual effect on a certain outcome, one has to define and observe two groups. One that receives the certain treatment (special education, medicine, income level) and one that does not. If the individuals assigned to treatment were chosen randomly, the estimation of the treatment effect would have been easy. However, most assignments include a certain selection bias. Patients get a medicine prescription because of their previous health condition, people attend work training programs because of their unemployment status and employees have higher income due to previous education and experience. The problem of selection bias makes it impossible to observe a randomised sample and much attention have to be paid to find a good comparison group.

This thesis will assess three different methods for estimating treatment effects, these will all be presented in the theory part in chapter 2. The purpose of the empirical part, starting with chapter 5, is to assess the effect of corporate social responsibility on the financial performance. Due to this, the theory part places an extra focus on the treatment effect method of propensity score matching used in the estimations. Chapter 3 and chapter 4 additionally highlight the theories surrounding corporate social responsibility and the methods for measuring company performance. The results of the empirical part will be presented, analysed and summarised in chapter 6 and chapter 7.

[1] Meyer (1995), pp. 1-2

2. Treatment effects

The main objective of a good research design is to find the variation in the explanatory variables that is exogenous. Hence, it should try to find the variables that are independent in relation to other variables and that can explain some of the variation in the dependent variable.[2] When estimating the effect of a treatment it is of equally great importance to find a comparable comparison group and to test for the effects with a correct and customised estimation method.

The problems that may arise estimating a casual relationship between a treatment and an outcome y can be divided in to internal threats and external threats. The internal threats show the possibilities that the difference in the dependent variable y is not caused by the difference in the explanatory variable x.[3] Such Internal threats are:

Omitted variables - Variables excluded from the estimation model that provide an alternative explanation for the observed difference. I.e. the excluded variables are also explanatory and should be included in the model.

Mismeasurements - Variations in the gathering of the data by using different surveys, measurements and definitions can cause changes in the measured variables. It is of great importance to observe how a certain variable has been measured and to assure that the definition and the measurement of the variable are unified and unchanged over time.

Trends in outcome - Time dependent variables such as wage, age and inflation change the explanatory variables over time.

Political Economy - Political changes, decisions and policies might influence the effect of an explanatory variable on the dependent variable.

Simultaneity - A relationship between depending and explanatory variable can be influenced by a second variable jointly influencing both the dependent and

[2] Mitchell et al. (2007), pp. 36-37
[3] Meyer (1995), pp. 5-6

the independent variable.

Selection - Non-random assignment to treatment leads to property differences between the treated and the non treated group even in a pre-treatment state.

Omitted interactions - Actions only effecting either treatment group or comparison group.[4]

The external threats highlight the possibility of interaction between treatment and individual characteristics, location or time. External threats can be due to:

Selection - Systematic selection bias in the assignment to a treatment group makes the outcome unrepresentative for the population it aims to represent.

Settings - The effect of a treatment on output might differ depending on circumstances such as geography and institutional settings.

History - The effects of a treatment may differ across time.[5]

The difficulty of finding the right variables and the right comparison group makes the examination of the comparability and exogeneity even more important. This thesis will highlight methods for estimating and assessing treatment effects and finding suitable control groups. The methods presented in this chapter are in research and literature commonly used methods for estimating treatment effects but they can not be interpreted as the only solution to a treatment problem. The chapter to follow describes the methods used to assess and evaluate treatment effects and possible procedures for selecting a comparison group.

2.1 Differences

The approach of differences is the basic step of treatment assessment but it is not very likely to lead to a valid conclusion. However, the examination of the differences often serve as a baseline for further assessment of causality between a dependent

[4] Gravetter et al. (2006), pp. 166-170 and Meyer (1995), pp. 5-7
[5] Gravetter et al. (2006), pp. 170-174 and Meyer (1995), pp.5-8

and an explanatory treatment variables.[6] The method of differencing can be described with the equation:

$y_{it} = \alpha + \beta d_t + \varepsilon_{it}$

Where y_{it} is the outcome of the variable of interest in period t, where t = 1, …, N, d_t is the dummy variable that adopts the value 1 if the individual is in the treatment group and value 0 if it is not. β represents the true value of the treatment effect on the output. For the model to hold, it must be assumed that the conditional mean does not depend on the value of the treatment dummy, i.e. if no treatment occur then βd_t is zero and the difference in mean for treatment and comparison group is the same. This is in statistics also known as the ignorable treatment assignment. The estimate of β can be expressed as[7]:

$$\hat{\beta}_{\mathrm{d}} = \Delta\bar{y} = \Delta\bar{y}_1 - \Delta\bar{y}_0$$

Where $\bar{y}_1$ expresses the output for the treated group and $\bar{y}_0$ the outcome for the comparison group.

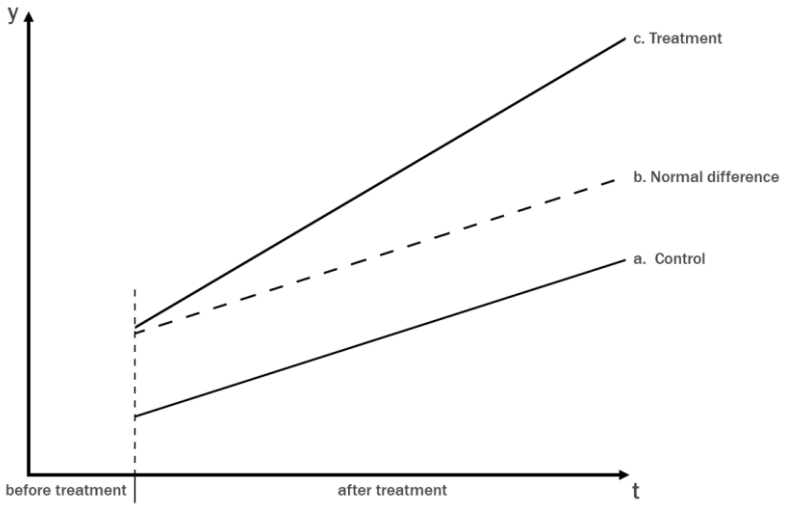
When estimating a treatment effect by assessing two different outcomes, one for a treatment group and one for a comparison group, it must be clear that the two groups can be comparable over time in the absence of the treatment. The problem by using the difference method is that it strictly compares the treatment group after treatment with a comparison group before treatment. To use such a model, strong assumptions that are easily validated must be made. Despite the lack of applicability, it gives the basics for the next more useful method: difference-in-difference.[8]

[6] Meyer (1995), p. 9
[7] Meyer (1995), p. 15
[8] Ibid., p. 16

2.2 Difference-in-difference

The difference-in-difference method is an extension of the basics of the difference approach described above. Its big advantage is that it takes the different group differentials into account. The method is considered useful when assessing differences between two groups, one which receives treatment and another which does not. Together with an additional method for finding two similar comparison groups the difference-in-difference is a popular approach to the non-experimental (non-randomised) evaluation of a causality relationship.[9] The estimation compares outcome changes over time for treatment groups and non-treatment groups.

Graph 1. The treatment effect illustrated as the deviation from the normal difference.



If one can assume that the comparison group and the treatment group only differ in the assignment to treatment then the idea is that the pre-treatment difference is the normal difference in output between the two groups. This is illustrated in the graph above as the distance between a and b. The changes in the normal difference after the treatment can therefore be accredited to the casual effect of the treatment and the difference between the actual output and the normal difference is the difference-in-difference (distance b to c).

[9] Bertrand et al. (2004), pp. 249-251

The difference-in-difference method can be described in three steps. First, one needs a baseline that covers both the comparison group and the treatment group before the treatment. Second, the researcher needs to make sure that the data used to compare the *before* and *after treatment* effect comes from the same survey and/or are produced in the same way. Third, the mean difference in output before and after the treatment must be calculated for both the treatment and the non-treatment group. After these three steps, one can calculate the differences in outcome variation between the treated and the comparison group. The model for the outcome variable is:

$$y_{it}^{j} = \alpha + \alpha_1 d_t + \alpha_1 d_j + \beta_{djt} + \varepsilon$$

Where i represent the time aspect and t is the indicator of receiving treatment. t can adopt the number 1 and 0, where t=1 it indicates that the group received treatment i.e. "after treatment". t=0 indicates *before treatment*. The group aspect is described by j, where j=1, …, k describes the group participation and j=1 indicates that the individual or the unit belongs to the treatment group. j=0 indicates that the observation is assigned to the treatment group. [10]

If one, for example, were to test the effect of a heart medicine on health for a risk group of men in ages between 50-65. The researcher would compare the risk group of men to a sample of men with the same type of previous health condition and age. These two groups would in absence of the medical treatment be subject to the same changes over time. This is expressed in the term α_1 in $\alpha_1 d_t$, where d_t=1 since after the time of treatment, even though no treatment was performed. Would t equal 0, meaning if one assess the two groups before the treatment, then the assignment to the groups, risk or comparison group, would have no effect on the outcome, since d_t=0 and d_j=1. The time invariant differences are captured in α_1 in $\alpha_1 d_j$. This indicates that when j=1 the effects of being in the treatment group is invariant of time.

The key assumption in difference-in-difference is that β=0 whenever t=0. I.e. that the treatment effect on the outcome equals zero if the groups still has not received

[10] Meyer (1995), p. 15 and Bertrand et al. (2004), p. 250

treatment. Hence, the variable of interest y is not depending on the group participation in the absence of a treatment. This can also be written as $E[\varepsilon_{itj}|d_{ij}]=0$.

Based on the assumption β=0, the estimator can be expressed as:

$$\hat{\beta}_{dd} = \Delta\bar{y}_{11} - \Delta\bar{y}_{00}$$

or:

$$\Delta y = (\bar{y}_{10} - \bar{y}_{11}) - (\bar{y}_{00} - \bar{y}_{01})$$

The problem by using difference-in-difference to measure the treatment effect is that even though the treatment group and the non-treatment group are exact matches the problem with selection bias will make the result inconsistent. That is, underlying reasons that affect the decision to enter the treatment group, or receiving treatment, make the estimation unable to account the whole post-treatment difference to the treatment. If the probability to receive treatment differs to begin with then using only the difference-in-difference method will not be a good estimate for causality. Hence, to find the comparison group that resembles the treatment group the result from a propensity score matching estimation can be used. [11]

2.3 Propensity score matching

The method of Propensity Score Matching (PSM) is increasingly being used in empirical studies to estimate casual treatment effects. It often serves as estimation method when evaluating labour market policies but can also be found in a wide range of other empirical work. The PSM can be used in all situations where one specific group receive treatment and others do not. Treatment can be anything from aid-program participation, medicine prescriptions, special school attendance and labour force attendance, anything that can possibly be separated in to two distinguished groups. One who gets the treatment and the ones who does not. For example; Hitt and Frei (2002) analysed the effect of online banking on the profitability of costumers. Jones and Richmond (2006) assessed the casual effects of alcoholism on earnings. Brand and Halaby (2006) analysed the effects of elite college participation

[11] Meyer (1995), pp. 16-17

on career outcome. Decision makers who want to evaluate (or estimate) the result of a program need to examine (estimate) the full social cost for receiving treatment per individual. This must be done by subtracting the cost from the average gross gain for the treatment in the program, i.e. $E(y1 - y0|D=1, X)$, where X represents the characteristics of the participants. The outcome shows the total social output for participation in a program compared with no participation and can give decision makers feedback upon the result of their decision.

Every individual can be categorised into a treated or a non-treated group, i.e. with a binominal variable that adopts the value zero and one depending on the assignment state. Each state is associated with an outcome; y_1 and y_0 depending on if the individual receives treatment or not. The gain from getting a treatment can on individual level therefore be described as $\Delta=y_1-y_0$, i.e. the individual difference in output for being treated to not being treated.[12] The problem is that one can not observe Δ since it is impossible for one individual to be in both the participating and the non-participating group simultaneously. This fundamental evaluation problem can therefore not be solved on an individual basis.[13]

The PSM is based on the idea of finding a way to compare outcomes of program participants (y_1) with the outcomes of "comparable" non-participants (y_0) so that differences in output can be attributed to the program. This can only be done if the treated and the non-treated samples are matched in such a way that the sample resembles the characteristics and design of a randomised sample. The matching is based on a propensity score estimation that tries to estimate the true propensity score, i.e. the probability that a treatment group receives treatment given their observed characteristics.[14] Since participants and non-participants usually differ from each other in other aspects besides only the treatment itself, using the mean from the non-participants group as an approximation for the treated group "if not treated" is not commendable. Thus, the mean for the treated group after treatment;

[12] Heckman et al. (1998), p. 263
[13] Austin (2008), pp. 31-33
[14] Jo et al. (2009), p. 2862

$E(y_1|D=1, X)$ can be calculated on the data at hand, but the counterfactual mean $E(y_0|D=1,X)$ can not be observed. The problem is called selection bias and can be expressed as:

$$B(X) = E(y_0|D=1, X) – E(y_0|D=0, X)$$

I.e. the expected outcome for the non-treatment group when treated subtracted by the expected outcome for the same group when not treated.[15] Also expressed as:

$$E(y_1|D=1, X) – E(y_1|D=0, X)$$

I.e. the expected output when treated, for the treatment group, subtracted by the non-observable value of what output would have been for the treatment group if they were not to receive treatment. Only a study where the probability of treatment assignment is related completely to the observed characteristics can be free from hidden bias.[16]

The basic idea with PSM is to try and overcome the issue of selection bias. Hence, by finding a group of non-participants with maximal similarity to the participants in the treatment group in all relevant pre-treatment characteristics X, one can assign the differences of outcome to the treatment.[17]

When using the Propensity Score Matching a number of decisions have to be made. First, one has to agree on the model to be used for estimation, and also decide on the variables that should be included in the model. Second, one should choose the matching algorithm.[18] Third one needs to control for the overlap and baseline balance between treatment and comparison group. The fourth step is the assessment of the

[15] Heckman et al. (1998), pp. 262-265
[16] Ibid., pp. 264-265
[17] Heckman et al. (1998), pp. 261-262, Caliendo et al. (2008), pp. 33-35 and Austin (2008), pp. 2037-2038
[18] An algorithm is the systematic procedure that in finite number of steps implement an estimation or solves a specific problem.

matching quality and evaluation of the sensitivity in the matching.[19] Lastly, after assessing the goodness of the model, the treatment effect can be calculated and results implemented. The remaining part of chapter 2.3 will in detail assess these procedures.

2.3.1 Roy-Rubin Model and selection bias

Measuring the individual outcome effect from the treatment includes as mentioned speculation about the outcome for the same individual if it had not received treatment. Roy (1951) and Rubin (1974) developed the Roy-Rubin model which in case of a binary treatment ($D_i = 0$ if no treatment , $D_i=1$ if treatment, where i=1, ..., N) defines the potential outcome as $Y_i(D_i)$ for individual i, where i=1, ..., N. The Roy-Rubin model describes the treatment effect for an individual as $\tau_i=Y_i(1)–Y_i(0)$. The fundamental evaluation problem arise because only one possible outcome, Yi(1) or $Y_i(0)$, can be observed for each individual. That is, either they receive the treatment or they do not.[20]

The problem evaluating the individual treatment effect by using the Roy-Rubin model can be helped by estimating parameters. In the literature, the two most commonly used parameters are the population average treatment effect (ATE) and the average treatment effect on the treated (ATT). The ATE is the difference between the expected output before and after the treatment:

$$E(\tau) = E[Y(1) – Y(0)]$$

The ATE helps to assess the effect of the treatment if the individuals in the population were randomly assigned to the treatment.[21] Thus it also includes individuals for which the treatment never intended to occur. To illustrate this problem, again imagine a researcher that wants to examine the effect of a medicine, but the group chosen as the treatment group is chosen randomly and therefore consists of people with and without a diagnosis. The effects of the medicine on

[19] Caliendo et al. (2008), pp. 31-66
[20] Roy (1951), pp. 135-145 and Rubin (1974), pp. 688-701
[21] Hirano et al. (2003), p. 1162

patients with diagnosis can therefore never be examined.[22]

The average treatment effects on the treated (ATT) focus, as revealed by its name, only on the treatment group. It is given by:

ATT= E[Y(1)|D=1] – E[Y(0)|D=1]

Implementing that the ATT is defined as the expected output *if treated* for the treatment group subtracted with the expected output for control group when treated. In this sense the ATT parameter represents the realised gross gain of the treatment and can be compared with its cost helping in the evaluation process when to examine whether a program was successful or not. E[Y(0)|D=1] is the counterfactual mean for those being treated and can, as mentioned above, *not* be observed. Hence, one needs to use a substitute in order to be able to estimate the ATT.

The true parameter τATT is only identified if E[Y(0)|D=1]-E[Y(0)|D=0]=0, if (E[Y(0)|D=1]-E[Y(0)|D=0])≠0 then we have a selection bias. For an experimental study where the treatment is chosen randomly, the true parameter is identified and the true effect of the treatment can be assessed. In non-experimental studies one has to make some identifying assumptions to form the same statement.[23]

2.3.2. Assumptions and implementation

The fist assumption made are the assumption of unconfoundedness.

Unconfoundedness: Y(0), Y(1) ⫫ D|X

Where ⫫ denotes independence and Y(0), Y(1) the output for the treated and the non-treated group. Uncounfoundedness assumes that, given a set of observable covariates X which are independent from treatment D_i, the output is independent of the treatment assignment D_i. Meaning that if the covariates are independent of the

[22] Caliendo et al. (2008), p. 34
[23] Caliendo et al. (2008), p. 34

participation in the treatment group, then the total outcome should also be independent of the treatment participation. In applied terms, this assumption means that for common values of covariates, the choice of treatment is not based on the benefits of alternative treatments, but randomly assigned. This means that among individuals with the same characteristics used for matching, the model assumes that these individuals are sorted into different treatments as if randomly assigned. This assumption is rather strong; its plausibility varies with the specific treatment effect involved and depends on the range of covariates included in the model.[24]

According to the assumption of unconfoundedness it is therefore clear that all the variables that influence the treatment participation have to be observed simultaneously with the potential outcome. This assumption restricts the usage of the propensity score matching since many data sets lack in information about both treatment participation incentives and output covariates. The assumption of unconfoundedness can not be tested but it can be intuitively violated if X includes variables that are post-affected by the treatment. For example if one would add post training schooling attendance as an explanatory variable when explaining the treatment effect of a job training program. The post-training schooling could have been chosen as a result of the job training participation.[25]

Overlap: $0 < P(D = 1|X) < 1$

The assumption of overlap tells us that person with the same value on X has the same positive probability to be a part of the treatment and non-treatment group. This assumption rules out the phenomenon of perfect predictability of D given X. If this assumption would not hold then matching D=1 with D=0 can not be performed. Some degree of randomness guarantees that individuals with the same characteristics can be observed in both the treatment and the non-treatment group.[26]

[24] Caliendo et al. (2008), p. 35
[25] Jo et al. (2009), p. 2862
[26] Heckman et al. (1998), p. 154

Uncounfoundedness for controls: Y(0) ⫫ D | X

The assumption of unconfoundedness for controls is a weakening of the first unconfoundedness assumption. In this way one only estimates the ATT, i.e. that the outcome after the treatment is independent of the treatment assignment D_i, given the covariates X.[27]

Weak overlap: P (D=1|X) <1

Weakening of the second assumption.[28]

If a model with the variable vector **X** consists of k covariates, the number of matching possibilities is 2^k. Thus, the number of matches increase exponentially, which increase the difficulty finding a proper match to the treated units.[29] To ease this dimensionality problem Rosenbaum and Rubin (1983) suggested the use of balanced scores. If, as assumed in unconfoundedness, the outcome is independent of the treatment given the covariates X, then the outcome is also independent on the balancing score b(X). This means that the propensity score P(D=1|X)=P(X) can be used as a balancing score.[30]

2.3.3 Choice of model

To estimate the treatment effect, one has to start by estimating the propensity score. There are two primary choices to be made when estimating the propensity score. The first is the model to use and the second is the variables to use in the model. The choice of model varies because of the properties of the treatment. In the case of a binary treatment, logit and probit models are the most commonly used. These two normally yield quite similar result due to the nature of the treatment. If estimating the effect of a multiple treatment, meaning if the possible alternative choices of treatment are larger than two, the multinomial probit model is build on weaker assumptions than the corresponding logit model. Therefore the logit regression model is considered as preferable. There are some extended discussions about

[27] Caliendo et al. (2008), p. 36
[28] Ibid.
[29] Dehejia et al.(2002), pp. 53-55
[30] Caliendo et al. (2008), p. 36

robustness and alternative multinomial models but since this thesis has its focus on a binominal treatment effect, it will not further assess the effects of a multinomial model.[31]

When estimating the propensity scores for a binominal model, both a probit and a logit model can be used. It is important to remember that the propensity score is estimated to reduce the dimensions of the matching problem; it has as an estimation method no behavioural assumptions attached to it.[32] Most researchers in the econometric literature focus on the logit model for propensity score estimation. The logit model that estimates the probability to receive treatment is expressed as:

$$\Pr(D_i = 1|X_i) = \frac{e^{\lambda h(Xi)}}{1 + e^{\lambda h(Xi)}}$$

Where D_i is the treatment status and h(Xi) the linear and high-order terms of the covariates. When estimating the propensity score the choice of which variable to include depends solely on the conditioning on the observable characteristics that determine treatment participation.[33]

2.3.4 Choice of variables

The matching strategy builds on the two main assumptions of unconfoundedness and overlap expressed in section 2.3.2. One requirement is that the outcome variables must be independent of the treatment conditioning on the propensity score. This fulfils the assumption of unconfoundedness and implies that only variables that, at the same time, influence both the treatment participation and the outcome should be included in the model. Important is also that the variables are independent of the treatment, i.e. that they are unaffected by the treatment execution. To control for this variables should either be fix over time or measured before and after the treatment. The location of these variables is made using economic theory, results of

[31] Caliendo et al. (2008), p. 36
[32] Dehejia et al. (2008), p. 152
[33] Ibid., p. 161

previous research and knowledge about institutional settings.[34]

Another important aspect of choosing the variables is that the variables should stem from the same source. By using data from e.g. the same questionnaire one insures that the data are gathered in a consistent way and that the data therefore are comparable. The better the data, the easier it is to meet the criterion under the unconfoundedness assumption and to find better matches. One point that is important to make here is that the data should not be too good. Probabilities that equal 0 or 1 make it impossible to use the values on X for matching. This since the probability that an individual with value X receive treatment or not is either one or zero, they either always get it or they never do. This invalidates the overlap assumption that assures randomness to guarantee that an individual with identical characteristics can be observed in both states.

If uncertainty arise about whether to include a variable in the model or not, there are pros and cons for doing both. Some research points that it's better to exclude an uncertain variable.[35] The reason is that it can increase the variance. Other researchers recommend to exclude variables with doubts only when the consensus of the variable is completely irrelevant to the outcome or is an inapplicable covariate according to economic theory.[36] In all other cases it should be included. Caliendo and Kopeing (2008) summarise their findings by saying that the choice of variables should be based on economic theory and previous empirical findings. There are also different methods to test whether a variable is proper or not. These methods can be used in addition to the economic theory and the findings of previous research. They include more technical methods on how to measure the relevance of a variable. Caliendo and Kopeing (2008) provide further information on the methods also known as hit or miss method, statistical significance and leave one out cross validation.[37]

[34] Caliendo et al. (2008), p. 38
[35] Augurzky et al. (2001), pp. 25-26
[36] Rubin et al. (1996), pp. 249-250
[37] Caliendo et al. (2008), pp. 39-40

2.3.5 Choice of matching algorithm

There are three commonly used ways to use the propensity score to estimate the treatment effect: *covariate adjustment using the propensity score*, *stratification on the propensity score* and *propensity score matching*.[38] The first method uses the propensity score together with the treatment exposure in a linear regression but it does not control for differences in treated and untreated individuals. The method with stratification places individuals with similar scores in the same group and calculates the treatment effect for each group. This method has similarities to the randomisation technique that the matching methods try to simulate, but is not the most efficient method. Austin (2008) suggests the use of propensity score matching since it can achieve a placebo randomisation and has proven to result in the greatest reduction of selection bias among the models.[39] Given the propensity score estimated by the logit model, one has to make the decision on how to best match the scores from the treatment group with the scores from the comparison group.[40] Different propensity score matching estimators use different methods to find a suitable nearest match to each treated individual. They also differ in the weight they assign to the methods.[41] Four commonly used methods for propensity score matching are described below.

Nearest Neighbour Matching (NN)

Propensity scores for the treatment group observations are matched with the control group observation with the nearest lying propensity score. This is being done for every observation in the treatment group. There can be nearest neighbour matching with or without replacement. Matching with replacement implies that an individual from the comparison group sample can be reused as an NN-match to another treated individual. By matching without replacement every individual from the comparison sample is only matched once. When the difference in propensity scores for the treated group and the comparison group is large, not allowing for replacement

[38] Austin (2008), p. 2038
[39] Ibin.
[40] Dehejia et al. (2002), p. 153
[41] Caliendo et al. (2008), pp. 40-43

increases bias.[42] By using the replacement method one will experience a trade-off between variance and bias. Hence, allowing replacement one will receive a higher matching quality, and a decrease in bias. This can be useful if there is a large difference between the propensity scores in the treated and the untreated group. Thus, if a sample contains many treated individuals with a high propensity score, but only a few non-treated individuals with scores at the same level, allowing for replacement will lead to better matches and therefore a reduction of bias. At the same time, using fewer numbers of units in the comparison group will increase the variance.[43] A problem that arises by using the method without replacement is that the order in which the matches are paired is of importance. This matching order has to be done randomly.

Both the nearest neighbour matching method and the caliper method described below can match a treated unit to one or many comparison units. By using more than one match to every treatment unit one will increase the precision of the estimates, but also increase the bias. By using only one comparison unit to each treated unit one can ensure that the closest possible match is always being used.[44]

Caliper and Radius Matching

The risk of using a neighbour far away from the treated individual's propensity score can be overcome with caliper matching. The caliper matching implies that one adopts a tolerance level, or *caliper,* in which treated and untreated individuals are matched. By this, bad matches such as a treated individual with low propensity score matched with the nearest untreated neighbour with a high propensity score, can be avoided. This is to prefer from a quality point of view since the quality of the matching rises. Another benefit of the caliper method is that it can include different numbers of matching units depending on the closeness of the fit. One caliper can contain three matches whereas other intervals might contain only one. The use of the caliper matching method increases the variance due to the fewer matches performed. Critics

[42] Dehejia et al. (2002), pp. 153-154
[43] Caliendo et al. (2008), pp. 41-42
[44] Dehejia et al. (2002), p. 153

against the method claim that it is difficult to know in advance which caliper that is best suitable. Researchers Dehejia and Wahaba (2002) present an extended method of radius matching that implies the use of a caliper and matches on all its consistent non-treatment members. Beneficial for this approach is that it allows for matching on more units when good matches are available. In this sense it combines the benefit of the caliper matching method but avoids the problem with the risk of bad matches.[45]

Stratification Matching

When matching with stratification one divides the treated and untreated individuals into interval groups based on the propensity score. For every interval group one calculates the difference in mean between the output of the treated and the untreated individuals. Previous research suggests the use of five intervals and claims that it can reduce the bias associated with the covariates to a 95 percent extent.[46]

Kernel and local linear matching

The non-parametric matching techniques of kernel matching and local linear matching calculate the weighted average on nearly all individuals in the non-treatment group. Previously discussed methods have only matched the treated observation with one or a few untreated observations. The kernel matching uses all the observations from the non-treatment group to compare them to the treatment group. Since more information are used in this method, the variance is lower than in previous matching methods. One disadvantage of the method is that individuals that are actually bad matches can be matched and therefore contribute to an increase in bias.[47]

To summarise the theory around the choice of matching algorithm, the decision needs to be based on the actual sample at hand. In a sample with only a few observations there is often a trade-off between bias and variance. Thus, the choice of algorithm is important. With larger sample sizes the matching is getting more exact,

[45] Caliendo et al. (2008), p. 42
[46] Ibid. pp. 42-43
[47] Ibid. pp. 43-44

and the choice of matching strategy is not of the same importance than with the smaller sample sizes. The best choice varies from case to case and for every situation at hand. When having few observables it makes no sense to use the matching method which involves no replacement. This would only imply that some treated observations do not get its best possible match or even close to the best possible match. When having many non-treated observations on the other hand, there might be a gain in the matching efficiency by using some of the methods that do not only use one nearest neighbour sorted by the propensity score.

For the implementation of the matching, the treated propensity scores and the comparison propensity scores are ranked from high-to-low, low-to-high, and random. When matching without replacement the unit with the highest rank is matched first, and its matched comparison unit is removed from further matching. When using matching with replacement both single nearest neighbour matching and caliper matching can be used.

2.3.6 Overlap and common support

To avoid any kind of evaluation bias, it is relevant to assess the common support of the two groups. Thus, controlling for incomparable observations in both groups improves the pseudo randomness of the estimation. Fulfilling the common support condition guaranties that all possible combinations of characteristics for the treatment group can, even if not perfectly corresponding, be seen in the control group.[48] The simplest way this can be done is by assessing the density of the propensity score distribution for the two groups. One other method is called the minima and maxima comparison and it controls for the common support by deleting all propensity scores that are smaller and larger than the propensity score of the other group. E.g. if the treatment group has a propensity score in the range between [0,06-0,88] and the control group [0,03-0,86], then the propensity scores included in the model has a range from [0,06-0,86]. The problem using the minima maxima comparison is that observations very close to the bounds are eliminated and therefore discarded form the analysis. Further problems can occur by gaps in the

[48] Caliendo et al. (2008), p.45

overlapping and when the number of excluded observation in relation to included observations is large. One then has to question whether the remaining individuals can be seen as representative for the whole sample.[49] When using the nearest neighbour matching the minima and maxima comparison is not as necessary. This since the observations anyhow will be matched with a control group observation of a similar propensity score.

2.3.7 Assessing the matching quality

The propensity scores are as previously mentioned estimated with a logit or a probit regression model. Most research with matching use scores from the logit regression function. These scores are then matched so that one set consists of one treatment unit together with one or many comparison units. When the matched samples have been constructed, one must assess the balance in the baseline. This because the matching is not done by conditioning on the covariates but on the propensity score. Balancing on the baseline means that the baseline covariates, for both treatment and control group, measured or observed before the treatment should correspond. If the comparison group and the treatment group are homogenous in the propensity score, even if they are heterogeneous in the covariates **X**, the covariates tend to balance on the baseline. This will make the estimation "look randomised" and reduce selection bias. Even in a randomised sample, the balance will not be perfect, therefore it is not to expect from the estimated propensity score matching either. But it should be a clear balance in the covariates when matching on the propensity scores. To assess the covariance balance, one common and often appropriate method is the use of standardised differences:

$$D = \frac{100 \times |\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}|}{\sqrt{\frac{s^2_{\text{treatment}} + s^2_{\text{control}}}{2}}}$$

The most previous research claim a standardised difference value between three to five percent to be adequate.[50] Another way to test the matching quality is by using a

[49] Caliendo et al. (2008), p. 45
[50] Austin (2008), p. 2039

two sample t-test. The test show the differences in covariates mean between the treated and non treated group. These are expected to exist before matching but hopefully be erased or minimised after the matching. The t-test can often be preferred to the standard differences if one is uncertain on the significance of the result. A shortcoming using the t-test is that it can not evaluate if the bias before and after the treatment has increased or decreased.[51] The problem when assessing the matching quality by the standardised differences is that the covariates X need to be in an interval scale of measurement. Data grouped or labelled with a nominal or ordinal value can not be tested. [52] If assessing the effect for a treatment that starts at a certain point in time and continues for some time, it is important to carefully assess the data used from the comparison group. The main objective is to ensure that the treated and the untreated group are being measured in the same economic environment.

2.3.8 Variance estimation

The problem by estimating the variance of the treatment effect is that the variance of the treatment effect should include the variance accredited to the estimation of the propensity score, the effects of the choice of matching algorithm and the common support together.[53] Caliendo (2008) discuss three different approaches for variance estimation. One of them is the variance estimator where the *sample* average treatment effect for the treated, previously known as (ATT) is given by:

$$\tau\mathrm{SATT} = \frac{1}{N_1} \sum_{i \in (D=1)} \left(\mathrm{Y_i}(1) - \mathrm{Y_i}(0)\right)$$

[51] Caliendo et al. (2008), pp. 46-47
[52] Stevens (1946), pp. 677-679 and Caliendo et al. (2008), p. 50
[53] Caliendo et al. (2008), p. 51

Only the average of the variance of the distribution is needed to estimate the variance of SATT:

$$\mathrm{Var_{SATT}} = \frac{1}{N_1} \sum_{i \in (D=1)} \left(\mathrm{D_i} - (1 - \mathrm{D_i}) \times \frac{\mathrm{K_M(i)}}{\mathrm{M}} \right)^2 \hat{\sigma}^2_{\mathrm{D_i}}(\mathrm{X_i})$$

Where M is the number of matches done by the sample merge and $K_M(i)$ is the number of times each individual i has been matched. If nearest neighbourhood without replacement has been used then $K_M(i)$ equals 1.

2.3.9 Sensitivity Analysis

A sensitivity analysis is done to show the robustness of the results if one or more of the identifying assumptions are not fulfilled. The propensity score matching is done on the assumption of unconfoundedness. Hence, if the covariates are independent of the participation in the treatment group, then the total outcome should also be independent of the treatment participation. Deviations from unconfoundedness, say if there is one variable outside the model which affects both the assignment to treatment and the outcome of the treatment simultaneously, then the model contains a hidden bias.[54] Since non-experimental data can not reveal the selection bias in the estimation, a robustness check of the model has to be done with a sensitivity analysis. One method to check the sensitivity of the estimated treatment effects is to assume that the participation probability is given by:

$$\mathrm{P_i} = \mathrm{P}(\mathrm{x_i}, \mathrm{u_i}) = \mathrm{P}(\mathrm{D_i} = 1 | \mathrm{x_i}, \mathrm{u_i}) = \mathrm{F}(\beta \mathrm{x_i} + \gamma \mathrm{u_i})$$

Where x_i represents the observed covariates for individual *i* and u_i the value of the unobserved variable. γ_i is the effect of u_i on the participation decision. If the estimated model is free from bias then the value of γ_i equals zero. If not, then two observations with the same covariates x will not have the same probability to receive

[54] Caliendo et al. (2008), pp. 56-58

treatment.[55] Altering the value on the unobserved parameter γ_i allows the researcher to examine the robustness of the results. From the reaction when altering the γ_i-value significant levels and confidence intervals can be derived. There are further methods to test the robustness used by researchers in the field of treatment effects. Irrespective of the model of choice one should keep in mind that the sensitivity analysis never gives 100 percent satisfying result justifying the assumption of unconfoundedness.[56]

2.4 Instrumental Variables

Solving a causality problem of a special treatment, using a simple ordinary least square method would be an option if one had perfect exogenous variables and all independent variables were observable. This is often not the case. The more common estimations are far more complex and include both endogenous and omitted variables. If $y=\beta_0+\beta_1 x_1+u$, where x_1 is a binary endogenous explanatory variable, then estimating the effect of a treatment x_1 on output y using OLS would imply a biased and inconsistent estimator of β_1 in the presence of omitted and endogenous variables. By using instrumental variables, that recognise and accounts for omitted variables, one can solve the problem of endogenity and excluded variables. The two main properties that an instrumental variable z has to fulfil is

1) Cov(z,u) =0
2) Cov (z,x) ≠0

That is, the instrumental variable z has to be uncorrelated with the error term u and at the same time correlated with the explanatory variable x. To test whether z is correlated with x, one can do a simple regression between the two:

$x= \pi_0 + \pi_1 z + v$

Where π_1= Cov(z,x)/Var(z) and since Cov(z,x) must be ≠0 to fulfil property (2) it holds

[55] Becker et al. (2007), p. 72
[56] Caliendo et al. (2008), p. 58

if, and only if, π_1 is larger than zero. If $\pi_1 > 0$ then one should be able to reject the null hypothesis: $H_0=\pi_1=0$. If that is the case then one can be fairly sure that the instrumental variable fulfils property (2). Testing for the independence between the instrumental variable z and the error term u is more difficult and conclusions on their correlation are based on economic theory and observed economic behaviour.
The instrumental variable estimation is similar to the OLS estimation and the IV has, just as the OLS when facing a larger sample size, a normal distributed estimator. One can also conditionally assume homoskedasticity on the instrumental variable z.[57]

$E(u^2|z) = \sigma^2 = Var(u)$

The consequence of the use of a bad instrumental variable can be a large standard error. This since the instrumental variable z and the explanatory variable x are only weakly correlated. Even if the correlation between the instrumental variable z and the error term u is weak, the lack of correlation between the explanatory variable x and z still makes the estimated IV deviate from the true beta, i.e. causing bias. In the case when the correlation between z and x is smaller than the correlation between z and u, it might be better to use OLS than to use an instrumental variable estimation.
If the model would include an instrumental variable that is in no way correlated with x, it would not only fail to fulfil one of the main properties of an IV but also give false parameter values in the estimation and show false standard errors. The main rule is therefore never to proceed with a variable that do not fulfil the two main properties. It is also notable that the R-square for the instrumental variable can be both positive and negative, since SSR for IV's can be larger than SST. The main objective of the instrumental variable estimation is not to provide the best goodness-fit but to provide a better estimate when the x variable is correlated with u.

2.4.1 IV estimation of a multiple regression model

If including additional variables in the model, the use of a multiple regression helps account for other factors affecting the output besides the treatment itself. By including more variables some of the variables previously present in the error term

[57] Wooldridge (2006), p. 512

can be included in the model. Thus, instrumental variables previously neglected because they were correlated to both the independent variable x and the error term u can be used again. This increases the possible number of IV's. If only one of the variables in the model is correlated with the error term, then:

$$y_1 = \beta_{ij} + \beta_1 x_1 + \beta_2 x_2 + u_1$$

Where x_1 is an endogenous variable and x_2 an exogenous explanatory variable, and where $E(u_1)=0$ by assumption.[58] Using an OLS when one or more explanatory variables are endogenous makes the estimation biased and inconsistent. The use of instrumental variables that fulfil the same properties as in the single instrumental variable case can help to improve the estimation. One therefore needs to find a variable that is correlated with the endogenous explanatory variable x_1 and uncorrelated with u. Even if the other variables in the model, in this example the exogenous explanatory variable x_2, is uncorrelated with u and correlated with x_1 it may not serve as an instrumental variable since it already appears in the model. Therefore an additional property of partial correlation needs to be fulfilled. The endogenous variable x_1 expressed as a linear function of the exogenous variable:

$$x_1 = \pi_o + \pi_2 x_2 + \pi_1 z_1 + v_2$$

Where $E(v_2)=0$, $Cov(x_2 v_2)=0$, $Cov(z_1 v_2)=0$, z_1 is the instrumental variable. v_2 is the error term and π is unknown parameters and $\pi_1 \neq 0$.[59] I.e. the instrumental variable z_1 must be correlated with the endogenous explanatory variable x_1. Once again, one can use an OLS to test the correlation between x_1 and z_1, but one can not test if the instrumental variable z_1 is uncorrelated with v_2. This assumption has to be build on economic reasoning and theory.

It is easy to extend the model described above with a number of additional exogenous explanatory variables:

[58] Wooldridge (2006), p. 521
[59] Ibid., p. 522

$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_{kxk} + u$

Where x_1 is correlated with u. The instrumental variable still has to be uncorrelated with the error term and correlated with the endogenous explanatory variable. An additional assumption is that no perfect linear relation between the exogenous variables exists. This assumption is similar to the OLS model assumption about non colinearity between the variables. For statistical inference it is also important that we assume homoskedasticity of u. [60]

2.4.2 Two stage least squares

If the model to be estimated has one or more exogenous variables that could work as an IV for the endogenous explanatory variable x, then the two stage least square estimator can find a linear combination of IV's that has the highest correlation with the endogenous explanatory variable.

Six assumptions have to hold for the two stage least square to have good enough sample properties for an estimation:[61]

1. The model in a population can be written as: $y=\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_{kxk}+u$.
2. y, x and u are all random samples.
3. There exists no perfect linear relationship between the IV's and the rank conditions for identification holds.
4. The error term has zero mean and every instrumental variable in the model is uncorrelated with the error term u.
5. $E(u2|\mathbf{z}) = \sigma^2$, Homoskedasticity.
6. No serial correlation, $E(u_t\, u_s|z_t\, z_s)=0$ for all $t \neq s$.

The best instrumental variable under the assumptions above is the linear combination of the exogenous variables not being included in the model, i.e. z_i when i=1,... ,k. This is given by the reduced form written as:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2 \qquad (1)$$

[60] Wonnacott (1997), pp. 727-730 and Wooldridge (2006), p. 23
[61] Wooldridge (2006), p. 550

Where $E(v_2)=0$, $Cov(z_1,v_2)=0$, $Cov(z_2,v_2)=0$, $Cov(z_3, v_2)=0$. Then the best IV for the endogenous explanatory variable y_2 is:

$$\tilde{y}_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 \qquad (2)$$

Where at least two of the IV's must be unequal to zero in order to avoid that y_2 is perfectly correlated with one single instrumental variable.[62] That is, at least on of π_2 and π_3 need to be different from zero.

$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$ is not identified if $\pi_2=0$ and $\pi_3=0$. One can test H_0: $\pi_2=0$ and $\pi_3=0$ against the alternative hypothesis H_A: $\pi_2\neq0$ and $\pi_3\neq0$ with a f-test. One needs to think of (1) as if it divides y_2 into two pieces. The first piece is $(\pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3)$. This is the part of y_2 which is uncorrelated with the error term u. The second piece v_2 is the part that is possibly correlated with u_1 and that makes y_2 a possible endogenous variable. Given the data on z_1, z_2 ,z_3 and given that π_j is known one can compute $\tilde{y}_2$ for each observation. But in practice the π_j can never be observed. To obtain the fitted values one can estimate y_2 by an OLS:

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi} z_2 + \hat{\pi} z_3 + \hat{\pi} z_1 \qquad (3)$$

Where z_2 and z_3 need to be jointly significant. Once one has an estimation of y_2 in $\hat{y}_2$ it can be used as an instrumental variable for y_2:

$$\sum_{i=1}^{n} \left(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}\right) = 0$$

$$\sum_{i=1}^{n} z_{i1}\left(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}\right) = 0$$

$$\sum_{i=1}^{n} \hat{y}_{i2}\left(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}\right) = 0$$

[62] Wooldridge (2006), p. 524

Solving these three unknowns gives the instrumental variables. The two stages of the two-stages least square refers to the regression of $\hat{y}_2$ (3) and a second step which includes a regression of y_1 on $\hat{y}_2$ and z_1. This is the OLS regression, but since one uses $\hat{y}_2$ instead of y_2 the two stage least square can differ from an OLS.[63]

[63] Wooldridge (2006), p. 526

3. Company performance

Measuring the financial performance of a firm may be done in several different ways. One might be interested in the manager specific impacts on the outcome whereas other researches distinctively focus on the financial operations. One can divide the performance measurements into financial and non-financial performance measurements. The financial measurements are so called backward looking measurements because they place their focus on the result from business operations already taken place. Such backward looking measurements are for example the Return on Assets (ROA), Net Profit, Profit Margin (PM) and cash flows. The financial figures and information used for the financial measurements are mainly gathered from the company's own quarterly and annually financial statements. The financial statements include the balance sheet together with the cash flow statement and the income statement and they express the cumulative results of a numerous of different business decisions.

The non-financial measurements are also so called forward looking measurements. They try to reflect the performance in the market such as market shares, market growth, product quality and customer satisfaction.[64] Previous research argues that improvements in a company's product quality, costumes satisfaction or innovation are assets that the company does not accredit for in their financial information.[65] At the same time Potter et al. (2000) and Moers (2000) suggest that the non-financial measurements serve a better long term predictors of financial performance. This since the non-financial measurements are drivers of the performance whereas the financial measurements show the outcome. Potter et al. and Moers further claim that using a non-financial measurement can increase the long-term incentives among managers. Measurements that try to valuate intellectual capital and intangible assets are Market-to-book value (MTB) and the Tobin's q.

The benefit of using accounting based backward looking measurements is its

[64] Moers (2000), p. 11
[65] Ittner et al. (1998), pp. 4-6

consequent measurement of the internal operational efficiency and financial capability in contrast to the non-financial measurements that uses measurements of market share, product quality or market growth. These measurements are influenced not only by external market and other macro elements but also by definition used and measurement methods.[66] The problems with the non-financial measurements are that they are difficult to measure in a way that makes them comparable to other companies. Till today there is no global unified way to measure customer satisfaction or market share. And even if there were, empirical studies of company performance would demand that companies compulsory reported numbers that included non-financial measurements. The two measurements approaching the non-financial assets in a somewhat unified way are the Tobin's q and the Market-to-Book measurement. The disadvantage and difficulties with the non-financial measurement are at the same time the benefits with the financial measurements. Due to the accounting standards and regulations unified measures of financial figures are made and large datasets can be created. These are to a large extent comparable and of great help when making empirical studies.[67]

Financial variables that influence or can be good indicators of the company performance are divided in to three main groups; Profitability, Solvency and Liquidity. A fourth group added are the non-financial measurements of MTB and Tobin's q, measuring and comparing market value and intangible assets.

A company's profitability shows the ability to earn profit from delivering goods and services. It reflects the company's competitive position on the market and the ability of its management. Ratios that can be calculated to assess a company's profitability are:

Gross profit margin (Operating Margin): Gross profit / Revenue. Describes the amount of gross profit (revenue-costs of goods sold before income tax, interest, provision and other expenses) that every unit of revenue has generated. A larger gross profit margin indicates a higher profitability, and a large profitability should

[66] Chang et al (2008), p. 370
[67] Bianchi et al. (2004), pp. 162-164

affect the company's total performance in a positive way.[68]

Operating profit margin: Earnings before interest and taxes (EBIT) divided by revenue.

Net result: Total Revenue – total expenses (i.e. including the income tax, interest, provision and other expenses) / Revenue. The net result is calculated as total revenue minus total expenses divided by revenue. It measures the amount of income that each unit of revenue was able to generate.

Solvency measures the amount of borrowed capital used in the business relative to the amount of owner's equity capital invested in the business.

Return on Equity (ROE): Net income / Shareholders equity. Measures the return earned by a company on its equity capital.

Equity ratio: The percentages of equity on the balance sheet. More then 40 percent is to be considered as good, between 20 – 40 percent as satisfactory and below 20 percent as poor.[69]

A company's liquidity or cash flow is important because it shows the company's ability to meet its payment obligations. Such obligations can be payments for new investments, increased orders or salaries. The ability to meet short term obligations, i.e. how fast assets can be converted into cash, is called liquidity.[70]

Quick Ratio: Cash + Short term marketable investments + receivables / Current liabilities. Is a stricter ratio than the current ratio since it only includes the quick assets, i.e. the more liquid assets. A higher quick ratio indicates a better liquidity.

Current Ratio: Current assets / Current liabilities. Assesses the assets to be consumed or converted into cash within one year in relation to the current liabilities. That is, liabilities falling due within one year. A high ratio indicates a better liquidity level in the company. A lower ratio makes it more difficult for the firm to meet its short-term obligations and make them more dependent on external financing to do so.

[68] Robinson et al. (2007), p. 515
[69] Ibid., pp. 506-511
[70] Ibid., p. 506

Market-to book value (MTB): Measures the market value of a company in relation to the book value of its assets. If there is a difference between the market value and the book value it's a signal that something is not accounted for in the balance sheet, i.e. intangible assets of the company.[71]

Tobin's q: The performance measure referred to as Tobin's q, after James Tobin (1969). It is measured as q = (equity market value + liability book value)/(Equity book value + liability book value). The nominator measures the market value of installed capital and the denominator the replacement cost of capital. It takes the future prospects of the company into consideration and tries to relate the market value with the replacement cost. If the value of Tobin's q = 1 then the cost of using an asset and its profit is equalised, meaning that the market value of capital is equalised with the book value of capital. If the q is higher than 1 this indicates that investments in the company can yield positive gains. This because the cost of capital (book value of liabilities + equity book value) is lower than the price you get on the market (equity market value + liability book value). The Tobin's q is also a common measure used as an explanatory variable in models assessing company performance.[72]

[71] Bianchi et al. (2004), p. 166
[72] Robinson et al. (2007), p. 534

4. Corporate Social Responsibility

When asking the question; *what is Corporate Social Responsibility*? The likelihood that one gets a unified answer is low. The question might sound simple, but the answer is complex and the conception about social responsibility varies. Therefore plenty of different definitions are obtained. One commonly used description of CSR is that it describes a company's commitment to operating in a way that takes into account not only the financial implications of business decisions, but also the social and environmental impact it has on the community. That is; Corporate Social Responsibility describes a company's commitment to be accountable to its stakeholders. This comprises not only the interest of customers and investors, but also suppliers, communities, employees, regulators, groups with special interests in the company and the society as a whole.

In 2001 the European Commission presented their work on Corporate Social Responsibility called *The Green Paper*. The European Commission's definition of CSR is "*a concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders on a voluntary basis".*[73]

The World Business Council for Sustainable Development (WBCSD) define CSR as: *"Corporate Social Responsibility is the continuing commitment by a company to behave ethically and contribute to economic development while improving the quality of life of the workforce and their families as well as of the local community and society at large"*.[74]

To be engaged to and undertake CSR activities demands that a company actively makes decisions that try to minimise the negative economic, social and environmental impact. The minimisation of the downsides helps to maximise the benefits of their actions towards their stakeholders. The numbers of activities that

[73] European Commission, COM (July 2002:347)
[74] Holme et al. (2000), p. 8

can be performed to increase benefits for stakeholders are infinite. Some CSR issues that are a common target for CSR strategies are labour standards, responsible sourcing, environmental management, human rights and community relations. The actions taken to maximise stakeholder benefits can also give positive side effects. It can for example give a company comparative advantage in terms of enhanced brand image, increase transparency and decrease risk in addition to attract and motivate employees. It might also highlight previously hidden commercial opportunities and improve access to capital due to improved investor value. Thanks to a higher awareness among firms and stakeholders, the demand for a sustainable actions and responsibility has increased. Companies with a previous short term view, trying to maximise short term profits, are starting to be aware of their ability to change and make a difference. This by managing their operations in such a way that the company and its employees together with additional stakeholders and society as a whole benefit in the long run.[75]

In theory many benefits are attached to CSR. In a globalised international economy, where business activity gets more spread and complex, acting with CSR as a benchmark can be a way to structure a whole organisation, its activities and increase transparency. Through globalisation and rapid information exchange the increased competition makes brands and brand reputation increasingly crucial. Showing stakeholders that the firm of concern takes social responsibility can be a way to distinguish the company brand and meet the stakeholders' expectations. Since CSR has become increasingly important in the public opinion, financial stakeholders also recognise and want to make sure that the public view of the company remains good. This helps reduce the risk of financial downturn due to changed customer perception.[76]

There are also some voices being heard from the CRS sceptics. One of them uses a classical economic argument from the traditional classical economic school led by Milton Friedman. The classical economic approach advocates that the sole

[75] European Commission, COM (July 2002:347), p. 5
[76] Ibid., p. 6

responsibility and main objective for a firm is to increase profit to maximise its owner interest. Thus, the view is that social problems can not be solved by companies or organisations but have to be done through regulations by governments.[77] Another objective is that companies are not equipped to solve social problems. It claims that managers have a focus on financial issues that makes it difficult to come up with good sustainable solutions. This would mean that CSR and activities for sustainability would dilute business purposes.[78] A further argument against CSR says that social responsibility increase the transparency and therefore uncovers some previously hidden costs. This can be costs that in the past have been carried by the society in terms of air pollution, toxic additaments to products and unsafe products. These cost will for the social responsible company be carried by the firm, which as a result of higher cost will increase the price of goods sold. The higher price level makes the company less competitive on the market and the company would therefore experience an economic downturn thanks to the sustainable actions.[79]

A related term to CSR is the Social Responsible Investment (SRI). Social Responsible Investment is the saving and investing in companies that consciously undertake sustainable actions. To measure and list firms that involve themselves in sustainable activities one can use methods of screening.[80] A second term that often figures in the same context as SCR is corporate governance. Corporate governance is a wider concept that also includes the corporate social responsibility. The idea of corporate governance gathers the set of activities e.g. regulations, institutions and other extern factors that together affect the way that a firm is managed.[81]

[77] Carroll et al. (2008), p. 49
[78] Ibid., pp. 49-50
[79] Ibid., p. 50
[80] Screening is the method of securing a selection of firms investing responsible. This can mean excluding companies that violate environmental laws, use child labor, discriminate applicants, produce goods that can be damaging to society e.g. drugs and weapons. (Source: Camejo et al. (2002), pp. 1-2)
[81] Logan et al. (2005), p. 3

4.1 Measurements of CSR

Taking responsible actions and sustainable decisions would not be the topic it is today if there were no possibility to measure the extent of sustainability. In order to distinguish a company from others in respect of social responsible actions taken, one needs a measurement to show the difference. Measuring something as complex as responsibility makes the measurement of corporate social responsibility somewhat problematic. The first problem occurs since the measurement of something often not quantitative requires innovation and a structured, unified framework. A large part of the listed companies have for years published information about their sustainable commitments through financial statements. Since the beginning of the 20th century the number of firms communicating their responsible action this way has largely increased. These reports are still important as information on the specific company and its actions, but using the same framework increases the transparency within and in between firms and improves domestic and global comparability. One of the most known and renowned measurements of CSR is the Global Reporting Initiative's standards.[82]

4.2 Global Report Initiative

The Global Report Initiative (GRI) is a benchmark that helps companies in their work towards behaving more social, environmental and economic responsible. It is created by the GRI organisation, a multi stakeholder organisation containing thousands of experts around the world who together created the guidelines for GRI.[83] Their framework is today one of the world's most renowned tool for reporting sustainable actions in a way so that they are informative and comparable. The demand for a benchmark that demonstrates the social responsible actions taken by a given firm and at the same time ease the comparability between companies over time brought the GRI development forth. The first version of the GRI framework was presented in year 2000 and resulted in more than 50 companies reporting according to GRI in the first year. The second version was presented two years later and after improvements

[82] Hedberg et al. (2003), pp. 153-154
[83] GRI Homepage (accessed 15.11.2009), www.globalreporting.org

of specific indicators and business specific supplement the third version was published in 2006. The interest for GRI has steadily increased since the start in year 2000 and the GRI-organisation could in 2008 report that over 1000 firms worldwide were reporting according to GRI standards. The aim for the GRI-organisation is to become an equally accepted reporting tool as today the financial reporting e.g. GAAP and IFRS. Policies and regulation on a national and international level would help GRI increase their importance. Some governments are already regulating on the use of sustainable reporting strengthening the GRI position even more.[84]

The GRI framework contains of guidance and detailed principles on how to define the needed containment and quality of the report. The reporting is divided in to two main parts; the first one includes the guidelines about the quality, containment and reporting principles whereas the second part contains the standard accounting principles. The standard accounting can be divided in to three areas: Strategy and profile, governance attitudes and indicators of achievement. The three areas are detailed frameworks and the reporting firm can independently choose to include the whole or only parts of each framework in their reporting. If a company choose to fully report according to the three frameworks it will be reporting according to level A. If some details in the frameworks are excluded, the firms will be reporting according to B or C, where C uses the least numbers of frameworks. It is important that the company clearly communicate according to which standard (A, B & C) that they are reporting.[85]

4.3 Global Compact

The United Nations Global Compact is an initiative from the UN that aims to help organisations and companies worldwide in their work towards social responsible actions and a sustainable development. The global compact consists of ten principles within human rights, environment, labour environment and corruption. These principles are recommended to follow and can be seen as the primary global policies for businesses and organisations around the globe, but they are not aiming to be

[84] GRI Homepage (accessed 15.11.2009), www.globalreporting.org
[85] Ibid.

used in a quantified way. Whereas the GRI provides very clear principles on how to report the responsible actions so that they can be compared with others, the global compact is more to be seen as a moral framework used for the intern use in the company. Today over 3500 firms has decided to affiliate the principles of the global compact and report accordingly.[86]

[86] UN Global Compact, (Accessed 09.09.2009), www.unglobalcompact.org

5. The Data

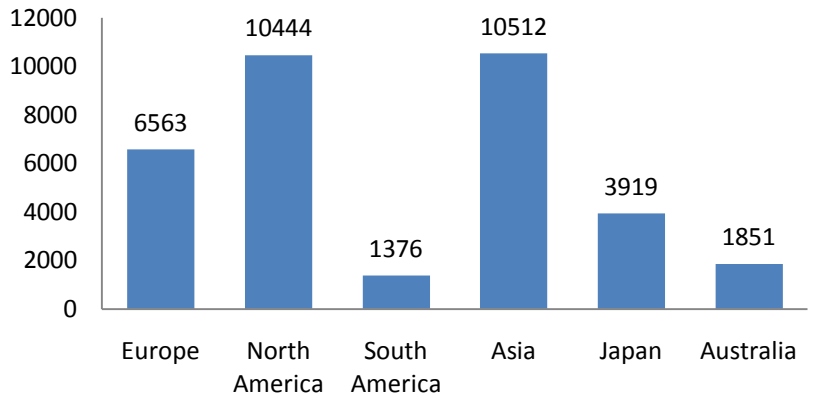
The data used in this estimation consists of 202000 observations spread over the time period 2003-2007. It describes 40400 stock market listed companies and contains information about company site, industry field and several financial figures. The panel data is balanced, which indicates that each individual is observed under an equal amount of years, in this data five years.

Panel data normally refers to data that are mostly cross sectional, i.e. where N>T, Where T is the numbers of time units t = 1, …, T and N the number of cross-sectional units with i = 1, …, N. The total number of observations is therefore N x T. Time series cross-sectional data usually refers to data that are relatively seen more like time series, i.e. data where T>N, but T is relatively high. With panel data there can be two types of variation; within the cross section (time varies) or between the cross section (differences in units measured). [87]

The data are gathered from observations in 139 countries and the largest countries and regions represented are presented in Graph 2 where one can see that North America and Asia are the largest groups both representing about 25 percent of the total sample. Worth to mention is that the in the table missing continent Africa counts for 1238 observations.

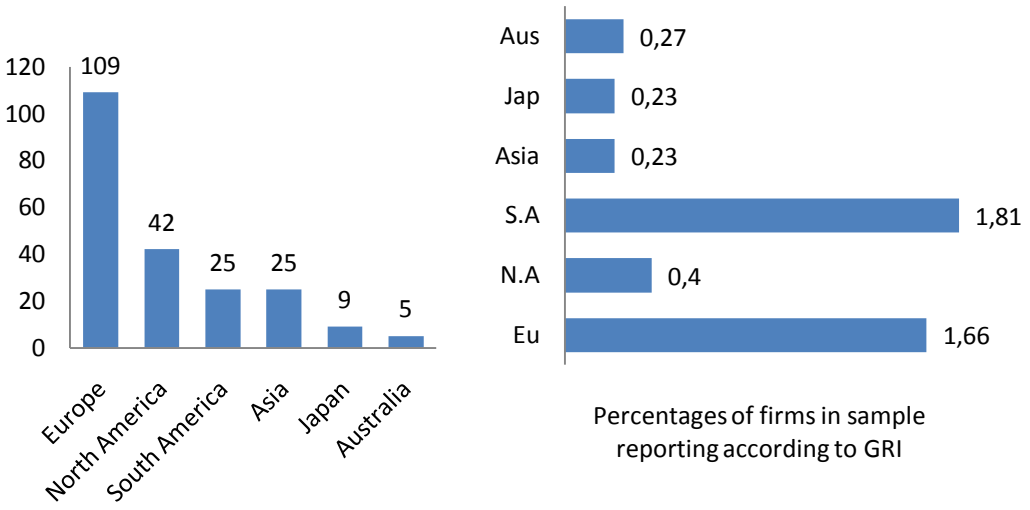
[87] Wooldridge (2006), p. 506

Graph 2. Total number of companies in sample according to geographic site.



The numbers of firms reporting according to GRI is 226, that is <0,6 percent of the total sample. Their geographical spread is presented in Graph 3. The group consisting of the 226 observed GRI reporting firms are from now on referred to as GRI group or treatment group. Graph 4 shows the percentages of firms reporting according to GRI in the data sample. Here one can see that South America has a quite small number of firms (25) reporting with GRI in this sample, but the highest relative number of firms reporting with GRI (1,81 percent).

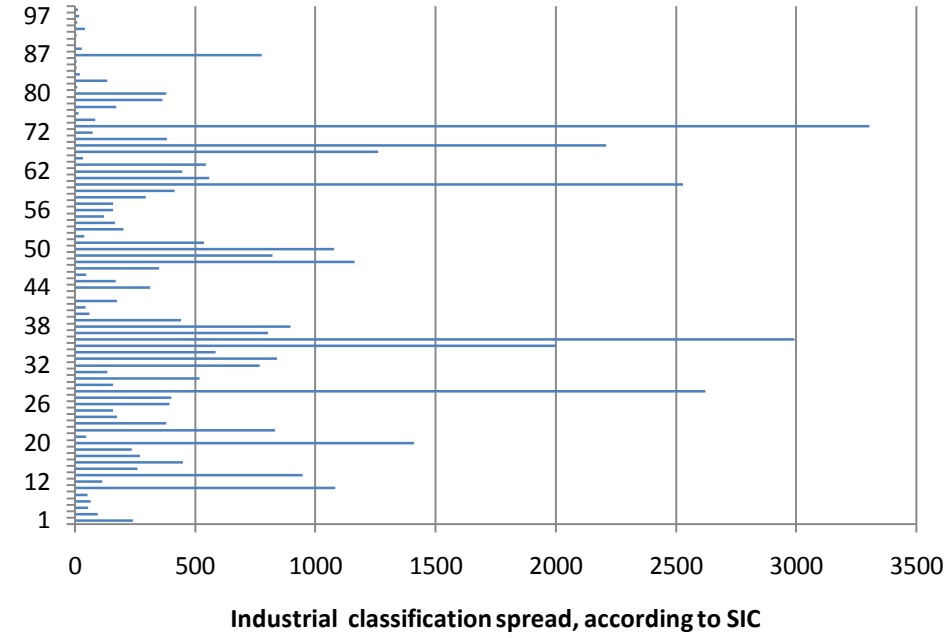
Graph 3 & 4



The data have an industrial classification using the Standard Industrial Classification (SIC), with two digits. Graph 5 shows the spread in industrial classification in the total sample. Each of the lines in Graph 5 represents an industry classified according to the

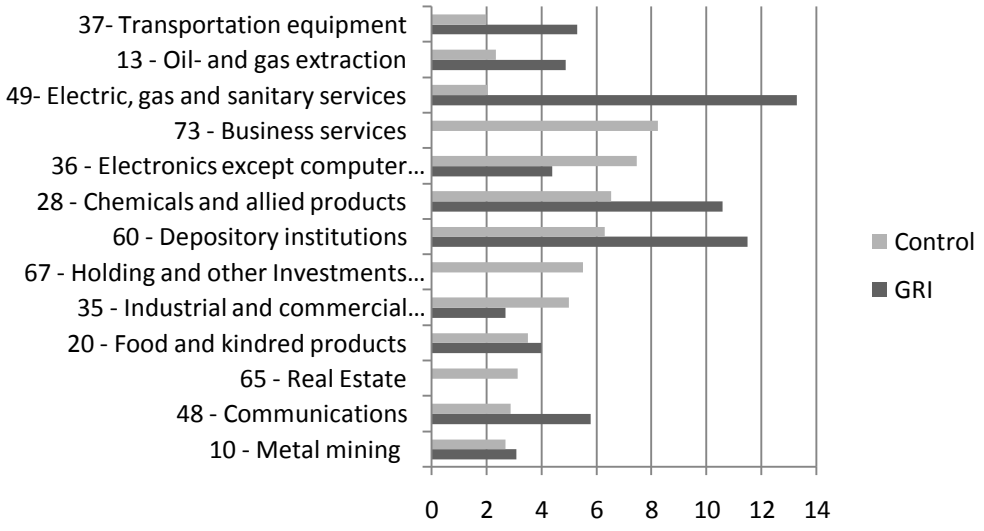
US Securities and Exchange Commission. The exact specification of the Standard Industrial Classification, two digit classification can be seen in appendix Table 1. The ten largest industrial groups in the sample are presented in Table 2 in appendix.

Graph 5. Industrial Classification spread



The whole sample consists of 80 industrial classifications where business services and electronics are the two largest groups.

Graph 6, 10 largest industry groups, relative number



Graph 6 shows the industrial spread among the GRI companies in the sample. Electric, gas and sanitary services has with 13 percent the largest relative amount of companies reporting with GRI. The same industry category only makes up for 2 percent of the total control group sample. Industry groups like business services and electronics, frequently occurring in the control group have a much lower representation among the GRI group.

5.1 Financial performance data

The data set contains a number of financial measurements commonly used to measure a firm's performance. Each captured from the year 2003 to 2007 for all 40400 observations.

Table A. Descriptive statistics, total sample.

Table A					
Variable	Obs	Mean	Std. Dev.	Min	Max
roa	149975	-5,7234	62,13068	-998,2	964,3
q	84295	11,1303	753,4844	0,0	175277,5
mtb	101597	8,6796	685,2225	0,0	175272,7
rosf	142202	1,3629	72,00685	-995,9	991,9
invest	81017	0,0727	1,531928	0,0	426,6
rds	43980	2,9422	136,0625	-303,4	25684,4
pm	138694	-4,1456	109,3257	-999,0	998,7
sh_return	118683	0,5727	20,99324	-1,0	4192,9

ROA = Return on assets, %
ROSV = Return on shareholder value, %
MTB = Market to book value
ROSF = Return on shareholder funds, %
Invest = Total Investments, th USD
RDS = Research & development, th USD
PM = Profit margin
SH_Return = Shareholder return

Observing the descriptive statistics for the raw data one can see that both the minimum and the maximum observations widely exceed the sample means for all the eight figures showed in Table A. At the same time both the values of the sample mean and the size of the standard deviations indicates that the data consist of some unusually large and small data, also known as outliers. This can be due to extreme economic situation and development for a single company but also due to

measurement- and type errors. To avoid that these extreme values influence the assessment of the treatment effect, the data ware cleared from possible outliers. After testing the data with different percentiles all observations below and above the 1st respective 99th percentile were deleted. This resulted in more economic plausible means and standard deviations together with radical changes in minimum and maximum values. Before the cleaning of the sample the mean of return on assets were -5,7 percent which indicates that an average firm would per invested cent (USD) accumulate a 5,7 percent loss annually, over a period of five years. This is not very plausible in a time where the global annual growth was 4,3 percent.[88] From now on, every reference to the data refers to the cleaned data, if nothing else is stated.

Table 3 in appendix shows the descriptive statistics for the raw data compared with the cleaned data for both the treatment and the control group. Here one can see that the mean return on asset value of the cleaned sample is -0,26. However, the size of ROA is highly dependent on the industry of the company. Companies and industries that requires a high initial investment level normally reports a lower ROA. The measure should therefore only be compared with firms in the same or similar industry. The SIC specific statistics are shown in Table 7 in the appendix where one can see that industries demanding physical machinery like *Transportation, Manufacturing* and *Mining* have a high average total asset value. Industries demanding less investment, e.g. *Financial Services,* have therefore a lower value of total assets. Assessing the ROA in the appendix, Table 7, one can see that industries like *Manufacturing* and *Mining* show a negative ROA, whereas the *Retail Trade* and *Financial Services* show a positive average return on asset.

Looking at the geographical differences in financial performance (Table 5 in appendix) one can see that Asia and Japan are the two regions with the best financial figures for GRI-reporting group respectively control group. The largest mean difference between GRI and control group has Australia and New Zeeland. Both America and New Zeeland had a negative overall return on assets. This indicates that certain regions, especially for the control group, contributes to a large extent to the negative ROA-

[88] CIA, The world Factbook (Accessed 23.11.2009), ww.cia.gov

value for the total control group. One can therefore conclude that the geographical aspect is important when determining the difference in company performance. Further descriptive statistics specifically describing industry groups and geographical classification are presented in the appendix Table 5 and Table 7.

Another measurement of financial performance is the Tobin's q. It measures the cost of replacing capital in relation to market value. A high figure indicates that the cost of capital is lower than the market value, this due to e.g. immeasurable assets. According to Table 3 in the appendix the companies reporting according to GRI also reports a slightly lower Tobin's q value. This could indicate that the firms actively reporting their social actions have a book value closer to the market value. This can be due to better communication, larger transparency or in general better company organisation. The market-to-book-value (MTB) is a simplified form of Tobin's q where only the market value is divided with total assets. Not surprisingly, the difference in MTB between the two groups is similar to the difference in Tobin's q value.

The companies reporting to GRI has an average incorporation year *(yoi)* of 1944 whereas the control group average firm was incorporated in 1975 with their minimum as early as 1399, see Table 6 in appendix.

By looking at the descriptive statistics (Table 3 in appendix) one can see that the treatment and control group varies not only in mean but also in the standard deviation. Worth notice is that both the standard error and the mean, with one exception each, are "better" in GRI sample than in the control group sample. This lead to the question if this difference in mean is statistically significant to the extent where one can to a 90-99% significant level say that the difference in mean between the two groups represents a "real" difference between the two groups. This will be tested by using the t-test. The null hypothesis implies that the mean for the treatment group (*GRI*) and the non-treatment group (*control*) are the same. It will be tested against the alternative hypothesis that the mean from the GRI group differs from the mean in the control group.

By calculating the t-value it is possible to assess if the differences in mean between the companies reporting with GRI and the observations in the control group are statically different.

Table C. T-values per continent/country

GRI vs. Control mean according to continent or country						
	t-value					
	Tobin's q		**DIFF**	**ROA**		**DIFF**
All	1,047		0,0603	-11,014	***	-9,1503
Europe	1,041		0,0593	-6,798	***	-6,4645
North America	1,058		0,1707	-6,941	***	-17,4034
South America	-0,662		-0,1111	-2,803	***	-5,6014
Asia	-1,051		-0,1250	-4,496	***	-6,4406
Japan	1,089		0,6046	-2,334	**	-20,7198
Australia	0,621		0,1146	-0,319		-0,4825
	ROSF		**DIFF**	**PM**		**DIFF**
All	-15,378	***	-18,7134	-9,723	***	-18,5530
Europe	-9,642	***	-16,6848	-5,947	***	-15,5229
North America	-8,402	***	-26,9216	-4,703	***	-25,1561
South America	-4,576	***	-13,3843	-2,849	***	-13,6768
Asia	-4,547	***	-13,1769	-3,818	***	-15,3153
Japan	-3,082	***	-40,3983	-2,862	***	-79,6990
Australia	-0,153		-0,5534	0,005		0,0137
	RDS		**DIFF**	**INV**		**DIFF**
All	2,949	***	0,1961	-5,649	***	-0,0190
Europe	2,741	***	0,3302	-2,247	**	-0,0110
North America	2,341	**	0,4517	-1,414		-0,0230
South America				0,171		0,0020
Asia	0,583		0,0213	-3,025	***	-0,0300
Japan	0,458		0,2117	-1,290		-0,0290
Australia	-0,679		-0,0279	-1,626		-0,0110
	MTB		**DIFF**	**SH-Return**		**DIFF**
All	0,986		0,0556	-1,776	*	-0,0400
Europe	1,960	**	0,1344	-1,220		-0,0320
North America	0,699		0,0990	-0,899		-0,0436
South America	-1,114		-0,1868	-0,344		-0,0269
Asia	-2,375	***	-0,2935	-0,284		-0,0216
Japan	1,441		1,0657	0,418		0,0938
Australia	0,630		0,1246	0,567		0,0489
*	90% significance level					
**	95% significance level					
***	99% significance level					

The t-values from the Table C show that the difference in mean between the GRI and the control group is significant for ROA, PM and ROSF. Australia and New Zeeland seem to be the exception, since the difference between their groups is insignificant for all financial measurements. Japan shows little or no significance in the difference in mean. Both Tobin's q and the closely related market-to-book-value are insignificant. Hence, one can not reject the null hypothesis that the Japanese means of the both groups might be equal given the spread in variability.

With a large sample like this, it is important to remember that a difference, even though it is statistically significant might be small in relation to measured size.[89] Investments are for example calculated as the capital expenditures in relation to total assets and the significant difference of 0,019 percent is in this context economically insignificant.

The Wilcoxon rank-sum test assesses the null hypothesis that two independent samples come from populations with the same distributions. The alternative hypothesis states that one sample is stochastically greater than the other. According to the test results (see Table D) the distribution of the samples for the GRI and the control group are significantly different with a large probability for ROA, ROSF, PM independent of geography. The exception in the first three measurements is Australia that indicates a similar sample distribution for the two groups. Tobin's q is the financial performance measurement that has the smallest differences between the both groups, this could therefore be seen as a further indicator that the Tobin's q mean from the sample does not represent the difference between the two groups if one would measure another sample.

[89] Acock (2002), p. 153

Table D. Wilcoxon rank-sum test, according to region.

Wilcoxon rank-sum test, according to region								
	z-value							
	Tobin's q	p-value	**ROA**	p-value	**ROSF**	p-value	**PM**	p-value
Europe	-0,86	0,38	-7,13	0	-14,02	0	-12,63	0
N.America	-1,85	0,06	-11,67	0	-12,83	0	-7,92	0
S. America	-1,14	0,25	-4,03	0	-6,05	0	-5,69	0
Asia	-2,77	0,00	-5,18	0	-6,64	0	-7,11	0
Japan	1,19	0,23	-2,88	0	-4,44	0	-5,34	0
Australia	-1,04	0,29	0,03	0,97	-0,12	0,89	0,02	0,98
	RDS	p-value	**INV**	p-value	**MTB**	p-value	**SH-Return**	p-value
Europe	6,88	0	-7,11	0	1,68	0,09	-5,17	0
N. America	2,51	0,01	-4,11	0	-1,79	0,07	-3,99	0
S. America			-1,13	0,25	-1,67	0,09	-2,39	0,01
Asia	1,56	0,11	-3,01	0	-2,81	0	-2,69	0
Japan	-0,25	0,79	-1,58	0,11	1,90	0,05	-0,55	0
Australia	-6,91	0	-4,02	0	-0,82	0,40	-0,56	0,57
	*		90% significance level					
	**		95% significance level					
	***		99% significance level					

6. Matching and results

6.1 First method

The first matching method simply assumes that controlling for general conditions, including size of the company, geography, industry and year of incorporation, one can find matches that are assumed similar. To do this, the control group observations with the same mean for size, region, industry and year are matched with the observations from the treatment group with exactly the same properties. The review of the data in chapter 5 has shown that size, geography, industry and age seem to matter when deciding whether to undertake sustainable and social responsible actions. By controlling for variables assumed to affect the choice of taking responsible actions and also report these by GRI one will get almost the same properties as in the propensity score matching with logit probabilities, only with another calculation method for the score. To match on size the variable *total assets* are grouped in to four equally large sample groups. Geography and industry uses the same classifications as previously in chapter 5 and year is the years since incorporation measured as 2007 minus YOI. The difference in financial output, in this case the return on assets, is expressed by the difference in output. Thus, the difference in mean between the ROA for the treated observation and the matched observation is expressed as the difference in mean in Table E below.

Table E. Differences and t-values, $X_{GRI} - X_{Non\text{-}GRI}$, X={q, ROA, ROSF, PM, RDS, INV, MTB, SH-return}

Matching results, Differences in mean.

	Obs	Diff. mean	Std. Err.	Std. Dev.	t-value	
Tobin's q	369	0,0052	0,0406	0,7804	0,128	
ROA	381	1,2977	0,3841	7,4966	3,379	***
ROSF	381	3,4486	1,0264	20,0354	3,360	***
PM	277	1,8640	0,7834	15,2116	2,379	**
RDS	224	-0,0121	0,0041	0,0616	-2,935	***
INV	273	0,0275	0,0063	0,1038	4,372	***
MTB	375	0,0749	0,0518	1,0038	1,446	
SH-Return	369	0,0579	0,0258	0,4959	2,244	**

*** 99% significance level
** 95% significance level
* 90% significance level

The H_0 hypothesis states that the difference in mean between the treatment and the control group is the same. This is tested against the alternative hypothesis that the differences between the two samples would be significantly different.

As one can see, the mean difference in return on assets (*ROA*) is nearly 1,3 percent. This indicates that a company reporting with GRI would yield a 1,3 percentages higher return on their assets compared to if the same company would not report their sustainable actions. The significance level of 99 percent also shows that the findings are significantly different. Other financial indicators can also be observed. For example is the return on shareholder funds (*ROSF*) 3,4 percentages larger for a company reporting with GRI compared to a similar firm if not reporting. The level of investment is higher and it is statistically significant to a 99 percent level, but the difference in mean is still to be considered as very small. Both the difference in Tobin's q and in market-to-book value is insignificant, therefore one can not state that there would be a difference in these measurements depending on the group participation.

6.2 Propensity score matching

When matching with propensity scores estimated by a logit model, one needs to start by specifying the variables to use in the model. Variables used in the model presented below are *ln_at*; a measurement calculated as the natural logarithm of total assets controlling for size and *industries;* 7 dummy variables (*ind_2 - ind_8*) defined according to SIC-codes as in the previous section. Additionally, dummy variables controlling for geography (*Europe, N. America, S. America, Asia, Japan, Australia*) is included in the model to estimate the propensity score.

An additional plausible assumption one can make is that a company searching for new capital from external investors have an incentive to report their complete actions including their social actions taking place. More information available for the investor will reduce the risk to invest, which can attract more funds to the company. Therefore a variable that measures a company's growth *(sgr)* is included in the model.

Table F. Logit estimates

Log estimates	**Number of observations =**	**74883**			
Log likelihood = -2272.0654	**Pseudo R2**	**=**	**0.3921**		

	Coef.	**Std. Err.**	**z**	**P>|x|**	**[95% Conf.**	**Interval]**
ln_at	.9878784	.0262488	37.64	0.000	.9364316	1.039325
sgr	.975194	.1983626	4.92	0.000	.5864105	1.363977
ind_3	.6967087	.2409895	2.89	0.004	.2243779	1.169039
ind_4	.6796337	.1441857	4.71	0.000	.397035	.9622325
ind_5	.3507545	.1608674	2.18	0.029	.0354602	.6660487
ind_6	.0854984	.2436763	0.35	0.726	-.3920983	.5630951
ind_7	1.386899	.4280956	3.24	0.001	.5478468	2.225951
europe	.1317433	.431421	0.31	0.760	-.7138263	.9773129
n_am	2.01599	.4444942	4.54	0.000	1.144798	2.887183
asia	.5956242	.4405838	1.35	0.176	-.2679043	1.459153
jpn	-.645267	.4671406	-1.38	0.167	-1.560846	.2703118
aus	-.0717168	.0362019	-1.98	0.048	-.1426712	-.0007625
_cons	-19.8069	.5888306	-33.64	0.000	-20.96099	-18.65282

ln_at = Size (measured as logarithm of total assets, th USD)
sgr = Growth (measured as share holder growth)
industries 2-8 = Industries (grouped 2-8)
c_groups = Geography (grouped: Europe, N. America, S. America, Asia, Japan, Australia)

The Pseudo R^2 value of the estimation is 0,3921. A higher score is always preferable but with the available information and with the problem and the data at hand, this is a satisfactory result. When calculating the probability values the logit values from Table F has been used and a probability value has been estimated for each and every observation, both in the GRI group and in the control group.

These propensity scores are then separated in to two parts; one for GRI individuals and one for the control groups. Both samples are then ranked by their propensity score and matched by their nearest neighbour.

Table G. Differences in mean between the matched observations and t-values

Variable	Obs	Mean	Std. Dev.	Min	Max
dif_q	487	.0506998	1.555137	-12.27645	8.455044
dif_mtb	536	.0637802	1.765483	-12.97371	8.739748
dif_sgr	636	.0263196	.8702399	-8.31041	4.213105
dif_invest	265	.0174852	.0952492	-.5472271	.4923689
dif_roa	635	7.08178	19.80662	-39.87	178.79
dif_rosf	610	14.54667	41.40029	-263.29	248.55
dif_rds	118	-.4044849	2.283986	-20.53203	.2208392
dif_pm	611	12.7847	52.71457	-107.23	488.15

	t-values
Tobin's q	0,7195
ROA	9,01***
ROSF	8,68***
PM	5,99***
RDS	-1.9238
INV	2,99***
MTB	0,8364
SGR	0,7627

By matching on propensity scores estimated with a logit model one can see that the differences in ROA between the GRI and the control group is 7,1 percent and statistically significant. This means that a company reporting with GRI would on average have a 7 percent higher return on their assets compared to firms that do not report with GRI. Statistically significant is also the profit margin which is nearly 13 percentages higher in the GRI company and the return on shareholder funds indicating 14,6 percentages higher return in the GRI reporting firm. Investment has an indifferently small difference in mean between the two groups with 0,017 percent.

By assessing the balance on the baseline one assures that the observations matched due to similar propensity score also have similar covariates after the matching. The aim is to compare the covariates before and after the matching on the propensity score. If the difference in covariates between the two groups diminishes then one can

state that the match was successful.[90] If differences still remain after matching it can depend on misspecification of the model or incomparability between the two compared groups.

To assess the balance on the baseline in this specific case a new likelihood estimation of the propensity that a company receives treatment is made. The variables used in the new estimation are controlling for *size, industry, geography* and *growth*. To assure that the model is balanced on the base line, i.e. that the covariates are alike, the Pseudo-R^2 value should decrease compared to the likelihood estimation with the unmatched sample. The result in Table H shows that the Pseudo-R^2 was reduced from 0,3921 to 0,1363. This indicates that the covariates X reduced their explanatory power of the probability for GRI usage. This because the covariates differ only minimal between the control group observations and the treatment group observations matched on propensity score, see Table H.

Table H. New Likelihood estimation, including only matched observations

Log estimates	**Number of observations = 1285**					
Log likelihood = -680.11018		**Pseudo R2 = 0.1363**				
	Coef.	**Std. Err.**	**z**	**P>|x|**	**[95% Conf. Interval]**	
ln_at	.6154307	.0393355	15.65	0.000	.5383345	.6925268
sgr	.2594166	.1171425	2.21	0.027	.0298216	.4890117
ind_3	-.3262412	.3170678	-1.03	0.304	-.9476828	.2952003
ind_4	-1.148069	.3368306	-3.41	0.001	-1.808245	-.4878935
ind_5	-.4499898	.2440459	-1.84	0.065	-.9283109	.0283313
ind_6	-.7462438	.2633232	-2.83	0.005	-1.262348	-.2301399
ind_7	-.0576714	.3694602	-0.16	0.876	-.7818001	.6664573
europe	-.5915486	.4260408	-1.39	0.165	-1.426573	.2434761
n_am	-.666083	.4368222	-1.52	0.127	-1.522239	.1900728
asia	.2280172	.4751752	0.48	0.631	-.7033091	1.159343
jpn	.1791539	.4639314	0.39	0.699	-.730135	1.088443
aus	-1.198596	.7996721	-1.50	0.134	-2.765925	.3687322
_cons	-8.37374	.7579795	-11.05	0.000	-9.859353	-6.888128

Ln_at = Size (measured as logarithm of total assets, th USD)
sgr = Growth (measured as share holder growth)
industries 2-8 = Industries (grouped 2-8)
c_groups = Geography (grouped Europe, N. America, S. America, Asia, Japan, Australia)

[90] Caliendo et al. (2008), p. 48

The covariate describing growth *(sgr)* and size (*ln_at*) has quantitative properties and is measurable on an interval scale. The further used covariates describing geography *(c_groups)* and industry (*industries)* consists of nominal scale values without any interpretation of rank order. This makes the assessment of the baseline covariates using standardised differences and t-tests a bit more complicated since any measure of mean, difference in mean and variance is misleading and/or impossible.

By using a t-test one would test the significant level of the difference in means between the two groups after matching. Once again only *size* and *growth* are of a quantitative interval scale and can therefore be tested by the mean. Hence, using the single observations again, the differences in size mean between GRI and non-GRI group can also be tested. The results are presented in Table I.

Table I. T-test for quantitative interval scaled variables.

T-test		**Size**		**Growth**
Before matching				
	Diff. in mean	4.63671		-1,9484
	t-value	60.6196	***	-0,2791
After matching				
	Diff. in mean	2,5136		-0,026
	t-value	22,43	***	-0,76

The outcome of the t-test in Table I show that the difference in mean does decrease after the matching but the differences in mean between the group of GRI and non-GRI are still statistically significant for the case of size. The difference in mean for the variable growth is after the matching unchanged insignificant. These results indicate that the model is not perfectly specified. At the same time the model is based on economic theory backed up with results from previous research. Continuing the search for the perfect model would not only mean excluding variables that are reasonable from a theory point of view but also to include variables that are not plausible or available. A perfect dataset could for example include information about previous sustainable actions taken, management guidelines, stakeholder attitudes

etc. This is difficult to measure and requires a systematic, well developed, global data gathering system for samples of this size.

For the assumption of unconfoundedness to hold there must not exist any external variable that simultaneously affects the financial outcome and the decision to report with GRI. This is a very strong assumption that cannot be validated with certainty. In this specific case, as well as in any case, there might be hidden bias present. Such hidden bias can for example be due to domestic or regional economic policies and regulations. Examples of other possible jointly influencing external variables are many and hard to exclude in a treatment model. With that said, it is hard to believe that the sample at hand would be free from similar external influences. One can still doubt that the deviations from the assumptions are large enough to disaffirm the whole model and its findings.

7. Conclusions

By working on the topic of corporate social responsibility and company performance it has become more and more clear how complex the theme is. Corporate social responsibility is convoluted because of its many definitions and the difficulty of measuring responsible actions. The results of the actions will first and foremost be shown in future value or gains in soft values difficult to observe in a balance sheet. To this, there is no unified and simple way to measure company performance since there are many aspects of valuation and different possibilities to analyse the financial information communicated by the companies.

Measuring a treatment effect using propensity score matching is often extensive. It includes several steps and many aspects to take into consideration. Even if one were to examine something maybe a bit more substantial, like the results on health of a medicine, the implication of assessment method is still not unified for every researcher. Trying to measure an effect of something so intangible and sometimes defuse as the social responsibilities of a company is difficult. It demands even more considerations regarding each of the steps, especially when it comes to forming the model and including variables. It is therefore pleasant to be able to present relatively clear results as the ones presented above. Considering the first estimation of the GRI treatment effect with matching, one can see that there is reason to believe that corporate social performance, here measured in the companies reporting with GRI, are able to show better financial figures, than similar companies that do not report. Five financial measurements (ROA, ROSF, PM, INV, SH-return) are statistically significant and they all indicate that the treated group shows higher average financial figures than the non-GRI companies. Of course the reporting itself is not the largest contributor to the gain, but the fact that there is an assumed correlation between firms reporting their actions with GRI and firms also acting social responsible.

Matching on propensity scores is a very useful but at the same time comprehensive method to assess treatment effects. With the right data at hand, profound theory knowledge and a suitable causality problem design the PSM is ideal. The causality

problem of corporate social responsibility and corporate performance is in many aspects a complex problem. With the results from the testing available one can see the results from the PSM estimation are pointing in the same direction as for the first matching model. The figures measuring financial performance seem to be higher for the GRI companies than for the non-GRI companies. The treatment group does for example show a 7 percent higher return on assets and a 13 percent higher profit margin. The assessment of the balance on the baseline shows that there's reason to believe that the covariates are somewhat similar, hence balanced, between the two propensity score matched samples. Never the less, one can not ignore the fact that basic assumptions are not completely fulfilled. Due to this it is therefore difficult to draw any definite conclusions from the PSM estimation. The results from the first matching still remain enjoyable, and even if PSM could not fully confirm the first findings, it is pleasant to see a somewhat positive linkage between the CSR and company performance. Main future incentives for firms to make greater sustainable efforts partly come from the possibility to gain, not only in soft values, but also in monetary values. The results can therefore be a small tap on the shoulder for those aiming in that direction.

Possible developments of these findings and an opening for future research are to further assess the direction of the causality between CSR and output. A question that remains unanswered is whether companies engaged to CSR do so because they can afford to, or if the sustainable actions itself lead to a higher performance. A classical chicken or the egg problematic. This thesis has assumed the latter, but the exact examination would deserve a research paper on its own. Additionally, one could assess the linkage between the incentive to report sustainable actions and ownership structure. There is reason to believe that state-owned organisations would have a smaller incentive to have an open communication with their stakeholders and hence possibly counteract the positive developments of CSR.

References

Acock, Alan C., *A gentle introduction to Stata,* 2nd edition, Stata Press (Texas, America), 2006.

Augurzky, Boris & Christoph M. Schmidt, *The Propensity Score: A means to an end*, IZA DP, Vol. 271 (March 2001), pp. 1-38.

Austin, Peter C., *A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003*, Statistics in Medicine, Vol. 27 (2008), pp. 2037-2049.

Barber, Brad M. & John D. Lyon, *Detecting abnormal operating performance: The impirical power and specification of test statistic*s, Journal of Financial Economics, Vol. 41 (1996), pp. 359-399.

Becker, Sascha O. & Marco Caliendo, *Sensitivity analysis for average treatment effects*, The Stata Journal, Vol. 1 (2007), pp. 71-83.

Bertrand, Marianne et al., *How much should we trust the difference-in-difference estimates?* The Quarterly Journal of Economics, Vol. 119 (February 2004), pp. 249-275.

Bianchi, Patrizio & Sandrine Labory, *The economic importance of intangible asset*, Ashgate Publishing Limited (Hants, England), 2004.

Brand, Jennie E. & Halaby N. Charles, *Regression and matching estimates of the effects of elite college attendance on education and career achievement*, Social Science Research, Vol. 35 (2006), pp. 749-770.

Caliendo, Marco & Sabine Kopeinig, *Some practical guidance for the implementation of propensity score matching*, Journal of Economic Surveys, Vol. 22(1) (2008), pp. 31-72.

Camejo, Peter et al., *The SRI advantage*, New Society Publishers (Gabriola Islands, Canada), 2002.

Carroll, Archie B. & Ann K. Buchholtz, *Business & society: Ethics and stakeholder management*, South-Western Cengage Learning (Mason, USA), 2008.

Chang, Dong-shang & Li-chin Regina Kuo, *The effects of sustainable development on firms' financial performance- an empirical approach*, Sustainable Development, No.16 (2008), pp. 365-380.

Commission of the European Communities, *Corporate social responsibility: A business contribution to sustainable development,* 2nd of July (2002), COM(2002) 347, (Brussels, Belgium).

Dehejia, Rajeev H. & Sadek Wahba, *Propensity score matching methods for nonexperimental causal studies*, The Review of Economics and Statistics, Vol. 48(1) (February 2002), pp 151-161.

Gravetter, Frederic J. & Lori-Ann B. Forzano, *Research methods for the behavioural science*, 3rd edition, Wadsworth Cengage Learning (Belmont, USA), 2006.

Heckman, James J. et Al., *Matching as an econometric evaluation estimator,* Review of Economic Studies, Vol. 65 (1998), pp. 261-294.

Hedberg, C-J & Fredrik Von Malmborg, *The Global Reporting Initiative and Corporate Sustainability Reporting in Swedish Companies*, Corporate Social Responsibility and Environmental Management, Vol. 10 (2003), pp. 153-164.

Hirano, Keisuke et al., *Efficient estimation of average treatment effects using the estimated propensity score*, Econometrica, Vol. 71(4) (2003), pp. 1161-1189.

Hitt L. & F. X. Frei, *Do better customers utilize electronic distribution channels? The case of PC banking*, Management Science, Vol. 48(6) (June 2002), pp. 732-748.

Holme, Lord, & Richard Watts, *Making good business sense,* The World Business Council for Sustainable Development, Publication 1 (January 2000). http://www.wbcsd.org/web/publications/csr2000.pdf

Ittner, Christopher D. & David F. Larcer, *Are nonfinancial measures leading indicators of financial performance? An analysis of customer satisfaction,* Journal of Accounting, Vol. 36 (1998), pp. 1-35.

Jo, Booil & Elizabeth A. Stuart, *On the use of propensity score in principal casual effect estimation*, Statistics in Medicine, Vol. 28 (2009), pp. 2857-2875.

Jones, Alison S. & David W. Richmond, *Casual effects of alcoholism on earning: estimates from the NLSY*, Health Economics, Vol. 15 (2006), pp. 849-871.

Logan, George W. et al. *What is Corporate Governance?* McGraw-Hill Companies Inc. (New York, USA), 2005.

Meyer, Bruce D., *Natural and quasi-experiments in economics*, Journal of Business and Economic Statistics, Vol. 13(2) (April 1995), pp. 151-161.

Mitchell, Mark L. & Janina M. Jolley, *Research Design Explained*, 7th Edition, Wadsworth (Belmont, USA), 2007.

Potter, Gordon et al., *An empirical investigation of an incentive plan that includes non-financial measurements.* The Accounting Review, Vol. 75(1) (January 2000), pp. 65-92.

Robinson, Thomas R. & Elaine R. Henry, *International financial statement analysis*, CFA Institute, Pearson Custom Publishing (Boston, USA), 2007.

Rosenbaum, Paul R. & Donald B. Rubin, *The central role of the propensity score in observational studies for casual effects*, Biometrica, Vol. 70 (1983), pp. 43-55.

Roy, A., *Some thoughts on the distribution of earnings*, Oxford Economic Papers, Vol. 3(2) (1951), pp. 135-145.

Rubin, D. & N. Thomas, *Matching using estimated propensity scores: Relating theory to practice*, Biometrics, Vol. 52 (1996), pp. 249-264.

Rubin, D., *Estimating causal effects to treatments in randomized and nonrandomized studies.* Journal of Educational Psychology, Vol. 66 (1974), pp. 688-701.

Stevens, S. S., *On the Theory of Scales and Measurements*, Science, Vol. 103(2684) (June 1946), pp. 677-680.

Wonnacott, Thomas H. & Ronald J. Wonnacott, *Introductory statistics for business and economics*, 4th Edition, John Wiley & Sons (New York, USA), 1997.

Wooldridge, Jeffrey M., *Introductory econometrics,* 3rd Edition, Thomson South-Western (Mason, USA), 2006.

Appendix

Table 1, Firms in the sample categorised according to Standard Industrial Classification (SIC)

SIC	Observations								
1	1205	**26**	1960	**44**	1565	**61**	2785	**86**	25
2	460	**27**	2000	**45**	835	**62**	2225	**87**	3875
7	260	**28**	13120	**46**	230	**63**	2715	**89**	135
8	315	**29**	785	**47**	1750	**64**	155	**92**	10
9	255	**30**	2590	**48**	5810	**65**	6300	**94**	25
10	5405	**31**	670	**48**	5810	**67**	11055	**95**	205
12	565	**32**	3840	**49**	4110	**70**	1905	**96**	35
13	4730	**33**	4200	**50**	5390	**72**	360	**97**	75
14	1295	**34**	2915	**51**	2685	**73**	16530	**99**	50
15	2235	**35**	10015	**52**	185	**75**	415	**Total**	**200**
16	1345	**36**	14970	**53**	1005	**76**	70		
17	1175	**37**	4010	**54**	830	**78**	850		
20	7045	**38**	4475	**55**	600	**79**	1810		
21	225	**39**	2195	**56**	790	**80**	1895		
22	4160	**40**	295	**57**	780	**81**	35		
23	1895	**41**	215	**58**	1470	**82**	665		
24	870	**42**	860	**59**	2060	**83**	95		
25	785	**43**	10	**60**	12650	**84**	30		

AGRICULTURE, FORESTRY, AND FISHING
01 - - AGRICULTURAL PRODUCTION - CROPS
02 - - AGRICULTURAL PRODUCTION - LIVESTOCK
07 - - AGRICULTURAL SERVICES
08 - - FORESTRY
09 - - FISHING, HUNTING, AND TRAPPING
MINING
10 - - METAL MINING
12 - - COAL MINING
13 - - OIL AND GAS EXTRACTION
14 - - NONMETALLIC MINERALS, EXCEPT FUELS
CONSTRUCTION
15 - - GENERAL BUILDLING CONTRACTORS
16 - - HEAVY CONSTRUCTION, EXCEPT BUILDING
17 - - SPECIAL TRADE CONTRACTORS
MANUFACTURING
20 - - FOOD AND KINDRED PRODUCTS
21 - - TOBACCO PRODUCTS
22 - - TEXTILE MILL PRODUCTS
23 - - APPAREL AND OTHER TEXTILE PRODUCTS
24 - - LUMBER AND WOOD PRODUCTS
25 - - FURNITURE AND FIXTURES
26 - - PAPER AND ALLIED PRODUCTS
27 - - PRINTING AND PUBLISHING
28 - - CHEMICALS AND ALLIED PRODUCTS
29 - - PETROLEUM AND COAL PRODUCTS
30 - - RUBBER AND MISC. PLASTICS PRODUCTS
31 - - LEATHER AND LEATHER PRODUCTS
32 - - STONE, CLAY, AND GLASS PRODUCTS
33 - - PRIMARY METAL INDUSTRIES
34 - - FABRICATED METAL PRODUCTS
35 - - INDUSTRIAL MACHINERY AND EQUIPMENT
36 - - ELECTRONIC & OTHER ELECTRIC EQUIPMENT
37 - - TRANSPORTATION EQUIPMENT
38 - - INSTRUMENTS AND RELATED PRODUCTS
39 - - MISC. MANUFACTURING INDUSTRIES

TRANSPORTATION, COMMUNICATIONS, ELECTRIC AND GAS
40 - - RAILROAD TRANSPORTATION
41 - - LOCAL AND INTERURBAN PASSENGER
42 - - TRUCKING AND WAREHOUSING
43 - - U.S. POSTAL SERVICE
44 - - WATER TRANSPORTATION
45 - - TRANSPORTATION BY AIR
46 - - PIPELINES, EXCEPT NATURAL GAS
47 - - TRANSPORTATION SERVICES
48 - - COMMUNICATION
49 - - ELECTRIC, GAS, AND SANITARY
WHOLESALE TRADE
50 - - WHOLESALE TRADE - DURABLE
51 - - WHOLESALE TRADE - NONDURABLE
RETAIL TRADE
52 - - EATING AND DRINKING PLACES
53 - - GENERAL MERCHANDISE STORES
54 - - FOOD STORES
55 - - AUTOMOTIVE DEALERS & SERVICE
56 - - APPAREL AND ACCESSORY STORES
57 - - FURNITURE AND HOMEFURNISHINGS
58 - - EATING AND DRINKING PLACES
59 - - MISCELLANEOUS RETAIL**FINANCE, INSURANCE, AND REAL ESTATE**
60 - - DEPOSITORY INSTITUTIONS
61 - - NONDEPOSITORY INSTITUTIONS
62 - - SECURITY AND COMMODITY
63 - - INSURANCE CARRIERS
64 - - INSURANCE AGENTS, BROKERS
65 - - REAL ESTATE
67 - - HOLDING AND OTHER INVESTMENT

SERVICES
70 - - HOTELS AND OTHER LODGING PLACES
72 - - PERSONAL SERVICES
73 - - BUSINESS SERVICES
75 - - AUTO REPAIR, SERVICES, AND PARKING
76 - - MISCELLANEOUS REPAIR SERVICES
78 - - MOTION PICTURES
79 - - AMUSEMENT & RECREATION SERVICES
80 - - HEALTH SERVICES
81 - - LEGAL SERVICES
82 - - EDUCATIONAL SERVICES
83 - - SOCIAL SERVICES
84 - - MUSEUMS, BOTANICAL, ZOOLOGICAL
86 - - MEMBERSHIP ORGANIZATIONS
87 - - ENGINEERING & MGMT SERVICES
88 - - PRIVATE HOUSEHOLDS
89 - - SERVICES, (OTHERS)
PUBLIC ADMINISTRATION
91 - - EXECUTIVE, LEGISLATIVE, AND GENERAL
92 - - JUSTICE, PUBLIC ORDER, AND SAFETY
93 - - FINANCE, TAX, & MONETARY POLICY
94 - - ADMIN. OF HUMAN RESOURCES
95 - - ENVIRONMENTAL QUALITY & HOUSING
96 - - ADMIN. OF ECONMIC PROGRAMS
97 - - NAT'L SECURITY, INT. AFFAIRS
NONCLASSIFIABLE ESTABLISHMENTS
99 - - NONCLASSIFIABLE ESTABLISHMENTS

Source:
US securities and Exchange Commission,
http://www.sec.gov/info/edgar/siccodes.htm
(update: 2010.01.13)

Table 2, Industrial classification for total sample, absolute numbers

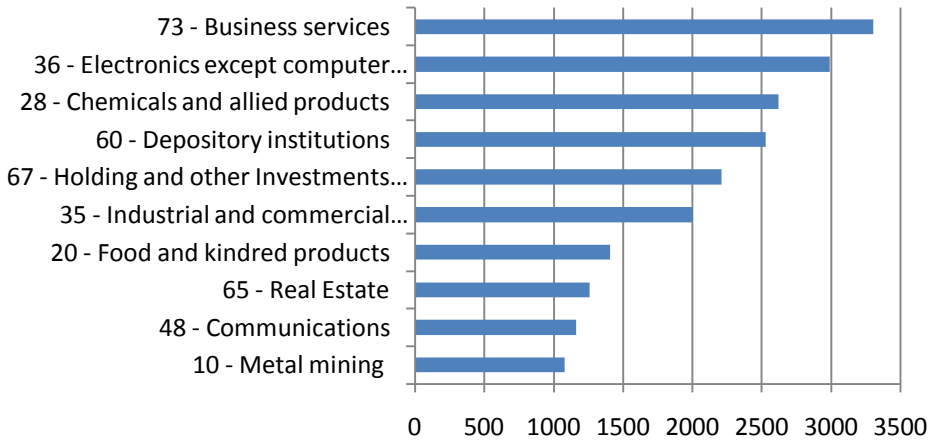


Table 3. Descriptive statistics for raw data (original) , cleaned data (All) , GRI group and control group

GRI vs. Control							
		Obs.	Mean	Med	Std.dev	Min	Max
roa	Original	149975	-5.723	3.52	62.130	-998.19	964.29
	Cleaned	138544	-.258	3,67	24,667	-270,45	42,1
	GRI	1015	7.982	6.28	8.087	-36.41	41.79
	Control	137529	-.3189	3.65	24.738	-270.45	42.1
q	Original	84295	11.13	1.1666	753.4844	0	175277.5
	Cleaned	77791	1,580	1,1503	1,598	0,3455	31
	GRI	826	1.519	1.2439	.992	.3673	9.2376
	Control	76965	1.581	1.1491	1.603	.3455	31
mtb	Original	101597	8.67	.7498	685.222	0	175272.7
	Cleaned	93344	1,18	0,7317	1,588	0	21,4115
	GRI	842	1.123	.7776	1.174	.0005	9.0464
	Control	92502	1.18	.7311	1.591	0	21.4115
rosf	Original	142202	1.362	10.32	72.006	-995.92	991.87
	Cleaned	132921	4,979	10,59	36,640	-299,18	100,55
	GRI	1009	22.32	21.95	21.255	-251.78	99.29
	Control	131912	4.847	10.5	36.701	-299.18	100.55
inv	Original	81017	.0727	.0302	1.531	6.68e-07	426.5735
	Cleaned	75877	.056	0,0309	0,075	0,0001	0,5701
	GRI	501	.075	.0519	.078	.0009	.5574
	Control	75376	.056	.0307	.075	.0001	.5701

		Obs.	Mean	Med.	Std.dev.	Min	Max
rds	Original	43980	2.94	.0156	136.062	-303.4483	25684.4
	Cleaned	40745	.225	.0142	1,312	0	22,9861
	GRI	416	.039	.0216	.049	0	.2234
	Control	40329	.227	.0141	1.319	0	22.986
pm	Original	138694	-4.145	6.21	109.325	-998.96	998.69
	Cleaned	129408	-.052	6,25	57,610	-536,47	189,79
	GRI	1006	16.63	12.745	16.501	-50.58	171.53
	Control	128402	-.182	6.2	57.798	-536.47	189.79
sh_	Original	118683	.572	.1136	20.993	-1	4192.867
	Cleaned	110585	.267	.1193	0,695	-0,8870	4,5
	GRI	1006	.308	.2353	.420	-.5787	3.1720
	Control	109579	.267	.1176	.697	-.8870	4.5

Table 4. Number of firms reporting according to GRI sorted by Industry and Region.

	Industry 1						**Industry 2**					
Region	1	2	3	4	5	6	1	2	3	4	5	6
Non-GRI	2335	5965	620	3425	331	1060	10840	15780	2170	31620	2000	8720
GRI	65	45	5	10	5		210	110	25	60		40
	Industry 3						**Industry 4**					
Region	1	2	3	4	5	6	1	2	3	4	5	6
Non-GRI	2795	3755	1015	3290	795	1125	2135	3755	385	3195	440	3555
GRI	125	20	55	25	10		35	5	15			5
	Industry 5						**Industry 6**					
Region	1	2	3	4	5	6	1	2	3	4	5	6
Non-GRI	7495	11745	1620	5665	1245	1805	6100	9420	445	5405	1415	3190
GRI	95	15	25	20	10		15	15		10		

Industry 1: Metal, Mining, Agricultural
Industry 2: Construction, Manufacturing
Industry 3: Communication, Transport Electric, Gas & Whole Sale, Retail
Industry 4: Financial Services, Insurance, Real Estate
Industry 5: Services
Industry 6: Health Services
Region: 1 Europe, 2: N America, 3: S America, 4: Asia, 5: Japan, 6: Australia

A large part of the Asian companies not reporting with GRI are in the 2nd industry. The same industry also has the highest relative amount of Asian GRI reporters. The Japanese companies have the largest part of the firms operating in the 1st industry with two firms each reporting with GRI in industry sector 3 & 5. The European firms are present in the 2, 4 & 6th industry and the companies most engaged to GRI are to find in 2nd industry group.

Table 5. The median and mean for GRI/control group according to region. 1=GRI, 0= Non-GRI

Mean/Median							
	GRI	**Europe**		**N. America**		**S. America**	
		Mean	Median				
q	0	1,50	1,20	2,03	1,38	1,29	0,97
	1	1,44	1,21	1,86	1,47	1,41	1,02
MTB	0	1,15	0,78	1,53	0,95	0,92	0,56
	1	1,02	0,70	1,43	1,05	1,10	0,67
rosf	0	5,75	11,89	-3,66	8,31	10,97	11,63
	1	22,43	21,34	23,26	23,34	24,36	21,06
roa	**0**	**0,72**	**4,04**	**-8,28**	**1,50**	**2,79**	**4,01**
	1	**7,18**	**5,82**	**9,12**	**8,76**	**8,39**	**5,93**
pm	0	-0,02	5,57	-10,58	5,75	6,94	8,63
	1	15,50	11,91	14,58	13,69	20,62	20,10
invest	0	0,06	0,03	0,07	0,03	0,09	0,05
	1	0,07	0,05	0,09	0,07	0,09	0,06
rds	0	1,06	0,04	5,39	0,07	0,87	0,02
	1	0,15	0,02	0,14	0,04	0,22	0,02
sh_return	0	0,27	0,15	0,19	0,13	0,46	0,17
	1	0,30	0,14	0,24	0,13	0,48	0,21
		Asia		**Aus/Nzl**		**JPN**	
		Mean	Median				
q	0	1,34	1,06	1,84	1,34	1,22	0,96
	1	1,47	1,20	1,23	1,13	1,11	0,99
MTB	0	0,93	0,61	1,89	1,18	0,78	0,47
	1	1,22	0,80	0,82	0,75	0,66	0,54
rosf	0	7,96	10,55	-12,20	-1,39	10,00	10,49
	1	21,13	21,47	28,20	23,61	10,56	9,43
roa	**0**	**3,69**	**4,49**	**-13,52**	**-1,74**	**4,50**	**4,21**
	1	**10,13**	**8,84**	**7,20**	**4,19**	**4,98**	**4,18**
pm	0	3,03	6,29	-31,93	4,29	5,48	4,51
	1	18,35	16,42	47,77	43,18	5,46	4,40
invest	0	0,06	0,04	0,04	0,02	0,04	0,03
	1	0,09	0,07	0,07	0,06	0,05	0,04
rds	0	1,27	0,01	2,49	0,00	0,11	0,01
	1	0,25	0,01	2,29	0,00	0,01	0,05
sh_return	0	0,35	0,16	0,36	0,22	0,19	0,02
	1	0,37	0,21	0,27	0,36	0,14	-0,01

Table 6. Descriptive statistics, financial figures, GRI (1) and Non-Gri (0)

Table B.						
Variable		**Obs**	**Mean**	**Std. Dev.**	**Min**	**Max**
yoi	0	145475	1975,53	32,35948	1399	2007
	1	756	1944.3	52,87365	1765	2007
sale	0	131461	963666,7	6417199	0	4E+08
	1	870	20200000	34000000	0	2E+08
emp	0	95667	4381,126	23749,54	0	2E+06
	1	699	58335	78512	81	490042

Yoi = Year of incorporation
sale = Total sales, th USD
emp = Number of employees

Table 7. ROA, ROSF and Total Assets according to industry

SIC	ROA	ROSF	AT
1-9	**-.7340563**	3.054403	197953.8
10-14	**-21.28311**	-17.07591	1451604
15-17	**2.483612**	8.786737	1278747
20-39	**-4.567239**	.5372111	1129823
40-49	**-3.728406**	4.52511	3602796
50-51	**-1.946766**	6.336487	727510.2
52-59	**1.595313**	11.44935	1417800
60-67	**2.064477**	11.95581	1006244
70-89	**-17.4675**	-8.927873	420428.7

ROA = Return on Assets, **ROSF** = Return on Shareholder Funds, **AT** = Total Assets
SIC 1-9 = Agriculture, Forstry, Fishing
SIC 10-14 = Mining
SIC 15-17 = Construction
SIC 20-39 = Manufacturing
SIC 40-49 = Transportation, Communications, Electric, Gas
SIC 50-51 = Wholesale Trade
SIC 52-59 = Retail Trade
SIC 60-67 = Finance, Insurance, Real Estate
SIC 70-89 = Services

Anhang

Abstract

The concept of treatment effects refers to the difference in output that occur when one group receives a certain treatment and another one does not, given that the observed groups are identical in the pre-treatment state. This thesis assesses various treatment effects and describes them in detail in the theory chapter 2. The purpose of the thesis is to apply the best suitable treatment effect method when assessing the effect of corporate social responsibility on corporate financial outcome. In order to succeed with a treatment effect estimation one needs a proper, consistent and comparable measurement of social corporate responsibility and financial output. This is found through the Global Reporting Initiative (GRI) assumed to be positively correlated with sustainable actions taking place. The measurements used for financial outcome are different backward- and forward looking financial measurements such as Return on Assets, Profit Margin and Tobin's q. The results of the empirical part show that there is a statistically significant difference in output between companies who do report with GRI and firms that do not. By using a simple matching model controlling for size, geography, industry and age one can see that the positive difference in ROA is 1,3 percent and for ROSF 3,4 percent. When applying the method of propensity score matching one can see that the positive difference in mean still remains. Although, one need to keep in mind that the assumptions made in the model are strong and that the balance on the baseline assuring the comparability between the two groups can not be completely confirmed. Still, the results all indicate in direction favoring higher financial outcome for GRI firms.

Zusammenfassung

Das Konzept des „Treatment effect“ bezieht sich auf die Differenz der Outputs zweier Gruppen, welcher resultiert, wenn eine Gruppe eine spezielle Behandlung erhält und die andere nicht, unter der Annahme, dass sich beide Gruppen in einem identischen Zustand vor der Behandlung befanden.

Die Zielsetzung dieser Arbeit ist es die optimale „Treatment effect“-Methode anzuwenden, für die Untersuchung des Effektes von Coroporate Social Responsibility auf finanzielle Ergebnisse eines Unternehmens. Um eine erfolgreiche Einschätzung des „Treatment effect“ zu erzielen, ist es notwendig passende, konsistente und vergleichbare Messgrößen von Corporate Social Responsibility sowie finanziellen Outputs zu generieren.

Diesen Anforderungen entspricht die „Global Reporting Initiative“ (GRI). GRI unterliegt der Annahme, dass sie positiv mit nachhaltigen Maßnahmen korreliert ist. Verwendete Messgrößen hinsichtlich des finanziellen Erfolgs unterteilen sich in u.a. Return on Assets, Profit Margin und Tobin’s q.

Die Ergebnisse des empirischen Teils deuten darauf hin, dass es statistisch signifikante Unterschiede zwischen den finanziellen Outputs von Unternehmen mit und ohne GRI gibt. Bei Anwendung eines einfachen Matching-Modells, unter Berücksichtigung von Größe, Geographie, Industrie und Alter, zeigen die Ergebnisse eine positive Differenz bei ROA von 1,3 Prozent und bei ROSF von 3,4 Prozent. Wenn die „Method of propensity score matching“ herangezogen wird, stellt sich ebenfalls heraus, dass die positive Differenz bestehen bleibt. Wenngleich man berücksichtigen muss, dass die dem Modell unterliegenden Annahmen sehr stark sind und die Balance der Baseline, welche die Vergleichbarkeit zwischen den Gruppen gewährleistet, nicht gänzlich bestätigt werden kann. Nichts desto trotz weisen die Ergebnisse darauf hin, dass GRI Unternehmen tendenziell höhere finanzielle Ergebnisse erzielen.

CURRICULUM VITAE

Name:	IDA MARGARETA NILSSON
Date of Birth:	31st of JULY 1982
Nationality:	SWEDEN

EDUCATION:

Date (start – end):	2007 – 2010
Type of education:	Master in Economics, University of Vienna, Vienna
Date (start – end):	2005 – 2007
Type of education:	Bachelor of Science in Business Administration and Economics, Umeå University, Umeå, Major in Economics
Date (start – end):	2004 – 2005
Type of education:	Business Administration, University of Zürich, Zürich
Date (start – end):	2002 – 2003
Type of education:	Linguistics, Leopold-Franzens-University Innsbruck, Innsbruck
Date (start – end):	1998 – 2001
Type of education:	Upper secondary school, Social Science program, Östra Gymnasieskolan, Umeå
Date (start –end):	1989 – 1998
Type of education:	Nine-year compulsory school, Umeå

WORK EXPERIENCE:

Date (start – end):	2008.06 – 2008.08
Employer:	Citigroup, Global Corporate Banking, Citi Stockholm
Position or function:	Internship, Corporate Banking
Date (start – end):	2007.06 – 2002.08 2006.06 – 2006.08 2004.06 – 2004.08 2003.06 – 2003.08 2002.06 – 2002.08
Employer:	Stadsledningskontoret, Umeå
Position or function:	Financial Assistant & Accounting Assistant
Date (start – end):	2002 – 2009
Employer:	Skischule Lech, Lech am Arlberg, Austria
Position or function:	Ski Instructor

LANGUAGE KNOWLEDGE:

German:	Fluent
English:	Fluent
Swedish:	Mother tongue