Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Business Analytics and Quantitative Marketing

# Talent Management and Firms' Sales: Seeking Causality with a Debiased Machine Learning Approach

Arianna Gambardella (647117)

| | |
|---|---|
| Supervisor: | Andrea A. Naghi |
| Second assessor: | Eoghan O'Neill |
| Date final version: | 27th June 2023 |

**Abstract**

This research investigates the existence of a causal effect of the quality of talent management practices on sales performance. Insights into talent management practices are of particular interest to managers who want to attract and develop talented employees in order to improve the performance of their firm. Using a diverse data set from 20 countries, I employ a Debiased Machine Learning (DML) approach to ensure unbiased estimates. I address multicollinearity issues through the pre-processing of the data. That is, variables that are highly correlated with each other are not included all together. Since these are variables that measure similar concepts, the exclusion of one of them should solve the multicollinearity issue without causing any loss of information. Subsequently, I use Post-Double Lasso Selection to determine optimal control variables.

Various machine learning methods are utilized, with Identification Strategy 3 showing statistically significant results using Random Forest. These suggest that a firm that is fully engaged in talent management and talent development exhibits sales 6-7 % higher than a company that does not invest in talent management practices at all.

The results indicate that firms should evaluate the return on investment in talent management practices by carefully considering their specific objectives. For example, better talent management practices may contribute to a better work environment and, based on the results from this thesis, there may be a pecuniary return if the implementation costs are not excessive. However, there exist more efficient, less expensive to implement, renowned strategies for a firm to increase sales, especially in case of necessity rather than ambitious views. In order to assess the dynamic effect of talent management practices on sales over time, future research may consider employing a longitudinal analysis. This thesis contributes to the existing body of literature on talent management practices and on DML. The findings are highly relevant for companies looking to make informed decisions regarding talent development strategies and resource allocation in such practices.

# Contents

# 1 Introduction

The increasing awareness surrounding talent management practices has led firms to seek top talent in order to be more successful (Morley, Scullion, Collings & Schuler, 2015). Many companies strive for improving talent management practices as it is generally believed that better talent management increases sales (Sparrow, Scullion & Tarique, 2014). However, it is particularly interesting to study the strength of this relationship when removing the bias that arises when simply fitting the model through a regression learner. The aim of this thesis is to analyze how talent management affects sales through Debiased Machine Learning (DML) on a diverse data set of companies across 20 countries. In this way, this research assesses whether there is a causal effect of the quality of talent management practices on sales performance. Although research studying this relationship exists, these do not make use of DML, which may indicate that their results are biased. Therefore, in this thesis, the DML approach ensures that the estimated coefficient is unbiased (Chernozhukov, Newey & Singh, 2022).

When estimating a regression through machine learning methods, one often encounters bias from regularization or model selection. Therefore, it seems necessary to seek a way to correct these biases. DML is a technique that aims at correcting the bias that arises from causal machine learning methods applied to high dimensional data. DML is flexible as it can be implemented to any regression learner (Chernozhukov et al., 2022). Therefore, this research aspires to find whether talent management practices have an impact on sales by applying DML to a large cross-sectional data set. This data set includes, among other variables, firms' talent management practices (those practices that a firm undertakes to retain and attract top talent employees) and sales data.

Therefore, the main purpose of this research is to detect the presence and the strength of a causal relationship between talent management practices and sales. If a strong causal relationship exists, firms should consider investing more in talent management in order to benefit from it through increased sales. However, if, after removing bias, this relationship becomes less credible, either because it is statistically insignificant or the effect size is small, companies might decide not to use talented employees as a competitive strategy. It is therefore important to assess whether the effect size is large enough to require action by firms.

Moreover, I repeat the analysis with two different instrumental variables. Firstly, I use whether a business is family-owned or not as an instrument for talent management. Although family businesses may be directly correlated with sales (e.g., through the size of the firm) (Gallo, Tàpies & Cappuyns, 2004), this instrument should satisfy the exclusion restriction because I will include control variables for sales performance (e.g., the size of the firm). However, an instrumental variable that is more likely to satisfy the exclusion restriction would be the number of colleges in the city where the headquarters of the firm is located. That is, if a firm is located in an area with many cultural and educational institutions, it is likely to attract more talent and, in turn, invest more in talent management practices. I provide an adaptation of the Hansen-Sargan test in order to further support the hypothesis that the instruments satisfy the necessary assumptions. Although a statistically significant result of the Hansen-Sargan test would indicate a violation of the assumptions, a statistically insignificant result does not ensure a correct specification of the model (Kiviet & Kripfganz, 2021). However, it can be

used as evidence supporting the unlikelihood of model misspecification. Finally, I include both variables in the vector of instruments in order to exploit over-identification. Although the DML technique should correct for bias in any of the three cases (Chernozhukov et al., 2022), I want to test whether a more valid instrument provides different results, even when using DML, or whether the final output is indeed equal across the two regressions. Similarly, I would like to test whether the machine learning method that is used impacts the final results when applying DML to it. Therefore, I primarily use Lasso and, subsequently, I will repeat the analysis using Gradient Boosting, Support Vector Machines, and Random Forests. The findings from this approach should provide insights on whether the results obtained from DML depend on the choice of the machine learning method used or on the choice of the instrumental variable.

Lastly, I assess the added value of using DML by comparing the results from this thesis to those of studies that investigate the relationship between talent management practices and sales performance through a traditional approach. In addition, I include a traditional instrumental variable approach as well as Ordinary Least Squares (OLS) applied to my data for comparative purposes. In case of discrepancies across the DML results and those obtained through traditional methods, companies that have relied upon findings of the latter may consider modifying their strategy accordingly. This investigation answers whether the results from this research suggest an alternative relationship between talent management practices and sales compared to previous research that does not use DML. Moreover, this thesis provides implications of such findings from the point of view of a company.

This study is highly relevant as it aims to answer the question of whether investing in talent management practices increases sales, which may be interesting for companies (Morley et al., 2015). This is especially true if the findings from previous studies that use traditional methods reveal to be subject to bias (Chernozhukov et al., 2022). In this case, companies would have to reconsider their internal organization around talent management. Overall, this research will aid firms in making informed decisions about whether and how much to invest in talent management practices to improve sales performance. Moreover, the use of DML in order to improve the accuracy of the results (Chernozhukov et al., 2022) might reveal important insights regarding machine learning theory, especially when comparing the results using different instruments and machine learning methods.

Firms looking to develop talented employees in order to increase sales generally have high expenditures on talent management practices. For this reason, understanding whether these practices effectively increase sales is highly important. This research is particularly relevant as its objective is to estimate a causal relationship, and not a correlation. Moreover, causal estimates from previous research may be biased (Chernozhukov et al., 2022) and potentially harm companies that follow advice based on the results of such studies. Hence, the practical relevance of this study that will provide accurate and unbiased estimates through the use of DML. Therefore, companies can consider these results reliable in order to have an idea on if and how much to invest in talent management practices. If a strong positive causal relationship is found, firms may decide to invest more in such practices. However, if the relationship turns out to be weak or statistically insignificant, companies might be better off allocating these funds on other competitive strategies that aim at increasing sales.

Secondly, DML is an innovative approach (Chernozhukov et al., 2022) and this thesis will provide insights on how powerful this method is in the instrumental variable setting. Therefore, this research may add onto the existing literature on machine learning and causal inference. Moreover, this study will further demonstrate how the use of traditional methods may sometimes be inapt to estimate causality when dealing with high-dimensional data. Therefore, this thesis will be of scientific relevance for researchers that wish to use machine learning for causal inference instead of traditional techniques.

This research is necessary in estimating whether and to what extent there exists a causal relationship between talent management practices and sales performance. Therefore, firms looking for a strong reason to invest in talent management practices may use this research in order to base their decision on whether to invest and how much to invest. The existing knowledge on this topic is not sufficient as it is mostly based on traditional regression analysis. Existing studies are likely to be biased due to a naive estimation of the causal effect (Chernozhukov et al., 2018) due to a variety of endogeneity concerns. For example, model selection may induce bias if the wrong control variables are included. Standard Lasso may lead to bias since it does not take into account correlation between the control variables and the treatment variable (Chernozhukov et al., 2018). In order to address these issues, this research uses DML, a technique that aims to obtain an unbiased estimate of the causal effect. Moreover, this study performs an accurate model selection in order to choose the best control variables from a large pool of options. By comparing this estimate to results of previous studies, it is possible to infer whether past research has overestimated or underestimated the effect of talent management on financial performance. In this way, firms will be able to weigh their expenditures on talent management practices accordingly.

## 2 Literature Review

Chernozhukov et al. (2022) and Chernozhukov et al. (2018) provide an overview of Double Machine Learning and of DML. In particular, Chernozhukov et al. (2022) address the general framework of DML and outline the two-step procedure to correct for bias that arises from endogeneity when estimating causal effects. The DML technique can be applied to many econometric models that aim to find causal and structural effects. Endogeneity occurs when the explanatory variable is correlated with the error term, leading to bias in the estimate of the causal effect. The DML approach first constructs the conditional expectation function of the dependent variable given the covariates through a machine learning method of choice. The causal parameter of interest is the expectation of a function of the conditional expectation that depends on an observation of the data and is linear in the conditional expectation. More precisely, denote the conditional expectation by

$$\gamma_0(x) = E[Y|X = x],$$

where $Y$ is the dependent variable and $X$ is a set of covariates. Then, Chernozhukov et al. (2022) rely on a regression learner estimator of $\gamma_0$, say $\hat{\gamma}$. Subsequently, denote the function $m(w, \gamma(x))$ as a function of $\gamma(x)$ (a functional of $\gamma_0(x)$), that is linear in $\gamma(x)$ and that depends on an observation of the data $w$. The causal estimand is then

$$\theta_0 = E[m(W, \gamma(x))],$$

where $W$ is an observation and $\gamma(x)$ is a regression. Chernozhukov et al. (2022) rely on a regression learner estimand of $\gamma_0(x)$, say $\hat{\gamma}(x)$, which is appropriately adjusted in order to satisfy specific conditions (e.g., orthogonality and moment condition). Moreover, the authors provide results on the validity of the technique as well as theoretical proof that the resulting causal estimated coefficient has been corrected from bias that arises from endogeneity concerns. Chernozhukov et al. (2022) use the DML framework to obtain an unbiased estimand of the average treatment effect on the treated for the National Supported Work Demonstration job training data, a job training program for underprivileged workers that were employed around 1970. They also apply this technique to estimate unbiased demand elasticities from Nielsen scanner data while allowing for individual preferences that are correlated with prices and total expenditure. Moreover, the authors compare different machine learning techniques and they provide guidance in the appendix for tuning the hyperparameters. It is clear that the DML technique is highly innovative and works well in evaluating causal effects while eliminating sources of endogeneity. The DML framework can be applied to different settings, including instrumental variable regressions, in order to obtain accurate and reliable causal effect estimates.

Chernozhukov et al. (2018) describe applications of DML in the instrumental variable context. This is a useful guide to the approach developed in this research. Consider the instrumental variable model

$$Y = D\theta_0 + g_0(X) + U, \quad E_P[U|X, Z] = 0,$$

$$Z = m_0(X) + V, \quad E_P[V|X] = 0,$$

where $Z$ denotes the instrument and the true value of the causal parameter is $\theta_0$. In order to estimate $\theta_0$, the authors use the Robinson-style score:

$$\psi(W; \theta, \eta) = (Y - l(X) - \theta(D - r(x)))(Z - m(X)), \quad \eta = (l, m, r),$$

where $W = (Y, D, X, Z)$ and $l, m, r$ are P-square integrable functions. This score function satisfies both the moment and the orthogonality condition, and it is therefore adequate for the analysis. Its advantage is that all of the nuisance parameters are conditional mean functions that can be directly estimated through machine learning methods. The authors also show how regularity conditions are satisfied in this case as well as asymptotic validity of the resulting estimand. The final estimand of $\theta_0$ is then obtained by averaging across the parameters of $\theta_0$ that are estimated through the Robinson-style score from cross-fitting samples. Cross-fitting is a data-splitting procedure where the estimated coefficient is constructed based on a partition of the data. The final estimand is then obtained by repeating this K times and averaging over the K estimands obtained. Chernozhukov et al. (2018) compare the results of the DML technique applied to their example across a variety of machine learning methods and hyperparameters. Furthermore, they use both two-fold and five-fold cross fitting.

The presence of many control variables may induce bias in the estimation of the causal effect, may lead to inflated standard errors, and might cause over-fitting issues. This is due to a higher chance of including bad control variables, which violate the unconfoundedness assumption behind DML. Hünermund, Louw and Caspi (2021) depict the situation in Figure 1.

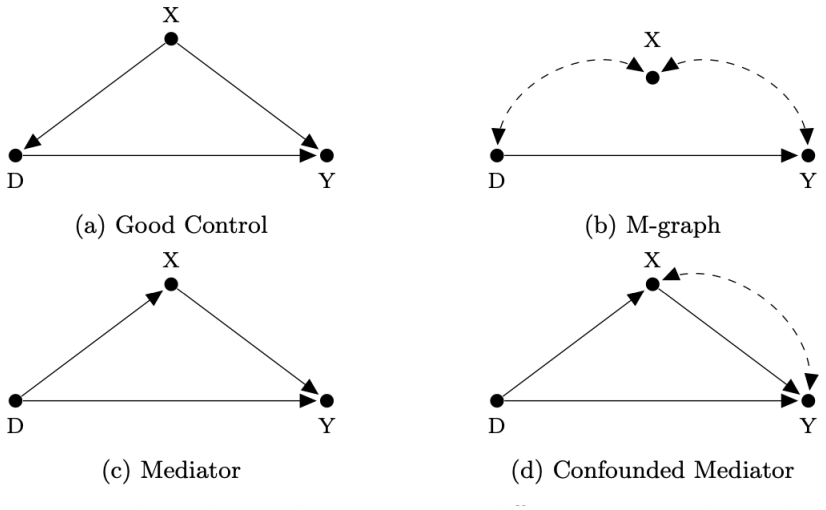


Figure 1: Directed acyclic graphs representing different structural causal models (Hünermund et al., 2021)

Hünermund et al. (2021) show that, if only good control variables are included, DML is able to obtain an unbiased estimate of the causal effect and to perform much better than Lasso. This is the situation that Hünermund et al. (2021) shows in the first row of Figure 2. However, if unconfoundedness does not hold and not all control variables are (conditionally) exogenous, DML may perform even worse than Lasso. This would suggest that one should include more controls

such that conditional exogeneity is likely to hold. However, if the covariate space is large, the probability of finding and including bad controls increases, especially when using purely data-driven techniques. Only a few bad control variables can produce this outcome because they are correlated with the treatment and/or the dependent variable just as good controls. Therefore, they are also likely to be selected by DML (Hünermund et al., 2021).

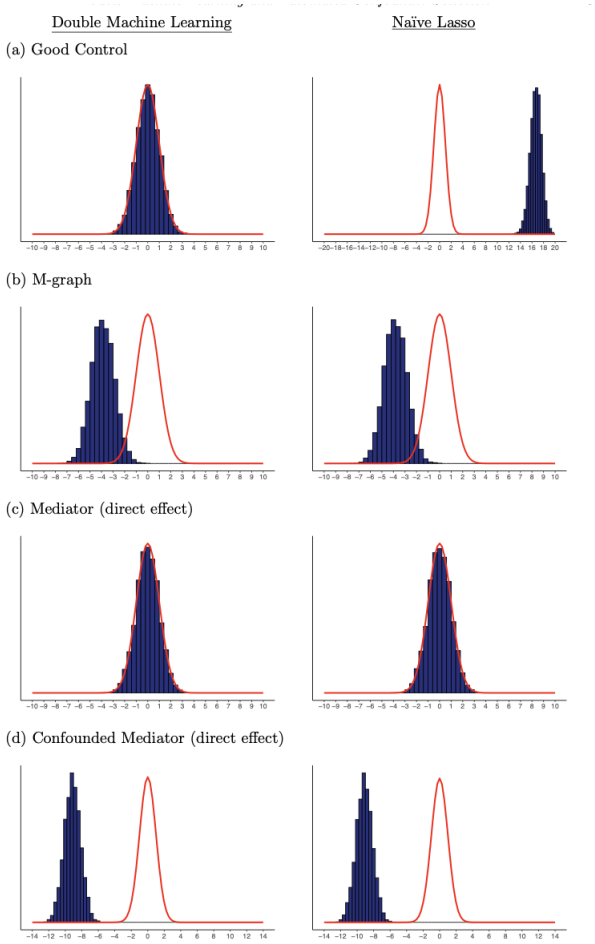


Figure 2: Performance of DML compared to naive LASSO for different causal models (Hünermund et al., 2021)

In order to obtain a reliable estimate of the causal effect when working with high-dimensions, it is important to find an appropriate way to exclude some of them without loosing too much information. Belloni, Chernozhukov and Hansen (2014) deal with control variable selection when considering a high number of covariates. They propose a procedure called Post-Double Lasso Selection in order to select the best control variables. Namely, their framework works with the regression

$$Y = D\theta_0 + X\alpha + U,$$

$$D = X\beta + V,$$

First, the approach of Belloni et al. (2014) selects a set of control variables for the treatment $D$. In this way, the authors find important potential confounding factors for the treatment $D$

that are strongly related to it since they are able to predict well the treatment variable. It is crucial to highlight that these are potential confounders as they affect both $D$ and $Y$, so they are not necessarily confounders. Secondly, Belloni et al. (2014) repeat the procedure for $Y$. This step ensures that the authors find controls that are useful in predicting $Y$. These controls might not be correlated with $D$ and, therefore, not be confounders. However, including such selected variables explains more variation in $Y$, potentially reducing standard errors. The final set of control variables is the union of the set of variables selected in the two steps. Given this final specification of the model that only includes a subset of the original set of covariates, one can then estimate the treatment effect $\theta_0$ using any suitable technique.

Tarique and Schuler (2010) give a comprehensive view of (global) talent management, i.e. those practices that a firm undertakes in order to attract and develop talent (in a global context). Specifically, they split talent management into three parts: strategic alignment, global mindset, and global talent pipeline. Strategic alignment is defined as aligning talent management practices with business goals. Global mindset refers to instilling a specific mindset into the top leaders and top performers of the firm, i.e. the talented employees. Global talent pipeline reflects the need to identify, develop, and keep talent through the operations of the firm. This includes, for example, the ability for talented employees to easily switch roles or geographical location within the company and to have access to valuable training opportunities. This thesis considers talent management practices as an average of indicators measured by the World Management Survey (WMS). These indicators are closely related to the definition of talent management given by Tarique and Schuler (2010). In particular, WMS measures the importance that a firm attributes to developing and making room for talent. That is comparable to a measure of how much talent management practices are aligned to the goals of a company, i.e. strategic alignment. Furthermore, WMS measures the ability of a company to instill a talent mindset, i.e. global mindset. Moreover, WMS outlines a set of talent management indicators that reflects the quality and quantity of promotion opportunities and of incentives/appraisal systems for talented employees, which can be directly relates to the definition of global talent pipeline. Therefore, this supports the choice of using the WMS data in this thesis.

Talent management is a highly relevant topic in nowadays business environment. Pagan-Castaño, Ballester-Miquel, Sánchez-García and Guijarro-García (2022) acknowledge that the shift in business culture due to globalization and technological advancement encourages talent management to align with the changing economy. In particular, firms should consider a more personalized approach to talent management, which addresses talented employees individually, rather than constructing talent management practices that fit all employees in the same way. A factor that positively affects talent development is, for example, the well-being of employees. Therefore, companies with a good work environment are able to build a strong work morale and job engagement in employees, which will further foster talent. According to Kwon and Jang (2022), there exists a *war for talent* amongst companies. Namely, firms focus on the scarcity of talent, competing for talented employees to work for their company. However, the authors highlight that talent should not be viewed as a scarce, finite resource. Rather, through talent management practices, firms may be able to foster and develop talent. Similarly, Kaliannan, Darmalinggam, Dorasamy and Abraham (2022) suggest that a key factor for embracing talent

development is inclusiveness. That is, companies should not only focus on a selected group of *talented* employees. Rather, by extending talent management practices to a broader, more inclusive group of people, firms may be able to further further develop talent. Talent is therefore not to be considered innate, but rather a skill that can be acquired. Although there exists plenty of research that examines talent management by itself and in relation to many factors such as innovativeness and job engagement, few investigate the relationship between talent management and sales performance. Ansar, Baloch et al. (2018) discuss the issues related to identifying and developing talent. That is, identifying talent is subject to bias when evaluating an employee. Moreover, concrete career opportunities and an engaging work environment are keys to developing talent. However, in order to effectively foster talent, companies must align talent management practices to their overall business goal. Lewis and Heckman (2006) claim that there is mixed empirical evidence regarding the effect of talent management on firm performance. Although talent acquisition is considered a competitive advantage, the effectiveness of talent management practices is not well-studied. Collings and Mellahi (2009) suggest that there are gaps to be filled in the body of literature on this topic. They find that talent management is associated to improved financial performance. However, Collings and Mellahi (2009) discuss the need for investigating causality between talent management and financial performance as well as the need for identifying those specific mechanisms through which talent management leads to success. In addition, the existing research on this topic is quite descriptive, highlighting correlation between talent management and firm performance, rather than causality. In this thesis, I aim to offer a more systematic approach to analyzing the effect of talent management practices on sales.

According to Rabbi, Ahad, Kousar and Ali (2015), talented employees are valuable because they are the ones that can actually lead a company towards success. Therefore, identifying, attracting, and developing talent is of high importance for a firm that aims at improving financial performance and sales. Moreover, Wright, Mcmahan and McWilliams (1994) investigate how good human resource management can be used as a competitive advantage. The authors claim that creating rare and unique resources, for example valuable talent management practices, can be a main driver of firm performance.

Kehinde (2011) studies how talent management practices impact firm performance. The author conducts a correlation analysis and finds that the quality of talent management practices, e.g., the ability to attract and develop skilled employees, is positively related to firm performance through innovativeness and employee satisfaction. Namely, talent management fosters innovation which, in turn, leads to better firm performance. Moreover, talent management has a positive impact on job engagement and satisfaction. More passionate employees strive for leading their company to success. Hence, talent management positively affects performance. Kafetzopoulos (2022) studies the effect of talent management on financial performance through a traditional approach. The author investigates the relationship between talent development and financial performance through a survey experiment. He conducts a mediation analysis in which he uses innovativeness and strategic flexibility as mediators. The findings of this study reveal a strong statistically significant positive effect of talent management practices on financial performance through strategic flexibility (with an estimate of 0.200) and innovativeness (with

an estimate of 0.216). This suggests that firms that invest more in talent management practices have better financial performance because they are more flexible and innovative. It may be interesting to compare the results of my research, which aims to correct for bias through machine learning methods, to the aforementioned papers. That is, one may compare whether my results show a strong positive effect of talent management practices on sales similarly to Kehinde (2011) and to Kafetzopoulos (2022), and whether this effect turns out to be causal, rather than a simple correlation.

# 3 Methods

## 3.1 Background Setting

Firstly, I will estimate the existence and the strength of a causal relationship between talent management practices and sales performance using DML. Better talent management practices may lead to higher sales. However, companies with better sales performance may decide to invest more on talent management practices. Therefore, the relationship is subject to simultaneity bias. This research proposes to use ownership of the firm as an instrument for talent management practices. If a firm is family-owned, it will be less prone to invest in talent management practices as it will be more likely to hire family members, regardless of talent. According to Bhalla and Bratton (2015), most family-businesses are likely to impose a glass-ceiling on employees that are not family members, which may prevent good talent management practices to be implemented. Therefore, the validity condition holds. After controlling for all factors (for example, size, market performance etc.) that may cause ownership type to be correlated with sales performance according to Wang and Shailer (2015), the fact of being a family business is not correlated with the sales itself, except through talent management practices. Therefore, the exclusion restriction holds if the appropriate controls are included. The model looks as follows:

$$Y = D\theta_0 + g_0(X) + U,$$

$$Z = m_0(X) + V, \tag{1}$$

where the outcome variable, $Y$, is (the logarithm of) sales and the endogenous explanatory variable, $D$, is a measure of talent management practices that averages over different indicators of such management practices. To deal with endogeneity, I use an instrumental variable $Z$ which is a dummy variable that equals 1 if the business is family-owned, and 0 otherwise. The original set of control variables, $X$, includes, amongst others: management performance indicators (unrelated to talent management), the age of the firm, the country of where the plant of the firm is located, the number of employees in the company, the return on capital employed (ROCE), the line of business of the firm, the number of managers with a college degree.

The analysis is also replicated with a different instrumental variable $Z$ which indicates the number of colleges in the city where the firm is located. If a firm is located in a city with lots of universities, it may attract more talented people which will encourage more investment in talent management practices. The number of universities in a city has been used as an instrumental variable for determining the likelihood of students to pursue higher-level education, for example by Proteasa and Crăciun (2020). As the likelihood for students to attend higher-level education directly speaks to their motivation and talent mindset, the number of colleges in the city is a good candidate for an instrumental variable in this research. In fact, a city with more universities and highly-educated students may increase the supply of talent. Therefore, a firm located in such a city might be better aware of talent management because they are more exposed to talent. However, the number of colleges in the city where the firm is located should not be correlated to sales performance, especially after having controlled for other factors (size, market performance etc). Finally, I exploit over-identification by including both instruments, i.e. ownership of the

firm and number of colleges in the area. This should increase the accuracy of the results as the instrument varies both by firm (ownership) and by city (colleges). Therefore, including more variation in the instrument should increase the reliability of the findings (Ionescu-Ittu, Delaney & Abrahamowicz, 2009). This answers whether varying the choice of the instrument provides different results.

Lastly, this research compares the results to previous results that do not use DML. For example, Kehinde (2011) finds evidence for a strong positive correlation between firm performance and talent management, through innovativeness and job satisfaction. Moreover, Kafetzopoulos (2022) follows a traditional mediation approach and finds a strong, positive relationship between talent management practices and financial performance due to strategic flexibility (with an estimate of 0.200) and innovativeness (with an estimate of 0.216). In case this thesis comes to a different conclusion (e.g., insignificance of talent management practices on sales or a lower impact than the one found in the study of Kafetzopoulos (2022)), it is likely that results from previous research is biased. Moreover, I compare the results using DML to the ones that I would obtain when running a simple instrumental variable model as well as when estimating a traditional OLS regression. In case the results do not match to those obtained through DML, this will bring about further evidence that DML corrects for bias.

### 3.2 Post-Double Lasso Selection

I first select the set of control variables $X$ by using Post-Double Lasso Selection, which is a renowned valid method for principled variable selection (Urminsky, Hansen & Chernozhukov, 2016). The procedure follows the approach of Belloni et al. (2014). Although DML should already be able to select the relevant variables, Post-Double Lasso Selection additionally ensures that the estimation is unbiased. Moreover, Post-Double Lasso Selection is not computationally expensive. However, Post-Double Lasso Selection is not strictly necessary when using DML. DML is indeed a generalization of Post-Double Lasso Selection. Rather, Post-Double Lasso Selection should serve as a validation of DML or as a pre-selection process that aids the DML procedure, in case the latter reveals to be extremely time-expensive in comparison to Post-Double Lasso Selection. Whereas in Post-Double Selection we assume the data generating process to be of the following type:

$$Y = D\theta_0 + X\beta + U,$$

$$D = X\alpha + V,$$

DML does not specify the relationship between $X$, $Y$, and $D$, which is instead modeled through the functions $g(\dot{)}$ and $r(\dot{)}$ (Chernozhukov et al., 2022). Therefore, DML works well in a setting where the relationship is not assumed to be linear, contrary to Post-Double Lasso Selection (Ahrens, Aitken & Schaffer, 2021). However, DML can produce inaccurate results in case I include bad covariates. That is because DML is based on unconfoundedness, which is the exogeneity of all control variables (Hünermund et al., 2021). Since, in this case, Post-Double Lasso Selection is not directly used for causal inference but, rather, for variable selection, bad controls should not be an issue and they will most likely be kicked out of the model. Therefore,

in order to have a complete overview of the model and of the relevant terms, Post-Double Lasso Selection is a useful preliminary step to DML. The double selection step indeed provides additional correction to reduce bias and improve the accuracy of the selected variables. This helps to avoid the inclusion of variables that may lead to biased estimates. It is crucial to highlight that DML itself should take care of variable selection. However, since DML is a generalization of Post-Double Lasso Selection, the latter could turn out to be useless, but not harmful. Applying this procedure does not guarantee better final results, but also not worse. Applying this additional step is to the discretion of the researchers, who should evaluate their decision based on the dimensions of the data set and the context of the study.

Post-Double Lasso Selection works as follows. I perform Lasso on the regression of all control variables in the set $X$ on the treatment variable $D$

$$D = X\beta + V,$$

which will yield a subset of $X$, say $S_1$, that contains only the variables in $X$ that are statistically significant in the regression setting above. Subsequently, I perform Lasso on the regression of all control variables in the set $X$ on the outcome variable $Y$

$$Y = X\alpha + U,$$

Similarly, this will result in a subset $S_2$. The union between $S_1$ and $S_2$ is the final set of covariates $X'$ that I will use when regressing $D$ on $Y$ by estimating equation 1.

## 3.3 Debiased Machine Learning

DML is particularly useful when considering causal inference applied to instrumental variable models with high-dimensional controls (Kreif & DiazOrdaz, 2019).

### 3.3.1 Overview of Debiased Machine Learning

Consider the partial linear model:

$$Y = D\theta_0 + g_0(X) + U,$$

$$D = r_0(X) + V,$$

The aim is to estimate the conditional expectations $l(X) = E[Y|X]$ and $r(X) = E[D|X]$ through machine learning and partial out the effect of $X$ (similarly to the Frisch-Waugh-Lovell theorem (Lovell, 2008)). That is, we retrieve $\theta_0$ through the simple equation:

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I}\hat{V}_i^2\right)^{-1}\frac{1}{n}\sum_{i\in I}\hat{V}_i \times \left(Y - \hat{l}\right),$$

where $V = D - \hat{r}$ (Ahrens, Hansen, Schaffer & Wiemann, 2023).

However, over-fitting issues may lead the error $l(X) - \hat{l}$ and $V$ to be correlated and, therefore, bad performance. In order to solve this, DML uses cross-fitting. This means that, in order to

estimate $\hat{l}$ and $\hat{r}$, DML relies only on a sub-sample if the data, whereas the remaining part of the observations are used to construct $l(X)$ and $m(X)$. More precisely, the algorithm for cross fitting works as follows (Ahrens et al., 2023):

- Split the sample in $K$ parts of equal sizes, and denote the sample that arises from the $k^{th}$ split by $I_k$.
- For $k = 1, ..., K$, construct $\hat{l}$ and $\hat{r}$ using sample $I_k$, whereas estimate $l(X)$ and $m(X)$ using the observations that are **not** contained in $I_k$.
- Finally, obtain $\theta_0$ as:

$$\hat{\theta}_0 = \left( \frac{1}{n} \sum_{i \in I} \hat{V}_i^2 \right)^{-1} \frac{1}{n} \sum_{i \in I} \hat{V}_i \times \left( Y - \hat{l} \right),$$

  where $V = D - \hat{r}$.

### 3.3.2 Debiased Machine Learning and Instrumental Variable Models

Applying DML to the instrumental variable model works in a similar fashion as described above, with the exception that I am interested in the conditional expectation $E[Z|X]$ (Ahrens et al., 2023) in addition to $E[Y|X]$ and $E[D|X]$.

I apply DML to the instrumental variable setting:

$$Y = D\theta_0 + g_0(X') + U,$$

$$D = r_0(Z, X') + V, \tag{2}$$

It is important to highlight that I now model the relationship between $X$, $Y$, and $D$ through the functions $g(\dot{})$ and $r(\dot{})$ (Chernozhukov et al., 2022), in contrast to Post-Double Lasso Selection, where the relationship is assumed to be linear (Ahrens et al., 2021) and, therefore, the coefficients $\beta$ and $\alpha$ are suitable in describing the aforementioned relationship. I begin by splitting the sample in a main part and an auxiliary part. Using the auxiliary sample, I estimate $\hat{g_0}(X')$ through the Lasso method from the equation

$$Y = D\theta_0 + g_0(X') + U,$$

as well as $\hat{r_0}(Z, X')$ through Lasso applied to the equation

$$D = r_0(Z, X') + V,$$

Then, I use the main sample and obtain the orthogonalized component $\hat{V}$ of $D$ on $(Z, X)$:

$$\hat{V} = D - \hat{r_0}(Z, X'),$$

In order to answer the research question regarding the strength of the causal effect, I obtain the debiased estimator of $\theta_0$ from the OLS formula:

$$\hat{\theta}_0 = \left( \frac{1}{n} \sum_{i \in I} \hat{V}_i \times D_i \right)^{-1} \frac{1}{n} \sum_{i \in I} \hat{V}_i \times \left( Y - \hat{g}_0(Z, X') \right),$$

The estimator is root-N consistent, which means that both the estimator and its standard deviation approach the true value as the the sample size increases. The procedure is repeated $K$ times by using different data proportions (cross-fitting) and the final estimate of $\theta_0$ is the mean of all results. Cross-fitting increases the rate of convergence (Newey & Robins, 2018). This is a valid procedure because the various estimates that arise from each cross-fitted sample are independent of each other (Williamson, Gilbert, Simon & Carone, 2021). This answers whether varying the choice of the machine learning method provides different results.

### 3.3.3 Over-identification in Instrumental Variable Models

One of the DML identification strategies that I consider in this thesis relies on an over-identified instrumental variable model. When the dimension of the excluded instruments vector is larger than one, the endogenous variable is over-identified. The derivation of the equations is slightly different compared to the case where the endogenous variable is just-identified. The proof exploits the Two-Stage Least Squares (2SLS) method and it is inspired by the procedure and explanations of Baum, Schaffer and Stillman (2003). Consider the model:

$$Y = D\theta_0 + g_0(X') + U, \tag{3}$$

and assume that I can find two valid instrumental variables, $Z_1$ and $Z_2$. The first-stage regression looks as follows:

$$D = r_0(Z_1, Z_2, X') + V,$$

Similarly to the just-identified case, I estimate $\hat{g_0}(Z_1, Z_2, X')$. Then, I estimate $\theta_0$ through Generalized Method of Moments (GMM) or efficient GMM. This is different than the just-identified case where I could estimate $\theta_0$ through a simple formula.

In order to implement GMM, I first define:

$$\iota_i(\hat{\theta_0}) = (Z_1, Z_2, X')_i^T \hat{U}_i = (Z_1, Z_2, X')_i^T (Y_i - D_i\theta_0 - \hat{g}_0(X')_i), \tag{4}$$

for all observations $i = 1, ..., n$. I define the set of all observations as the set $I$. Equation 4 arises from the exogeneity of the instruments and of the other covariates, which can be written as $E[(Z_1, Z_2, X')_i \dot{U}_i] = 0$. If $X'$ has dimension $L \times n$, then $\iota_i(\hat{\theta_0})$ has dimension $(L + 2) \times 1$. Therefore, there are $L+2$ moment conditions that will be satisfied at the true value of $\theta_0$ (Baum et al., 2003):

$$E[\iota_i(\theta_0)] = 0$$

Moreover, define the sample moments as the sample mean of these moments. That is:

$$\iota(\bar{\hat{\theta}_0}) = \frac{1}{n}\sum_{i\in I}\iota_i(\hat{\theta_0}) = \frac{1}{n}\sum_{i\in I}(Z_1, Z_2, X')_i^T(Y_i - D_i\theta_0 - \hat{g}_0(X')_i) = \frac{1}{n}(Z_1, Z_2, X')^T\hat{U}$$

The GMM method aims to find $\theta_0$ such that $\iota(\bar{\hat{\theta}_0}) = 0$ (Baum et al., 2003).

Subsequently, I construct the GMM objective function, which measures the difference between the sample moments and the moments obtained through the estimated $\hat{r_0}(Z_1, Z_2, X')$. The objective function looks as follows:

$$GMM(\theta_0) = n\iota(\bar{\hat{\theta}_0})^T W \iota(\bar{\hat{\theta}_0}), \tag{5}$$

where $W$ is a positive semi-definite weighting matrix. In order to find an estimate of $\theta_0$, the objective function in equation 5 needs to be minimized (Baum et al., 2003). Efficient GMM chooses the optimal weighting matrix, i.e. the weighting matrix that minimizes the asymptotic variance of $\hat{\theta_0}$ (Baum et al., 2003).

# 4 Data

I use (cross-sectional) data from 7094 firms in 20 countries used in the Bloom, Genakos, Sadun and Van Reenen (2012) survey paper. The data has been retrieved in the years from 2004 to 2010. The data set can be found on the World Management Survey website. The three most relevant variables that are included are: sales for each firm (dependent variable), six talent management practices indicators (variable of interest), the type of ownership of the firm, i.e. whether it is family-owned or other (instrument). The sales variable is provided in different formats and I will use the logarithm of sales as dependent variable due to scaling and interpretability reasons. Figure 3 shows the distribution of the variable *lsales*. The talent management practices indicators refer to different organizational aspects that are related to talent management. These are measured on a scale from 1 ( = worse) to 5 ( = best) based on their efficiency. I aggregate talent management indicators into one variable, which is an average over them. Therefore the resulting treatment variable *talent* is continuous and ranges from 1 to 5. Figure 4 displays the distribution of this variable. I aggregate 10 performance indicators and 5 indicators regarding lean management in a firm together in a similar way. These will be included in the set of control variables. Regarding the instrument, I aggregate all types of ownership that are not family-related. This results in two groups, one which contains 1288 firms and takes value 1 (family-owned, both with internal or external CEO), and another one which contains 5081 firms and takes value 0 (all other types of ownership such as dispersed shareholders, government, private equity etc.). Ownership data is missing for 725 firms. Moreover, other covariates that can be controlled for are included. For example, the data set provides information on general performance, size etc. of each firm which can also influence sales and must be accounted for.

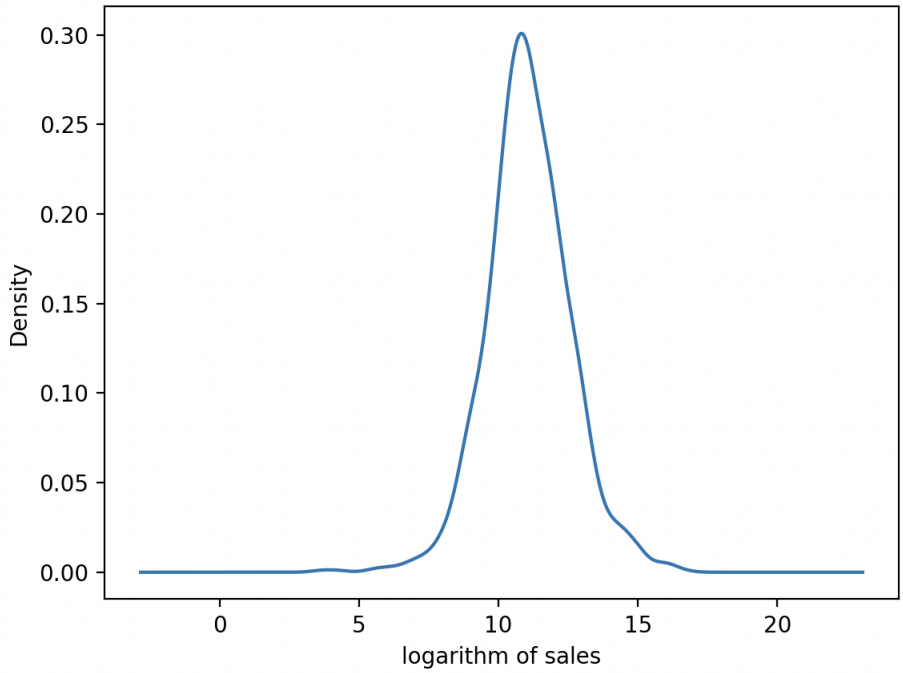


Figure 3: Distribution of lsales variable.

In order to construct the second instrument, i.e. the number of colleges in the city where the

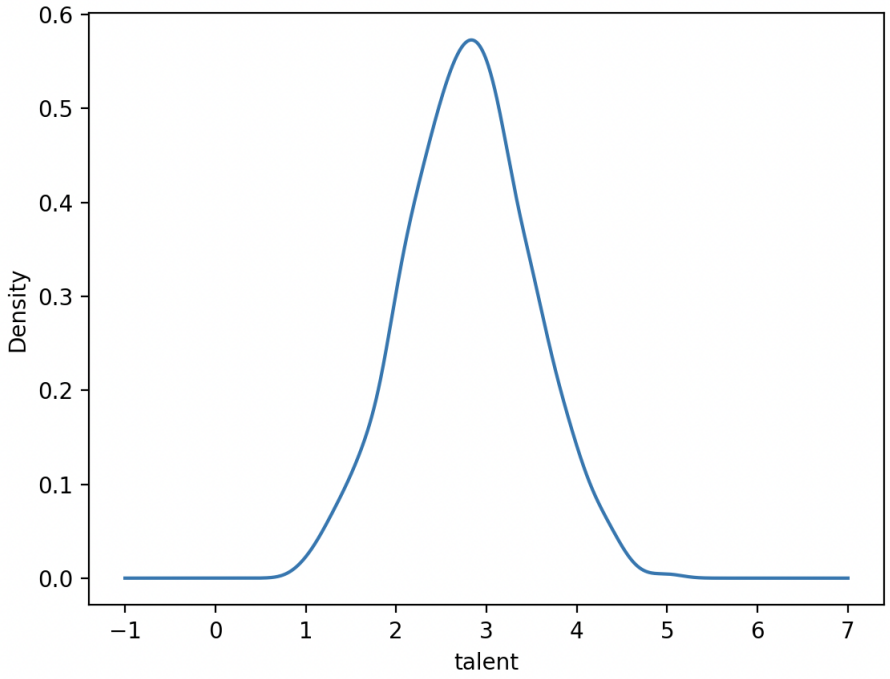


Figure 4: Distribution of talent variable.

firm is located, I use additional data that are obtained upon request to the World Management Survey. This data set contains 9231 firms across the same 20 countries. However, the firms included in this data do not fully match the ones in the original data set. Indeed, only a subset of this data set is equal to a subset of the original data. Therefore, since I want to use the same sample throughout the whole analysis, I merge the two data sets and I obtain the intersection of the two. The final sample is representative of 1512 firms that are common to both data sets. I then exclude those observations that include missing values, which results in 1189 observations remaining. I construct the second instrument by reporting, for each city, the number of universities in the area. This variable does not yield much variation as it is shown by plotting its distribution in Figure 5, where the range of this variable is from 0 to 120. In order to include more variation in this variable, I transform the number of universities into its corresponding squared value (Grissom, 2000). I hereby provide an overview of the data. Namely, I report the relevant variables (dependent, treatment, instruments) of the first five observations in Table 1 as well as descriptive statistics of these variables in Table 2[1]. The descriptive statistics of all variables are reported in Table 13 as well as an explanation of their meaning. The distribution of *lsales* in Figure 3 resembles a normal distribution with the mean centered around 11, which is also supported by the mean value given by the descriptive statistics. Similarly, the distribution of *talent* in Figure 4 resembles a normal distribution with the mean centered around 3, which is confirmed by the corresponding mean value in Table 13. Moreover, Figure 6 is a visual display of the amount of firms that are and are not family-owned. Figure 7 describes the distribution of the university instrument. In addition, I investigate the validity of the instrument by inspecting the correlation between the instrument (ownership) and the treatment (talent). The

[1]All numbers rounded to two decimals.

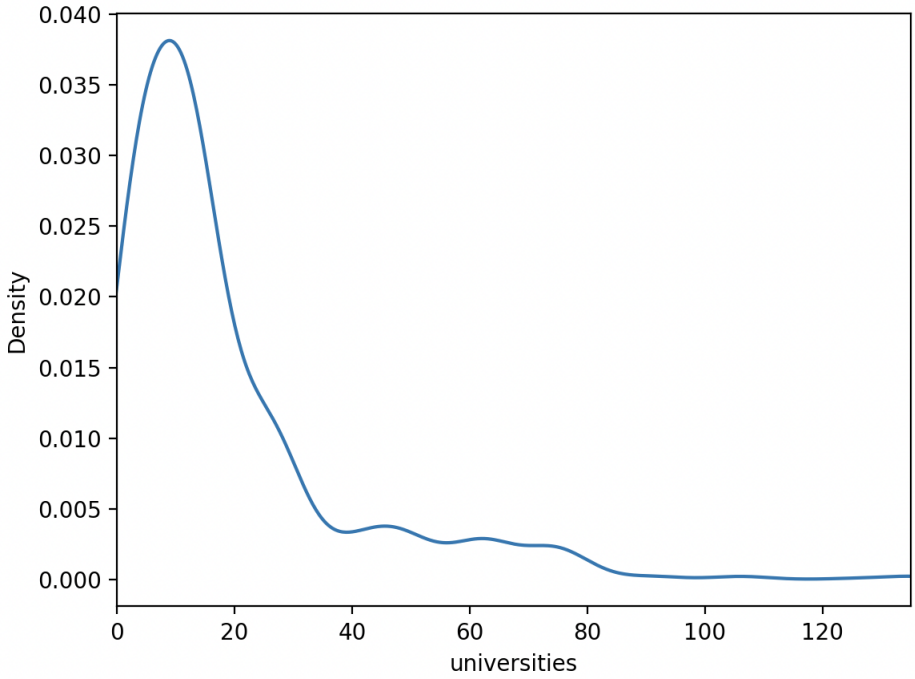


Figure 5: Distribution of university variable (before transformation).

| log sales | talent | ownership | universities |
|---|---|---|---|
| 10.85 | 1.33 | 0.00 | 225.00 |
| 11.66 | 1.00 | 0.00 | 121.00 |
| 12.01 | 1.16 | 0.00 | 4.00 |
| 8.84 | 1.00 | 1.00 | 576.00 |
| 9.59 | 1.833 | 0.00 | 100.00 |

Table 1: First five observations

| | log sales | talent | ownership | universities |
|---|---|---|---|---|
| **count** | 1189.00 | 1189.00 | 1189.00 | 1189.00 |
| **mean** | 11.08 | 2.80 | 0.21 | 720.74 |
| **std** | 1.51 | 0.69 | 0.41 | 1695.17 |
| **min** | 3.64 | 1.00 | 0.00 | 0.00 |
| **25%** | 10.2 | 2.33 | 0.00 | 36.00 |
| **50%** | 11.01 | 2.83 | 0.00 | 144.00 |
| **75%** | 11.96 | 3.20 | 0.00 | 576.00 |
| **max** | 16.59 | 5.00 | 1.00 | 18225.00 |

Table 2: Descriptive Statistics

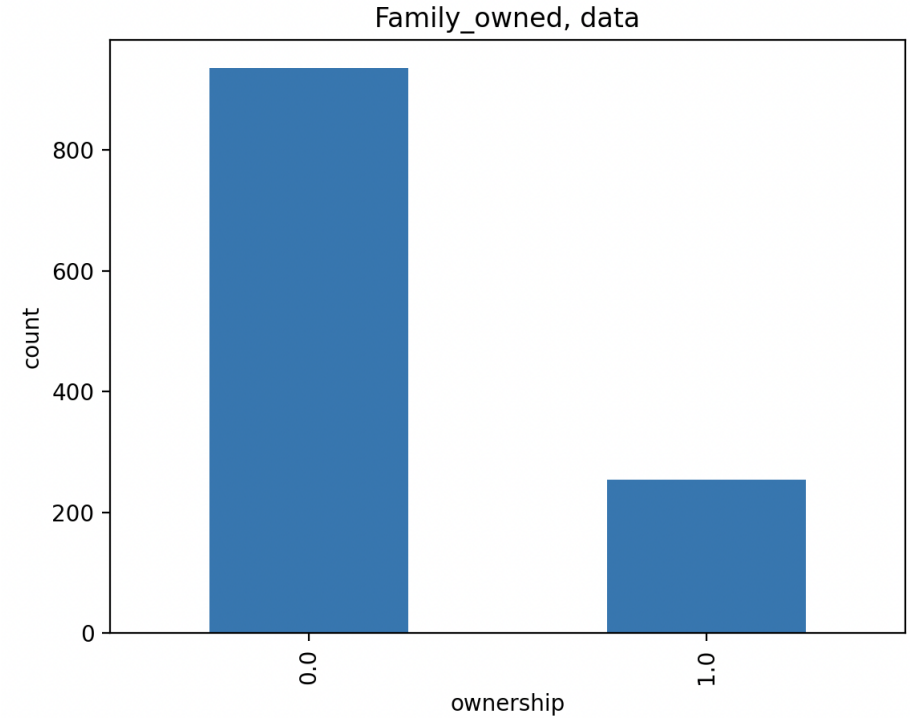


Figure 6: Distribution of ownership variable: number of firms that are family-owned (1) versus not family-owned (0)

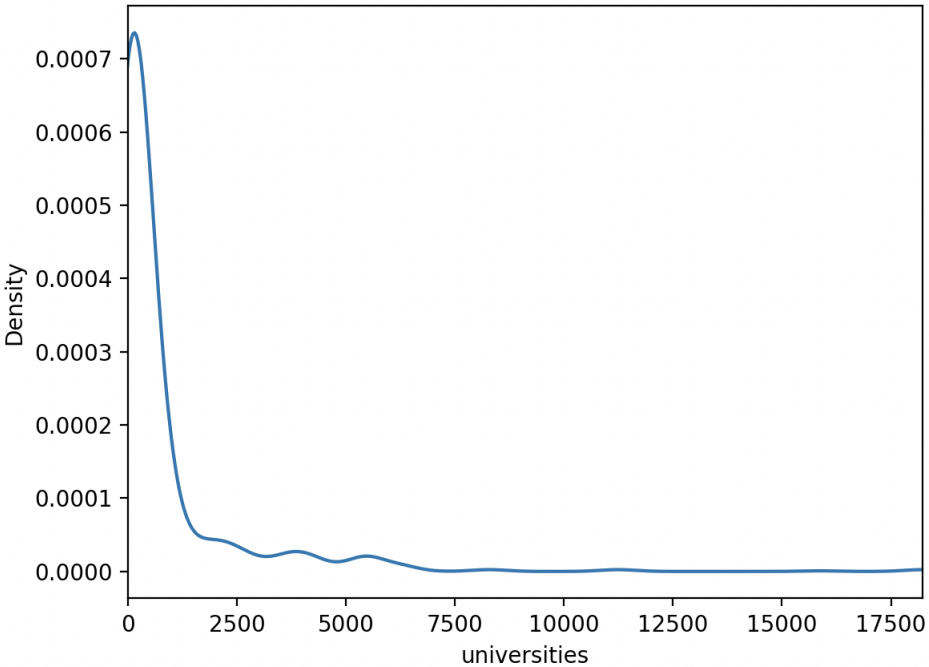


Figure 7: Distribution of university variable (after transformation).

value of the correlation coefficient is -0.15 and it is statistically significant at a 5% level (p-value: $1.59 \times 10^{-38}$). This confirms that, if a business is family-owned, it will be less inclined to invest in talent management practices. Figure 8 also suggests that talent and ownership are correlated since firms that are family-owned (1) have a lower density respective to the talent variable than those that are not (0). Therefore, the validity condition holds. Regarding the second instrument,

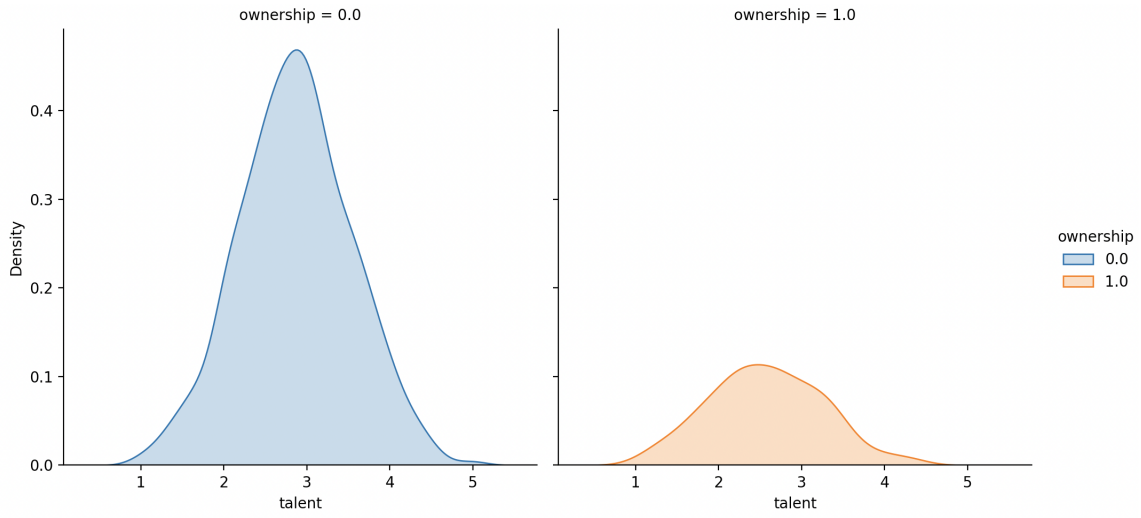


Figure 8: Density of talent for two groups of ownership: family-owned (1) versus not family-owned (0)

the correlation coefficient is 0.06 and it is statistically significant at a 5% level (p-value: 0.03). This confirms that, if a company is situated in an area with many educational institutions, it will attract more talented employees.

If the instrument impacts sales even without talent management practices, the instrument might not be valid. In order to check for this, I use a test conditionally on the other instrument. That is, the first model is misspecified if the variable *ownership* or the variable *university* directly affects sales through, for example, talented employees, irrespective of their development through talent management practices. If I control for *ownership* in the model that uses *university* as instrument and vice versa, I should be able to capture the effect of talent as the control would serve as a proxy for it. Therefore, if this control variable turns out to be statistically significant, the model is misspecified. This is similar to the Hansen-Sargan test conditional on one instrument (Kiviet & Kripfganz, 2021). The limitation of this approach is that, if the control variable is statistically insignificant, this does not guarantee that the model is correctly specified. Nonetheless, this would be a supporting evidence of the argument that the control variable at hand, when used as an instrument, is valid. I provide the results for the variable *ownership* when *universities* is employed as instrument and for the variable *universities* when *ownership* is instrumented in Table 3. Table 3 displays the coefficients as well as the p-value for the two cases.

| | **coefficient** | **std err** | **z** | $P \geq \|z\|$ | **95% conf interval** |
|---|---|---|---|---|---|
| *ownership* | -0.04 | 0.11 | -0.40 | 0.69 | [-0.25, 0.17] |
| *universities* | 0.00 | 0.00 | -0.33 | 0.74 | [0.00, 0.00] |

Table 3: Misspecification Test

As the coefficients are statistically insignificant at the 5% level in both scenarios, the instruments are more likely to be valid. However, one cannot formally test whether the model is correctly specified. Rather, I can state that this misspecification test does not reject the null hypothesis that the model is correctly specified.

# 5 Results

## 5.1 Control Variable Selection

Model selection is key to obtaining accurate results. The control variables are selected using Post-Double Lasso Selection. As Lasso works best in the absence of multicollinearity, I first analyse correlation amongst the variables in my model. However, since I do not know which variables are in the true DGP, dropping some of them does not necessarily solve the multicollinearity issue. Nonetheless, this data pre-processing step might still be useful if paired with nuisance on which variable can be safely excluded from a logical, rather than only mechanical, standpoint (Kozak, 2009). In fact, I consider both the magnitude of the pairwise correlation coefficient as well as the interpretation of the two variables at hand in order to make inference on which variable should be excluded at this stage. The correlation table is reported in Figure 10. The variable measuring the degree to which a firm makes use of lean practices (*lean*) and the variable measuring firm performance (*perf*) have a correlation of 0.61 that is statistically significant. Although not the scope of this thesis, it is interesting that firms adopting modern, lean strategies perform better than more tradition-oriented ones. Due to this being a source of multicollinearity, I only keep the variable *perf*. Moreover, I find almost perfect correlation that is statistically significant between the variables *management* and *factor management* as they measure the same concept using different specifications. Moreover, these variables have a statistically significant high correlation of, respectively, of 0.95 and of 0.96 with the variable *perf*. Therefore, I exclude both *management* and *factor management* from the model. In addition, I find a high correlation of 0.72 that is statistically significant between the variable measuring the number of employees in a firm (*lemp*) and the variable indicating the amount of long-term, tangible assets that a company owns (e.g., trucks, machinery) (*lppent*). Therefore, I only keep the variable *lemp* because it seems more fit to be a proxy for the size of the firm.

In order to apply Post-Double Lasso Selection, I perform Lasso to the regression:

$$talent = X\beta + V, \tag{6}$$

and, subsequently, on the regression:

$$lsales = X\alpha + U, \tag{7}$$

where $X$ represents the remaining control variables. The penalty coefficient is chosen through cross-validation. Figure 9 provides an overview of the selected control variables in equation 6 (*lasso talent*) and in equation 7 (*lasso lsales*). These variables consist of the final set of control variables that are included in the model, which will be referred to as $X'$. It is important to underline that this step is not necessary if the dimension of the covariate space is not extremely large, such as in the case presented in this thesis. That is because DML is a generalization of Post-Double Lasso Selection and, therefore, already incorporates good variable selection properties.

| | mylassotalent | mylassosales |
|---|---|---|
| firmid | x | x |
| firmage | x | x |
| mne_f | x | x |
| ldegree_t | x | x |
| perf | x | x |
| roce | x | x |
| sic | x | x |
| dow | x | x |
| china | x | x |
| france | x | x |
| japan | x | x |
| poland | x | x |
| portugal | x | x |
| sweden | x | x |
| lemp | | x |
| dead | | x |
| year | | x |
| reliability | | x |
| i_comptenure | | x |
| i_seniority | | x |
| australia | | x |
| brazil | | x |
| germany | | x |
| greece | | x |
| italy | | x |
| northernireland | | x |
| republicofireland | | x |
| unitedstates | | x |
| _cons | x | x |

Figure 9: Results from Post-Double Lasso Selection

### 5.2 Debiased Machine Learning Identification Strategy 1

The first identification strategy of DML that I consider uses ownership of the firm as an instrument for talent management practices. The setting is as follows:

$$lsales = \theta_0 talent + g_0(X') + U,$$

$$talent = r_0(ownership, X') + V, \tag{8}$$

where the aim is to obtain an unbiased estimate of $\theta_0$. I randomly split the data in two parts, the auxiliary sample and the main sample. First, I compute the residual of the dependent variable $lsales_a$ on the treatment variable $talent_a$ using the auxiliary sample, say $\hat{lsales}_a$. I distinguish the auxiliary sample and the main sample by introducing a lower case $a$ and $m$, respectively. Namely,

$$\bar{\theta_0} = \left({talent_a}^T talent_a\right)^{-1} {talent_a}^T lsales_a,$$

$$\hat{lsales}_a = lsales_a - \bar{\theta_0} talent_a.$$

I then retrieve an estimate of $g_0(X'_m)$ from:

$$\hat{lsales}_a = \theta_0 talent_a + g_0(X'_a) + U,$$

through the use of machine learning (i.e. Lasso) predictions on the main sample, say $\hat{g}_0(X'_m)$. I then obtain $\hat{U}$ using the main sample and $\hat{g}_0(X'_m)$. Namely,

$$\hat{U} = lsales_m - \hat{g}_0(X'_m),$$

Similarly, I retrieve $\hat{r}_0(ownership_m, X'_m)$ from:

$$talent_a = r_0(ownership_a, X'_a) + V,$$

and obtain $\hat{V}$ from:

$$\hat{V} = talent_m - \hat{r}_0(ownership_m, X'_m).$$

The estimation of $\hat{U}$ and $\hat{V}$ allows me to retrieve $\hat{\theta}_0$ from:

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I} \hat{V}_i \times talent_{mi}\right)^{-1} \frac{1}{n}\sum_{i\in I} \hat{V}_i \times \hat{U}_i,$$

I use 5-folds cross-fitting, meaning that the data is divided into 5 proportions when considering the cross-fitting algorithm. The Lasso penalty parameter is chosen through cross-validation. Moreover, in addition to Lasso, I repeat the analysis for different machine learning methods: Random Forest, Gradient Boosting, and Support Vector Machines. Table 4 outlines the results. The coefficients are statistically insignificant. This is the case for all the machine learning

| Method | $\hat{\theta}_0$ | std err | z | $P \geq \|z\|$ | 95% conf interval |
|---|---|---|---|---|---|
| Lasso | 1.37 | 1.16 | 1.48 | 0.14 | [-0.56, 4.01] |
| Forest | 2.00 | 1.41 | 1.43 | 0.15 | [-0.75, 4.76] |
| GB | 1.27 | 1.03 | 1.23 | 0.22 | [-.75, 3.30] |
| SVM | 1.85 | 1.17 | 1.58 | 0.11 | [-0.45, 4.15] |

Table 4: Results Identification Strategy 1

methods that I consider for which the estimates, although statistically insignificant, are similar across each other. This suggests that the accuracy of the results does not heavily depend on the machine learning technique that is applied and all methods lead to the same conclusion, i.e. talent management practices do not affect sales. Nonetheless, when considering the p-value or the standard error, the estimates are close to being statistically significant. Therefore, when using *ownership* as instrument, talent management practices do not seem to have an effect on sales. This may mean that firms should not be investing in talent management practices as it does not bring any benefit when considering sales. It may still be that talent management adds value to a company through different channels, for example better employee morale. However, ultimately, firms should only invest in those practices that have a positive pecuniary return as the net profit of the investment would then be positive. Although the validity condition of the instrument is satisfied, it may be that the exclusion restriction is violated. Indeed, when considering the p-value or the standard error, the estimates are close to being statistically significant. This suggests that a better instrument may yield statistically significant results. In the next section, I propose an alternative instrument, i.e. the number of educational institutions in the city where the firm is located.

### 5.3 Debiased Machine Learning Identification Strategy 2

The second identification strategy of DML exploits the number of universities that are in the area where the headquarters of the firm are located as instrument:

$$lsales = \theta_0 talent + g_0(X') + U,$$

$$talent = r_0(universities, X') + V, \tag{9}$$

where I want to find an unbiased estimate of $\theta_0$. The technique is similar to the one used in Identification Strategy 1 (section 5.2), with the only exception that the variable *ownership* is replaced by the variable *universities*. In the same fashion, I use 5-folds cross-fitting and the Lasso penalty parameter is chosen through cross-validation. Table 5 displays the results for Lasso, Random Forest, Gradient Boosting, and Support Vector Machines. The coefficients are

| Method | $\hat{\theta}_0$ | std err | z | $P \geq \|z\|$ | 95% conf interval |
|---|---|---|---|---|---|
| Lasso | 1.31 | 1.20 | 1.09 | 0.27 | [-1.04, 3.65] |
| Forest | 1.26 | 1.19 | 1.06 | 0.29 | [-1.07, 3.59] |
| GB | 1.45 | 1.24 | 1.17 | 0.24 | [-0.97, 3.87] |
| SVM | 0.99 | 0.86 | 1.15 | 0.25 | [-0.70, 2.68] |

Table 5: Results Identification Strategy 2

statistically insignificant and similar across all machine learning methods used. This confirms the findings of Identification Strategy 1 ((section 5.2). Although varying the instrument does not seem to yield different results, it may also be the case that both instrumental variables do not appropriately capture the endogeneity in *talent*. I investigate this further by exploiting a setting where the instrument over-identifies the endogenous variable in the next section.

### 5.4 Feature Selection for Lasso and Random Forest

Lasso and Random Forest allow for feature selection. Lasso drives the coefficients of irrelevant variables to zero through the L1 regularization penalty (Fonti & Belitser, 2017). Random Forest computes the importance of each feature by calculating the decrease in accuracy when that variable is removed. The larger the decrease in accuracy is, the more important that variable is (Hasan, Nasser, Ahmad & Molla, 2016). In order to check which variables are excluded in the DML framework and whether there is a logical, economical explanation for their exclusion, Table 6 and Table 7 provide an overview of which features have a coefficient or an importance larger than zero according to Lasso and Random Forest, respectively.

| Variable | Active (A) Identification Strategy 1 | Active (A) Identification Strategy 2 |
|---|---|---|
| **firm id** | A | A |
| **firmage** | A | A |
| **mne f** | | |
| **ldegree** | A | A |
| **perf** | A | A |
| **roce** | A | A |
| **sic** | A | A |
| **dow** | | |
| **lemp** | A | A |
| **dead** | | |
| **year** | A | A |
| **reliability** | A | |
| **i comptenure** | A | A |
| **i seniority** | A | A |
| **australia** | | |
| **brazil** | | |
| **germany** | | |
| **greece** | | |
| **italy** | | |
| **northern ireland** | | |
| **china** | | |
| **republic of ireland** | | |
| **france** | | |
| **japan** | | |
| **poland** | | |
| **portugal** | | |
| **sweden** | | |
| **US** | | |

Table 6: Lasso. Coefficients that are different from zero are defined as *active* and highlighted with an A.

Lasso excludes all country variables. This may be because the country effect is already captured by other variables such as performance. It makes economic sense that the variable

*dead* is excluded as the owner of the firm being dead or alive is not necessarily predictive of sales performance. Similarly, the variable *mne f* is not relevant. That is, whether the headquarters are located in the country where the firm originated is not a predictor of sales according to the Lasso procedure. Finally, the coefficient of the variable *dow* is shrunk to zero. That may be because the market performance indicator is strictly related to the performance variable (*perf*). Therefore, Lasso only keeps *perf* and excludes *dow*. Lasso in Identification Strategy 2 also shrinks to zero the coefficient of the variable *reliability*. This may also be because the performance of the firm already reflects the reliability of the company, e.g., a less reliable firm is probably going to perform worse.

Random Forest includes all variables that are selected by Lasso, except for *year* and *reliability*. Moreover, Identification Strategy 2 finds that *i comptenure* is not important. The variable *year* might not be considered as important because the time effect might already be captured by some other variable. Contrarily to Lasso, Random Forest finds that some of the country variables are indeed important. Moreover, the variable *mne f* has an importance different from zero in both identification strategies, whereas the variable *dead* is considered important in Identification Strategy 2. However, the variable importance criteria of Random Forest is much more inclusive than the feature selection of Lasso. That is, a variable may have a low importance even if different than zero.

The variables that are definitely predictive of sales performance are *firm id*, *firmage*, *ldegree*, *perf*, *roce*, *sic*, *lemp*, and *i seniority*. The variable *firmid* is a firm identifier, hence it is important to include it in the model. The variable *firmage* may be predictive of sales as a start up has much less experience than a well-established firm. Therefore, this may impact sales. The variables *ldegree* and *lemp* are indicators of the quality and quantity of the employees, respectively. Therefore, a firm with many employees of which a large percentage possess a university degree might perform better than one that does not. The variables *perf* and *roce* are performance indicators, respectively for production and Return on Capital Employed. Therefore, performance indicators that are not directly related to sales are still indirectly connected to sales performance. Controlling for them allows to identify the effect of talent management practices on sales performance rather than overall performance. It is crucial to control for the variable *sic* as the industry in which a company operates affects sales performance, irrespective of talent management practices. Finally, the variable *i seniority* may affect performance as a company that promotes earlier based on seniority rather than merit has different characteristics than one that does not agree with such a promotion policy, e.g., less innovative or less capable employees at high levels, which would negatively affect sales.

### 5.5 Debiased Machine Learning Identification Strategy 3

The third identification strategy of DML uses over-identification in order to obtain better results. Namely, the instrument includes both the ownership of the firm and the number of universities that are located near the city where the headquarters of the firm are. Because *ownership* is firm-specific while *universities* is city-specific, the instrument that includes both variables varies in two directions and is therefore more specific. This may lead to the instrument being able to capture more variability of the endogenous variable which possibly yields better results

| Variable | Important (I) Identification Strategy 1 | Important (I) Identification Strategy 2 |
|---|---|---|
| **firm id** | I | I |
| **firmage** | I | I |
| **mne f** | I | I |
| **ldegree** | I | I |
| **perf** | I | I |
| **roce** | I | I |
| **sic** | I | I |
| **dow** | | |
| **lemp** | I | I |
| **dead** | | I |
| **year** | | |
| **reliability** | | |
| **i comptenure** | I | |
| **i seniority** | I | I |
| **australia** | | |
| **brazil** | | |
| **germany** | | |
| **greece** | I | I |
| **italy** | | |
| **northern ireland** | I | I |
| **china** | I | I |
| **republic of ireland** | I | I |
| **france** | | |
| **japan** | | |
| **poland** | I | I |
| **portugal** | I | I |
| **sweden** | | I |
| **US** | | |

Table 7: Random Forest. Variables importance that are different from zero are defined as *important* and highlighted with a I.

(Wooldridge, 2010, Chapter 15). The setting is the following:

$$lsales = \theta_0 talent + g_0(X') + U,$$

$$talent = r_0(ownership, universities, X') + V, \tag{10}$$

and I aim to retrieve an unbiased estimate of $\theta_0$. The procedure is mirrored to the previous identification strategies. I use 5-folds cross-fitting and the Lasso penalty parameter is cross-validated. Table 8 outlines the results for Lasso, Random Forest, Gradient Boosting, and Support Vector Machines.

| **Method** | $\hat{\theta}_0$ | **std err** | **z** | $P \geq \|z\|$ | **95% conf interval** |
|---|---|---|---|---|---|
| Lasso | 1.78 | 0.95 | 1.88 | 0.06 | [-0.07, 3.64] |
| Forest | 1.67 | 0.84 | 2.00 | 0.05 | [0.03, 3.31] |
| GB | 1.15 | 0.75 | 1.53 | 0.12 | [-0.32, 2.63] |
| SVM | 1.64 | 0.87 | 1.89 | 0.06 | [-0.06, 3.34] |

Table 8: Results Identification Strategy 3

The estimates are still statistically insignificant at the 5 % level, except for Random Forest. However, the statistical significance has greatly increased. Namely, if we consider the threshold to be 10 %, we obtain estimates of $\theta_0$ that are statistically significant across all machine learning methods, except for Gradient Boosting. Namely, in this case, a firm that abundantly invests in talent management (i.e. scoring 5) increases sales by 6.68% according to Random Forest, 7.12% according to Lasso, and 6.56% according to Lasso [2] compared to a firm that does not (i.e. scoring 1), *ceteris paribus*. It is possible that taking into consideration a larger data set, a different or expanded set of control variables would greatly improve the statistical significance of the estimates, leading all of them to be statistical significant at the 5 % level, and not only at the 10 % level.

There are two important considerations to be made. Firstly, these findings validate the fact that results obtained from different machine learning methods agree with each other. Not only the p-values or standard errors that arise from the different techniques are close to each other, but also the magnitude of the coefficients is similar. Overall, Gradient Boosting seems to consistently underestimate the standard error (Identification Strategy 1 and Identification Strategy 3), although the result is not far from the other estimates of the causal effect. Therefore, I conclude that the choice of the machine learning method does not heavily affect the results. However, it is suggested to try different techniques when applicable. Secondly, these findings are evidence that the choice of the instrument affects the estimation, even when using DML. Therefore, DML cannot replace the choice of an appropriate instrument, or, at least, not entirely. Including both instruments seems the most fitting choice.

### 5.6 Results with DML and without Instruments

In order to further validate the choice of implementing an instrumental variable setting, I repeat the analysis with DML using a model that does not account for *talent* to be exogenous. Table

[2] This comes from the fact that there are 4 score points separating a firm with a score of 1 to that with a score of 5. Therefore, $1.67 \times 4 = 6.68$, $1.78 \times 4 = 7.12$, and $1.64 \times 4 = 6.56$. Hence, the effect lies between 6 and 7%

9 displays the estimated effect of talent on sales based on such model for the different machine learning techniques.

| **Method** | $\hat{\theta}_0$ | **std err** | **z** | $P \geq \|z\|$ | **95% conf interval** |
|---|---|---|---|---|---|
| Lasso | 0.11 | 0.05 | 2.44 | 0.01 | [0.02, 0.20] |
| Forest | 0.11 | 0.04 | 2.54 | 0.01 | [0.03, 0.20] |
| GB | 0.11 | 0.04 | 2.45 | 0.01 | [0.02, 0.20] |
| SVM | 0.12 | 0.04 | 2.59 | 0.01 | [0.03, 0.20] |

Table 9: Results Model without Instruments

The coefficient of *talent* is statistically significant at the 5 % level in all cases. This suggests that *talent* is indeed endogenous. As these results are different from those obtained through the instrumental variable setting, the instruments control for endogeneity.

### 5.7 Comparison and Reflections on Traditional Research

It is generally believed that better talent management practices increase sales. For example, Kafetzopoulos (2022) finds a positive effect of talent development on financial performance. However, the author uses a mediation analysis, which does not account for causality. Rather, through innovation and strategic flexibility, talent development increases sales. However, investing in talent management practices does not necessarily increase sales. In fact, if talent development is not correlated with innovation and strategic flexibility, the findings of Kafetzopoulos (2022) are not valid anymore. Similarly Kehinde (2011) studies the correlation, rather than causality, between talent management and performance. Although he finds that the two are correlated through innovativeness and employee satisfaction, the results are biased. That is, the study does not prove that talent management by itself increases sales. Rather, the study provides evidence for the hypothesis that high performant companies are also usually those firms that have high innovation and in which employees are satisfied by their job. Therefore, a company should not rely on such results in order to increase sales since the exogeneity of the relationship between talent development and financial performance is not guaranteed. This research finds evidence of a causal relationship between the two through DML. Although my research and the one of Kafetzopoulos (2022) both measure talent development on a 1 to 5 scale, the author constructs the financial performance variable through sales growth, return on investment, and profitability based on the scale of Iqbal, Ahmad, Nasim and Khan (2020), Iqbal (2020), and Khan and Quaddus (2015). Therefore, it is difficult to make a direct comparison of the magnitude of the estimates obtained by Kafetzopoulos (2022) and those of my thesis. In order to capture the strength of DML, it is interesting to compare the findings of this thesis to the results that the exact same analysis would have yield without applying DML. For this reason, I hereby provide the estimation of three models that are equal to the ones employed in Identification Strategy 1 (section 5.2), Identification Strategy 2 (section 5.3), and Identification Strategy 3 (section 5.5). Finally, I estimate the model using standard OLS for comparative purposes. Table 10 displays the first stage estimates of the instrument(s) for the models of Identification Strategy 1, Identification Strategy 2, and Identification Strategy 3 under the traditional instrumental variable setting, i.e. $\hat{r_{10}}$. Table 11 reports the results from the F-tests of excluded instruments, i.e. the F statistic and the corresponding p-value. Table 12 outlines the estimated effect of talent on

sales based on such models as well as traditional OLS.

| | $\hat{r_{10}}$ | **std err** | **z** | $P \geq \|z\|$ | **95% conf interval** |
|---|---|---|---|---|---|
| Identification Strategy 1 | -0.07 | 0.04 | -1.78 | 0.08 | [-0.15, 0.01] |
| Identification Strategy 2 | 0.00 | 0.00 | 1.73 | 0.08 | [-2.63$e^{06}$, 0.00] |
| Identification Strategy 3 (*ownership*) | -0.07 | 0.04 | -1.73 | 0.08 | [-0.14, 0.01] |
| Identification Strategy 3 (*universities*) | 0.00 | 0.00 | 1.69 | 0.09 | [-3.17$e^{06}$, 0.00] |

Table 10: First Stage

| | **F-statistic** | **Prob ¿ F** |
|---|---|---|
| Identification Strategy 1 | 32.62 | 0.00 |
| Identification Strategy 2 | 32.61 | 0.00 |
| Identification Strategy 3 | 31.67 | 0.00 |

Table 11: F-Test of Excluded Instruments

| | $\hat{\theta}_0$ | **std err** | **z** | $P \geq \|z\|$ | **95% conf interval** |
|---|---|---|---|---|---|
| Identification Strategy 1 | 1.88 | 1.30 | 1.44 | 0.15 | [-0.68, 4.43] |
| Identification Strategy 2 | 1.27 | 1.10 | 1.16 | 0.25 | [-0.88, 3.43] |
| Identification Strategy 3 | 1.58 | 0.86 | 1.84 | 0.07 | [-0.10, 3.27] |
| OLS | 0.11 | 0.04 | 2.43 | 0.01 | [-0.02, 0.20] |

Table 12: Second Stage and OLS

Firstly, the first stage estimates are not statistically significant at the 5 %. However, the instruments are indeed valid as the F-tests of excluded instruments suggest. The F-statistic column is an F-statistic for the (joint) significance of the additional instruments. The column that displays the p-value indicates whether the F-statistic is significant. If the F-statistic is not statistically significant, it suggests that the additional instruments do not provide significant explanatory power when considering the influence of the other covariates. In other words, these instruments do not contribute significantly to explaining the variation in the outcome variable once the effects of the other variables have been taken into account (Angrist & Pischke, 2009, Chapter 4). In this case, the F-statistics are all statistically significant at the 5 % level (p-values: 0.00). Hall, Rudebusch and Wilcox (1996) have shown, through Monte Carlo simulation, that the F-statistic should not only be statistically significant but, as also Stock, Wright and Yogo (2002) suggest, the F-statistic should be larger than 10 for the excluded instruments to be considered strong enough (Hayashi, 2011, Chapter 5). In this case, the F-statistics are all larger than 30. Therefore, I can conclude that the instruments are strong even if the first-stage p-value does not account for statistical significance at the 5 % level, but only at the 10 % level (Cameron, Trivedi et al., 2010, Chapter 7). The fact that the coefficient of the instrument, especially that of *universities* (Identification Strategy 2 and Identification Strategy 3) which is equal to 0 when rounded to two decimals, is low does not invalidate these results. Rather, this means that the instrument does not have a large correlation with the endogenous variable . However, the instrument is still strong as both the F-test and the correlation coefficient (section 4) demonstrate. Moreover, Identification Strategy 3 uses both instruments and the corresponding first-stage coefficient of each instrument, which is displayed in Table 10, remains unchanged compared to the ones in Identification Strategy 1 and Identification Strategy 2, where

the instruments are considered separately. This further validates the over-identifying procedure in Identification Strategy 3 (Greene, 2003, Chapter 11) as it suggests that the two instruments do not interfere with each other when taken all together (Wooldridge, 2010, Chapter 15).

When comparing the results from DML to those of standard the instrumental variable setting, the the results are very similar. However, the traditional instrumental variable setting overestimates the standard error which may lead to incorrect inference as one may not consider an effect as causal when it actually is. In Identification Strategy 1 and Identification Strategy 2, DML and the standard instrumental variable setting agree with each other as both find that the effect of talent is not significant. Nevertheless, Identification Strategy 3 finds a statistically significant causal effect of talent on sales when considering DML in one out of the four machine learning techniques applied, whereas the standard instrumental variable setting does not yield such result. Naturally, this is because I have established *a priori* the significance level to be 5 %, whereas setting it to 10 % would have resulted in the same conclusion across DML and standard instrumental variable regression, namely that talent has a positive effect on sales. However, this suggests that DML, because of its bias correction procedure and because it allows for a complete overview through the comparison across different machine learning techniques, is superior to the traditional method. Finally, the results of the OLS regression are extremely different from those of DML as well as those of the standard instrumental variable setting. As expected, the variable *talent* is endogenous. For this reason, OLS finds a much lower, positive effect of talent on sales performance which is statistically significant at the 5 % level. This result is inconclusive and it is evidence that endogeneity must be taken care of in this scenario.

Overall, companies should indeed consider investing in talent management in order to increase sales. There seems to be support for the hypothesis that better talent management practices increase sales performance. This increase is around 6-7 % if a company that does not have any talent management practice in place decides to fully undertake the investment in talent development. However, there may exist less expensive and more efficient strategies in order to increase sales. Moreover, the relationship should be further investigated in order to validate the strength of the results of Identification Strategy 3 (section 5.5). In general, this advice would be of use for firms that aim at increasing sales because of ambition rather than necessity.

# 6 Conclusion

## 6.1 Limitations

Although this thesis has provided valuable insights into DML as well as into the relationship between talent management practices and sales, there are some limitations that are worth recognizing. In particular, the small sample size, potential omitted control variables affecting exogeneity, and the inconsistency in results amongst the machine learning methods are part of such limitations.

The data set that I use is a merge of two data sets, which, if considered singularly, contain a large number of observations (7094 and 9231 observations). The first data set contains information regarding the ownership of the firm and no information regarding the city where the headquarters of the firm is located, which I need in order to create the variable that describes the number of universities in such city. The second data set has no ownership information, but does contain the city where the headquarters of the firm are located. Therefore, the merge of the data sets is fundamental because it allows me to create both the *ownership* instrument and the *universities* instrument. In fact, using two different samples, one for each instrument, would not allow me to easily compare results across Identification Strategy 1 (section 5.2) and Identification Strategy 3 (section 5.3. Moreover, Identification Strategy 3 (section 5.5), which includes both instruments, would not be possible if the observations across the two data sets do not match. However, the intersection between the two data sets only includes 1512 common observations. Furthermore, excluding the observations that contain missing values results in 1189 observations left, which is the final sample that I use. Therefore, the sample size is relatively small. This induces the external validity of the findings to be less credible and it may reduce the statistical power of the resulting estimates (Cohen, 1992). Nevertheless, DML and machine learning techniques such as Lasso should be able to partially correct for the small sample size when dealing with a high-dimensional covariate space as in this case (Chernozhukov et al., 2018). Moreover, I have included the results for Identification Strategy 1 when using the full original sample in Table 15.

Although the control variables that I include seem sufficient, it is possible that additional factors influencing the relationship between talent management practices and sales have not been considered. This might induce bias and affect exogeneity, which DML should be able to partially correct for (Chernozhukov et al., 2018). However, the analysis would be strongly deteriorated if important confounding factors have not been accounted for (Angrist & Pischke, 2009, Chapter 2). Despite the fact that the omitted variable bias does not seem to be an issue in this case, additional covariates could have been gathered since the Post-Double Lasso selection of control variables would have ignored them if they eventually turned out to be unimportant (Belloni et al., 2014). In addition, it is possible that some of controls, for example the size of the firm, may also mediate the impact of talent management practices on sales. In this case, I would not be estimating the full effect of talent management practices on sales.

In this thesis, I use four machine learning methods: Lasso, Random Forest, Gradient Boosting, and Support Vector Machine. Although this is a strength of this research, it also reveals an inconsistency among the machine learning methods in the conclusions that are drawn about the

effect of talent management practices on sales performance in Identification Strategy 3 (section 5.5). Indeed, Identification Strategy 3 shows that only two of the four machine learning methods (Lasso and Random Forest) find a statistically significant effect of talent management practices on sales. Therefore, this research could have exploited replication of the analysis through an appropriate simulation method in order to obtain a more comprehensive overview of the relationship between talent management and sales.

Acknowledging the limitations of this thesis when evaluating the results is extremely important. Issues such as the small sample size, the potential omitted variable bias, and the inconsistency amongst the machine learning methods may introduce inaccuracy in the results. Despite my attempt to deal with such potential problems, further research in DML may confirm and expand upon the accuracy of the methods and the findings presented in this thesis.

## 6.2 Future Research

The research conducted in this thesis is of added value to the existing work on DML as well as to the body of literature regarding the relationship between talent management practices and sales performance. Nevertheless, future research can address the limitations presented in section 6.1 in many ways.

In order to improve the external validity as well as the statistical power of the results, future work may consider expanding the sample size. This implies considering a different data set or even collecting the data first hand in order to retrieve a single data set for both instruments instead of merging two data sets. Another possibility would be to use one of the two data sets that are employed in this research while varying one of the instruments in such a way that the instrument can be retrieved from the same data set as the first one. This would result in more robust and, in turn, reliable findings.

Considering more control variables would also be of benefit to the robustness and reliability of the results. Future research may further investigate whether there is omitted variable bias in the model that I use in this research as well as retrieve additional factors that may be included in the set of control variables. Overall, the exogeneity of the analysis should be taken into account and, possibly, improved. Moreover, it would be of interest to study whether some of the control variables mediate the effect of talent management practices on sales and, if this is the case, to correct for it. In this way, one would be estimating the full effect of talent management practices on sales. This would be a great addition, especially if the results differ greatly compared to the ones obtained in this thesis.

This thesis applies DML to the following machine learning techniques: Lasso, Random Forest, Gradient Boosting, and Support Vector Machine. Because the results from these four methods only partially agree with each other, future work may aim at replicating the analysis in order to find more coherent results across these methods. This would greatly improve the robustness of the findings. Another possibility would be to further investigate the assumptions behind each machine learning method and decide upon which ones provide more creditable results.

Moreover, this research uses Post-Double Lasso in order to choose the most appropriate control variables. However, this is not the only available method that can be employed for control variables selection. Future work may explore the application of alternative techniques, evaluate

the different models, and choose the best performing one. This would add onto the research on control variables selection, which is both fundamental in general and specifically to obtaining unbiased estimates when applying DML. For example, one may employ forward or backward step-wise selection which consists in iteratively adding or removing control variables from the model based on their statistical significance (Harrell, 2017). Another possibility would be to use Bayesian Model Averaging (BMA), a technique that computes a weighted average over different models, each considering a specific combination of control variables. The weights are computed based on the fit of the model to the data (Hoeting, Madigan, Raftery & Volinsky, 1999). In order to enhance variable parsimony, information-based criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used. These penalize models with a larger number of covariates. Therefore, researchers can choose the set of control variables that provides with the best trade-off between fitness to the data and complexity (Burnham & Anderson, 2004). In addition, one may substitute Lasso with Elastic Net in Post-Double selection. This might yield slightly different results since Elastic Net is a mixture of Ridge and Lasso (Hastie, Tibshirani & Wainwright, 2015, Chapter 3). This would require choosing two, rather than one, penalty parameters through cross validation. However, the overall approach would be similar to the one outlined in this thesis.

Lastly, future research might develop a longitudinal analysis in order to investigate the relationship between talent management and sales performance through different time points. This way, researchers would be able to study the dynamics of this relationship over time and its long-term effects (Singer, Willett, Willett et al., 2003, Chapter 1).

These extensions would be of great impact from both a scientific and a managerial perspective. By increasing the sample size, ensuring exogeneity, and, overall, expanding the research in the aforementioned directions, researchers would add onto the innovative body of literature on DML. Moreover, future research might find additional reliable, valuable insights that are useful for companies that are considering different managerial strategies to increase sales.

## 6.3 Concluding Remarks

In conclusion, this thesis is a valuable application of DML to the research on talent management as a way to increase sales. Despite the results being value-adding, it is crucial to be aware of the limitations of this analysis. When interpreting the results, one should acknowledge that the small sample size, the potential omitted variable bias, and the inconsistency issues outlined in section 6.1 might affect the robustness of the study. Nevertheless, the efforts made at correcting such issues provide a strong motif to consider the findings as reliable. Although the external validity of the results is to be further investigated, this thesis suggests that better talent management practices do increase sales. Future research might be able to confirm such finding by considering a large sample size, exploring different techniques, and comparing the performance of various models. Moreover, a longitudinal analysis can study how talent management practices evolve and impact the sales of a company over time. Therefore, despite the limitations, this research adds onto the body of literature regarding DML and talent management practices. Although one should be cautious in drawing definitive conclusions from these findings, this thesis is surely a promising start to a topic that is recently growing in popularity. Therefore, researchers should

continue investigating this topic as it is of great importance to businesses looking for a solid strategy to increase sales. Shedding light on the value of talent management is of benefit both to employers who strive for improved performance and employees who want their talent to be cultivated. Finally, DML provides a reliable approach for firms to make informed decisions on budget allocation to talent development.

## References


Ahrens, A., Aitken, C. & Schaffer, M. E. (2021). Using machine learning methods to support causal inference in econometrics. *Behavioral predictive modeling in economics*, 23–52.

Ahrens, A., Hansen, C. B., Schaffer, M. E. & Wiemann, T. (2023). ddml: Double/debiased machine learning in stata. *arXiv preprint arXiv:2301.09397*.

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Ansar, N., Baloch, A. et al. (2018). Talent and talent management: definition and issues. *IBT Journal of Business Studies (JBS)*, *1*(2).

Baum, C. F., Schaffer, M. E. & Stillman, S. (2003). Instrumental variables and gmm: Estimation and testing. *The Stata Journal*, *3*(1), 1–31.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Bhalla, V. & Bratton, J. (2015). Managing talent in a family business. *Boston Consulting Group. Available online: https://www. bcg. com/publications/2015/leadership-talent-human-resources-managing-talent-familybusiness. aspx (accessed on 10 September 2019)*.

Bloom, N., Genakos, C., Sadun, R. & Van Reenen, J. (2012). Management practices across firms and countries. *Academy of management perspectives*, *26*(1), 12–33.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, *33*(2), 261–304.

Cameron, A. C., Trivedi, P. K. et al. (2010). *Microeconometrics using stata* (Vol. 2). Stata press College Station, TX.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1-C68.

Chernozhukov, V., Newey, W. K. & Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, *90*(3), 967–1027.

Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychol. Bull.*, *112*, 1155–1159.

Collings, D. G. & Mellahi, K. (2009). Strategic talent management: A review and research agenda. *Human resource management review*, *19*(4), 304–313.

Corthoud, M. (2022). Post-double selection. *Machine Learning for Economics*.

Fonti, V. & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam research paper in business analytics*, *30*, 1–25.

Gallo, M. Á., Tàpies, J. & Cappuyns, K. (2004). Comparison of family and nonfamily business: Financial logic and personal preferences. *Family Business Review*, *17*(4), 303–318.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of consulting and clinical psychology*, *68*(1), 155.

Hall, A. R., Rudebusch, G. D. & Wilcox, D. W. (1996). Judging instrument relevance in instrumental variables estimation. *International Economic Review*, 283–298.

Harrell, F. E. (2017). Regression modeling strategies. *Bios*, *330*(2018), 14.

Hasan, M. A. M., Nasser, M., Ahmad, S. & Molla, K. I. (2016). Feature selection for intrusion detection using random forest. *Journal of information security*, *7*(3), 129–140.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hayashi, F. (2011). *Econometrics*. Princeton University Press.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science*, *14*(4), 382–417.

Hünermund, P., Louw, B. & Caspi, I. (2021). Double machine learning and automated confounder selection–a cautionary tale. *arXiv preprint arXiv:2108.11294*.

Ionescu-Ittu, R., Delaney, J. A. & Abrahamowicz, M. (2009). Bias–variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental variables: a simulation study. *Pharmacoepidemiology and drug safety*, *18*(7), 562–571.

Iqbal, Q. (2020). The era of environmental sustainability: Ensuring that sustainability stands on human resource management. *Global Business Review*, *21*(2), 377–391.

Iqbal, Q., Ahmad, N. H., Nasim, A. & Khan, S. A. R. (2020). A moderated-mediation analysis of psychological empowerment: Sustainable leadership and sustainable performance. *Journal of Cleaner Production*, *262*, 121429.

Kafetzopoulos, D. (2022). Talent development: a driver for strategic flexibility, innovativeness and financial performance. *EuroMed Journal of Business*(ahead-of-print).

Kaliannan, M., Darmalinggam, D., Dorasamy, M. & Abraham, M. (2022). Inclusive talent development as a key talent management approach: A systematic literature review. *Human Resource Management Review*, 100926.

Kehinde, J. S. (2011). Talent management: Effect on organizational performance. *Journal of Management Research*, *4*, 178-186.

Khan, E. A. & Quaddus, M. (2015). Development and validation of a scale for measuring sustainability factors of informal microenterprises–a qualitative and quantitative approach. *Entrepreneurship Research Journal*, *5*(4), 347–372.

Kiviet, J. F. & Kripfganz, S. (2021). Instrument approval by the sargan test and its consequences for coefficient estimation. *Economics Letters*, *205*, 109935.

Kozak, M. (2009). What is strong correlation? *Teaching Statistics*, *31*(3), 85–86.

Kreif, N. & DiazOrdaz, K. (2019). Machine learning in policy evaluation: new tools for causal inference. *arXiv preprint arXiv:1903.00402*.

Kwon, K. & Jang, S. (2022). There is no good war for talent: A critical review of the literature on talent management. *Employee Relations: The International Journal*, *44*(1), 94–120.

Lewis, R. E. & Heckman, R. J. (2006). Talent management: A critical review. *Human resource management review*, *16*(2), 139–154.

Lovell, M. C. (2008). A simple proof of the fwl theorem. *The Journal of Economic Education*, *39*(1), 88–91.

Morley, M. J., Scullion, H., Collings, D. G. & Schuler, R. S. (2015). Talent management: A capital question. *European Journal of International Management*, *9*(1), 1–8.

Newey, W. K. & Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric

estimation. *arXiv preprint arXiv:1801.09138*.

Pagan-Castaño, E., Ballester-Miquel, J. C., Sánchez-García, J. & Guijarro-García, M. (2022). What's next in talent management? *Journal of Business Research*, *141*, 528–535.

Proteasa, V. & Crăciun, D. (2020). The use of instrumental variables in higher education research. In *Theory and method in higher education research.* Emerald Publishing Limited.

Rabbi, F., Ahad, N., Kousar, T. & Ali, T. (2015, Oct.). Talent management as a source of competitive advantage. *Journal of Asian Business Strategy*, *5*(9), 208–214. doi: 10.18488/journal.1006/2015.5.9/1006.9.208.214

Singer, J. D., Willett, J. B., Willett, J. B. et al. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Sparrow, P., Scullion, H. & Tarique, I. (2014). Multiple lenses on talent management: Definitions and contours of the field.

Stock, J. H., Wright, J. H. & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, *20*(4), 518–529.

Tarique, I. & Schuler, R. S. (2010). Global talent management: Literature review, integrative framework, and suggestions for further research. *Journal of world business*, *45*(2), 122–133.

Urminsky, O., Hansen, C. & Chernozhukov, V. (2016). Using double-lasso regression for principled variable selection. *Available at SSRN 2733374*.

Wang, K. & Shailer, G. (2015). Ownership concentration and firm performance in emerging markets: A meta-analysis. *Journal of Economic Surveys*, *29*(2), 199–229.

Williamson, B. D., Gilbert, P. B., Simon, N. R. & Carone, M. (2021). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 1–14.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wright, P. M., Mcmahan, G. C. & McWilliams, A. (1994). Human resources and sustained competitive advantage: a resource-based perspective. *International Journal of Human Resource Management*, *5*, 301-326.

# A Additional material

This appendix includes additional information on the data. Namely, Table 13 provides descriptive statistics for all the control variables included in the data set. Moreover, I provide a description of the meaning of each variable in Table 14. In addition, Figure **??** is the full pair-wise correlation table of the variables, which motivates why I have excluded some control variables from the model *a priori*.

Table 15 displays the results of Identification Strategy 1 when using the original data set. As the original data set for Identification Strategy 2 does not include data on sales, I cannot replicate the analysis only using this full data set. In fact, I can only use in my model the observations that are in common with the original data set for Identification Strategy 1 because the latter includes the sales variable.

| | **count** | **mean** | **std** | **min** | **max** |
|---|---|---|---|---|---|
| **firm id** | 1189 | 1712.138 | 778.335 | 1 | 2927 |
| **perf** | 1189 | 3.160667 | 0.7367801 | 1 | 5 |
| **factor management** | 1093 | -0.0037842 | 1.041617 | -3.117611 | 2.761346 |
| **ldegree** | 1189 | 1.948448 | 1.32308 | -2.995732 | 4.510859 |
| **firmage** | 1189 | 50.23465 | 48.77487 | 0 | 361 |
| **lean** | 1189 | 2.965459 | 1.052614 | 1 | 5 |
| **lppent** | 1189 | 9.420793 | 1.633486 | 2.944439 | 15.2441 |
| **lemp** | 1189 | 5.672413 | 1.120009 | 1.098612 | 10.95209 |
| **management** | 1189 | 3.018285 | 0.662965 | 1.111111 | 4.888889 |
| **mne d** | 1189 | 0.2809083 | 0.4496319 | 0 | 1 |
| **mne f** | 1189 | 0.3246426 | 0.4684382 | 0 | 1 |
| **roce** | 1189 | 14.80801 | 16.17594 | -25 | 50 |
| **dead** | 1189 | 0.0210261 | 0.1435315 | 0 | 1 |
| **year** | 1189 | 2007.173 | 1.79372 | 2004 | 2010 |
| **sic** | 1189 | 304.4609 | 59.10506 | 201 | 399 |
| **dow** | 1189 | 2.948696 | 1.388344 | -1 | 6 |
| **reliability** | 1189 | 8.124474 | 1.623579 | 4 | 10 |
| **i comptenure** | 1189 | 13.81665 | 10.56987 | 0 | 50 |
| **i seniority** | 1189 | 2.926829 | 0.9038983 | 0 | 5 |
| **argentina** | 1189 | 0.000841 | 0.0290007 | 0 | 1 |
| **australia** | 1189 | 0.0042052 | 0.0647384 | 0 | 1 |
| **brazil** | 1189 | 0.0016821 | 0.040996 | 0 | 1 |
| **chile** | 1189 | 0.000841 | 0.0290007 | 0 | 1 |
| **china** | 1189 | 0.0227082 | 0.1490342 | 0 | 1 |
| **france** | 1189 | 0.1236333 | 0.3293012 | 0 | 1 |
| **germany** | 1189 | 0.058873 | 0.2354859 | 0 | 1 |
| **UK** | 1189 | 0.2068966 | 0.4052511 | 0 | 1 |
| **greece** | 1189 | 0.0908326 | 0.2874919 | 0 | 1 |
| **italy** | 1189 | 0.0824222 | 0.2751226 | 0 | 1 |
| **japan** | 1189 | 0.0563499 | 0.2306931 | 0 | 1 |
| **northern ireland** | 1189 | 0.0218671 | 0.1463111 | 0 | 1 |
| **poland** | 1189 | 0.0866274 | 0.2814067 | 0 | 1 |
| **portugal** | 1189 | 0.0807401 | 0.2725502 | 0 | 1 |
| **republic of ireland** | 1189 | 0.0252313 | 0.1568929 | 0 | 1 |
| **sweden** | 1189 | 0.0916737 | 0.2886862 | 0 | 1 |
| **US** | 1189 | 0.0445753 | 0.2064562 | 0 | 1 |

Table 13: Descriptive Statistics

| variable | description |
|---|---|
| **firm id** | firm identifying ID |
| **perf** | average of 10 performance indicators (score from 1 to 5) |
| **factor management** | management quality indicator (factorized) |
| **ldegree** | logarithm of the number of managers in the firm that have a university diploma |
| **firmage** | age of the firm |
| **lean** | average of 6 indicators relative to the extent to which lean practices have been adopted by the firm |
| **lppent** | logarithm of the amount of long-tern, tangible assets (e.g., trucks, machinery) that the firm owns |
| **lemp** | logarithm of the number of employees in a the firm |
| **management** | management quality indicator |
| **mne d** | dummy variable that equals 1 if the firm has headquarters located in the same country as the location where the firm originated |
| **mne f** | dummy variable indicating that equals 1 if the firm has headquarters located in a foreign country compared to the location where the firm originated |
| **roce** | Return On Capital Employed |
| **dead** | dummy variable that equals 1 if the owner of the firm is alive |
| **year** | year in which the firm has begun implementing lean practices |
| **sic** | code indicating the line of business of a firm (Standard Industrial Classification) |
| **dow** | market performance indicator |
| **reliability** | indicator of reliability of the firm |
| **i comptenure** | number of years that the current CEO has been in place |
| **i seniority** | number of years before an employee can be automatically promoted, not based on merit |
| *country name* | dummy variable that equals 1 if the origin of the firm is in this country |

Table 14: Variable Description

| | firmid | firmage | nanagement | mne_d | mne_f | nanagement | ldegree_t | lemp | lppent | lean | perf | roce | dead | year | sic | dow | reliability | _comptenure | i_seniority |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| firmid | 1 | | | | | | | | | | | | | | | | | | |
| firmage | -0.11** | 1 | | | | | | | | | | | | | | | | | |
| nanagement | -0.05 | 0.03 | 1 | | | | | | | | | | | | | | | | |
| mne_d | -0.14** | 0.04 | 0.07** | 1 | | | | | | | | | | | | | | | |
| mne_f | 0.01 | -0.09** | 0.27** | -0.43** | 1 | | | | | | | | | | | | | | |
| nanagement | -0.05 | 0.03 | 1** | 0.06** | 0.27** | 1 | | | | | | | | | | | | | |
| ldegree_t | 0.13** | -0.01 | 0.17** | -0.01 | 0 | 0.17** | 1 | | | | | | | | | | | | |
| lemp | -0.23** | 0.11** | 0.26** | 0.2** | 0 | 0.26** | 0.04 | 1 | | | | | | | | | | | |
| lppent | -0.07** | 0.13** | 0.27** | 0.12** | 0.06 | 0.28** | 0.1** | 0.72** | 1 | | | | | | | | | | |
| lean | -0.06** | 0.03 | 0.71** | 0.07** | 0.22** | 0.74** | 0.08** | 0.22** | 0.2** | 1 | | | | | | | | | |
| perf | -0.04 | 0.04 | 0.95** | 0.06** | 0.26** | 0.96** | 0.16** | 0.25** | 0.27** | 0.61** | 1 | | | | | | | | |
| roce | -0.08** | 0 | 0.1** | 0.04 | 0.06** | 0.1** | -0.01 | 0.07** | -0.05 | 0.09** | 0.08** | 1 | | | | | | | |
| dead | -0.02 | -0.03 | -0.05 | 0.01 | -0.03 | -0.04 | -0.09** | -0.06** | -0.05 | -0.03 | -0.04 | -0.09** | 1 | | | | | | |
| year | 0.14** | 0.01 | 0.04 | -0.08** | 0.04 | 0.02 | 0.1** | -0.08** | 0 | 0.04 | 0.04 | -0.14** | -0.07** | 1 | | | | | |
| sic | -0.08** | -0.12** | 0.13** | 0.11** | 0.05** | 0.12** | 0.12** | 0.06** | -0.13** | 0.12** | 0.09** | 0.14** | -0.02 | -0.02 | 1 | | | | |
| dow | 0 | 0.01 | 0.03 | 0 | -0.01 | 0.04 | 0.01 | 0.02 | 0 | 0.05 | 0.04 | 0.02 | 0.01 | -0.04 | 0.03 | 1 | | | |
| reliability | 0.13** | -0.02 | 0.28** | 0.04 | 0.07** | 0.29** | 0.12** | 0.04 | 0.05 | 0.2** | 0.3** | 0.06** | -0.05 | -0.02 | 0.04 | 0.01 | 1 | | |
| _comptenure | 0.07** | 0.09** | 0.01 | -0.01 | 0.01 | 0.03 | 0.05 | 0.05 | 0.04 | -0.02 | 0.01 | 0 | -0.08** | 0.08** | 0.08** | -0.05 | 0.02 | 1 | |
| i_seniority | -0.02 | 0.06** | -0.02 | -0.01 | 0.07** | -0.01 | -0.04 | -0.01 | 0.01 | 0.01 | -0.02 | -0.01 | 0.02 | 0.1** | 0.03 | -0.05 | -0.13** | 0.01 | 1 |

Figure 10: Pairwise Correlation Table

| Method | $\hat{\theta}_0$ | std err | z | $P \geq \|z\|$ | 95% conf interval |
|---|---|---|---|---|---|
| Lasso | 1.48 | 1.16 | 1.27 | 0.20 | [-0.80, 3.76] |
| Forest | 2.91 | 1.75 | 1.66 | 0.10 | [-0.53, 6.34] |
| GB | 1.15 | 1.12 | 1.03 | 0.30 | [-1.04, 3.35] |
| SVM | 1.48 | 1.19 | 1.24 | 0.21 | [-0.85, 3.82] |

Table 15: Full Sample Results Identification Strategy 1

# B Programming code

This appendix includes the code that I use. I use Python (section B.1 and section B.2) for the data pre-processing as well as for DML. The DML code in python serves as an example on how the procedure works in detail. It covers Application 1 and Application 2. This code is an adaptation of the procedure outlined by Corthoud (2022). For the DML results outlined in the paper, I use Stata (section B.3), which conveniently includes a DML package. Post-Double Lasso selection has also been conducted in Stata.

## B.1 Python code Instrument 1

```
import numpy as np
import pandas as pd
import random
from sklearn.preprocessing import PolynomialFeatures
from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LassoCV, Lasso, Ridge
import matplotlib.pyplot as plt
from sklearn import svm
from xgboost import XGBRegressor
import io
from scipy.stats import pearsonr

# Data exploration
with open('AMP_table.csv', 'r', encoding='utf-8', errors='replace') as f:
    content = f.read()

data = pd.read_csv(io.StringIO(content), delimiter=",")

print(data.head())

print(data.describe())

#print(data.columns)
#distinct_values = data['ownership'].unique()
```

```
#print(distinct_values)

# Create ownership dummy
data['ownership'] = data['ownership'].replace({'Dispersed Shareholders': int(0),
'Founder': int(0), 'Private Equity': int(0), 'Private Individuals': int(0),
'Managers': int(0), 'Government': int(0), 'Other': int(0)})
data['ownership'] = data['ownership'].replace({'Family owned, family CEO': int(1),
'Family owned, external CEO': int(1)})

# Use merged data because you will use the same data set when using different IV
with open('thesismerge.csv', 'r', encoding='utf-8', errors='replace') as f:
    content = f.read()

data = pd.read_csv(io.StringIO(content), delimiter=";")
data = data[(data["_merge"] == "Matched (3)")]

print(data.head())

print(data.describe())

# Count the number of family owned versus not
value_counts = data['ownership'].value_counts()
print(value_counts)
data['ownership'].value_counts().plot(kind='bar')
plt.title('Family_owned, data')
plt.xlabel('ownership')
plt.ylabel('count')
#plt.show()

# Create dummies for countries (baseline: UK)
value_countscountry = data['country'].value_counts()
#print(value_countscountry)
dummies = pd.get_dummies(data["country"])
data = pd.concat([data, dummies], axis=1)

print(data.head())

print(data.describe())

# Aggregate indicators
data["talent"] = data[['talent1', 'talent2', 'talent3', "talent4", "talent5",
"talent6"]].mean(axis=1)
```

```
data["lean"] = data[['lean1', 'lean2']].mean(axis=1)
data["perf"] = data[['perf1', 'perf2', 'perf3', "perf4", "perf5", "perf6", "perf7",
"perf8", "perf9", "perf10"]].mean(axis=1)

sns.displot(data, x="talent", hue="ownership", col="ownership",
            kind="kde", fill=True)
#plt.show()

print(data[['ownership', 'lsales']].groupby('ownership').mean().diff())
data = data.reset_index(drop = True)
data['previous_B'] = data['uni'].shift(1)
mask = data['country'] == data['country'].shift(1)
data.loc[mask, 'uni'] = data['previous_B']
data = data.dropna(subset=['uni'])
data = data.drop('previous_B', axis=1)
print(data.head())
data['previous_B'] = data['firmage'].shift(1)
data.loc[mask, 'firmage'] = data['previous_B']
data = data.dropna(subset=['firmage'])
data = data.drop('previous_B', axis=1)
data['previous_B'] = data['plantage'].shift(1)
data.loc[mask, 'plantage'] = data['previous_B']
data = data.drop('previous_B', axis=1)
data = data[data["ownership"].notnull()]
print(data.head())

print(data.describe())

# Validity of instrument
corr_coef, p_value = pearsonr(data['talent'], data['ownership'])

print('Correlation coefficient:', corr_coef)
print('p-value:', p_value)

print("Sales:")
print(data["lsales"].head())
print(data["lsales"].describe())
print("Talent:")
print(data["talent"].head())
print(data["talent"].describe())
print("Ownership:")
print(data["ownership"].head())
```

```
print(data["ownership"].describe())

data.to_csv("datathesis22.csv")

# Plot distribution of talent and sales
data['talent'].plot(kind='kde')
plt.xlabel('talent')
plt.show()

data['lsales'].plot(kind='kde')
plt.xlabel('logarithm of sales')
plt.show()

# control variables found through stata: did a first screening myself to avoid correlation
# to create issues in lasso.
# then used post double lasso selection on stata
# ddml: NOTE: THE OFFICIAL RESULTS THAT I USE ARE GENERATED IN STATA, but this is a nice
# overview of how a manual ddml algorithm would work
# Generate variables
# Add constant term to dataset
#data = data.fillna(0) #NaN corresponds to a 0 score
data['const'] = 1
D = data['talent'].values.reshape(-1,1)
X = data[['const', 'firmid', 'firmage', 'mne_f', 'ldegree_t', 'perf', 'roce', 'sic', 'dow',
'lemp', 'dead', 'year', 'reliability', 'i_comptenure', 'i_seniority', 'Australia', 'Brazil',
'Germany', 'Greece', 'Italy', 'Northern Ireland', 'China', 'Republic of Ireland', 'France',
'Japan', 'Poland', 'Portugal', 'Sweden', 'United States']].values
y = data['lsales'].values.reshape(-1,1)
Z = data[['const', 'firmid', 'firmage', 'mne_f', 'ldegree_t', 'perf', 'roce', 'sic', 'dow',
'lemp', 'dead', 'year', 'reliability', 'i_comptenure', 'i_seniority', 'Australia', 'Brazil',
'Germany', 'Greece', 'Italy', 'Northern Ireland', 'China', 'Republic of Ireland', 'France',
'Japan', 'Poland', 'Portugal', 'Sweden', 'United States', 'ownership']].values

thetas = np.zeros(shape=[1000,1])
coefs = np.zeros(shape=[1000,29])
for i in range(1000):
 I = np.random.choice(1189, int(1189/2),replace=False)
 I2 = [x for x in np.arange(1189) if x not in I]
 reg = LassoCV(max_iter=1000)
 G1 = reg.fit(X[I], y[I]).predict(X[I2])
 G2 = reg.fit(X[I], y[I]).predict(X[I])
 for j in range(29):
```

```
    coefs[i][j] = reg.coef_[j]
 M1 = reg.fit(Z[I], D[I]).predict(Z[I2])
 M2 = reg.fit(Z[I], D[I]).predict(Z[I])
 V1 = D[I2] - M1
 V2 = D[I] - M2
 theta1 = np.mean(np.dot(V1,(y[I2]-G1)))/np.mean(np.dot(V1,D[I2]))
 theta2 = np.mean(np.dot(V2,(y[I]-G2)))/np.mean(np.dot(V2,D[I]))
 thetas[i][0] = 0.5*(theta1+theta2)

coef2 = np.mean(coefs, axis = 0)
print("Coefficients: ", coef2)

print("Lasso: ", np.mean(thetas))

thetas = np.zeros(shape=[1000,1])
imp = np.zeros(shape=[1000,29])
for i in range(1000):
 I = np.random.choice(1189, int(1189/2),replace=False)
 I2 = [x for x in np.arange(1189) if x not in I]
 reg = RandomForestRegressor(max_depth = 2)
 G1 = reg.fit(X[I], y[I]).predict(X[I2])
 G2 = reg.fit(X[I], y[I]).predict(X[I])
 for j in range(29):
     imp[i][j] = reg.feature_importances_[j]
 M1 = reg.fit(Z[I], D[I]).predict(Z[I2])
 M2 = reg.fit(Z[I], D[I]).predict(Z[I])
 V1 = D[I2] - M1
 V2 = D[I] - M2
 theta1 = np.mean(np.dot(V1,(y[I2]-G1)))/np.mean(np.dot(V1,D[I2]))
 theta2 = np.mean(np.dot(V2,(y[I]-G2)))/np.mean(np.dot(V2,D[I]))
 thetas[i][0] = 0.5*(theta1+theta2)

imp2 = np.mean(imp, axis = 0)
print("Variables Importance: ", imp2)

print("Forest: ", np.mean(thetas))

thetas = np.zeros(shape=[1000,1])
for i in range(1000):
 I = np.random.choice(1189, int(1189/2),replace=False)
 I2 = [x for x in np.arange(1189) if x not in I]
 G1 = svm.SVR().fit(X[I], y[I]).predict(X[I2])
```

```
 G2 = svm.SVR().fit(X[I], y[I]).predict(X[I])
 M1 = svm.SVR().fit(Z[I], D[I]).predict(Z[I2])
 M2 = svm.SVR().fit(Z[I], D[I]).predict(Z[I])
 V1 = D[I2]-M1
 V2 = D[I] - M2
 theta1 = np.mean(np.dot(V1,(y[I2]-G1)))/np.mean(np.dot(V1,D[I2]))
 theta2 = np.mean(np.dot(V2,(y[I]-G2)))/np.mean(np.dot(V2,D[I]))
 thetas[i][0] = 0.5*(theta1+theta2)

print("SVM: ", np.mean(thetas))

thetas = np.zeros(shape=[1000,1])
for i in range(1000):
 I = np.random.choice(1189, int(1189/2),replace=False)
 I2 = [x for x in np.arange(1189) if x not in I]
 G1 = XGBRegressor().fit(X[I], y[I]).predict(X[I2])
 G2 = XGBRegressor().fit(X[I], y[I]).predict(X[I])
 M1 = XGBRegressor().fit(Z[I], D[I]).predict(Z[I2])
 M2 = XGBRegressor().fit(Z[I], D[I]).predict(Z[I])
 V1 = D[I2]-M1
 V2 = D[I] - M2
 theta1 = np.mean(np.dot(V1,(y[I2]-G1)))/np.mean(np.dot(V1,D[I2]))
 theta2 = np.mean(np.dot(V2,(y[I]-G2)))/np.mean(np.dot(V2,D[I]))
 thetas[i][0] = 0.5*(theta1+theta2)

print("GB: ", np.mean(thetas))
```

### B.2 Python code Instrument 2

```
import numpy as np
import pandas as pd
import random
from sklearn.preprocessing import PolynomialFeatures
from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LassoCV, Lasso, Ridge
import matplotlib.pyplot as plt
from sklearn import svm
from xgboost import XGBRegressor
import io
from scipy.stats import pearsonr
```

```
# Data exploration
with open('thesismerge.csv', 'r', encoding='utf-8', errors='replace') as f:
    content = f.read()

data = pd.read_csv(io.StringIO(content), delimiter=";")
data = data[(data["_merge"] == "Matched (3)")]

print(data.head())

print(data.describe())

#print(data.columns)
#distinct_values = data['uni'].unique()
#print(distinct_values)

# Count the number of family owned versus not (before merge)
value_counts = data['uni'].value_counts()
print(value_counts)
data['uni'].value_counts().plot(kind='bar')
plt.title('Number of universities, data')
plt.xlabel('number of universities')
plt.ylabel('count')
plt.show()

# Create dummies for countries (baseline: UK)
value_countscountry = data['country'].value_counts()
#print(value_countscountry)
dummies = pd.get_dummies(data["country"])
data = pd.concat([data, dummies], axis=1)

# Aggregate indicators
data["talent"] = data[['talent1', 'talent2', 'talent3', "talent4", "talent5",
"talent6"]].mean(axis=1)
data["lean"] = data[['lean1', 'lean2']].mean(axis=1)
data["perf"] = data[['perf1', 'perf2', 'perf3', "perf4", "perf5", "perf6", "perf7",
"perf8", "perf9", "perf10"]].mean(axis=1)

#data = data.fillna(0) #NaN corresponds to a 0 score/0 unis

data = data.reset_index(drop = True)
data['previous_B'] = data['uni'].shift(1)
mask = data['country'] == data['country'].shift(1)
```

```
data.loc[mask, 'uni'] = data['previous_B']
data = data.dropna(subset=['uni'])
data = data.drop('previous_B', axis=1)
data["uni"] = data["uni"].astype(int)
data["uni"] = data["uni"]**2
data['previous_B'] = data['firmage'].shift(1)
data.loc[mask, 'firmage'] = data['previous_B']
data = data.dropna(subset=['firmage'])
data = data.drop('previous_B', axis=1)
data['previous_B'] = data['plantage'].shift(1)
data.loc[mask, 'plantage'] = data['previous_B']
data = data.drop('previous_B', axis=1)
data = data[data["ownership"].notnull()]
print(data.head())

print(data.describe())

# Validity of instrument
corr_coef, p_value = pearsonr(data['talent'], data['uni'])

print('Correlation coefficient:', corr_coef)
print('p-value:', p_value)

data['uni'].plot(kind='kde')
plt.xlabel('universities')
plt.xlim(0,18225)
plt.show()
np.sqrt(data['uni']).plot(kind='kde')
plt.xlabel('universities')
plt.xlim(0,np.sqrt(18225))
plt.show()

print("Sales:")
print(data["lsales"].head())
print(data["lsales"].describe())
print("Talent:")
print(data["talent"].head())
print(data["talent"].describe())
print("Number of Universities:")
print(data["uni"].head())
print(data["uni"].describe())
```

```
data.to_csv("datathesis22.csv")

# Count the number of family owned versus not (merge)
value_counts = data['ownership'].value_counts()
print(value_counts)
data['ownership'].value_counts().plot(kind='bar')
plt.title('Family_owned, data')
plt.xlabel('ownership')
plt.ylabel('count')
plt.show()

sns.displot(data, x="talent", hue="ownership", col="ownership",
            kind="kde", fill=True)
plt.show()

# control variables found through stata: did a first screening myself to avoid correlation
# to create issues in lasso.
# then used post double lasso selection on stata
# ddml: NOTE: THE OFFICIAL RESULTS THAT I USE ARE GENERATED IN STATA, but this is a nice
# overview of how a manual ddml algorithm would work
# Generate variables
# Add constant term to dataset
data['const'] = 1
D = data['talent'].values.reshape(-1,1)
X = data[['const', 'firmid', 'firmage', 'mne_f', 'ldegree_t', 'perf', 'roce', 'sic', 'dow',
'lemp', 'dead', 'year', 'reliability', 'i_comptenure', 'i_seniority', 'Australia', 'Brazil',
'Germany', 'Greece', 'Italy', 'Northern Ireland', 'China', 'Republic of Ireland', 'France',
'Japan', 'Poland', 'Portugal', 'Sweden', 'United States']].values
y = data['lsales'].values.reshape(-1,1)
Z = data[['const', 'firmid', 'firmage', 'mne_f', 'ldegree_t', 'perf', 'roce', 'sic', 'dow',
'lemp', 'dead', 'year', 'reliability', 'i_comptenure', 'i_seniority', 'Australia', 'Brazil',
'Germany', 'Greece', 'Italy', 'Northern Ireland', 'China', 'Republic of Ireland', 'France',
'Japan', 'Poland', 'Portugal', 'Sweden', 'United States', 'uni']].values

thetas = np.zeros(shape=[1000,1])
coefs = np.zeros(shape=[1000,29])
for i in range(1000):
 I = np.random.choice(1189, int(1189/2),replace=False)
 I2 = [x for x in np.arange(1189) if x not in I]
 reg = LassoCV(max_iter=1000)
 G1 = reg.fit(X[I], y[I]).predict(X[I2])
 G2 = reg.fit(X[I], y[I]).predict(X[I])
```

```
 for j in range(29):
     coefs[i][j] = reg.coef_[j]
 M1 = reg.fit(Z[I], D[I]).predict(Z[I2])
 M2 = reg.fit(Z[I], D[I]).predict(Z[I])
 V1 = D[I2] - M1
 V2 = D[I] - M2
 theta1 = np.mean(np.dot(V1,(y[I2]-G1)))/np.mean(np.dot(V1,D[I2]))
 theta2 = np.mean(np.dot(V2,(y[I]-G2)))/np.mean(np.dot(V2,D[I]))
 thetas[i][0] = 0.5*(theta1+theta2)

coef2 = np.mean(coefs, axis = 0)
print("Coefficients: ", coef2)

print("Lasso: ", np.mean(thetas))

thetas = np.zeros(shape=[1000,1])
imp = np.zeros(shape=[1000,29])
for i in range(1000):
 I = np.random.choice(1189, int(1189/2),replace=False)
 I2 = [x for x in np.arange(1189) if x not in I]
 reg = RandomForestRegressor(max_depth = 2)
 G1 = reg.fit(X[I], y[I]).predict(X[I2])
 G2 = reg.fit(X[I], y[I]).predict(X[I])
 for j in range(29):
     imp[i][j] = reg.feature_importances_[j]
 M1 = reg.fit(Z[I], D[I]).predict(Z[I2])
 M2 = reg.fit(Z[I], D[I]).predict(Z[I])
 V1 = D[I2] - M1
 V2 = D[I] - M2
 theta1 = np.mean(np.dot(V1,(y[I2]-G1)))/np.mean(np.dot(V1,D[I2]))
 theta2 = np.mean(np.dot(V2,(y[I]-G2)))/np.mean(np.dot(V2,D[I]))
 thetas[i][0] = 0.5*(theta1+theta2)

imp2 = np.mean(imp, axis = 0)
print("Variables Importance: ", imp2)

print("Forest: ", np.mean(thetas))

thetas = np.zeros(shape=[1000,1])
for i in range(1000):
 I = np.random.choice(1189, int(1189/2),replace=False)
 I2 = [x for x in np.arange(1189) if x not in I]
```

```
 G1 = svm.SVR().fit(X[I], y[I]).predict(X[I2])
 G2 = svm.SVR().fit(X[I], y[I]).predict(X[I])
 M1 = svm.SVR().fit(Z[I], D[I]).predict(Z[I2])
 M2 = svm.SVR().fit(Z[I], D[I]).predict(Z[I])
 V1 = D[I2]-M1
 V2 = D[I] - M2
 theta1 = np.mean(np.dot(V1,(y[I2]-G1)))/np.mean(np.dot(V1,D[I2]))
 theta2 = np.mean(np.dot(V2,(y[I]-G2)))/np.mean(np.dot(V2,D[I]))
 thetas[i][0] = 0.5*(theta1+theta2)

print("SVM: ", np.mean(thetas))

thetas = np.zeros(shape=[1000,1])
for i in range(1000):
 I = np.random.choice(1189, int(1189/2),replace=False)
 I2 = [x for x in np.arange(1189) if x not in I]
 G1 = XGBRegressor().fit(X[I], y[I]).predict(X[I2])
 G2 = XGBRegressor().fit(X[I], y[I]).predict(X[I])
 M1 = XGBRegressor().fit(Z[I], D[I]).predict(Z[I2])
 M2 = XGBRegressor().fit(Z[I], D[I]).predict(Z[I])
 V1 = D[I2]-M1
 V2 = D[I] - M2
 theta1 = np.mean(np.dot(V1,(y[I2]-G1)))/np.mean(np.dot(V1,D[I2]))
 theta2 = np.mean(np.dot(V2,(y[I]-G2)))/np.mean(np.dot(V2,D[I]))
 thetas[i][0] = 0.5*(theta1+theta2)

print("GB: ", np.mean(thetas))
```

### B.3 Stata code

```
import delimited datathesis22.csv

*descriptive stats
summarize

*correlation table
pwcorr firmid firmage management mne_d mne_f factor_management ldegree_t
lemp lppent lean perf roce dead year sic dow reliability i_comptenure
i_seniority argentina australia brazil canada chile china france
germany greece italy japan northernireland poland portugal republicofireland
sweden unitedstates, star(.05)

*controls selection -- manual
```

```
*first reg
lasso linear talent firmid firmage mne_f ldegree_t
lemp perf roce dead year sic dow reliability i_comptenure
i_seniority argentina australia brazil canada chile china france
germany greece italy japan northernireland poland portugal republicofireland
sweden unitedstates, stop(0)
estimates store mylassotalent
lassocoef
cvplot

*second reg
lasso linear lsales firmid firmage mne_f ldegree_t
lemp perf roce dead year sic dow reliability i_comptenure
i_seniority argentina australia brazil canada chile china france
germany greece italy japan northernireland poland portugal republicofireland
sweden unitedstates, stop(0)
estimates store mylassosales
lassocoef
cvplot

*compare
lassocoef mylassotalent mylassosales
lassocoef mylassotalent mylassosales, display(coef, postselection)
*to keep:
*firmid firmage mne_f ldegree perf roce sic dow lemp dead year reliability i_comptenure
i_seniority australia brazil germany greece italy northernireland china
republicofireland france japan poland portugal sweden unitedstates

*control selection -- package
dsregress lsales talent, controls(firmid plantage mne_d mne_f ldegree_t lemp perf
roce dead year sic dow reliability i_comptenure i_seniority argentina australia
brazil canada chile china france germany greece italy japan northernireland poland
portugal republicofireland sweden unitedstates) selection(cv)

*ddml packages
ssc install ddml
ssc install pystacked

*defining iv setting: IV: ownership (application 1)
global Y lsales
global X firmid firmage mne_f ldegree_t perf roce sic dow lemp dead year reliability
i_comptenure i_seniority australia brazil germany greece italy northernireland china
```

```
republicofireland france japan poland portugal sweden unitedstates
global D talent
global Z ownership

*results: random forest, lasso (cv), gradient boosting, support vector machine
ddml init iv
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(rf))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(lassocv))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(gradboost))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(svm))

*defining iv setting: IV: number of unis (application 2)
global Y lsales
global X firmid firmage mne_f ldegree_t perf roce sic dow lemp dead year reliability
i_comptenure i_seniority australia brazil germany greece italy northernireland china
republicofireland france japan poland portugal sweden unitedstates
global D talent
global Z uni

*results: random forest, lasso (cv), gradient boosting, support vector machine
ddml init iv
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(rf))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(lassocv))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(gradboost))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(svm))

*defining iv setting: IV: overidentification: ownership and number of unis (application 3)
global Y lsales
global X firmid firmage mne_f ldegree_t perf roce sic dow lemp dead year reliability
i_comptenure i_seniority australia brazil germany greece italy northernireland china
republicofireland france japan poland portugal sweden unitedstates
global D talent
global Z uni ownership

*results: random forest, lasso (cv), gradient boosting, support vector machine
ddml init iv
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(rf))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(lassocv))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(gradboost))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(svm))

*without instruments
```

```
ddml init partial
qddml $Y $D ($X), model(partial) cmdopt(method(rf))
qddml $Y $D ($X), model(partial) cmdopt(method(lassocv))
qddml $Y $D ($X), model(partial) cmdopt(method(gradboost))
qddml $Y $D ($X), model(partial) cmdopt(method(svm))

*standard iv
ivregress 2sls lsales firmid firmage mne_f ldegree_t perf roce sic dow lemp dead year
reliability i_comptenure i_seniority australia brazil germany greece italy northernireland
china republicofireland france japan poland portugal sweden unitedstates
(talent = ownership), first

*hansen sargan
ivregress 2sls lsales firmid firmage mne_f ldegree_t perf roce sic dow lemp dead year reliab

ivregress 2sls lsales firmid firmage mne_f ldegree_t perf roce sic dow lemp dead year
reliability i_comptenure i_seniority australia brazil germany greece italy northernireland
china republicofireland france japan poland portugal sweden unitedstates (talent = uni)
, first

*hansen sargan
ivregress 2sls lsales firmid firmage mne_f ldegree_t perf roce sic dow lemp dead year reliab

ivregress 2sls lsales firmid firmage mne_f ldegree_t perf roce sic dow lemp dead year
reliability i_comptenure i_seniority australia brazil germany greece italy northernireland
china republicofireland france japan poland portugal sweden unitedstates
(talent = uni ownership), first

*OLS
reg lsales talent firmid firmage mne_f ldegree_t perf roce sic dow lemp dead year
reliability i_comptenure i_seniority australia brazil germany greece italy northernireland
china republicofireland france japan poland portugal sweden unitedstates

*with original data (appendix)
clear
import delimited datathesisown.csv

*defining iv setting
global Y lsales
global X firmid mne_f ldegree_t perf roce sic dow lemp dead year reliability i_comptenure i_
global D talent
global Z ownership
```

```
*results: random forest, lasso (cv), gradient boosting, support vector machine
ddml init iv
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(rf))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(lassocv))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(gradboost))
qddml $Y ($X) ($D = $Z), model(iv) cmdopt(method(svm))
```