RESEARCH REPORT
STUDY PAPER 220
SEPTEMBER 2024

Uncovering sources of heterogeneity in the child penalty: Conventional versus data-driven approaches

Laura Weile, Caroline Rygaard, Nabanita Datta Gupta,
Astrid Würtz Rasmussen and Nina Smith

THE ROCKWOOL FOUNDATION
RESEARCH

STUDY PAPER 220 SEPTEMBER 2024

Uncovering sources of heterogeneity in the child penalty: Conventional versus data-driven approaches

Published by:

Address:
The ROCKWOOL Foundation Research Unit
Ny Kongensgade 6
1472 Copenhagen, Denmark

Telephone +45 33 34 48 00
E-mail: kontakt@rff.dk
en.rockwoolfonden.dk/research/

September 2024

Uncovering sources of heterogeneity in the child penalty: Conventional versus data-driven approaches

Laura Weile[1], Caroline Rygaard[2] Nabanita Datta Gupta[3], Astrid Würtz Rasmussen[4] and Nina Smith[5]

September 16, 2024

---

[1]Danmarks Nationalbank, [2]Rambøll Management Consulting,[3,4,5]Department of Economics and Business Economics, Aarhus University.

Note: The views expressed in the paper are solely those of the authors and do not necessarily reflect the views of the authors' employers, companies, institutions or other associated parties.

This paper is a revised and extended version of the (unpublished) master thesis "A study of child penalties – heterogeneity across labour market characteristics", by Laura Weile & Caroline Rygaard, 2023. Cecilie Berg Kristensen provided very competent research assistance. We thank Hans Henrik Sievertsen, VIVE, Chen Huang, Aarhus University, and Søren Albeck Nielsen, Aarhus University, for their valuable comments and suggestions.

Abstract

This paper explores sources of heterogeneity in the estimated child penalty in the medium run in Denmark using both conventional and data-driven approaches. First, we analyse heterogeneity in the child penalty in the conventional way by estimating a model within specific subgroups *(hypothesized subgroups)*, which enables us to zoom in on specific characteristics, identified beforehand as being relevant. Second, we apply a data-driven machine learning approach *(causal forest)*, which explores treatment heterogeneity by combining and jointly analyzing multiple features in the covariate space. The results of our analyses confirm the existence of child penalties five years after childbirth in Denmark. The conventional analysis finds an overall medium-run penalty equal to 19.8%, while the causal forest analysis finds a significantly lower medium-run child penalty, between 6% and 8%. However, both conventional and machine-driven methods point to the pre-childbirth share of couples' wage income earned by the mother as the most important source of heterogeneity in the child penalty. In addition, the causal forest analysis also identifies industry type and firm financial variables as being influential for generating heterogeneity in the medium-run child penalty.

1 Introduction

The slowdown in the convergence of the gender wage gap over time is an issue that concerns both researchers and policymakers alike (Goldin, 2014; Blau and Kahn, 2017, OECD, 2022). At the same time, it is becoming increasingly clear that the disproportionate burden of childbirth and child rearing likely accounts for a large part of the remaining inequality in earnings between women and men, i.e., the so-called 'child penalties' faced mainly by women (Kleven et al., 2019b). Such penalties have been found to be present in almost every country, although precise magnitudes may vary (Kleven et al., 2019a; Sieppi and Pehkonen, 2019; de Quinto et al., 2021; Artmann et al., 2022; Fontenay et al., 2023). Apart from measuring the overall child penalty, it is important to know which sources contribute most to heterogeneity in these penalties for enhancing our understanding of their underlying structural causes and the policy tools needed to address them.

In this study we investigate sources of heterogeneity in the child penalty in Denmark, including both labour market characteristics and pre-market factors. Our interest is in determining whether differences in industries, firm-specific factors or individual background characteristics drive the child penalty. We employ detailed Danish administrative data and follow two distinct strategies for estimating child penalties by source. First, we follow previous literature and estimate child penalties within various hypothesized subgroups (Kleven et al. 2019b, de Quinto et al., 2021; Andresen and Nix, 2022; Artmann et al., 2022; Casarico and Lattanzio, 2023; Fontenay et al., 2023). Second, we follow a data-driven approach based on causal machine learning, and apply the causal forest estimator to identify relevant sources of heterogeneity in the child penalty (Athey and Imbens, 2016, Athey and Wager, 2019, 2021, Tibshirani et al., 2022). The machine learning approach frees us from having to focus only on a few predetermined subgroups considered one at a time. Instead, it allows for full interactions between all the relevant characteristics, thus allowing for novel hypothesis generation which researchers may not have considered at the outset (Ludwig and Mullainathan, 2023).

We first replicate the results in the study by Kleven et al. (2019b) on a more recent Danish sample, using the event study framework. In this framework, the child penalty is measured as the percentage by which women's outcome falls behind men's due to children born at a certain event time. The replication exercise yields a long-run child penalty in earnings of 21.3% compared to 19.4% in the original study. As in Kleven et al. (2019b), apart from the overall child penalty in earnings, we find sizable penalties along the dimensions of hours worked, labour market participation and the wage rate.

We then identify various sources of heterogeneity in the aggregate child penalty in earnings through our two approaches. First, the hypothesized subgroups approach allows us to focus singly on each characteristic that is important for creating variation in the child penalty, and second, the causal forest approach, which conducts a machine-led analysis of the child penalty through the lens of a large set of characteristics and interactions between them. The goal is to gain a better understanding of the underlying structure of the child penalty by comparing and contrasting results from both approaches.

The hypothesized subgroup analysis reveals that the child penalty in Denmark varies across labour market and individual characteristics, e.g., industry, employment level, female representation, but only significantly so when it comes to the share of wage income earned relative to partner. The application of the causal forest approach offers a comprehensive analysis of the sources of heterogeneity through the lens of multiple features (potential outcome predictors) and their interactions. The basic intuition behind this approach is that if the treatment effect is believed to be moderated by the covariates, it is possible to uncover this heterogeneity by splitting the sample into subsamples (leaves) according to the values of each covariate such that the difference in the treatment effect is maximized across these subgroups. The splitting procedure is repeated until there is no more treatment effect heterogeneity left. More details are provided in the methodological section (section 3), but a very appealing feature of causal trees is "honesty", meaning that the sample used to make the splits (training sample) is different from the sample used to predict the treatment effect (Athey & Imbens, 2016).

On the basis of such an analysis, we find that the share of household wage income earned pre-childbirth is very influential for heterogeneity in the child penalty, but also the worker's industry and some financial measures of the worker's firm. Thus, both conventional and machine-learning analysis point to the household division of labour following childbirth as a key determinant. Both methods find that a lower share of pre-childbirth household wage income earned by the mother is associated with a higher child penalty for mothers. However, the machine learning method uncovers additional factors associated with heterogeneity in the child penalty such as type of industry and firm financial aspects. The analysis finds that a higher penalty is associated with firms with high revenue, and lower penalties are associated with being employed in Manufacturing/Utilities/Construction or Primary/Building/Construction industries. Thus, it points to other sources of heterogeneity in the child penalty in Denmark which the conventional analysis fails to capture.

The rest of the paper proceeds as follows: Section 2 reviews recent literature relating to gender inequality in earnings focusing on the role played by childbirth. Section 3 presents the machine learning method (causal forest) applied in this paper and contrasts it to the basic event study approach from Kleven et al. (2019b). Section 4 describes the comprehensive Danish administrative data we access to build the sample and estimations. Section 5 presents the heterogeneity analyses employing hypothesized subgroups. Section 6 presents heterogeneity analysis but based on data-driven methods to uncover sources of heterogeneity in the child penalties. Section 7 discusses potential limitations, offers suggestions for policy and concludes.

2 Literature review

Gender differences in earnings have been declining in most advanced nations of the world. This convergence is generally attributed to women becoming more similar to men over time with respect to education, labour market participation and due to women working more hours (Blau and Kahn, 2017). Gallen et al., 2019 find that the gender earnings gap in Denmark has been halved between 1980 and 2010, mainly due to women increasing their hours worked. However, in Denmark and in other developed countries there has been a slowdown in the rate of convergence of the gender wage gap since the mid-1990s/2000s (Datta Gupta, Oaxaca and Smith, 2006, Goldin, 2006, Blau and Kahn, 2017). As women continue to obtain more education and increase their labour market experience, these factors are now declining in importance for the gender pay gap. Instead, occupation and industry explain a larger share of the gender wage gap, i.e. they now constitute more than 50%, as found by Blau and Kahn (2017). Furthermore, large differences still exist at the top of the distribution and within top positions (Blau and Kahn (2017), Smith, Smith & Verner, 2011).

In the well-known paper by Kleven et al. (2019b) analyzing gender differences in earnings in Denmark, the authors show that a large fraction of the remaining gender inequality in the labour market can be attributed to children. Using Danish administrative data, Kleven et al. (2019b) estimate the child penalty using an event study approach with the birth of the first child as the defining event and show that the long-run child penalty in earnings (i.e. the % by which women's earnings fall behind men's earnings after entering parenthood) in Denmark is 19.4% over the period 1980-2013. Furthermore, decomposing gender inequality in earnings according to whether its source is child-related, related to educational levels or due to some residual source, Kleven et al. (2019b) show that while the latter two decline in importance over time, child-related inequality grows over time from 40% in 1980 to 80% in 2013. A similar result is obtained by Cortes and Pan (2020) for the U.S. As in the paper by Kleven et al. (2019b), we will also focus on the 'child penalty' in this study, by which we mean the effect on earnings of the first child and all subsequent children born in the considered period as opposed to the broader concept of the 'motherhood penalty', which captures all labour market consequences of motherhood, not just earnings related.

Event studies have discovered sizable, long-run child penalties (i.e. at least 10 years after childbirth) for many developed countries, which nonetheless differ in terms of their labour market institutions and the provision of parental leave schemes and family-friendly policies (see Table 1). Kleven et al. (2019a) show quite a bit of heterogeneity across Western countries with the German-speaking countries having the largest child penalties, i.e., 51-61%, and the Scandinavian countries having the smallest, 21-26%. Additionally, Fontenay et al. (2023) find a penalty of around 32% for Belgium while Artmann et al. (2022) find that the Netherlands has a very high penalty of around 64%, comparable to the German-speaking countries. A study by Sieppi and Pehkonen (2019) estimates a child penalty of 25% for Finland, while de Quinto et al. (2021) estimate a child penalty of 25% in Spain. All these papers use a similar event study approach as in Kleven et al. (2019a).

Table 1: Cross-country comparisons of child penalties from event studies

Country	Child Penalty	Event time	Cohorts	Study
Austria	51%	t = 5, ..., 10	1985-2007	Kleven et al. (2019a)
Belgium	32%	t = 6, ..., 8	2002-2013	Fontenay et al. (2023)
Denmark	21%	t = 5, ..., 10	1985-2003	Kleven et al. (2019a)
Finland	25%	t = 5, ..., 10	1992-2007	Sieppi and Pehkonen (2019)
Germany	61%	t = 5, ..., 10	1989-2005	Kleven et al. (2019a)
Spain	25%	t = 5, ..., 10	1994-2009	de Quinto et al. (2021)
Sweden	26%	t = 5, ..., 10	1995-2011	Kleven et al. (2019a)
The Netherlands	64%	t = 10	2004-2013	Artmann et al. (2022)
United Kingdom	44%	t = 5, ..., 10	1991-2008	Kleven et al. (2019a)
United States	31%	t = 5, ..., 10	1967-2006	Kleven et al. (2019a)

Notes: Cohorts denote the time periods in which parents have their first child.

These and other studies have now begun to look further into sources of heterogeneity in child penalties. A first potential source of heterogeneity is mother's education or skill level. Early work by Anderson et al. (2002) on NLSY data, estimating simple OLS and fixed effects models of the effect of children on women's wages, found that controlling for unobserved heterogeneity, being married, (potential) time out of the labour market, education and actual experience (or alternatively, age), child penalties were higher for college educated women compared to women with no college diploma. They concluded that the costs of temporarily exiting the labour market were much higher for highly skilled mothers because of the higher return they obtain on their human capital investments. The same is argued by England et al. (2016) who also find that high penalties are strongly related to high returns on experience, making even short periods of lost experience very costly for these women. A theoretical explanation could be the importance of firm-specific human capital and/or high depreciation rates of human capital in such jobs.

More recent studies estimating event study models in different contexts find that the labour supply response following the birth of a child can be quite different depending on mothers' education level. For instance, de Quinto et al. (2021) show evidence from Spain that non-college educated women react more on the extensive margin by reducing working days, while college-educated women react more on the intensive margin by reducing working hours.

However, education can embody more than just the level of individual skills. Due to assortative mating, high-educated women are normally married to high-educated, high-wage men, meaning that family income may not necessarily suffer seriously if intrafamily specialization increases after childbirth. Still, potential future earnings loss is likely to be larger for high-educated women, even after taking into account that they may enjoy a faster rate of wage growth after childbirth (England et al., 2016). Therefore, a question is whether high-educated women perfectly anticipate these future losses of child-related leaves?

Kuziemko et al. (2018) bring empirical evience from the US and UK showing that mothers from the US and UK tend to underestimate the employment costs of motherhood, which include costs of child-minding measured in terms of both time, effort, and money. This underestimation of the costs of motherhood is in fact larger for highly educated women. Not only do highly educated women in particular underestimate the resources that need to be allocated towards child rearing following the birth of the first child. They also tend to experience a shift in their beliefs after childbirth and become less supportive of women being in the labour market.

Type of education may also matter. Bütikofer et al. (2018) estimate, using an event study approach, a higher child penalty 10 years after first childbirth among top earners with MBA and law degrees (30%) compared to those with STEM and medicine degrees (10%). They interpret this finding as resulting from differences in the wage structure. Therefore, as a source of heterogeneity, we explore differences arising from the wage structure or demand side as proxied by sectors, industries, occupations, firms, and workplace characteristics.

Starting with sector, in many European countries the public sector is both a large employer of women and also the more family-friendly sector (Nielsen et al., 2004). Using Danish administrative data Pertold-Gebicka et al. (2019) show that mothers are much more likely to be employed in the public sector after the birth of their first child compared to non-mothers. As in Goldin (2014), they find that mothers who switch to the public sector do so from previous positions in the private sector characterized with convex pay schemes. Fontenay et al. (2023) use data from Belgium and estimate child penalties across industries in the public sector equaling 33% for health and social work, 15% for public administration and defense, and 15% for education. They find a positive correlation between the long-run child penalty in earnings for a given industry and survey-reported measures such as poor work-life balance, atypical work schedules with irregular hours, and evening, night, and weekend hours.

Goldin (2014) suggests that occupations and how they reward hours may be the key source of heterogeneity in child penalties. This is due to compensating differentials, which penalizes low hours and rewards high hours. For women to participate on equal terms with men, occupations need to offer temporal flexibility i.e., workers do not have to be around during fixed intervals but have flexibility to adjust their schedules. Goldin and Katz (2016) give the example of pharmacy, which is an occupation that has become more flexible and family-friendly over time due to its increased use of information technology. Thus, pharmacy can now be characterized as an occupation where earnings is a linear function of hours worked, which is less penalizing for women with small children.

Penner et al. (2023) exploit employer-employee data from 15 countries over a ten-year period starting from 2005 to investigate the gender earnings difference that arises between similar job types in the same firms among workers aged 30-55. Their results show that even within the same establishment and controlling for age, age-squared, education and degree of attachment (full time or part-time), men's earnings exceed women's in all 15 countries in their dataset. Thus, within-job differences are important, and for Denmark, the study finds this difference accounts for about 40% of the overall gender pay gap. In terms of workplace characteristics, a study based on employer-employee data from West Germany by Bruns (2019) explores the effect of workplace characteristics on the gender gap in firm-specific wage premiums. The drivers of the growth in premiums are high-wage firms and firms with a higher male share of employees. Furthermore, they argue that both being employed in a relatively low-wage firm and lower female bargaining power can explain these differences in wage premiums. Using matched employer-employee data from Italy, Casarico and Lattanzio (2023) find that mothers are more likely to select into firms with lower value added and lower wages. They also explore heterogeneous effects at the individual and firm levels and find that among individual factors, child penalties are larger among those employed in blue-collar or low-wage firms, among younger individuals and those taking longer maternity leaves. At the level of the firm, child penalties are largest in firms with many female workers.

Of course, gender differences in demand-side characteristics may also reflect sorting in the labour market. Women after childbirth may change their sorting behavior and may re-locate in jobs, occupations, industries, or sectors that more easily accommodate their new function as primary parent. This was first recognized by Becker (1985), who argued that women take their expected fertility into account and therefore choose different types of jobs, leading to the observed occupational segmentation by gender and hence, the differences in labour market outcomes. This idea was elaborated further by Blau and Kahn (2017) who argue that motherhood might affect women's choice of workplace, inducing them to shift to more family-friendly positions, as well as their effort and productivity. Goldin et al. (2017) furthermore gives a rationale for an increasing child penalty over the life cycle. Women and men diverge over time in terms of the type of employment they seek out, with men pursuing careers with higher earnings and promotion possibilities and women pursuing careers with lower earnings but greater stability and flexibility. For example, several previous studies have shown that mothers in Denmark self-select into the public sector where non-pecuniary benefits such as flexible working arrangements are larger, which in turn has implications for the family gap in wages (Nielsen et al., 2004; Simonsen and Skipper, 2006; Pertold-Gebicka et al., 2016).

Other than sorting, bargaining differences may also produce different outcomes by gender. Using Portuguese employer-employee data, Card et al. (2016) explore how much bargaining and sorting channels combined contribute to the gender wage gap in Portugal. Sorting is captured by women working in firms that give smaller surpluses, while bargaining implies that within the same firm, women earn smaller wage premiums than men. Their estimates show that bargaining and sorting together account for 20% of the gender wage gap in Portugal. Their results show that three quarters of this gap can be attributed to the sorting effect and one quarter to the bargaining effect. Furthermore, the sorting effect itself is heterogeneous along the distribution of skills and is more salient for low-skilled and middle-level labour. Gallen et al. (2019) apply the same methods to decompose the gender wage gap in Denmark, using matched employer-employee data. By 2010, they find a gender wage gap equal to 20.2%. They also find that sorting explains about 15% of the gap in 2010. Their results further show that by 2010, observable gender differences account for 7.5 percentage points of the total gender wage gap of 20.2%, and that the narrowing of the wage gap by 50% since 1980 is mainly due to a reduction in the gender gap in hours and labour market experience.

A few papers have explored the importance of family policies for child penalties, including childcare and leave policies. Kleven et al. (2022b) use Austrian administrative data and exploit changes in family policies and leave reforms over time for investigating gender inequality in earnings. The authors find that these historic changes in childcare policies have little impact on earnings differences between men and women. Expansions in parental leave schemes have short-run negative impacts but hardly any long-run impacts on women's labour market outcomes. On the other hand, Casarico and Lattanzio (2023) find that child penalties are lower (higher) in regions in Italy that are considered more (less) progressive, as measured by childcare provision. According to the authors, both gender norms and legal institutions do matter for child penalties. Similarly, Weile (2022) also finds large regional differences in the child penalty between eastern and western Denmark, regions, which differ according to conservative values, for instance proxied by the share voting for far-right parties.[1]

Progressivity in part reflects norms relating to gender. Kleven et al. (2019a) draw data from the International Social Survey Program (ISSP) about individuals' attitudes towards employment of mothers with young children to elicit gender norms. They uncover a positive correlation between the elicited gender norms and the estimated child penalties. Carrying on this line of enquiry, Kleven et al. (2022a) find a negative relationship between the progressivity of the birth canton in Switzerland (measured as the fraction voting yes to the female right to vote in 1971) and child penalties even when conditioning on the current area of residence. Kleven (2023) finds similar results using pseudo-event studies to estimate the child penalties in the US. Estimated child penalties are significantly higher for parents originating from high-penalty states than for parents originating from low-penalty states, conditional on the current state of residence. When considering immigrants, they find a similar mechanism at play, i.e., that estimated child penalty reflects the child penalty in the country of origin. The importance of culture and social factors is convincingly shown by Rosenbaum (2021), too, who finds a child penalty also among adopting mothers (although not as large as that within biological mothers). This set of papers concludes that norms and culture drive the differences in child penalties more than policies do (Kleven et al., 2019a). However, as we discuss in Section 7, policies may play an important role in influencing norms.

Digging deeper into the black box of gender norms, it is primarily the norm concerning who will be the primary provider of childcare and housework that is salient for gender gaps and slow convergence (Giuliano, 2020). However, there is also an economic explanation for this: it is generally assumed that the partner with the lowest opportunity cost of working in the labour market takes on more of the home production, e.g. Gronau (1977), and typically women tend to have lower wage rates and earnings than their spouses.

Data on Danish households from 2002 confirm that opportunity costs matter and show that women devote more time to unpaid work compared to men, especially to more time inflexible activities like childcare, picking-up children from leisure activities etc, Bonke et al. (2004). Moreover, men allocate more time to paid work, and the difference only increases with children in the household (Craig and Mullan, 2010). Cortés and Pan (2020) model time allocation within households and assume that the productivity in child-rearing differs for

[1] See Landkort viser forskelle i partiernes stemmeandele - Danmarks Statistik (dst.dk).

mothers and fathers. Thus, even though mothers may earn more than fathers, mothers might still reduce their hours in the labour market if they are more productive at raising children. Gender norms are modelled by introducing differences in preferences. Women may try to fulfill their traditionally ascribed role of homemaker if society judges them for that, and similarly, men may try to live up their traditionally ascribed role of primary provider for the family. When taking the model to PSID data, the authors find that comparative advantage alone cannot explain the child penalty in earnings, which indicates the presence of gender norms/preferences.

In a similar vein, Andresen and Nix (2022) develop a household decision-making framework that incorporates gender norms and preferences regarding childcare, specialization within households, and the biological consequences of giving birth. The authors compare same-sex to different-sex couples in order to tease out the mechanisms using Norwegian administrative data, in a framework that is similar to Rosenbaum (2021). Similarly to Rosenbaum (2021), the results show that gender norms and preferences drive the persistent child penalty experienced by women in different-sex couples while biology seems to explain a minor part of it.

Focusing more directly on the period around childbirth, Andersen (2018) exploits Danish administrative data to investigate the effects of five parental leave reforms within a household exploiting paternity leave extensions among others. Her findings suggest that the father's engagement in paternity leave and childcare positively impacts the mother's labour market outcomes both in the short and long run, as the father increases his household capital. This is because mothers develop a greater attachment to the labour market and improved negotiations in the household, e.g., when deciding who should stay home with a sick child. Furthermore, the change in the distribution of parental leave yields a more equal allocation of household responsibilities even when the father's paternity leave increases. These findings suggest that the size of the child penalty is influenced by the allocation of parental leave between the parents. Thus, extending paternity leave may break the intrahousehold specialization and increased investments made in the specialized spheres by each parent. We return to the role of public leave policy in breaking down gender norms in Section 7.

Finally, we cannot rule out that labour market discrimination against mothers could lead to heterogeneity in the observed child penalty, if e.g., some employers are more discriminating than others. A study by Correll et al. (2007) experimentally show that mothers are subject to discrimination due to their parental status, whereas fathers do not experience such discrimination and even experience a fatherhood bonus. The authors expose employers to job applications from candidates of the same sex and qualifications who only differ according to their parental status. The study finds that among female applicants, non-mothers are considered more competent and committed than mothers, and that employers recommend a higher salary for non-mothers compared to mothers and are more likely to recommend non-mothers for promotion. On the other hand, there is no fatherhood penalty and in fact, employers even consider fathers to be more committed.

While we do not have a way of measuring discriminating attitudes and practices of employers, on the basis of the literature reviewed here, we will search for heterogeneity in the child penalty in our data according to individual background characteristics such as age and the education obtained prior to entering the labour market, as well as sector, industry, occupation, firm-level factors, variables representing the pre-childbirth household division of labour and the area and municipality of residence. In the next section we present both the event study and causal forest methodologies used in this paper and discuss strengths and limitations of each approach.

3 Methodology

Our aim is to study how the impact of children on female and male earnings differs across various labour market characteristics. First, we follow Kleven et al. (2019b) and use the conventional event study approach to estimate the child penalty. We segment the analysis based on hypothesized subgroups to explore heterogeneity in the child penalty. Next, we propose to use a machine learning (ML) approach, causal forest, to conduct a data-driven exploration of heterogeneity in child penalties. While the hypothesized subgroup approach only allows focusing on specific pre-defined characteristics at a time, the ML approach allows us to study how the interaction of various characteristics and thresholds relates to the child penalty.

3.1 Event study approach

Kleven et al. (2019b) popularized the event study approach to estimate the effects of children on parents' labour market outcomes. Since then, a growing number of studies have applied it to study the child penalty in different contexts (Andresen and Nix, 2022; Artmann et al., 2022; Fontenay et al., 2023; Casarico and Lattanzio, 2023; Sieppi and Pehkonen, 2019). The main outcome of the model is the child penalty, which measures the difference in the event-specific impact of children between fathers and mothers' earnings relative to the counterfactual earnings (i.e., absent children) outcome of the mothers. Thus, the child penalty can be interpreted as the percentage by which women's earnings are falling behind that of men’s due to children.

The approach builds on several assumptions for identifying a causal effect of the child penalty. First, the event study approach requires that the birth of a child creates sharp changes in labour market outcomes in the precise year in which the child arrives (Kleven et al., 2019b). It also assumes that these changes are orthogonal to unobserved determinants affecting labour market outcomes, which are assumed to evolve smoothly over time. For example, families must not have children in response to anticipated labour market outcomes. However, even if the smoothness assumption is satisfied, the event study results will be most credible close to the event time.

One could be concerned that other events occurring close to the first childbirth such as marriage or pregnancy also produce discontinuous effects which confound with the child penalty effect. However, Berniell et al. (2022) does not find that pregnancy and marriage create sharp changes in labour outcomes. This lends more support to the birth of the first child being the event that causes a sharp change in women’s labour market outcomes. Also, the smoothness assumption may not hold the further one moves away from the event time, $t = 0$; it is plausible that factors other than children also affect labour market outcomes in the long run. Trusting the existence of a long-run penalty requires thus stronger assumptions (Kleven et al., 2019b). To reduce bias due to other concurrent changes, Kleven et al. (2019b) controls for age and year effects non-parametrically. It is further recommended to include a control group of individuals who never become parents or use an instrumental variables approach to account for endogeneity (Kleven et al., 2019b). Studies that adopt an IV approach are Lundborg et al. (2017, 2024) and Bensnes et al. (2023) who estimate the child penalty by exploiting IVF treatments as an instrument for having a child.

In this paper, we rely on the smoothness assumption, compare mothers to fathers only, and assume that conditional on age and year, the birth of the first child is exogenous to labour market outcomes. Our aim is to compare the standard way of investigating heterogeneity to a machine-led method. Future analyses can generalize the analyses here by relaxing the above assumptions.

The event study method also relies on the parallel trend assumption. The assumption is typically assessed by looking at the pre-period coefficients from five years prior to having children which must be visually similar between mothers and fathers. However, parents’ earnings in the five years before parenthood may be a poor indicator of lifetime earnings, e.g., if some of the included parents are still enrolled in education. Thus, pretrends may not necessarily capture parallel trends even though age and year fixed effects are accounted for (Artmann et al., 2022). Sun and Abraham (2021) argue that there could be contamination of the pre-period coefficients from other periods, including the excluded period, $t = -1$, leading to treatment effect heterogeneity over time.

Another case where the exogeneity assumption is violated is if women with high earnings prospects tend to have children later than women with low earnings prospects (see e.g., Bensnes et al., 2023). In fact, they find a small estimate of the child penalty around 2% in the long-run using the 2SLS (with IVF as instrument for fertility) event study approach on Norwegian data while the authors estimate a long-run estimate of around 20% when using the OLS event study approach as Kleven et al. (2019b). Therefore, the authors conclude that the child penalties estimated by the event study approach from Kleven et al. (2019b) might be upward biased. Similarly, Lundborg et al. (2024) use Danish data and find significantly higher event study estimates of child penalties than child penalties estimated using an IV-IVF approach.

In addition to the assumptions presented above, another implicit assumption is that of a common trend across cohorts of first-time parents. This means that women who give birth in different calendar years would have the same annual changes in labour market outcomes in the counterfactual scenario where they did not have children (Andresen and Nix, 2022; Artmann et al., 2022). If this is not the case, then a causal interpretation cannot be given either.

The choice of a relevant control group is important for interpreting child penalties and interpreting the results in the heterogeneity analysis in the event study setup. Throughout this paper, our sample consists of individuals who have entered parenthood, similar to previous studies on child penalties (Kleven et al., 2019a; Artmann et al., 2022; Fontenay et al., 2023; Casarico and Lattanzio, 2023; Sieppi and Pehkonen, 2019). We argue that there is selection into parenthood as we assume that most parents voluntarily choose to have children.

Consequently, the causal interpretation of the estimated child penalties applies to parents and not the entire population. Alternatively, we could estimate the child penalty by comparing mothers with non-mothers and fathers with non-fathers, similar to a robustness check by Kleven et al. (2019b). They assign placebo births to non-parents, e.g., comparing never-mothers with mothers by assigning births to non-parents using the empirical age distribution and the cohort of first-time parents in the sample (see also Pertold-Gebicka et al., 2016).

However, we have chosen to limit our study to look at the difference in the impact of children between mothers and fathers and thereby do not apply this type of analysis.

Finally, our sample is sensitive to the sample restrictions imposed in Table 2. For example, we include only individuals who enter parenthood in the age span between 20-45 years old, like in Kleven et al. (2019b). This restriction implies that the youngest individuals in our sample are 15 years old at event time $t = -5$ as we observe all individuals in the five years prior to childbirth. Therefore, the pretrend coefficients in earnings growth for these young individuals may not be the best indication for the trend in future earnings. This can potentially bias our results, as the counterfactual outcome of no children is based on unreliable pretrends. On the other hand, we want to estimate a representative average child penalty, so it would lead to non-representative results if we excluded young parents. When estimating child penalties, our study is also subject to the limitations discussed above.

3.2 Identification of heterogeneity

We aim to investigate the heterogeneity in the child penalty across various labour market and individual characteristics. One approach is to divide the sample into hypothesized subgroups and estimate the child penalty separately for each of these sub-groups. This is the method widely used in the literature (Fontenay et al., 2023; Andresen and Nix, 2022; Casarico and Lattanzio, 2023; de Quinto et al., 2021; Artmann et al., 2022). For example, de Quinto et al. (2021) investigate heterogeneity in the child penalty by education level by conditioning on the highest education level obtained pre-birth while Casarico and Lattanzio (2023) investigate the child penalty in different subgroups based on labour market characteristics such as occupation, the share of female workers, and firm size.

The approach in this paper is twofold. First, we investigate the heterogeneity in child penalties by splitting our sample into subsamples based on categorical and continuous variables, which we hypothesise to be a source of heterogeneity in child penalties. Given motherhood in $t = 0$, we condition on the relevant variables in $t = -1$, i.e., prior to motherhood, to avoid conditioning on post-treatment effects related to the outcome of the birth of the first child (Casarico and Lattanzio, 2023). Second, we investigate various interactions between the included labour market, firm, and individual characteristics by applying a causal tree-based machine learning method known as causal forest.

Machine learning methods are now entering the labour economists' toolkit, used often to complement and compare to conventional methods. The great advantage of ML methods is that they can account for multiple features and interactions between them which may not have been envisioned by the researcher as being important at the outset. Applying ML methods (the post-double LASSO procedure) to the gender pay gap (GPG) in Germany, Bonaccolto-Töpfer and Briel (2022) find important differences in findings compared to conventional methods. Davis and Heller (2017) apply the causal forest to data originating from two RCT's of the same summer job program and analyze treatment heterogeneity in predicting crime and employment outcomes.

In this paper we apply a method known as honest causal forest, which is a tree-based supervised ML method. Tree-based ML methods are a way of uncovering heterogeneity in the conditional average treatment effect (CATE), (Athey & Imbens, 2015, 2016). They partition the sample into subsets through binary splits of the covariates at threshold values where the thresholds are determined through a data-driven approach. Afterwards, the method fits a model within each of the final subsamples, called the leaves of the tree. A pitfall of using the causal tree method, a method relying on a single tree, is that the estimates may be invalid if the training set contains a high degree of noise. This happens as the causal tree aims to maximize the variance between the leaves and minimize the variance within the leaf (Brand et al., 2021), i.e., the regression trees split the data iteratively to minimize mean squared error (MSE) within each partition. Consequently, the estimates may be unstable and be strongly dependent on the initial seed, which determines how individuals are allocated between training and estimation samples.

For this reason, we may want to estimate a treatment effect in each leaf and then average across multiple trees, growing what is known as a causal forest, see e.g. Athey, Tibshirani, and Wager, 2019; Wager and Athey, 2018, Hastie et al., 2009, Breiman et al., 1984 and Breiman, 2001. The advantage of the causal forest over the causal tree is that it randomly draws multiple samples, s, from the original sample, n, using subsampling where the subsamples are of a smaller order than the original sample, i.e. $s < n$, and takes an average across these to smooth predictions. As mentioned already, the conditional average treatment effect is then predicted for each unit (individual) in the dataset, rather than producing a group level average effect or the predicted outcome of the dependent variable. Causality is based on the unconfoundedness assumption, i.e., the treatment effect is a conditional average treatment effect (CATE). The causal forest method is honest because a subset of the data is withheld (training sample) for learning the features of the model and cross-validating them, which are then applied to another subset of the data (the holdout or estimation sample) for estimating effects. These subsets of data can be drawn independently from the same distribution. The honesty feature prevents overfitting and minimizes the false discovery problem (Athey et al., 2019; Athey & Imbens, 2016; Kattenberg et al., 2023).

In the first part of our heterogeneity analysis, we estimate the baseline event study model by subgroups to identify heterogeneity in child penalties. In other words, we allow for different age and year effects for parents in each subgroup like Fontenay et al. (2023). In the second step, we adopt the causal forest approach. However, we agree with Bütikofer et al. (2018), that both the hypothesized subsample analysis and the ML method may be prone to a selection bias when we condition on labour market characteristics. The bias arises because men and women might select different industries, occupations, sectors, and firms, and we do not account for this type of sorting. Therefore, we cannot claim to identify the causal effect of labour market characteristics on child penalties due to possible endogeneity bias. Instead, we interpret the child penalties as causal conditional on the given characteristics we split on. Such an analysis is still relevant because any differences we find compared to the event study method gives new information on sources of heterogeneity in the child penalty, and allows us to compare and contrast our findings to those from a growing literature using event study methods to study heterogeneity (Bütikofer et al., 2018; de Quinto et al., 2021; Andresen and Nix, 2022; Casarico and Lattanzio, 2023;

Fontenay et al., 2023).

Following Kleven et al., (2019), we write the basic event study equation to be estimated as

$$Y_{ist}^{g} = \sum_{j \neq -1} \alpha_j^g \cdot \mathbf{1}\,[j = t] + \sum_{k} \beta_k^g \cdot \mathbf{1}\,[k = age_{is}] + \sum_{y} \gamma_y^g \cdot \mathbf{1}\,[y = s] + v_{ist}^g, \;\; g \in \{m, f\} \tag{1}$$

where $g \in \{m, f\}$ is the gender of individual i, t is the event time relative to the birth of the first child, and s is the year of the observation. This equation is estimated separately for each gender for each of the hypothesized subgroups. In each case, child penalties are then calculated as the gender difference in the event time coefficients in any period scaled by the counterfactual female income. The counterfactual female income is the predicted female income from the equation above minus the event time estimates. These results appear in Figures 1-12. Thus, the child penalty in event time t is given by

$$P_t = \frac{\hat{\alpha}_t^m - \hat{\alpha}_t^f}{E[\tilde{Y}_{ist}^w | t]} \tag{2}$$

where $\tilde{Y}_{ist}^w$ is the counterfactual labour market outcome of females when excluding the event-specific contribution of children. Next, we turn to the implementation of causal forest in this setup. The causal forest algorithm is designed to estimate the conditional treatment effect $\tau(X)$, given 1 indicates assignment to a treatment state, i.e. being a father in event time t, and 0 not, and where Y measures outcome. Thus, $\tau(X) = E[Y(1) - Y(0)|X = x]$, where $Y(1)$ and $Y(0)$ are the outcomes under the treated and control state, respectively, captures the relative treatment effect of parenthood for fathers in event time t compared to mothers. By training the model to identify subgroups in which the Conditional Average Treatment Effect (CATE) differs from the Average Treatment Effect (ATE), the method yields consistent estimates of heterogeneity in the treatment effects.

To be able to estimate the equation above using causal forest, we first need to convert it to a fully gender-interacted one-equation model, as causal forest estimation is based on a single equation.

Thus, we rewrite equation (1) as:

$$Y_{ist} = \sum_{j \neq -1} \left(\alpha_j^f + \alpha_j^m \cdot \mathbf{1}[g = m]\right) \cdot \mathbf{1}\,[j = t] + \sum_k \left(\beta_j^f + \beta_j^m \cdot \mathbf{1}[g = m]\right) \cdot \mathbf{1}\,[k = age_{is}]$$
$$+ \sum_y \left(\Upsilon_j^f + \Upsilon_j^m \cdot \mathbf{1}[g = m]\right) \cdot \mathbf{1}\,[y = s] + \tau(X_i) \cdot \sum_{j \neq -1} \left(\delta_j^f + \delta_j^m \cdot \mathbf{1}[g = m]\right) \cdot \mathbf{1}\,[j = t]$$
$$+ \left(v_{ist}^f + v_{ist}^m \mathbf{1}[g = m]\right) \tag{3}$$

As the causal forest is a cross-sectional model, we need to further reduce the dimensionality of the problem. We do this by considering only event time = -1 and event time = 5 and considering the difference between the two as our dependent variable in the analysis. We select event time = 5 because as we will show, the gender-specific impacts appear to stabilize by that time. This choice ensures that the identifying assumptions of a causal child penalty are stronger, as moving further away from the event time may weaken these assumptions. Furthermore, because studies such as Bensnes et al. (2023) and Lundborg et al. (2024) find little evidence of long-run child penalties, our focus will be on the medium run, i.e., 5 years after childbirth. We thus redefine equation 2 as the difference in earnings between event time -1 and event time 5 giving the estimation equation below:

$$Y_{is5} - Y_{is-1} = \sum_{j \neq -1} \left(\alpha_j^f + \alpha_j^m \cdot \mathbf{1}[g = m]\right) \cdot \mathbf{1}\,[j = 5] + \sum_k \left(\beta_j^f + \beta_j^m \cdot \mathbf{1}[g = m]\right) \cdot \mathbf{1}\,[k = age_{is}]$$
$$+ \sum_y \left(\Upsilon_j^f + \Upsilon_j^m \cdot \mathbf{1}[g = m]\right) \cdot \mathbf{1}\,[y = s] + \tau(X_i) \cdot \sum_{j \neq -1} \left(\delta_j^f + \delta_j^m \cdot \mathbf{1}[g = m]\right) \cdot \mathbf{1}\,[j = 5] + v_{ist} \tag{4}$$

The treatment effect is thus male*t = 5, while the main effects of male, age and calendar time are absorbed in the covariate vector X (dropping age and year in t = 5 to avoid perfect collinearity) and with a combined error term. A key assumption for the validity of causal forest is that of positivity or overlap. Say W is the binary variable that indicates assignment to treatment, then we want that $P(W_i = w|X_i) > 0, w = 0{,}1$. This means that this assignment probability cannot be zero for some values of X. Causal forest methodology is described further in Section 6. The main settings we will work with are 2000 trees aggregated into the forest. We will randomly sample 50% of the data to build each tree. Since the honesty criterion is used, these subsamples will again by split in half to, respectively, determine the splits and obtain predictions.

4 Description of Data

We exploit extensive administrative register data from Statistics Denmark to construct our sample of first-time parents. Our sample of parents is based on all individuals in Denmark from the population register, BEF, from Statistics Denmark. We follow the sample selection from Kleven et al. (2019b), Rygaard and Weile (2023), and Weile (2022) closely and construct a relevant sample of first-time parents (see Table 2 for details on the sample selection process). Therefore, we keep only individuals registered as biological (and not adoptive) parents and with unique information about gender and date of birth. Additionally, we exclude parents in same-sex couples and parents whose children have more than one registered mother or father.

One difference compared to Kleven et al. (2019b), is that our study examines a more recent sample of parents with children born in the period 2002-2010. The reason we start the analysis from 2002 is that a major reform of the parental leave system came into force from March 7th, 2002. The reform extended parental leave from 10 to 32 weeks, dropped 2 weeks of earmarked paternity leave that could be taken after the 10 weeks of maternity leave and replaced the previous one year of parental leave with compensation at 60% UI benefit by 28 weeks (14 weeks each) parental leave with no compensation. The reform was announced in January 2002 and applied to all parents whose children were born on or after March 7th, 2002 (see Lassen (2022) for more details). However, parents whose children were born between January 1st, 2002, and March 7th, 2002, could choose whether to take up the new scheme or the old one and almost all chose the new scheme (Beuchert et al., 2016).

At the other end of the data period, to ensure that we have least 10 years of post-birth information on labour market outcomes, and since information on employment and employer contributions (ATP) is only available until 2020, we stop at the 2010 cohort. We also require, like Kleven (2019b) parents to be between 20 and 45 years old when they enter parenthood. This reduces the likelihood of including parents who have yet to enter the labour market or who have already withdrawn themselves via retirement. Finally, we restrict our sample to be balanced such that both parents of the child must be present in our sample, and both parents must have available income information for all 16 years in our event window. We show the sample selection criteria in Table 2. Our final sample consists of 372,798 individuals.

Table 2. Sample selection for main sample

Sample Restriction	Individuals
All individuals from the BEF population register	8,734,928
Excluding individuals with ambiguous or missing gender and birth year information	8,729,908
Excluding individuals who have not yet entered parenthood	3,623,493
Excluding same-sex couples and children with more than two registered parents	3,621,298
Keeping individuals who enter parenthood in 2002-2010	459,206
Keeping individuals who enter parenthood between 20-45 years of age	450,496
Excluding individuals with missing income information in one or more years in the event window $t = \{-5, \ldots, 10\}$ around the first childbirth	404,257
Excluding individuals for whom their child does not have both parents observed in the sample	380,388
Deleting couples for whom labor market status is missing for either one or both members	376,582
Deleting couples for whom education is missing for either one or both members	372,798

We include income in all years, even when income in a given year is zero. This is because some individuals may decide to stay home with their children, and this is also a part of the child penalty. Indeed, while observations of zero earnings are almost equally distributed between mothers and fathers in $t = -1$, this is not the case for $t > 0$. Unlike Kleven et al. (2019b), we deflate earnings using Statistics Denmark's annual change in the consumer price index, PRIS112, however, this does not change results compared to using non-deflated earnings.[2] Similar to Kleven et al. (2019b), we define the three margins of earnings: hours worked, hourly wage, and participation in the labour market. We construct the labour market outcomes using the

[2] Another difference to that paper is that we use the loenmv13 variable from the income register to measure the yearly income for the parents in our sample compared to the loenmv variables used by Kleven et al. (2019b). The loenmv13 measure was added as an update to the income registers in 2013 and includes employer-paid sick leave, remuneration for non-competition clauses, and researcher salary in addition to the income components in the loenmv variable. Again, we can show that the two variables produce nearly identical results.

individual ATP contribution, like the approach of Kleven et al. (2019b).[3]

4.1 Employer and employee data

We include information about the individual's firm of employment in t-1, and all labour market characteristics are measured in t = −1 to avoid conditioning on decisions made after the arrival of the first child. We are interested in the overall performance of the firm and therefore we measure characteristics at the firm level and not the workplace level. Firm characteristics include the firm's age and a dummy variable indicating whether the firm is Copenhagen-based and whether it is considered a new firm (< five years of tenure, as most bankruptcies occur when the firms are between 3-5 years old (The Danish Foundation for Entrepreneurship, 2018). Further, we include firm financial information including selected information from the firms' accounting statements, i.e., revenue, balance, equity, and net income.

Additionally, we include two different financial indicators to compare firms' financial efficiency, return on assets (ROA), which measures how well a firm uses its assets to generate profit, and return on equity (ROE), which measures profitability by comparing a firm's net profit with its equity. Lastly, we calculate the firms' revenue growth to allow heterogeneity in growth (Harrison Jr et al., 2014). The firm-specific variables are summarised in Appendix Table A.1.1. In addition to financial and structural information about the firms of employment, we include additional available information about the individual's employment and industry in Appendix Tables A.1.2 and A.1.3. We construct the industry variable using relevant industry information from the workplace register, covering the relevant period.

Note that the means of the firm-specific variables reported in Table A.1.1 such as revenue or share of women in each employment position is measured as the mean across individuals employed in these firms, i.e., firms with many employees might occur in the sample several times. Consequently, the reported means are not the average of firm characteristics in 2002-2010 but the mean characteristics of the companies in which the individuals in our sample are employed.

Statistics Denmark reports detailed workplace statistics, such as the firm's share of employees at different employment levels. The proportion employed at each level is determined based on

[3] Since 1964, the ATP contribution has been mandatory for all employees, including the self-employed, aged 16 or older, who work more than 39 hours per month. The individual contribution is a discrete function of hours worked and depends on the pay period. Thus, the ATP variable and the defined labour market outcomes will limit the scope of measuring the marginal change in hours worked due to the discrete nature of the variable. This limitation may result in underestimating the child penalty in hours worked and overestimating the child penalty in wage rates (Kleven et al., 2019b).

their primary job title. In addition, Statistics Denmark reports the proportion of employees in the individual categories who are women. Finally, we include additional variables relating to the individual's employment, e.g., experience in the labour market, and tenure in their present employment position.

4.2 Individual characteristics

We also include individual background information on mothers and fathers such as age, marital status and relative wage income, see Appendix A.1.4. Again, we measure these individual characteristics in $t = -1$ to avoid conditioning on post-treatment decisions. The average age at first birth of mothers (fathers) increases from 28 (29) to 29 (30) over our sample period. We do not use any post-birth characteristics in the model as they can be affected by the treatment. However, we compare the completed family size ten years after the birth of the first child and find no difference across parents - the average number of children of fathers ten years after birth is 2.127 (0.713) while that of mothers is 2.132 (0.706).

As a measure of specialization/gender norms within the family, we utilize the couple's relative wage income (mother's wage/(mother's wage + father's wage)), measured at $t = -1$. While in most countries, men earn a larger share of the household income, women's share has been increasing over time. In the U.S. it was 26% in 2010, up from 11% in 1970 (Bertrand et al., 2015). A sharp "cliff" in this share at 50% is associated with a lower match likelihood and can be taken to represent gender identity norms within the family. When women's share exceeds that of men, it may signal a more egalitarian gender norm prevailing within the couple. At the same time, such women are highly selected and may behave differently from other women. For example, studies show that women who earn more than their husbands often compensate for going against the traditional norm by doing a larger share of household activities (Bittman et al. 2003, Bertrand et al., 2015). Note that we use the actual share of household (wage) income in the estimation and not the predicted or instrumented share as in Bertrand et al. (2015). First, we already control for demographic characteristics such as age and education in our models, and second, given the homogeneity of the setting we are analyzing, marriage markets do not differ as much as in the U.S. setting analyzed in previous papers. Also, Bertrand et al. (2015) find largely similar estimates when using the actual share versus the predicted or instrumented share.

The distributions of relative wage income for the mother and father appear in Appendix A6. Although not as dramatically different as in the U.S., in Denmark too, fathers are the main earners in the family, and the mother's distribution is steeper to the right of 0.5. The average mother's (father's) share of household wage income seen from Table A.1.4 is 0.458 (0.542) at $t = -1$. When comparing over cohorts, the mother's share varies from 0.452 for the oldest cohort to

0.461 for the youngest cohort in our sample, indicating stability in this share over time.

Finally, we add information about individuals' highest completed level of education and their place of residence. These variables are summarized in Appendix Tables A.1.5 and A.1.6. Statistics Denmark classifies municipalities into five distinct categories based on two factors, the density of available workplaces and the number of inhabitants in the largest city within the municipality. The five categories include capital, commuter, metropolitan, provincial and rural municipalities. We include the municipality groups variable to our sample and report its distribution in Appendix Table A.1.7.

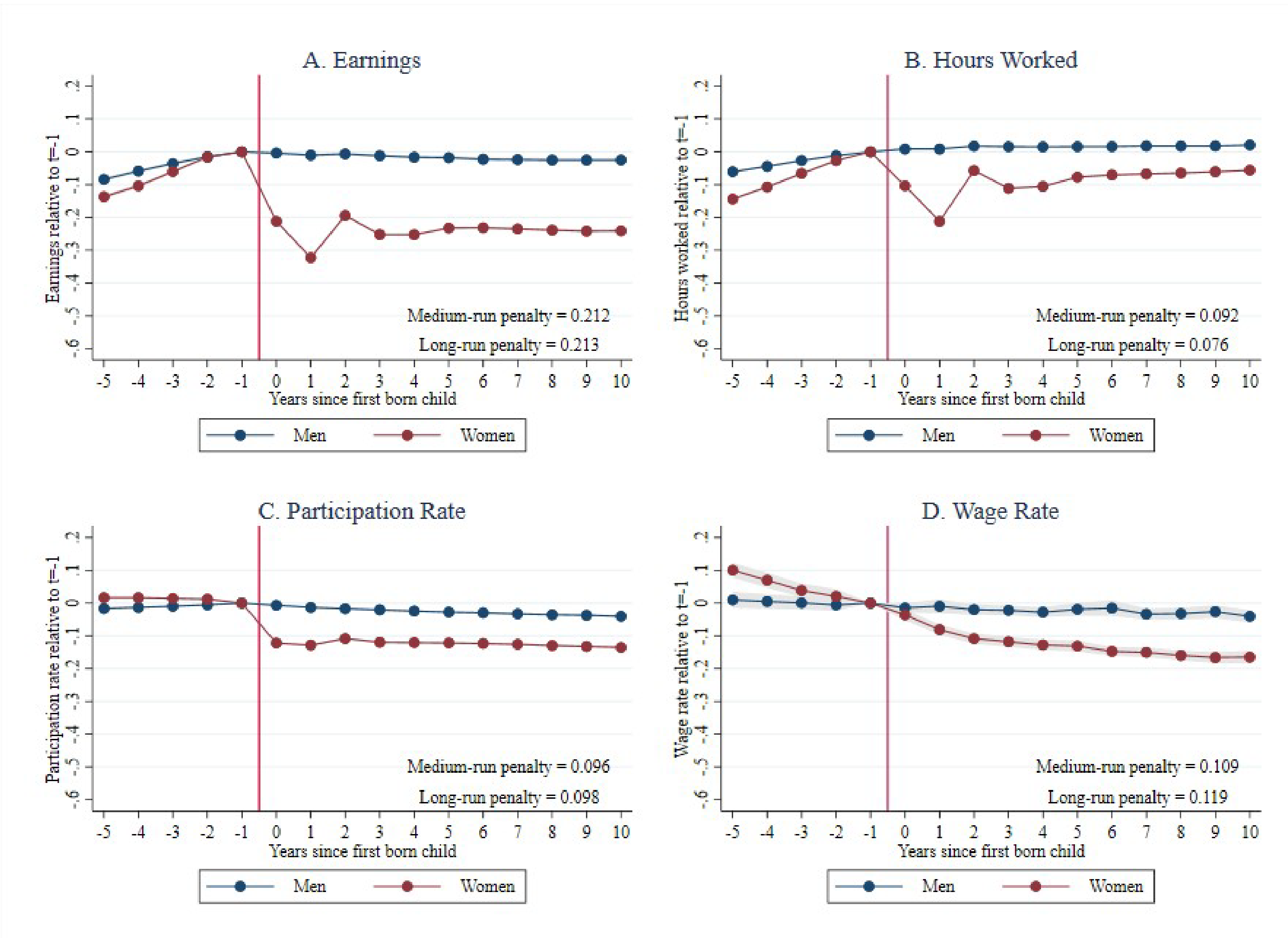
5 Heterogeneity in the impact of children – event study

In this section, we first evaluate the impacts of children on first-time parents' labour market outcomes using the event study approach by Kleven et al. (2019b) presented in Section 3. As in Kleven et al. (2019b), we estimate the gender-specific impact of children on earnings and the three margins of earnings: hours worked, participation rate, and wage rate.

Figure 1 displays the results of the event study approach. More specifically, it captures the event time coefficient for men and women relative to the counterfactual outcome of having no child. Thus, the figure captures how the impacts of having a child differ between the two genders for five years before and ten years, after childbirth. As discussed earlier, we also show the medium-run child penalty (between -1 and 5), and as can be seen from the figure, the gender gap in earnings becomes more prevalent mainly during this event time interval. Similar to Kleven et al. (2019b), we estimate a substantial child penalty in earnings both in the short-run, medium-run and long-run, i.e., the child penalty in earnings is estimated to 31.0% in t = 1, 21.2% in t = 5 and 21.3% in t = 10. If anything, we find slightly larger child penalties than Kleven et al. (2019b). This is remarkable as we analyze more recent cohorts, indicating that the size and trend of child penalties are persistent over time in Denmark.

In panels B-D in Figure 1, we estimate the impact of children on the three margins and display the dynamics, again, like Kleven et al., (2019b). From the panels, it is clear that all three variables contribute to explaining the overall child penalty in earnings, as all panels display a significant impact of children on women. Moreover, we estimate a positive child penalty for all three margins, which are comparable in size with the findings of Kleven et al. (2019b), even though it is based on a more recent sample. We note that while the estimated child penalty in our study is lower in hours worked and participation rate compared to Kleven et al. (2019a), the estimated (long-run) child penalty for the wage rate is approximately 4 percentage points higher compared to the estimate in Kleven et al. (2019a).

Figure 1: Child penalties in earnings, hours worked, participation and the wage rate



Notes: The panels reproduce Figure 1 in Kleven et al. (2019b) for a more recent sample. Instead of using the 1985-2003 cohorts, we use children born in 2002-2010 and a slightly different measure of earning (see Table 2 for more details on sample selection). We plot the gender-specific impacts of children over the event window for earnings and the three margins of earnings. The vertical line indicates the first birth, and the reported medium-run and long-run child penalties are measured in t = 5 and t = 10, respectively, by Equation 2. We follow Andresen and Nix (2022) and include confidence intervals in all four panels constructed by bootstrapping 500 times.

Figure 1 includes confidence bands for the gender-specific dynamics. The confidence bands are estimated using 500 bootstrap replications similar to the approach of Andresen and Nix (2022), with clusters on the identifier of each parents, and by stratifying the sample by gender to ensure balance. The confidence bands displayed in Figure 1 are the 1st and 99th percentiles in the distribution of the 500 bootstrap estimates. Note however, that in subsequent event study figures we follow Kleven et al. (2019b) and construct 99% confidence bands using robust estimates of the standard errors (i.e., not bootstrapped confidence bands) due to computational limitations.

Identifying the long-run child penalty builds on the fact that we use men as controls for women, similar to previous literature on child penalties (Kleven et al., 2019b; de Quinto et al., 2021; Artmann et al., 2022; Fontenay et al., 2023). Note also that this strategy becomes less informative in the long run as we move further away from the event time, t = 0, as it becomes arguable whether men constitute good controls for women in the long run, since the two genders might not necessarily follow the same trends in the counterfactual case of no children. Nevertheless, for the sake of comparison to Kleven et al. (2019b), we retain this strategy and investigate how the impact of children differs for women relative to men rather than the absolute effect of children.

The key identifying assumption in the event study is the absence of a trend in the period before the event. The presence of a pretrend may indicate anticipation of the event or even the presence of a third factor affecting the development of the outcome (Miller, 2023). Visually comparing pre-period coefficients between the two genders to see if they show the same pretrends (overlapping confidence intervals) reveals that this does not hold in several cases. In many of these cases, even though there is pre-event gender difference in trends, there is also a clear reversal in the female trend after the event. That is, the female outcome was rising faster than male outcome before the event, but the pattern reverses itself after. However, it cannot be said that the pretrends are the same, even in these cases.

The pretrends are highly problematic for 3 out of 46 event studies we have estimated. This occurs for the wage rate in periods -5 and -4, though not for -3 or -2 (see Figure 1, panel D) and for the two lowest levels of education (Figure 7, panels A and B). In such cases, it is difficult to conclude that the effect found is the result of motherhood. Testing for pretrends visually in this way is not without its problems, however. Partly, the tests itself are low-powered, party they test only for linear violations and, furthermore, statistical distortions can be large (Roth, 2022). Therefore, Roth (2022) suggests more pragmatic tests allowing for wider confidence bands or based on applying more context-specific economic intuition to the problem (Rambachan and Roth, 2023).

In the next subsections, we study heterogeneity in the child penalty in earnings across various subgroups hypothesized *a priori* as being relevant. The hypothesised subgroups are based on labour market characteristics presented earlier. First, we investigate differences in industries and employment levels, followed by an investigation of firm-specific characteristics. Finally, we also investigate differences in individual characteristics. The covariate set is large therefore we show only one example from each of the above categories in Section 5, and instead refer the reader to Appendix A2 for the rest of the subgroups. However, we present a summary of all the child penalties found through hypothesized subgroup analysis in Tables 3 and 4.

Within each subgroup, it is crucial to remember that the assignment to a subgroup is not random. Consequently, the estimated child penalties depend on the self-selection of individuals into these specific subgroups. Furthermore, the long-run child penalties also include the potential effects of subsequent children, although robustness tests on the sample of parents with two children yield the same findings (available on request). Thus, the child penalties uncovered here cannot be viewed as a causal effect of the given characteristic on the impact of children, but instead should be interpreted as the causal effect of children conditional on a specific characteristic and conditional on the assumptions in the event study setup. Casarico and Lattanzio (2023) and Bütikofer et al. (2018) also caution against the same when they perform hypothesized subgroup analyses to uncover heterogeneous effects of the child penalty.

5.1 Heterogeneity by industry

Goldin (2014) finds that the gender wage gap varies by industry due to job structures, the level of flexibility, and the convexity of pay schemes affecting how easy it is to accommodate children in one's work life. Furthermore, studies also find evidence of sorting with women selecting more family-friendly firms or sectors due to children or in anticipation hereof (Nielsen et al. (2004), Casarico and Lattanzio, 2023; Pertold-Gebicka et al., 2016). Fontenay et al. (2023) also hypothesise that industries requiring large sacrifices to balance work and child responsibilities lead to large child penalties. However, there might also be selection in that women who are very committed to the labour market select into less family-friendly industries, which would result in a smaller child penalty. Consequently, the child penalties would reflect differences in the type of parents rather than industries.

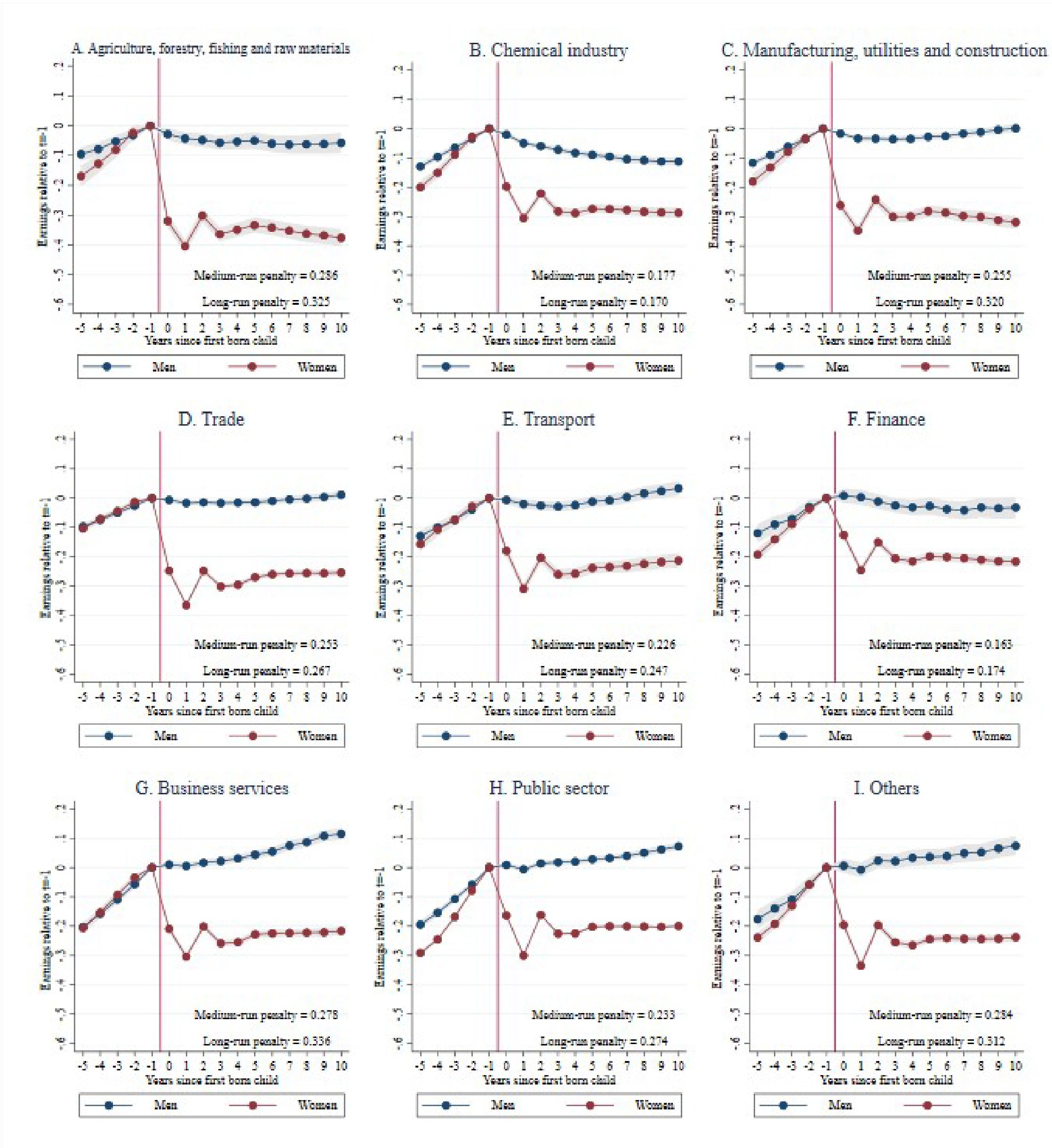
We argue that such selection might not be a concern based on the empirical findings of Kuziemko et al. (2018), who find evidence that women do not fully anticipate the effects of motherhood and the associated costs of having children beforehand. An important implication is that women will choose education and implicitly the industry while they are still young and, according to Kuziemko et al. (2018), be overly optimistic about their post-childbirth labour supply. Despite these findings, there could be some self-selection into specific industries, which could bias the results as the child penalties would capture inherent differences between parents rather than sector-specific differences. With these conflicting dynamics in mind, we investigate the heterogeneity in the child penalty across industries using a similar approach to Fontenay et al. (2023), who investigate how child penalties vary across industries in Belgium.

Figure 2 displays the event study results capturing the effects of children on earnings for each of nine industries in our data. There is some noise in definitions of industries, as, especially, the public sector industry is complex as we do not have a variable directly indicating whether an individual is employed in the public sector. For instance, being employed in health does not necessarily mean being employed in the public sector. Consequently, we can only view the public sector variable as a proxy for the actual public sector.

Figure 2 shows that a similar pattern emerges across all industries, i.e., a distinct drop in women's earnings after childbirth. However, the size of the penalty and the trend for men vary across industries (e.g., flat in agriculture, rising in business services). In addition, focusing on the pretrends of the two genders in each industry, they all display similar positive trends in earnings. The earnings growth for the two genders start to diverge after the birth of the first child. Panel A, which includes individuals employed in agriculture, forestry, fishing and raw materials exhibits the largest long-run and medium-run child penalty, equal to 32.5% and 28.6%, respectively. We observe that women, on average, have more children in this industry compared to other industries (available on request). Thus, the size of the child penalty aligns with Kleven et al. (2019b), who find that more children lead to a higher long-run child penalty.

Performing another sensitivity check, where we condition on the number of children within the event window to be exactly two, we still find a substantial child penalty in this industry which is significantly higher than other industries (available on request). Aside from this, the industry might be assumed to be more traditional, which could further explain the large child penalty. However, we consider the causal dynamics outside the scope of this paper and leave this for further research. Child penalties are lowest in Finance and Chemicals (0.163 and 0.170 in the long run; 0.174 and 0.177 in the medium run). These are traditionally male dominated spheres of work and may therefore indicate a positive selection of women choosing to work in these industries.

Figure 2: Child penalties in earnings, by industry



Notes: The figure shows the impact of children on earnings conditional on industry information at $t = -1$. In addition to the sample restrictions in Table 1, we condition on available industry information. The vertical line indicates the birth of the first child, and the reported medium-run and long-run child penalties are measured in $t = 5$ and $t = 10$, respectively, by Equation 2. Similar to Kleven et al. (2019b), we construct 99% confidence bands using robust estimates of the standard errors.

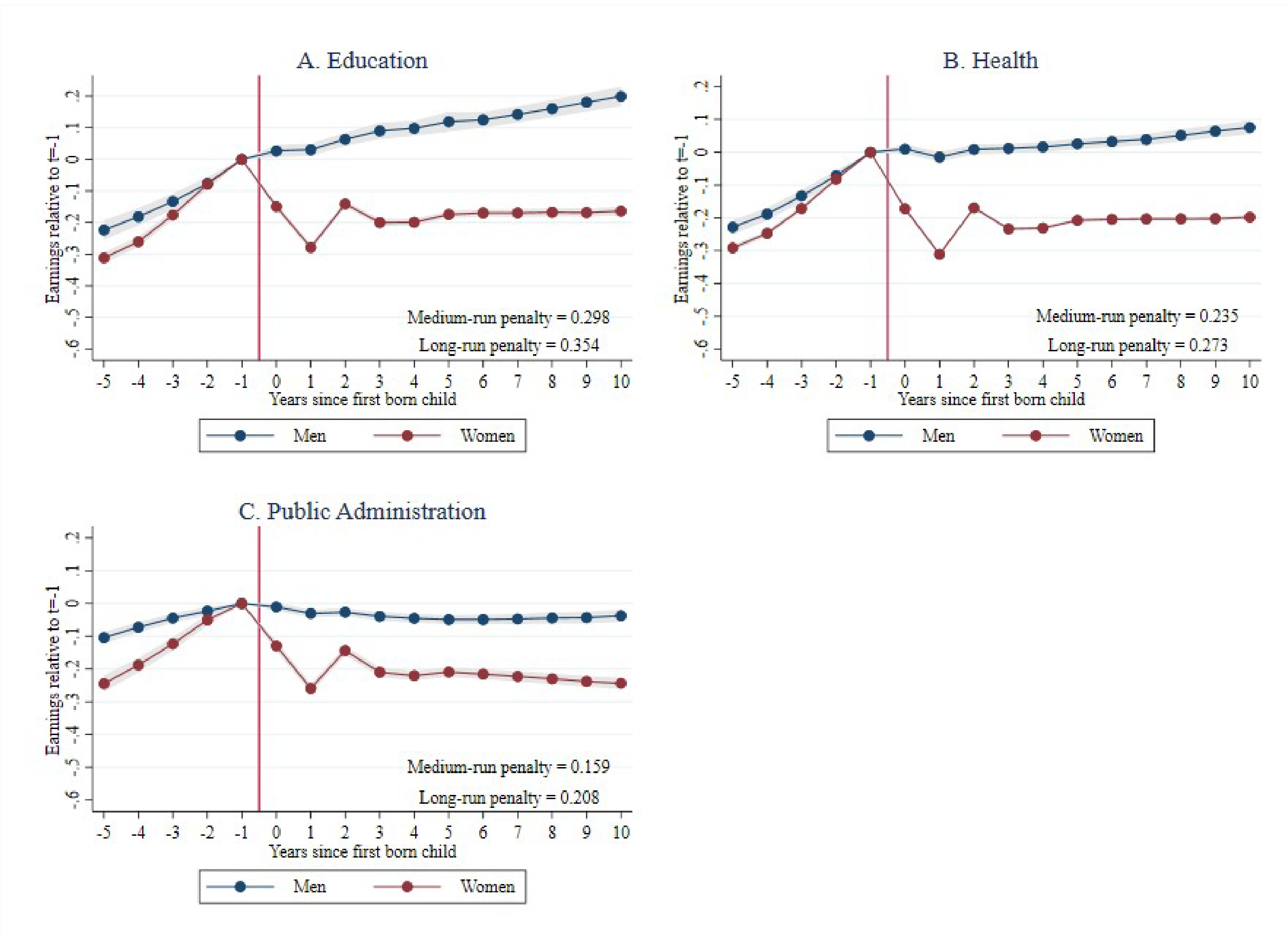
5.2 Heterogeneity within the public sector

Figure 2 showed a large, long-run (medium-run) child penalty equal to 27.4% (23.3%) in the public sector, which was larger than the overall child penalty equal to 21.3% (21.2%). The results are surprising as the public sector is generally considered to be more family-friendly e.g., full salary during parental leave. Nielsen et al. (2004) find that women in the public sector encounter a modest negative effect on wages due to birth-related leaves in the short run. However, the effect is temporary, and the women experience full catch-up over time. On the other hand, women in the private sector face a more significant penalty for birth-related interruptions, and experience limited catching up in the long run. These policies could also have the opposite effect on child penalties, as Nielsen et al. (2004) find that women in the public sector take more parental leave than their peers in the private sector. During these longer breaks, women fail to accumulate experience, which increases the child penalty. However, as described earlier, in a robustness test we found that the total number of children had by parents in different industries (including the public sector) is not driving the child penalty differences between industries.

Work intensity may vary across industries and could help explain why balancing work and child responsibilities as a mother might be challenging in certain sectors. Such challenges may result in a larger child penalty in the health industry where working conditions are particularly challenging with high-speed work, emotionally stressful situations etc. (Eurofound, 2017). Thus, in another analysis we divide our subsample of individuals working in the public sector into three subcategories: education, health, and public administration and estimate the child penalty for each of the three subsamples individually. The estimation results are presented in Figure 3. In all three panels the pretrends are not the same for the gender-specific impacts of children, although the trend for mothers is rising faster than that for fathers before the event and the pattern is reversed after the event. This is suggestive of child penalties but, obviously, not conclusively so. If we are to attribute the post-trend difference to child penalties, then surprisingly, we estimate a child penalty for health which is significantly lower than the child penalty for education. The estimated long-run (medium-run) child penalties are 27.3% (23.5%) for health and 35.4% (29.8%) for education. Thus, this would imply that Denmark has a lower child penalty in the health sector compared to the one in Belgium equal to 33% (Fontenay et al., 2023). Instead, the child penalty in the educational sector pulls up the total child penalty in the public sector in Denmark.

One may expect that women would leave jobs where working conditions are particularly stressful. Strikingly, we observe that 80% of women employed in the public sector in t = −1 are still employed in the public sector in t = 10. This is much larger than the share of individuals who remain in other industries. Furthermore, we see a shift towards the public sector for women in all industries. On average, around 30% shift to the public sector from other industries. The pattern may indicate self-selection into the public sector induced by parenthood, most likely due to its more family-friendly characteristics (Nielsen et al., 2004, Pertold-Gebicka et al., 2016). The same pattern is absent for men, where only around 8% shift to the public sector from other industries after childbirth.

Figure 3: Child penalties in earnings, by public sector division



Notes: The figure shows the impact of children on earnings conditional on being employed in the public sector at t = −1. In addition to the sample restrictions in Table 1, we condition on available industry information. The vertical line indicates the birth of the first child, and the reported medium-run and long-run child penalties are measured in t = 5 and t = 10, respectively, by Equation 2. Similar to Kleven et al. (2019b), we construct 99% confidence bands using robust estimates of the standard errors (i.e., not bootstrapped confidence bands) due to computational limitations.

5.3 Heterogeneity by level of employment

We investigate heterogeneity also by level of employment, by dividing our sample into the following four categories: management, top-level employment, mid-level employment, and base-level employment. Balancing work and home responsibilities may be more challenging for those working in management jobs which require longer working hours and high pace of work. The results are displayed in Figure 4. All four panels display similar, though not the same, pretrends for the two genders. Taking a pragmatic approach and for the reasons mentioned earlier, we do not consider the common trend assumption violated even though the pretrends in Panel C differ. As mentioned previously, the female trend rises faster than the male trend in all cases before the event, and this pattern reverses sharply after, suggestive of an effect of the event. However, this is not conclusive evidence, of course. Interpretating the effects nonetheless as child penalties, overall, the same pattern emerges across all four categories, but the size of the penalty differs, significantly. Panel B, i.e., mid-level employment, displays larger long-run and medium-run child penalties than panels A, C and D. i.e., base-level, top-level and managers. Mothers employed in management are subject to the lowest child penalty, as we estimate the long-run (medium-run) child penalty to be 10.7% (16.5%), which is significantly lower than the population average of 21.3% (21.2%). Thus, the observed pattern might seem counterintuitive, as occupations within management are generally seen as more demanding due to the difficulties in achieving work-life balance (Eurofound, 2017).

This may indicate that the low child penalty arises from the characteristics of managers rather than working conditions i.e., selection into management is a possible explanation for the low child penalty. As can be seen from Table A.1.2 in the Appendix, only 0.82% of women are employed in managerial positions in the year before entering parenthood, whereas this holds for more than twice as many men in $t = -1$. This may indicate that women in management positions prior to having children may exhibit higher levels of commitment to the labour market and their careers. We also investigate and find that female managers give birth to their first child later in life and have a lower number of children in total compared to the three remaining employment categories, again supporting the hypothesis that female managers are more committed to their careers (available on request).

Figure 4: Child penalties in earnings, by employment level

Notes: The figure shows the impact of children on earnings conditional on employment level at t = −1. In addition to the sample restrictions in Table 2, we condition on available employment information. The vertical line indicates the birth of the first child, and the reported medium-run and long-run child penalties are measured in t = 5 and t = 10, respectively by Equation 2. Similar to Kleven et al. (2019b), we construct 99% confidence bands using robust estimates of the standard errors (i.e., not bootstrapped confidence bands) due to computational limitations.

5.4 Heterogeneity by firm characteristics

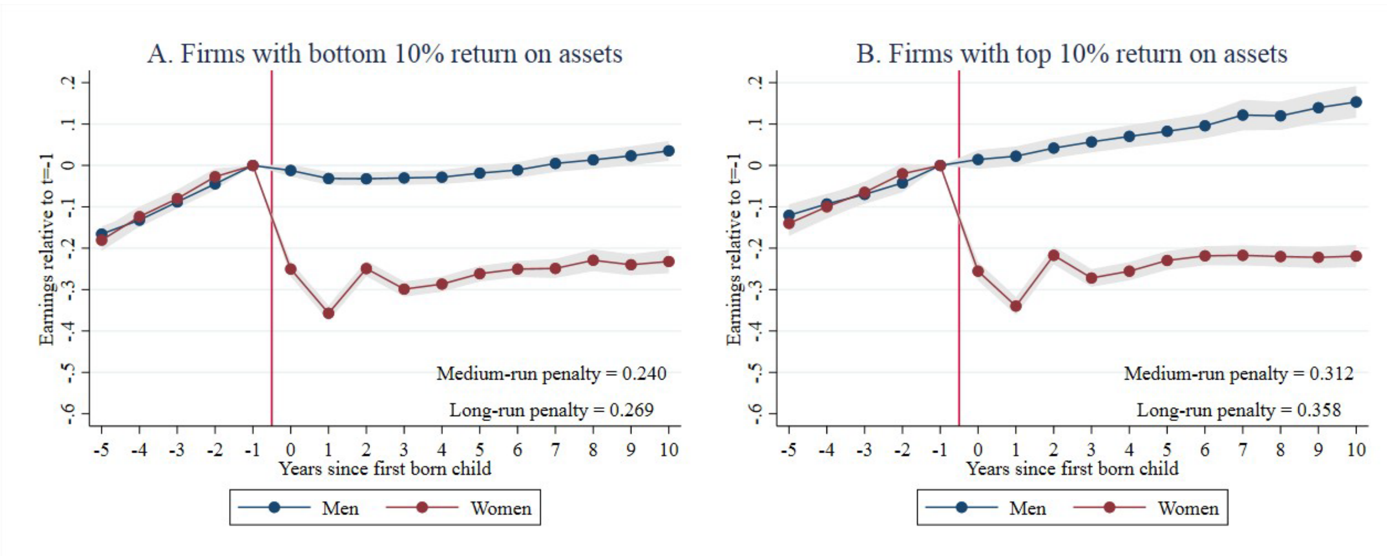
In this section we examine whether differences in firm characteristics can be a possible source of heterogeneity in the child penalty. Our hypotheses are primarily motivated by previous literature showing that firm differences can help describe the size of the estimated gender pay gap between men and women, as argued in Section 2. We thus examine differences in child penalties between individuals employed in firms with high and low returns on assets (ROA). This subsample analysis is motivated by Casarico and Lattanzio (2023), who, using Italian data, find that mothers are more likely to select into firms with a lower value added compared to non-mothers. This complements Brunns (2019) who finds evidence of sorting with women self-selecting into less-productive, low-wage firms and that childbirth is an important driver of these patterns. We also have available a variable that measures the value added in the firm, and we hypothesize that individuals who seek employment in firms with high and low value added differ. [4]

Here, we use ROA as a proxy and hypothesise that the ROA might be a source of heterogeneity. Figure 5 shows the estimated child penalties conditional on the size of pre-birth ROA in the firm. We use two measures of ROA – firms with above/below median ROA (see appendix A.2.2) and in Figure 5, firms with top 10% and bottom 10% ROA, to isolate extremes. We note that the estimated child penalties are similar for the individuals employed in firms with high ROA and firms with low ROA. If there is any difference, the estimated child penalty is slightly smaller for firms with low ROA.

A firm's ROA most likely depends on the business cycle, and to test the sensitivity of the estimated child penalties, we create alternative subsamples which account for the business cycle by using the median in each year as the relevant threshold (available on request). We find that the estimated long-run child penalties are insensitive to the alternative splitting that accounts for fluctuations in the business cycle, as the threshold change does not alter the estimated child penalties by much.

[4] Economic value added and ROA are different measures. Value added measures the generated value from investments and accounts for the cost of capital, while ROA measures the efficiency of a firm's assets to generate profits. Therefore, value added might provide a more detailed description of a firm's value creation (Harrison Jr. et al., 2014).

Figure 5: Child penalties in earnings, by the firm's financial efficiency



Notes: The figure shows the impact of children on earnings conditional on information about the size of a firm's ROA at t = −1. High ROA is defined as firms with top 10% ROA, while low ROA are defined as firms with bottom 10% ROA. In addition to the sample restrictions in Table 2, we condition on available accounting information. The vertical line indicates the birth of the first child, and the reported medium-run and long-run child penalties are measured in t = 5 and t = 10, respectively by Equation 2. Similar to Kleven et al. (2019b), we construct 99% confidence bands using robust estimates of the standard errors.

Turning to other firm-specific differences, Casarico and Lattanzio (2023) has shown, using Italian data, that the child penalty is higher in firms with a high share of women. To investigate this hypothesis in a Danish context, we exploit the information about the share of women at the firm level and construct a variable that indicates whether a majority of the employees in a firm are female or male. Figure 6 shows the gender-specific impact of children on the two subsamples, and we estimate a positive relationship between the long-run child penalty and the share of women, similar to Casarico and Lattanzio (2023). The long-run (and medium-run) child penalty for individuals employed in firms with a majority of male employees is around 18%, and the long-run child penalty for individuals employed in firms with a majority of female employees is 25.5%, while the medium-run penalty is 23.3%

Figure 6: Child penalties in earnings, by the share of female employees

Notes: The figure shows the impact of children on earnings conditional on female representation in the firm at t = −1. High female representation is defined as a majority of women. In addition to the sample restrictions in Table 2, we condition on available information about female representation. The vertical line indicates the birth of the first child, and the reported medium-run and long-run child penalties are measured in t = 5 and t = 10, respectively by Equation 2. Similar to Kleven et al. (2019b), we construct 99% confidence bands using robust estimates of the standard errors.

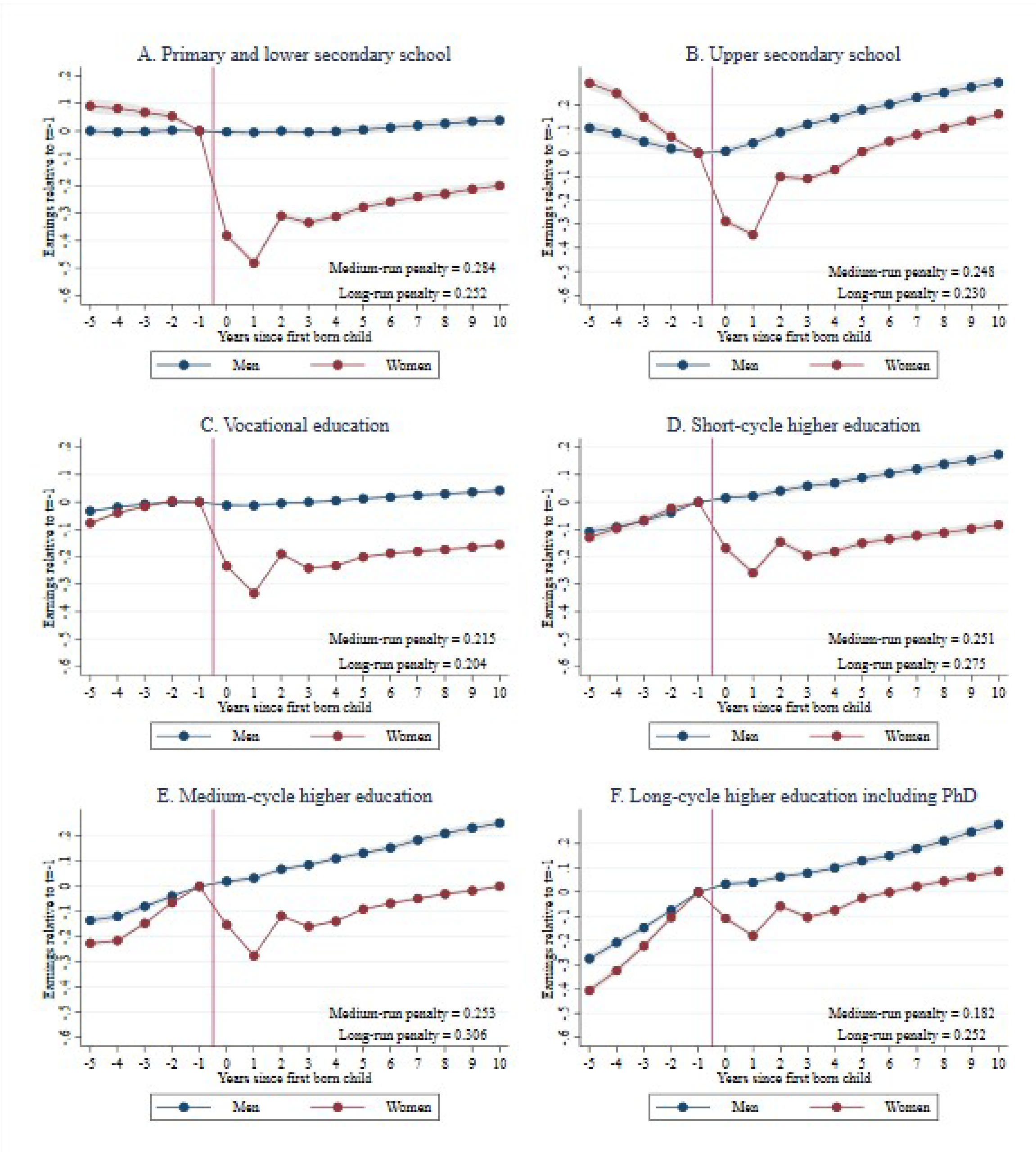
To investigate the sensitivity of the results presented in Figure 6, we explore two additional divisions of subgroups to investigate the relationship between the share of female employees and the child penalty further. First, we inspect the child penalty for a more detailed division of the share of females, 0% − 25%, 26% − 50%, 51% − 75%, 76% − 100%, respectively. Second, we hypothesise that not only does the female representation matter for the child penalty, but also the female representation at each employment level. The results for the first hypothesis reveal a concave relationship between the share of female employees and the estimated child penalties ranging from 15% for individuals in firms with less than 25% female employees to 18% for firms with more than 75% female employees (available on request). We also find significantly higher penalties when interacting female share with employment level. Thus, the results from Figure 6 fail to capture the heterogeneity across the distribution of female representation.

Consequently, our results indicate that both the share of women in similar positions and share of women in the firm overall are sources of heterogeneity in the child penalty. Again, this is not necessarily a causal effect of the representation of women and the revealed patterns might simply arise from selection into specific firms. Nevertheless, the share of women at each employment level reveals a source of heterogeneity in child penalties, but the effect is sensitive to the definition of female representation.

5.5 Heterogeneity by education level

We now turn to individual characteristics and investigate whether there is heterogeneity in the child penalty depending on the individual's education level. As mentioned earlier, child penalties have been found to be higher for the college educated compared to women with no college diploma (Anderson et al., 2002). This is because the costs of temporarily exiting the labour market are higher for the highly educated because of the higher return they obtain on their human capital investments. Because of assortative mating, highly educated women are often married to highly educated (and high wage) men, meaning that family income may not necessarily suffer much if highly educated women exit the labour market following childbirth. However, this may not be fully internalized as Kuziemko et al. (2018) show evidence that highly educated women may underestimate the costs of motherhood. Thus, we may expect higher penalties for highly educated women, but as the evidence from Figure 7 shows, the long-run child penalty does not vary substantially with education. However, as mentioned earlier, the pretrends cannot be compared in panels A and B so we cannot conclude that we have identified a child penalty in these two cases. In the other cases, except for panel D (short-cycle higher education) and, to some extent, panel C (Vocational education), pretrends are not the same either, but as argued earlier, the reversal in the female trend in these cases may be more amenable to a causal effect interpretation as the female trend was growing faster than the male trend pre childbirth.

Figure 7: Child penalties in earnings, by education level



Notes: The figure shows the impact of children on earnings conditional on the parent's highest education obtained in the firm at t = −1. The vertical line indicates the birth of the first child, and the reported medium-run and long-run child penalties are measured in t = 5 and t = 10, respectively by Equation 2. Similar to Kleven et al. (2019b), we construct 99% confidence bands using robust estimates of the standard error5.6

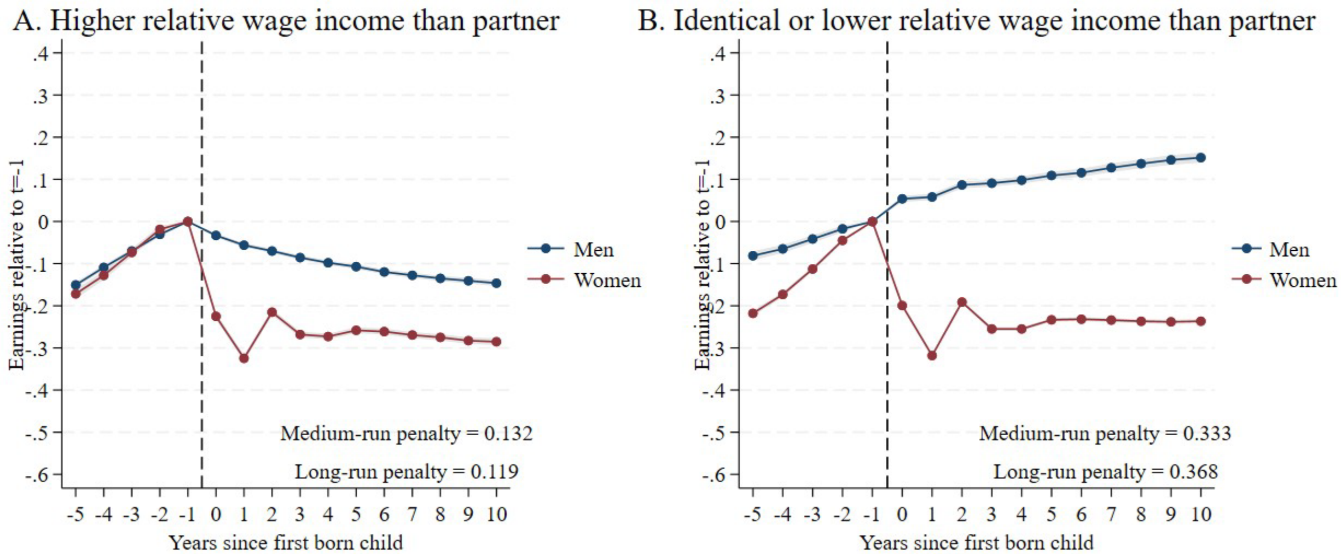
5.6 Heterogeneity by intrahousehold dynamics

Another individual dimension that can affect the child penalty is how individuals interact in a household, i.e., intrahousehold dynamics, which can provide information on gender roles, decision-making, and the division of responsibilities between the mother and father within a household. We examine whether there is heterogeneity in the child penalty across indicators of intrahousehold dynamics. To do this, we aim to examine the heterogeneity across different opportunity cost dynamics within a household.

As discussed in Section 2, the household decisions regarding the allocation of time between the labour market and home production depend on various factors, e.g., opportunity costs and gender norms. Household decision-making models explain this by incorporating the role of gender norms and specialisation. Empirical evidence using American and Norwegian data suggests that comparative advantage in the labour market or home production cannot fully explain the child penalties in earnings, indicating a potential role of gender norms (Cortés and Pan, 2020; Andresen and Nix, 2022). Gender norms are captured here via the relative wage income earned by each partner. When the mother's share exceeds 50%, this may indicate a more egalitarian gender norm operating within the household (Bittman et al., 2003; Bertrand et al., 2015).

We estimate the long-run child penalty for two subsamples conditional on whether the partner (mother or father) earns more than 50% of the total wage income of both partners within the household. The results of the estimation are displayed in Figure 8. Here, for the first time, we see a considerable difference in the size of the penalty between the two groups. For the sample of mothers and fathers who earn more than 50% of total wage income, the long-run and medium-run child penalties are equal to 11.9% and 13.21%. This is significantly lower than the population average. For the sample of mothers and fathers who earn less than 50% of household income, the corresponding child penalties are 36.8% and 33.3% in the long run and medium run, respectively. Note that while the pretrend is the same for the first group, for the second group the pretrends are both rising with the female trend rising faster than the male trend, but they are not the same. However, appealing to the same argument that the female trend reverses after birth, we can cautiously interpret the difference in effects as a child penalty. Therefore, our analysis supports the finding that there is heterogeneity in the child penalty based on the relative wage income before birth, which proxies for the within-household gender norm. However, it is important to point out again that we cannot identify a causal relationship between the prebirth relative wage income and child penalties. Mothers who earn a larger proportion of total wage income may differ in preferences or job types compared to mothers who earn less than their partners. This potential source of selection limits our ability to draw causal conclusions.

Figure 8: Child penalties in earnings conditional on relative wage income



Notes: The figure shows the impact of children on earnings conditional on earning a higher/lower share of household wage income than one's partner, i.e., we divide our sample into two subsamples based on the individual's share of the household wage income being >/<= 50%. In addition to the sample restrictions in Table 2, we condition on available leave information for both parents. The vertical line indicates the birth of the first child, and the reported medium-run and long-run child penalties are measured in t = 5 and t = 10, respectively by Equation 2. Similar to Kleven et al. (2019b), we construct 99% confidence bands using robust estimates of the standard errors.

The child penalties in Figures 1-8 and in Appendix A2 are computed at t = 5 (medium run) and at t = 10 (long run), i.e., two specific event dates. To align the event study findings with those from the machine learning part in which we look at the change between t = -1 and t = 5, in Tables 3-5 we show average penalties calculated over these two time periods, from -1 to 5 in the first column and from -5 to 10 in the second column.

Averaging penalties from t = -1 to t = 5, and from t = -5 to t = 10, respectively, do not give very different results from those calculated at t = 5 or t = 10 (Figure 1). In both cases, the penalties associated with the main labour market outcomes are substantial, do get smaller over time, but not by much.

Table 3. Medium-run and long-run child penalties in main labour market outcomes; mean and 95% CI

Outcome	Medium-run child penalty t = –1 to t = 5	Long-run child penalty t = –5 to t = 10
Earnings	0.198	0.155
	(0.127-0.269)	*(0.103, 0.207)*
Hours worked	0.107	0.084
	(0.058, 0.156)	*(0.059, 0.108)*
Participation	0.088	0.062
	(0.059, 0.118)	*(0.033, 0.091)*
Wage rate	0.067	0.055
	(0.036, 0.098)	*(0.018, 0.092)*

In Tables 4-5 we present the heterogeneity uncovered by the hypothesized subgroup method for all the variables we consider, again considering between the two event intervals, medium run and long run. It can be easily seen from these tables that penalties do not differ much between these two periods. Medium-run penalties are larger in size, but convergence over time is slow. Of course, we treat fertility as exogenous here, and more recent studies that instrument fertility (e.g. using successful IVF interventions) find convergence in the long-run penalty (Lundborg et al., 2024).

Although the figures showed evidence of some degree of heterogeneity across the many dimensions studied, at the 95% level of confidence, the only significant factor that generates heterogeneity in the child penalty in the hypothesized subgroup analysis is the prebirth relative wage income, which is associated with significant heterogeneity in the long-run child penalty.

Table 4: Heterogeneity in medium-run and long-run child penalties, by hypothesized subgroups; industry, sector, and firm characteristics; mean and 95% confidence intervals

Subgroup	Medium-run child penalty, t = –1 to t = 5	Long-run child penalty, t = –5 to t = 10
Industry		
Agric., forestry, fishing & raw mat.	0.255 *(0.168, 0.341)*	0.211 *(0.142, 0.279)*
Chemical	0.165 *(0.107, 0.223)*	0.128 *(0.087, 0.170)*
Manufacturing, utilities & construc.	0.221 *(0.145, 0.297)*	0.191 *(0.127, 0.255)*
Trade	0.232 *(0.151, 0.312)*	0.173 *(0.104, 0.243)*
Transport	0.189 *(0.121, 0.257)*	0.153 *(0.095, 0.211)*
Finance	0.147 *(0.091, 0.203)*	0.119 *(0.080, 0.158)*
Business Services	0.230 *(0.151, 0.309)*	0.183 *(0.100, 0.265)*
Public Sector	0.196 *(0.125, 0.267)*	0.175 *(0.123, 0.226)*
Other	0.231 *(0.149, 0.313)*	0.196 *(0.128, 0.263)*
Public sector division		
Education	0.228 *(0.144, 0.312)*	0.211 *(0.145, 0.277)*
Health	0.198 *(0.127, 0.269)*	0.172 *(0.117, 0.227)*
Public Administration	0.137 *(0.084, 0.190)*	0.137 *(0.105, 0.169)*
Employment level		
Base-level	0.186 *(0.116, 0.255)*	0.135 *(0.085, 0.184)*
Mid-level	0.209 *(0.134, 0.285)*	0.170 *(0.098, 0.242)*
Top-level	0.145 *(0.092, 0.199)*	0.132 *(0.092, 0.171)*
Management	0.144 *(0.093, 0.196)*	0.106 *(0.067, 0.145)*
Firm's financial efficiency		
Firms with top 10% ROA	0.267 *(0.176, 0.358)*	0.215 *(0.133, 0.298)*
Firms with bottom 10% ROA	0.217 *(0.142, 0.292)*	0.162 *(0.093, 0.232)*
Firm's financial efficiency		
Firms with higher than median ROA	0.207 *(0.135, 0.279)*	0.153 *(0.091, 0.215)*
Firms with lower than median ROA	0.193 *(0.124, 0.261)*	0.139 *(0.083, 0.195)*
Firm's net income		
Firms with high net income	0.175 *(0.113, 0.236)*	0.127 *(0.074, 0.179)*
Firms with low net income	0.234 *(0.153, 0.315)*	0.172 *(0.103, 0.241)*
Share of female employees		
Female majority among employees	0.204 *(0.131, 0.276)*	0.163 *(0.104, 0.222)*
Male majority among employees	0.175 *(0.113, 0.237)*	0.133 *(0.087, 0.179)*

Significant difference in bold

Table 5: Heterogeneity in medium-run and long-run child penalties, by hypothesized subgroups; individual characteristics, intrahousehold allocation, and region; mean and 95% confidence intervals

Subgroup	Medium run child penalty, t = –1 to t = 5	Long run child penalty t = –5 to t = 10
Age		
Above 30 years	0.164 *(0.105, 0.224)*	0.129 *(0.084, 0.175)*
Below 30 years	0.213 *(0.135, 0.291)*	0.148 *(0.094, 0.201)*
Education		
Primary and lower secondary school	0.297 *(0.190, 0.405)*	0.194 *(0.102, 0.285)*
Upper secondary school	0.250 *(0.158, 0.342)*	0.163 *(0.083, 0.243)*
Vocational education	0.202 *(0.129, 0.275)*	0.156 *(0.104, 0.208)*
Short-cycle higher education	0.206 *(0.133, 0.278)*	0.168 *(0.105, 0.232)*
Medium-cycle higher education	0.213 *(0.135, 0.290)*	0.192 *(0.135, 0.250)*
Long-cycle higher education including PhD	0.155 *(0.010, 0.211)*	0.147 *(0.105, 0.188)*
Relative wage income		
Higher than partner	0.145 *(0.088, 0.201)*	**0.091 *(0.045, 0.137)***
Identical or lower than partner	0.275 *(0.180, 0.371)*	**0.252 *(0.185, 0.319)***
Region		
Bornholm	0.203 *(0.127, 0.279)*	0.152 *(0.108, 0.196)*
Copenhagen	0.206 *(0.131, 0.280)*	0.160 *(0.102, 0.218)*
Funen	0.208 *(0.134, 0.282)*	0.168 *(0.112, 0.224)*
Urban area of Copenhagen	0.185 *(0.118, 0.252)*	0.142 *(0.094, 0.190)*
Northern Jutland	0.203 *(0.131, 0.275)*	0.164 *(0.111, 0.216)*
North Zealand	0.169 *(0.107, 0.230)*	0.130 *(0.090, 0.170)*
South Jutland	0.206 *(0.133, 0.279)*	0.167 *(0.113, 0.222)*
West and South Zealand	0.202 *(0.131, 0.273)*	0.163 *(0.111, 0.216)*
West Jutland	0.194 *(0.125, 0.263)*	0.161 *(0.113, 0.209)*
Eastern Jutland	0.220 *(0.142, 0.300)*	0.174 *(0.114, 0.234)*
Eastern Zealand	0.183 *(0.117, 0.249)*	0.147 *(0.097, 0.198)*

Significant difference in bold

In the next section we apply the causal forest algorithm to search for heterogeneity in the child penalty within the same data setup. It is possible that the machine learning framework can find additional sources of heterogeneity which were not hypothesized from the outset as being potentially important, for instance by taking into account interactions and trying different threshold levels.

6 Heterogeneity in the impact of children - causal forest

Given that the hypothesized subgroup analysis showed significant heterogeneity in the child penalty in at least one factor as summarized in Tables 4 and 5, we now perform a similar exercise using a machine learning algorithm applying the causal forest approach (Athey and Wager, 2019). While the hypothesized subgroups presented in the previous section allowed us to concentrate on specific characteristics one at a time, the causal forest allows us to study how different characteristics interact when considered all together. The causal forest approach allows for testing for thresholds/non-linearities that the researcher could not have envisioned beforehand and without running into multiple hypothesis testing issues.

The causal forest method extends tree-based machine learning procedures for the purpose of predicting treatment heterogeneity. A regression tree is built up by calculating mean outcome in each "leaf", where a leaf is selected by grouping together observations which have similar values of a covariate, X. To identify similar values, the tree splits the data at some threshold value of X, sending observations which have $x<X$ to the left, and observations with values $x>X$ to the right. Further splits are made until some stopping criterion is reached and the final prediction of Y made (see Davis and Heller, 2017). In regression trees, the split is chosen to minimize the mean squared error within the sample, while in the causal trees the criterion is to prioritize heterogeneity and penalize high variance in leaf estimates. Because the single-tree method can generate an unstable and high variance prediction, the causal forest method averages over many trees to generate a smooth and stable prediction.

We estimate the honest causal forest using the causal forest package in R created by Tibshirani et al. (2022), see [Causal forest — causal_forest • grf (grf-labs.github.io)](). As in Athey and Wagner (2019), we follow a two-step procedure suggested by Basu et al. (2018) for obtaining greater precision. In the first step we estimate a pilot forest on all the available features, while in the second step we only retrieve the features that showed a reasonable number of splits (based on variable importance) and use them to train the final causal forest.

We apply the defaults, and grow 2000 trees in the forest, randomly using half the sample for building each tree and applying half of this again for determining splits and the other half for prediction. Since we apply honest splitting, each estimation sample tree is pruned so that no leaves remain empty. See Appendix A6 for the chosen parameter settings.

All categorical variables are converted to 0/1 dummies. Some categories are less frequently observed for some variables in our sample. For example, we observe few women in management positions and in the agriculture, forestry, fishing and raw materials industry. To address any problems that may arise due to there being too few observations, we redefine our industry dummies at a more aggregate level resulting in new industry dummies.

First, we define a primary, building, and construction category which covers agriculture, forestry, fishing, raw materials, utilities, and construction industries. Next, we keep the manufacturing industry as described in Table 4 but combine the trade and transport industries. Then, we keep the large industry consisting of public administration, health, and education industries. Lastly, we collect the finance and business services industries into one category. We include these broader categories and the more detailed industry categories from Table 4 in our sample. Likewise, we define broader educational categories, i.e., low educational level, vocational education, and high educational level. We define high educational level as medium-cycle or long-cycle higher education, including PhD level. Further, low educational level is defined as the residual educational categories from Table 5 excluding vocational education.

6.1 Omnibus test

As a first test of the presence of treatment heterogeneity, we present the results of the Omnibus test of calibration of the forest, which is essentially a test of the goodness of fit of the forest (see e.g. Chernozhukov et al., 2020). As regressors, it uses the best linear fit of the forest predictions on the held-out data and the mean forest prediction[5]. Standard errors reported are one-sided and are heteroskedasticity-robust. The test yields the estimates of the mean forest prediction and the differential forest prediction from this regression. A value equal to 1 for the mean forest prediction means that the prediction is ‘correct’. A value equal to 1 for the differential forest prediction means that the heterogeneity of the forest estimates is well captured. Also, as long the differential forest prediction coefficient is significantly greater than zero as indicated by its p-value, we can reject the null hypothesis of no treatment heterogeneity.

Table 6: Omnibus test of the causal forest estimation

Variable	Estimate	Standard error	T-value	Prob>t
Mean forest Prediction	1.185600	0.287224	4.1278	1.832e−05***
Differential forest prediction	0.864422	0.077059	11.2177	2.2e−16***

Note: Best linear fit which uses as regressors the forest prediction on held-out data and mean forest prediction and reports the one-sided heteroskedasticity robust standard errors (jackknife covariance type HC3). P-values are shown for H_0: Estimate = 0, H_a: Estimate>0.

The results reveal that the mean prediction is no different from 1 (mean forest prediction estimate = 1.186, CI = [0.623, 1.749]). Heterogeneity estimates captured by the differential forest prediction appear well-calibrated, too, since the estimated coefficient is close to 1, CI = [0.713, 1.015] showing evidence of treatment heterogeneity.

Some of our variables contain many missing values, e.g., the firm financial variables. Missing values in the covariates is in general not a problem for the grf algorithm, except that by default missing values are sent to the left of a given threshold. Therefore, in cases where there are many missing values, we should use caution when making interpretations. Another aspect to remember is that we only estimate the numerator of equation (1) in Section 3, and not the child penalty, which is equation (1) scaled by the counterfactual level of female earnings, i.e., the predicted earnings of females minus the event-specific coefficients, which measures what females would have earned had they not had children in any year t.

6.2 Variable importance

Variable importance shows how many times a variable in the covariate set is selected in the causal forest estimation. While it is tempting to conclude that the variable which is chosen most often is the most important one, Javier et al. (2023) argue that splitting on a variable does not imply that the variable is important in the data generating process. The same

[5] See Omnibus evaluation of the quality of the random forest estimates via calibration. — test_calibration • grf (grf-labs.github.io). We started by giving the forest all 114 features but this yielded a mean forest prediction > 1, which indicated to us a problem with calibration due to overfitting/excessive complexity. Therefore, following Athey and Wager (2019), we trained a pilot forest on all the features and then selected only those features that gave a reasonable number of splits for the main forest estimation.

argument is made by Athey and Wager (2021), who argue that, in the case of high correlation between two or more variables, the algorithm splits on one of the variables without regard to economic theory. Consequently, if the algorithm does not split on a variable, it does not necessarily imply that the variable is not important for the heterogeneity in child penalties, as there may be several ways to construct similar subgroups using other splitting variables.

Because the causal forest averages across many trees, this may mitigate the problem somewhat. Furthermore, there is no evidence of strong correlation between variables in our setup. Appendix A3 Figure A.3.1 shows correlations between all the variables used in the causal forest estimation. The darker the colour, the greater the strength of the correlation. The only noteworthy correlations are that some of the firm financial variables appear to be somewhat positively correlated with each other. The grey boxes indicate undefined correlations. For instance, the industry categories are naturally mutually exclusive and thus, by definition, are uncorrelated.

In Appendix A5, we show variable importance for all 24 features in the forest, selected from the pilot analysis. In Figure 9 we graphically show variable importance for features with variable importance > 0.1. The variable with the greatest importance for treatment heterogeneity is mother's relative wage income, with importance equal to nearly 70%.

Table 7 shows numerically variable importance for features with importance > 0.1. Other than mother's relative wage income, these include employment in the manufacturing, utilities and construction industry, share of employees at base level within the firm, female representation at the base level within the firm, worker experience, being employed in the primary, building and construction industry and firm revenue.

One issue with the machine learning algorithm is that it gathers more information from continuous variables in the covariate set as opposed to discrete or binary variables. Thus, if relative wage income appears to be important, it may be because it is a continuous variable. Another issue is that in the case of many missing values the algorithm by default chooses to go left, rendering interpretation of the finding difficult. Variables with high shares of missing values are the firm financial variables such as ROE, ROA, balance, equity, net income and revenue growth. We correlate the original variable importance measure and the variable importance measure residualized for both continuous/discrete variable type and percentage missings and find a correlation of 0.981 between the two. Thus, it is unlikely that either the nature of the variable (discrete/continuous) or missing values are driving the findings.

Figure 9. Variable importance for selected features

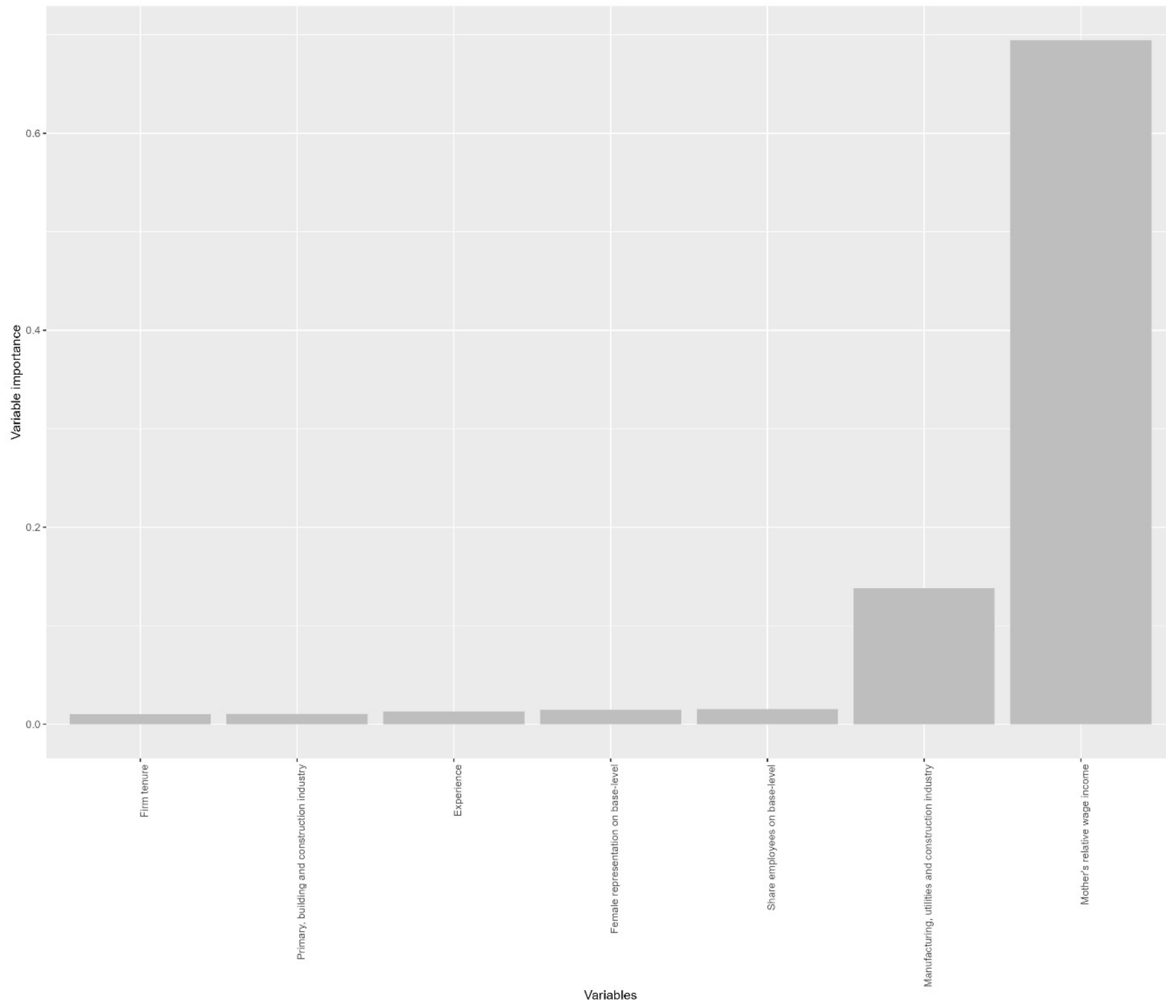


Table 7: Variable importance for features with importance > 0.1

Variable	Importance %
Mother's relative wage income	0.695
Manufacturing, utilities and construction industry	0.138
Share employees at base level	0.015
Female representation at base level	0.014
Experience	0.013
Primary, building and construction industry	0.011
Firm tenure	0.010
Total	0.895

6.3 CATE distribution

The CATE or conditional average treatment effect is the estimate of τ from equation (5) in the Methodology section and is the equivalent of the estimated child penalty in the conventional analysis. The difference between this and the child penalty is that, here, we do not scale by counterfactual income. In Figure 11, distributions of CATE are shown for each decile, where we subdivide each decile into sub-deciles for clarity.

CATE means ranges from -12,000,000 DKK at the bottom to +15,000,000 DKK at the top, see Figure 11. Recall that this is the estimate of the gender difference in earnings between t = -1 and t = 5. The CATE turns positive from the mid-5th decile and on. Nearly all estimates are significant, although noisy at the very bottom and at the very top, most likely due to a paucity of observations at these two extremes in the distribution. Taking the average of 5th and 6th deciles, the mean estimated CATE is a 126713 DKK shortfall between men and women over the 6-year interval, as a result of first childbirth. In the conventional analysis we found a medium run child penalty in earnings equal to 19.8% between t = -1 and t = 5 (see Table 3). Dividing the 126713 DKK by, respectively, the sample average father's and mother's total earnings over the 6-year period (approximately 2,201,324 DKK and 1,572,361 DKK) gives a penalty bounded between 6% and 8%. This is somewhat lower than the penalty from the event study analysis but closer to more recent child penalty estimates found in Bensnes et al. (2023) and Lundborg et al. (2024), who account for non-randomness in fertility choice and timing. Recall that the child penalty in the event study analysis is scaled by the mother's counterfactual income, had she not had children. As we have not modified the algorithm in causal forest to estimate the counterfactual maternal income, we provide instead bounds on the child penalty in a pragmatic way here, by assuming that the counterfactual income for mothers, had they not had children, is bounded by the observed income of mothers and the observed income of fathers.

We also compare CATE estimation to the AIPW or Augmented Inverse-Probably Weighted estimation method, which is doubly robust and models the propensity to be treated together with the outcome equation. It does so by first estimating propensity scores, which are derived by predicting treatment (in our case, male*t = 5) from the X covariate vector. However, AIPW also uses another set of information, which is the counterfactual outcome predicted from X and, thus, it is able to predict the treatment effect based on X. Combining these two approaches, AIPW uses the propensity scores from the first step to weigh the outcomes under treatment and control from the second step, thereby making it possible to derive a weighted average outcome. This is shown in Figure 12. Since CATE estimates do not differ much from AIPW estimates, this reassures us that our CATE estimates are free from model misspecification.

Figure 11. CATE distributions, per decile and subdecile

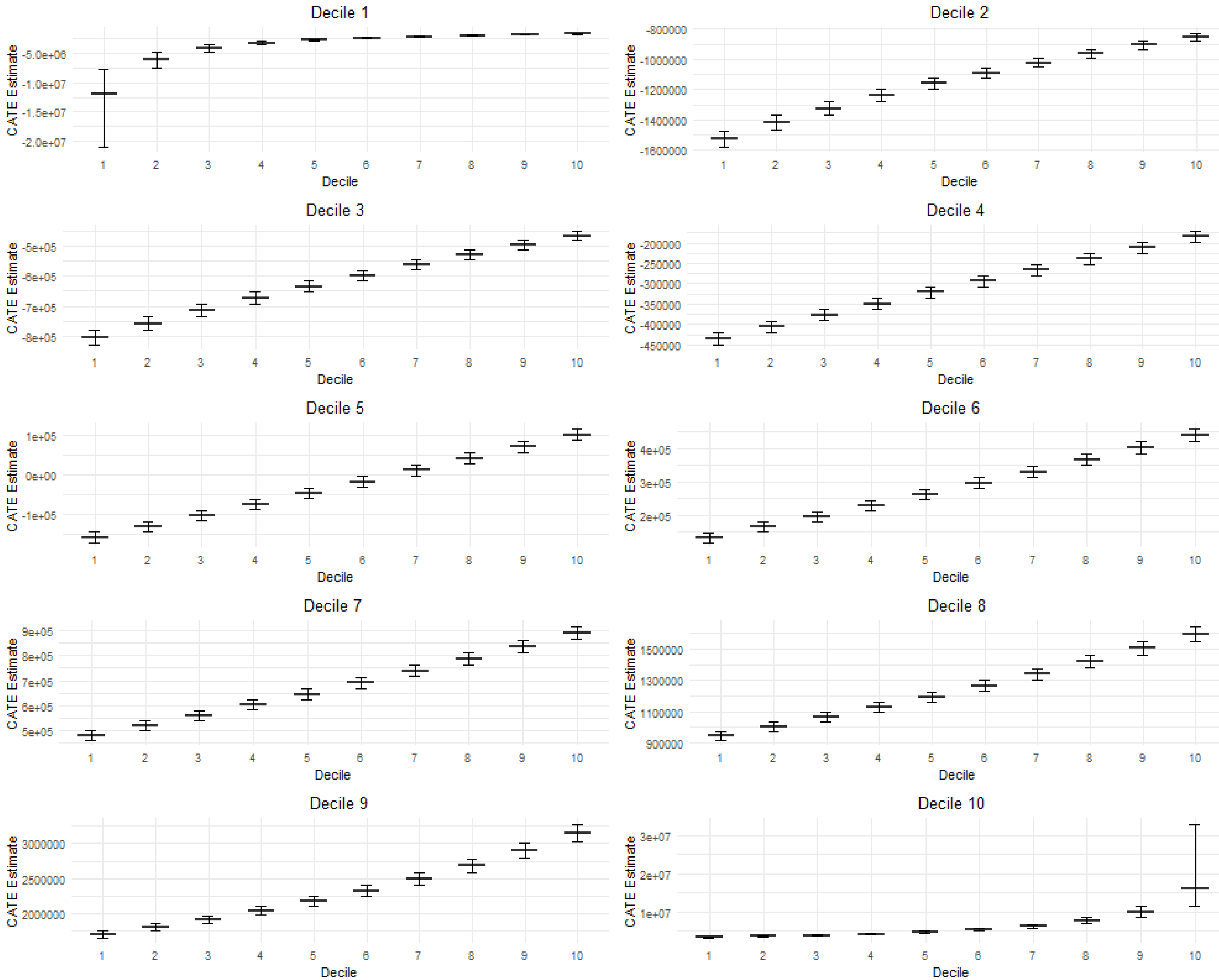
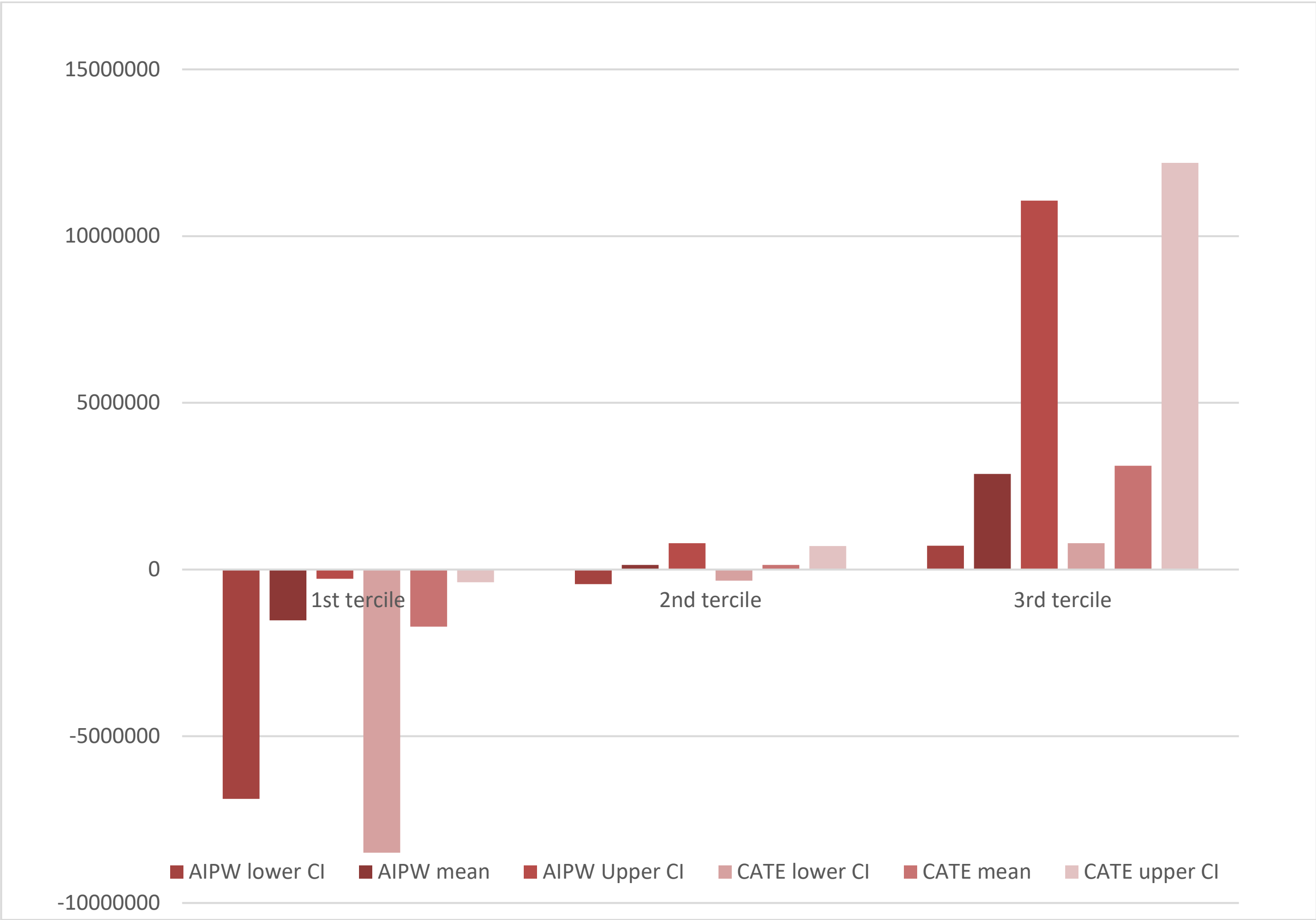


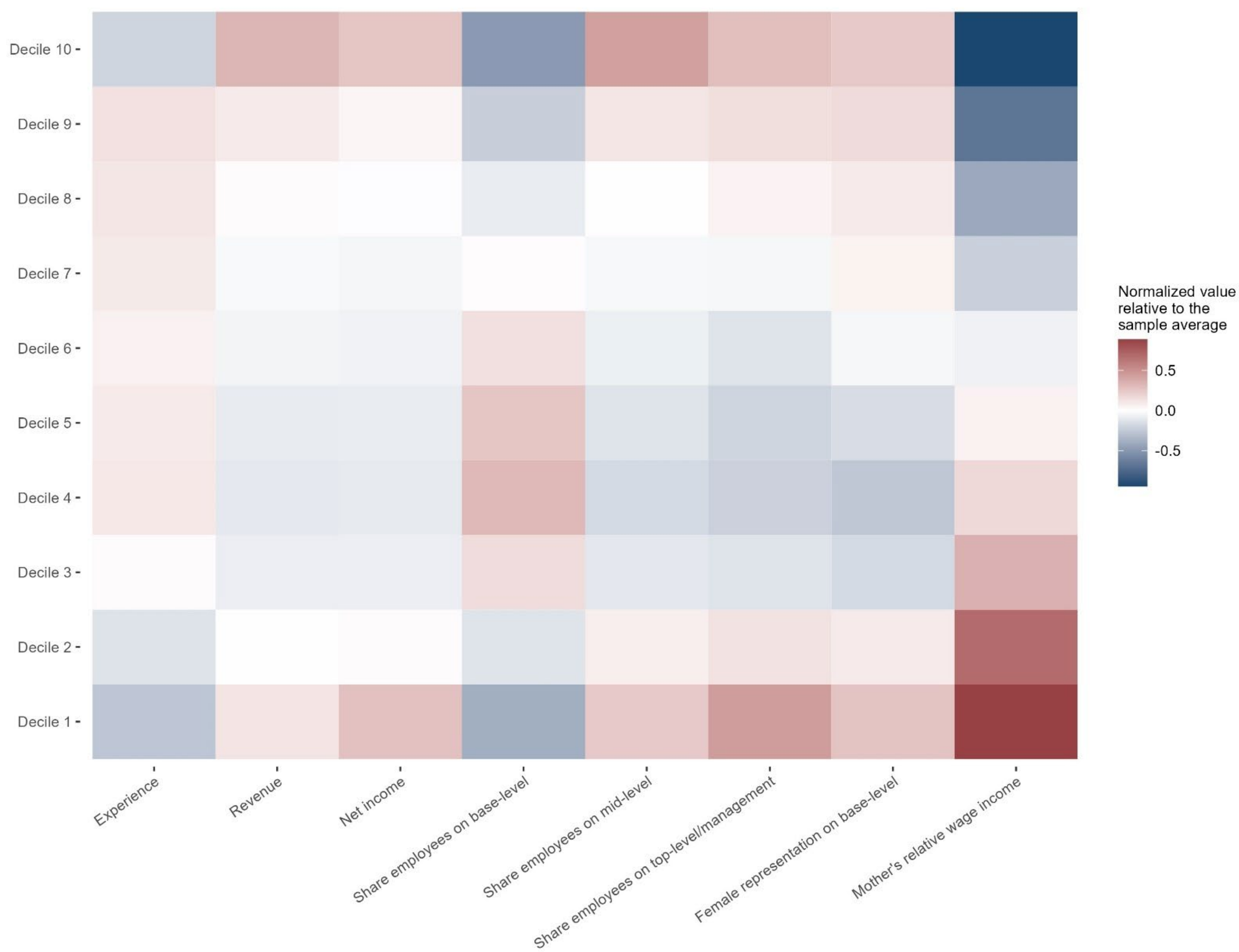
Figure 12: CATE estimates vs. AIPW estimates



6.4 CATE distribution by features

We show the deciles of the distribution of the conditional average treatment effect (CATE) obtained from the causal forest estimation over each of the features in our data by way of heat maps. The covariates include characteristics of parents (experience, municipality type, employment, mother's relative wage income etc.), characteristics of firms, industries, employee shares and female representation at various levels. In Figure 12, we show the CATE distribution over individual background characteristics and firm characteristics. The heterogeneity analysis by other relevant features is shown in Appendix Figures A.4.1-A.4.2.

Figure 12: CATE distribution by parents' experience, firm financial variables, share of employees and female representation and mother's relative wage income



It is clear from Figure 12 that both in the first decile of the CATE and in the 10th (topmost) decile, mother's relative wage income is at least 0.5 times higher and lower, respectively, than the average relative wage income. This means that mothers with high (low) relative wage income are associated with a lower (higher) child penalty. This is by far the strongest effect among the individual background characteristics and firm characteristics. Other interesting results from the heat maps can be seen in Appendix A4. From Figure A.4.1, it can be seen that among the firm financial variables, firm revenue is 0.3 times higher than average at the top decile of the CATE distribution. Thus, high revenue firms seem to be associated with higher child penalties. When it comes to firm net income, equity and balance, however, results are more ambiguous, as firms with higher-than-average values in these characteristics are represented both at the top and the bottom of the CATE distribution. From Figure A.4.2 we can see that firms more likely to be located in Primary, building and construction and in Manufacturing, utilities and construction are overrepresented in the 4th decile of CATE, i.e., having relatively lower penalties.

Thus, in a similar fashion to the conventional analysis, mother's relative wage income is found to be a factor generating heterogeneity in estimated penalties. However, using causal forest methodology we also find that certain firm financial variables and industry types tend to be associated with higher or lower penalties. Of course, we cannot scale the estimated CATE by the mother's counterfactual level of income in the causal forest algorithm. This may of course lead to some difference in the findings compared to the hypothesized subgroups analysis. We find the median estimated medium-run penalty (earnings difference) between mothers and fathers to be around 127,000 DKK, which when scaled by the average mother's and father's incomes over the 6-year period (rough bounds on the mother's counterfactual income) yields a child penalty of between 6% and 8%, which is significantly lower than what is found in the conventional analysis (19.8%).

7 Discussion and conclusion

The aim of the paper was to investigate sources of heterogeneity in the child penalty. We did this using two independent approaches. First, the hypothesized subgroups approach allowed focusing on specific characteristics one at a time, which we surmised at the outset would be a relevant source of heterogeneity. Second, the causal forest approach allowed us to analyse the child penalty conditional on multiple different characteristics in combination with each other. The two approaches, while not being causal, provide independent sources of empirical evidence that can complement each other in achieving a better understanding of the child penalty. Clearly, there can be multiple sources of potential bias due to selection into length of education, type of industry and firm and workplace type. However, in the absence of a randomized setup or exogenous variation we rely on a conditional independence assumption to identify heterogeneous treatment effects.

In the conventional analysis, we investigate differences in child penalties across individual, labour market, industry and firm-specific characteristics such as financial indicators but find little evidence of heterogeneity in any of these dimensions. The only significant dimension of heterogeneity we find is by the pre-childbirth division of household income. I.e., the child penalty is found to be substantially lower when the individual's relative income is higher than their partner's.

In the machine-led analysis we also find that mother's relative wage income is an important source of heterogeneity. However, we also uncover heterogeneity by industry and a few firm-level characteristics. Since both methods - conventional and machine-driven - agree that the relative contribution to household income is an important source of heterogeneity in the child penalty, this indicates to us the robustness of this finding. Assuming exogeneity, this can indicate the importance of gender norms within the household for the child penalty in Denmark.

We do not have any other measure of gender norms and commitment to egalitarian principles in the register data, although we note that the effect of prebirth relative wage income arises even after parents' experience and education measured pre-birth is included in the covariate set. Thus, this finding can imply that gender norms regarding time allocation and specialization in market work vs. housework could be the source of heterogeneity in the child penalty and, thus, of the remaining gender inequality in earnings in Denmark. This aligns well with existing literature in the field of gender inequality who also point to the transmission of norms over generations (Kleven et al., 2019b; Kleven et al., 2022a; and Datta Gupta et al., 2023). Also, Andresen and Nix (2022) point to gender norms as a persistent driver of the child penalty within same-sex couples. Correspondingly, Blau and Kahn (2017) also suggest gender roles and the gender division of labour as important factors in accounting for the gender wage gap.

Recent literature has debated whether policy changes are at all needed when it comes to the child penalty if comparative advantage and specialization arising from differences in preferences and utility optimization underlie the child penalty and not necessarily a lack of possibilities, Kleven et al. (2019b). Kleven et al. (2021) estimate the child penalty in reported life satisfaction. The authors find no significant difference in the gender-specific impact of children on life satisfaction between mothers and fathers after childbirth. Thus, female life satisfaction is somehow compensated even though the mothers are subject to stark changes in labour market outcomes compared to the fathers. On the other hand, using Austrian and Danish data and the event study framework, Ahammer et al. (2024) find that parenthood places a greater mental burden on mothers than fathers, leading to a long-term gap in antidepressant use.

Bertrand (2002) argues that the generation of norms is a complex process - norms can be the outcome of sticky gendered stereotypes within the family and society about the relative skills and roles of women and men (i.e., “women are better at caregiving” etc.). In fact, Bertrand argues that gender norms can become self-fulfilling if they become internalized and determine gender stereotypical choices made with respect to education, family leave or work (Bertrand, The Richard T. Ely Lecture, 2020). This debate also feeds into the role of public policies targeting maternity and paternity leave. Empirical research from Austria by Kleven et al. (2022b), which we presented in Section 2, found that governmental interventions such as parental leave reforms and expansions of childcare had no significant effect on reducing the long-run child penalty. The authors even found a negative effect of parental leave reforms on the short-run child penalty. For instance, Sweden was one of the first countries to implement earmarked paternity leave, which led to a small dip in the earnings trajectories of men in the short run. However, the dip eventually levels off, and the effect of earmarked paternity leave disappears in the long run (Kleven et al., 2019a). Andresen and Nix (2022b) also find small and insignificant effects of paternity leave use on the child penalty in Norway. However, unlike in the Austrian case, they find that a large expansion of publicly provided childcare in Norway reduced the child penalty by 25%. Thus, there is still a case for universal childcare provision.

When it comes to paternity leave, Bertrand (2020) again argues that there *is* a role for public policy for enforcing counter-stereotypical behavior such as polices to encourage fathers to take more leave. She cites research from Japan and Saudi Arabia showing that the low takeup rate of men of such policies can either be due to too low a financial incentive or due to "plurastic ignorance" – i.e., even though the individual fathers may not prescribe to a gender-stereotypical norm themselves, they may incorrectly believe that many others in society believe in the norm, and that they will be sanctioned if they depart from it (Miyajima and Yamaguchi, 2017; Bursztyn, González, and Yanagizawa-Drott, 2018). For Germany, Unterhofer and Wrohlich (2017) have analysed the impact on gender norms of the introduction of the German father quota in 2007. They found that the reform implied more gender equal norms in the grandparents' generation.

In countries like Denmark where maternity and paternity leave has been equalized since August 2022 and childcare is universal and subsidized, there may still be a role for other types of policies. First, flexible working arrangements and more inclusive workplace cultures could create a better setting for mothers to balance child and labour market responsibilities. This perspective is consistent with Goldin (2014), who argues that a reduction in gender differences must be driven by a development of features in the labour market, such as temporal flexibility. Indeed, our results also showed that certain firm and industry features generated heterogeneity in the child penalty. Second, pay transparency requiring firms to report statistics on gender pay gaps will help address pay disparities. This policy has been implemented in the European Union in March 2023 with the hope that it will discourage firms from maintaining a gender pay gap within the organisation. Pay transparency can also strengthen women's bargaining power, which has been found to be a significant part of the gender wage gap (Bruns, 2019; Card et al., 2016). Third, increased public awareness on the unequal wage consequences of having children can potentially improve the societal focus on gender gaps and gender inequality also along the career ladder and in leadership positions (Diversity Council Report, 2022). Finally, as argued by Bertrand (2020), public information policy can complement family policy combating "plurastic ignorance" when it comes to paternity leave by breaking down sticky gender stereotypes.

To our knowledge, this is the first paper to apply a causal machine learning method to study heterogeneity in the child penalty. Consequently, it contributes to the extensive and emerging literature on child penalties. The traditional method to investigate heterogeneity in the child penalty using the event-study approach is easy to use but is dependent on hypothesized subgrouping. In contrast, the machine learning method provides a data-driven approach to exploit a data setup with many covariates and flexibly uncover important features, even though it comes with some algorithmic limitations. Therefore, we suggest that future research investigates how to mitigate some of the limitations associated with utilizing this novel method. For instance, there is a need to alter the algorithm to account for differences in child penalties instead of what is estimated in the model, which is the absolute difference between the event-specific impacts of children for mothers and fathers.

Overall, the results from both conventional and machine-driven methods indicate the existence of child penalties across various labour market characteristics in the medium run, although child penalties estimated conventionally appear substantially larger. This may indicate that the machine learning approach, by combining features flexibly, can better capture the effects of omitted variables compared to the conventional analysis. However, both methods are based on assumptions of unconfoundedness, particularly that the fertility decision is random. More recent studies that take the fertility decision into account find that child penalties do not persist in the long run. Combining event studies and causal machine learning with IVs or experimental data will overcome some of these limitations.

References

Ahammer, A., Glogowsky, U., Halla, M. and Hener, T. (2024). The Parenthood Penalty in Mental Health. Evidence from Austria and Denmark. IZA Discussion Paper no. 16459.

Andersen, S. H. (2018). Paternity leave and the motherhood penalty: New causal evidence. Journal of Marriage and Family, 80(5):1125–1143.

Anderson, D. J., Binder, M., and Krause, K. (2002). The motherhood wage penalty: Which mothers pay it and why? American Economic Review, 92(2):354–358.

Andresen, M. E. and Nix, E. (2022). What causes the child penalty? Evidence from adopting and same-sex couples. Journal of Labor Economics, 40(4):971–1004.

Andresen, Martin and Nix, Emily, (2022b), Can the child penalty be reduced?. Evaluating multiple policy interventions, Discussion Papers, Statistics Norway, Research Department, https://EconPapers.repec.org/RePEc:ssb:dispap:983.

Artmann, E., Oosterbeek, H., and van der Klaauw, B. (2022). Household specialization and the child penalty in the Netherlands. Labour Economics, 78:102221.

Athey, S. and Imbens, G. (2015). Machine learning methods for estimating heterogeneous causal effects. Conference Proceedings. 1504.01132v2 (arxiv.org)

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27):7353–7360.

Athey, S., Imbens, G., Kong, Y., and Ramachandra, V. (2016). causalTree. R package.

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. Observational Studies, 5(2):37–51.

Athey, S. and Wager, S. (2021). Online lecture series: Machine learning & causal inference. Stanford Graduate School of Business.

Basu, S., Kumbier, K., Brown, J.B. and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. Proceedings of the National Academy of Sciences, 115 (8) 1943-1948.

Becker, G. S. (1985). Human capital, effort, and the sexual division of labor. Journal of Labor Economics, 3(1):S33–S58.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives, 28(2):29–50.

Bensnes, S., Huitfeldt, I., and Leuven, E. (2023). Event studies, endogenous timing and the child penalty. Working paper, Statistics Norway, Research Department.

Berniell, I., Berniell, L., de la Mata, D., Edo, M., Fawaz, Y., Machado, M. P., and Marchionni, M. (2022). Motherhood, pregnancy or marriage effects? Economics Letters, 214:110462.

Bertrand, M. (2020). Gender in the 21st Century. Richard T. Ely Lecture. AEA Papers and Proceedings 2020, 110: 1–24. https://doi.org/10.1257/pandp.20201126

Bertrand, M., Goldin, C., and Katz, L. F. (2010). Dynamics of the gender gap for young professionals in the financial and corporate sectors. American Economic Journal: Applied Economics, 2(3):228–255.

Bertrand, M., Kamenica, E. & Pan, J. (2015). Gender Identity and Relative Income within Households. The Quarterly Journal of Economics, 130(2):571–614.

Beskæftigelses Ministeriet (2023). Ny orlovsmodel gælder for børn født fra 2. august 2022. Press release. Accessed: May 1, 2023.

Beuchert, L. V., Humlum, M. K., and Vejlin, R. (2016). The length of maternity leave and family health. Labour Economics, 43:55–71.

Bittman, M., England,P., Sayer, L., Folbre, N. and Matheson, G. (2003). When does gender trump money? Bargaining and time in household work. American Journal of Sociology, 109:186–214.

Blau, F. D. and Kahn, L. M. (1997). Swimming upstream: Trends in the gender wage differential in the 1980s. Journal of Labor Economics, 15(1, Part 1):1–42.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. Journal of Economic Literature, 55(3):789–865.

Bonaccolto-Töpfer. M. and Briel, S. (2022). The gender pay gap revisited: Does machine learning offer new insights? Labour Economics, 78(2022), 102223.

Bonke, J., Datta Gupta, N.and Smith, N. (2004). The Timing and Flexibility of Housework and Men and Women's Wages, in Hamermesh D. S. and G. A. Pfann (eds.), The Economics of time Use, Contributions to Economics Analysis, Vol 271, Elsevier.

Brand, J. E., Xu, J., Koch, B., and Geraldo, P. (2021). Uncovering sociological effect heterogeneity using tree-based machine learning. Sociological Methodology, 51(2):189–223.

Breiman, L. (2001). Random forests. Machine learning 45 (1), 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

Bruns, B. (2019). Changes in workplace heterogeneity and how they widen the gender wage gap. American Economic Journal: Applied Economics, 11(2):74–113.

Budig, M. J. and England, P. (2001). The wage penalty for motherhood. American Sociological Review, pages 204–225.

Bütikofer, A., Jensen, S., and Salvanes, K. G. (2018). The role of parenthood on the gender gap among top earners. European Economic Review, 109:103–123.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott. 2018. "Misperceived Social Norms: Female Labor Force Participation in Saudi Arabia." NBER Working Paper 24736.

Card, D., Cardoso, A. R., and Kline, P. (2016). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. The Quarterly Journal of Economics, 131(2):633–686.

Casarico, A. and Lattanzio, S. (2023). Behind the child penalty: understanding what contributes to the labour market costs of motherhood. Journal of Population Economics, pages 1–23.

Chan, F. and Mátyás, L. (2022). Econometrics with Machine Learning, chapter 3.4, 9.3. Springer.

Chernozhukov, V., Demirer, M., Duflo, E., and Fern´andez-Val, I. (2020). Generic machine learning inference on heterogenous treatment effects in randomized experiments. NBER WP 24678.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2013). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Routledge.

Correll, S. J., Benard, S., and Paik, I. (2007). Getting a job: Is there a motherhood penalty? American Journal of Sociology, 112(5):1297–1338.

Cortés, P. and Pan, J. (2020). Children and the remaining gender gaps in the labor market. Working paper, National Bureau of Economic Research.

Cortés, P., Pan, J., Pilossoph, L., Reuben, E., and Zafar, B. (2023). Gender Differences in Job Search and the Earnings Gap: Evidence from the Field and Lab. The Quarterly Journal of Economics.

Craig, L. and Mullan, K. (2010). Parenthood, gender and work-family time in the United States, Australia, Italy, France, and Denmark. Journal of Marriage and Family, 72(5):1344–1361.

Datta Gupta, N., Oaxaca, R.L., and Smith, N. (2006). Swimming Upstream, Floating Downstream: Comparing Women's Relative Wage Progress in the United States and Denmark. Industrial & Labor Relations Review, 59(2):243-266.

Datta Gupta, N., Koch Gregersen, M., Würtz, A., and Smith, N. (2023). Intergenerational transmission of parental leave choices. Study Paper 180, The ROCKWOOL Foundation Research Unit.

De Quinto, A., Hospido, L., and Sanz, C. (2021). The child penalty: Evidence from Spain. SERIEs, 12:585–606.

England, P., Bearak, J., Budig, M. J., and Hodges, M. J. (2016). Do highly paid, highly skilled women experience the largest motherhood penalty? American Sociological Review, 81(6):1161–1189.

Eurofound (2017). Sixth European working conditions survey – overview report 2017. Publication, Publications Office of the European Union, Luxembourg.

European Parliament (2023). Gender pay gap: Parliament adopts new rules on binding pay transparency measures. Press release. Accessed: April 15, 2023.

Fontenay, S., Murphy, T., and Tojerow, I. (2023). Child penalties across industries: Why Job Characteristics matter. Applied Economics Letters, 30(4):488–495.

Gallen, Y., Lesner, R. V., and Vejlin, R. (2019). The labor market gender gap in Denmark: Sorting out the past 30 years. Labour Economics, 56:58–67.

Giuliano, P. (2020). Gender and culture. Oxford Review of Economic Policy, 36(4):944–961.

Goldin, C. (2006). The quiet revolution that transformed women's employment, education, and family. American Economic Review, 96(2):1–21.

Goldin, C. (2014). A grand gender convergence: Its last chapter. American Economic Review, 104(4):1091–1119.

Goldin, C. and Katz, L. F. (2016). A most egalitarian profession: Pharmacy and the evolution of a family-friendly occupation. Journal of Labor Economics, 34(3):705–746.

Goldin, C., Kerr, S. P., Olivetti, C., and Barth, E. (2017). The expanding gender earnings gap: Evidence from the LEHD-2000 census. American Economic Review, 107(5):110–114.

Goldin, C. and Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. American Economic Review, 90(4):715–741.

Gronau, R. (1977). Leisure, home production, and work–the theory of the allocation of time revisited. Journal of Political Economy, 85(6):1099–1123.

Harrison Jr, W. T., Horngren, C. T., and Thomas, C. W. (2014). Financial Accounting, chapter 12. Pearson Education.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). The Elements of Statistical Learning: Data mining, inference, and prediction, volume 2, chapter 2 and 9.2. Springer.

Indenrigs og Socialministeriet, Danske Regioner, and KL (2009). Status for kommunalreformens gennemførelse. Publication.

Javier, K. K., Keleher, N., Klosin, S., Mühlbach, N. S., Nie, X., and Schaelling, Matt, E. (2023). ML-based Causal Inference. Bookdown.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., and Zweimüller, J. (2024). Do family policies reduce gender inequality? Evidence from 60 years of policy experimentation. American Economic Journal: Economic Policy 2024, 16(2), 110-149.

Kleven, H. (2023). The geography of child penalties and gender norms: Evidence from the United States. Working paper, National Bureau of Economic Research.

Kleven, H., Landais, C., and Leite Mariante, G. (2022a). The child penalty atlas [forthcoming]. Presented at the COSME Gender Economics Workshop 2022.

Kleven, H. (2021). The Child Penalty. Lecture 1, Zeuthen lectures. Copenhagen, September 2021.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., and Zweimüller, J. (2019a). Child penalties across countries: Evidence and explanations. In AEA Papers and Proceedings, volume 109, pages 122–126. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Kleven, H., Landais, C., and Søgaard, J. E. (2019b). Children and gender inequality: Evidence from Denmark. American Economic Journal: Applied Economics, 11(4):181–209.

Kuziemko, I., Pan, J., Shen, J., and Washington, E. (2018). The mommy effect: Do women anticipate the employment effects of motherhood? Working paper, National Bureau of Economic Research.

Lassen, A. S. (2022). Gender norms and specialization in household production: Evidence from a Danish parental leave reform. Working paper, Copenhagen Business School, Department of Economics.

Le Barbanchon, T., Rathelot, R., and Roulet, A. (2021). Gender differences in job search: Trading off commute against wage. The Quarterly Journal of Economics, 136(1):381–426.

Lundborg, P., Plug, E., and Rasmussen, A. W. (2017). Can women have children and a career? IV evidence from IVF treatments. American Economic Review, 107(6):1611–1637.

Lundborg, P., Plug, E., and Rasmussen, A. W. (2024). Is there really a child penalty in the long run? New evidence from IVF treatments. IZA Discussion Paper No. 16959

Ludvig, J. and Mullainathan, S. (2023). Machine learning as a tool for hypothesis generation. NBER Working Paper 31017.

Mühlbach, N. (2020). Essays in Applied Econometrics and Causal Machine Learning. PhD paper.

Miyajima, Takeru, and Hiroyuki Yamaguchi. (2017). "I Want to but I Won't: Pluralistic Ignorance Inhibits Intentions to Take Paternity Leave in Japan." *Frontiers in Psychology* 8: 1508.

Nielsen, H. S., Simonsen, M., and Verner, M. (2004). Does the gap in family-friendly policies drive the family gap? Scandinavian Journal of Economics, 106(4):721–744. OECD (2022). oecd.org/coronavirus/en/data-insights/the-gender-pay-gap-continues-to-close-but-slowly.

Penner, A. M., Petersen, T., Hermansen, A. S., Rainey, A., Boza, I., Elvira, M. M., Godechot, O., Hällsten, M., Henriksen, L. F., Hou, F., et al. (2023). Within-job gender pay inequality in 15 countries. Nature Human Behaviour, 7(2):184–189.

Pertold-Gebicka, B., Pertold, F., and Datta Gupta, N. (2016). Employment Adjustments around Childbirth. IZA Discussion Papers 9685, Institute of Labor Economics (IZA).

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. Review of Economic Studies 90: 2555–2591.

Reuben, E., Sapienza, P., and Zingales, L. (2014). How stereotypes impair women's careers in Science. Proceedings of the National Academy of Sciences, 111(12):4403–4408.

Rosenbaum, P. (2021). Pregnancy or motherhood cost? A comparison of the child penalty for adopting and biological parents. Applied Economics, 53(29):3408–3422

Roth, J. (2022). Pretest with caution: Event study estimates after testing for parallel trends.

AER: Insights 2022, 4(3): 305–322 https://doi.org/10.1257/aeri.20210236

Rygaard, C., Weile, L. (2023). A study of child penalties – heterogeneity across labour market characteristics. Master's thesis, Aarhus University.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5):612–634.

Sieppi, A. and Pehkonen, J. (2019). Parenthood and gender inequality: Population-based evidence on the child penalty in Finland. Economics letters, 182:5–9.

Simonsen, M. and Skipper, L. (2006). The costs of motherhood. An analysis using matching estimators. Journal of Applied Econometrics. 21: 919–934.

Smith, N., Eriksson, T., and Smith, V. (2021). Gender stereotyping and self-stereotyping among Danish managers. Gender in Management: An International Journal, 36(5):622–639.

Smith, N., Smith, V. and Verner, M. (2011). The gender pay gap in top corporate jobs in Denmark: Glass ceilings, sticky floors or both? International Journal of Manpower, Vol. 32 No. 2, pp. 156-177. https://doi.org/10.1108/01437721111130189

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics, 225(2):175–199.

The Danish Foundation for Entrepreneurship (2018). Iværksætterindblikket:Tager temperaturen på iværksætteri og entreprenørskab i Danmark. Publication.

The Diversity City Council (2022). The Danish gender equality paradox in leadership roles. Publication.

Therneau, T. M. and Atkinson, E. J. (2022). An Introduction to Recursive Partitioning Using the RPART Routines. R package.

Tibshirani, J., Athey, S., Sverdrup, E., and Wager, S. (2022). Grf: Generalized Random Forests. R package version 2.2.0.

Unterhofer, U. and Wrolich, K. (2017). Fathers, parental leave and gender norms. DIW Berlin Discussion Paper no. 1657.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242.

Weile, L. (2022). The motherhood penalty: Heterogeneity across education and geography in Denmark. Topics in Economics (P), Aarhus University. Manuscript

Zhao, Q., Small, D. S., and Ertefaie, A. (2022). Selective inference for effect modification via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(2):382–413.

A. Appendix

A1. Descriptive Statistics

Table A.1.1 Descriptive statistics of labour market & firm-specific variables

	Father				Mother			
	Mean	SD	Count	Missing (%)	Mean	SD	Count	Missing (%)
Copenhagen firm	0.254	0.435	165539	11.191	0.249	0.433	160211	14.049
Age of firm	19.754	20.201	156810	15.874	21.190	21.200	133967	28.129
New firm	0.306	0.461	165539	11.191	0.406	0.491	160211	14.049
Revenue (100t DKK)	18657	78491	156810	15.874	16979	68638	133968	28.128
Balance (100t DKK)	24243	114524	119121	36.094	30811	106826	68249	63.386
Equity (100t DKK)	11376	64851	119121	36.094	14785	57930	68249	63.386
Net income (100t DKK)	1845	11737	119121	36.094	2518	11805	68249	63.386
ROA	−11.080	1945.3	119085	36.113	−4.453	1286.7	68227	63.397
ROE	−21.528	2909.9	118993	36.162	−1.777	557.14	68160	63.433
Revenue growth	2.657	295.55	120997	35.087	4.470	380.61	73706	60.458
Share employees – base level	42.245	30.532	154784	16.961	37.998	27.522	149073	20.025
Share employees – mid level	16.696	20.545	154784	16.961	24.048	22.980	149073	20.025
Share employees top/mgmt.. level	16.930	23.424	154784	16.961	19.641	23.923	149073	20.025
Share employees – other levels	23.776	27.080	154784	16.961	18.047	22.597	149073	20.025
Female share – base level	35.470	33.195	154784	16.961	68.874	29.781	149073	20.025
Female share – mid level	31.015	33.282	154784	16.961	58.225	38.168	149073	20.025
Female share – top/mgmt.. level	17.668	25.182	154784	16.961	40.304	35.483	149073	20.025
Female share – other levels	21.874	32.085	154784	16.961	41.110	39.379	149073	20.025
Experience	8.065	5.278	186399	0.000	5.469	4.177	186399	0.000
Firm tenure	3.140	2.769	165539	11.191	2.583	2.146	160211	14.049

Notes: The sample is composed of mothers and fathers who had their first child between 2002-2010 and is subject to the sample selections described in Table 2. All variables are measured in $t = -1$. The share of missing includes unemployed individuals and non-participants in the labour market, e.g., due to studying or receiving social benefits.

Table A.1.2: Descriptive statistics of employment level of parents

	Father	**Mother**	**Total**
Base-level employment	36.685	34.251	35.468
Mid-level employment	15.119	24.730	19.924
Top-level employment	15.918	15.471	15.695
Management	2.018	0.819	1.418
Self-employed	5.043	1.455	3.249
Unemployed	2.709	3.814	3.261
Other	16.666	10.172	13.419
Out of labour market	5.842	9.288	7.565
Total	100.00	100.00	100.00

Table A.1.3: Descriptive statistics of industry representation of parents

	Father	**Mother**	**Total**
Agriculture, forest, and fishing industry	4.554	2.067	3.311
Business services industry	12.940	9.478	11.209
Chemical industry	10.670	4.282	7.476
Finance, estate, and rental industry	4.284	4.003	4.143
Manufacturing, utilities & construction industry	14.652	2.559	8.606
Public administration incl. eduction & healthcare	13.798	35.249	24.523
Trade industry	18.198	16.452	17.325
Transport industry	5.538	3.080	4.309
Other industries	3.888	8.379	6.134
Missing	11.478	14.451	12.964
Total	100.00	100.00	100.00

Table A.1.4: Descriptive statistics of individual background characteristics of parents

	Father				**Mother**			
	Mean	SD	Count	Missing (%)	Mean	SD	Count	Missing (%)
Age	29.730	4.507	186399	0.000	27.792	4.202	186399	0.000
Married	0.299	0.458	186399	0.000	0.299	0.458	186399	0.000
Single	0.123	0.329	186399	0.000	0.123	0.328	186399	0.000
Cohabiting	0.575	0.494	186399	0.000	0.576	0.494	186399	0.000
Relative wage income	0.542	0.109	156843	15.856	0.458	0.109	156843	15.856

Table A.1.5: Descriptive statistics of parents' educational categories

	Father	Mother	Total
Long-cycle higher education incl. PhD	11.803	11.024	11.413
Medium-cycle higher education	14.335	26.496	20.416
Short-cycle higher education	7.333	5.692	14.969
Upper secondary	10.083	13.683	11.883
Vocational education	40.374	29.241	34.807
Primary and lower secondary school	16.073	13.865	14.969
Total	100	100	100

Table A.1.6: Descriptive statistics of area of residence of parents

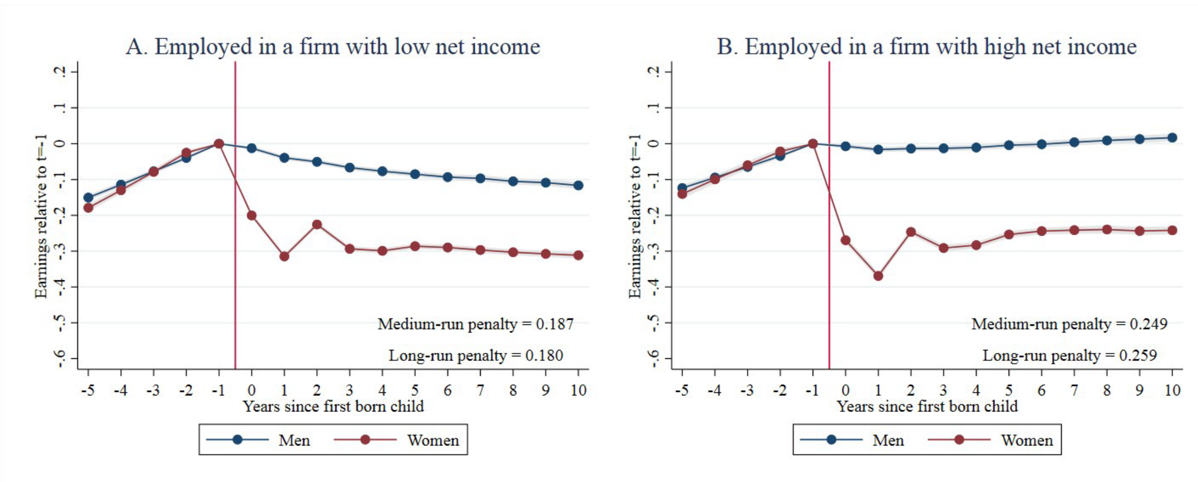
	Father	Mother	Total
Bornholm	0.413	0.403	0.408
Copenhagen	21.706	21.894	21.800
Funen	8.105	8.082	8.093
The Copenhagen Area	7.740	7.758	7.749
North Jutland	9.651	9.622	9.637
North Zealand	5.574	5.520	5.547
South Jutland	11.701	11.704	11.703
West and South Zealand	8.113	8.048	8.080
West Jutland	7.277	7.255	7.266
East Jutland	16.152	16.179	16.166
East Zealand	3.565	3.533	3.549
Missing	0.003	0.001	0.002
Total	100	100	100

Table A:1.7: Descriptive statistics of municipality of residence of parents

	Father	Mother	Total
Capital municipality	32.473	32.648	32.561
Commuter municipality	13.077	12.868	12.972
Metropolitan municipality	15.846	15.943	15.894
Provincial municipality	21.808	21.835	21.822
Rural municipality	16.792	16.706	16.749
Missing	0.004	0.001	0.002
Total	100.00	100.00	100.00

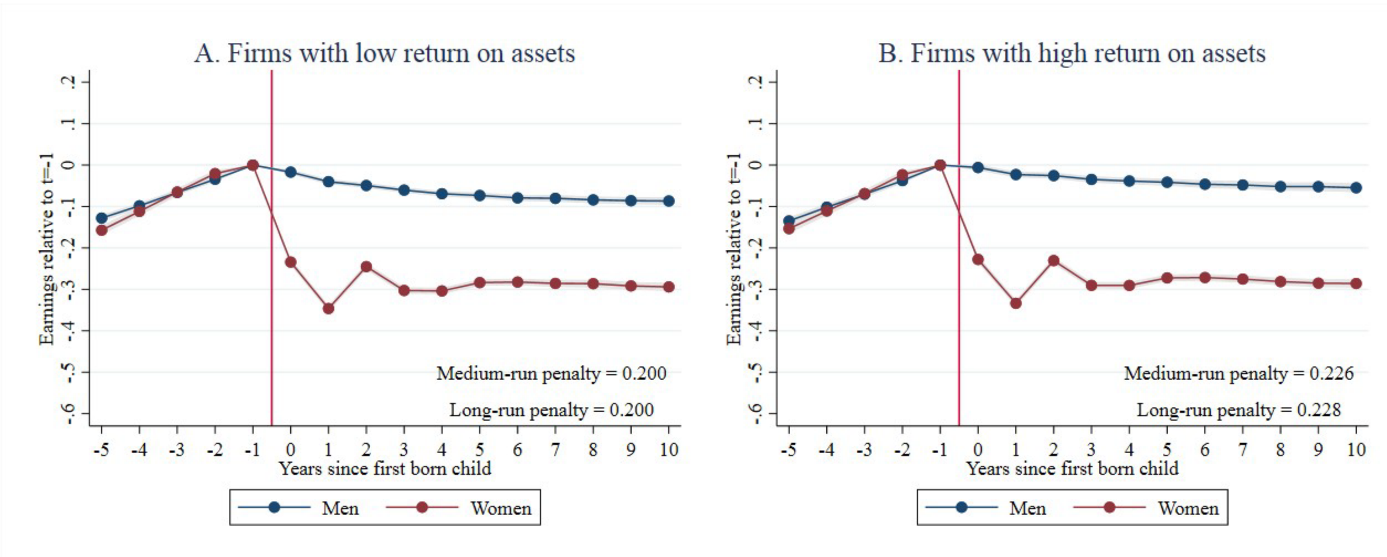
A2. Hypothesized subgroup analysis – other features

Figure A.2.1: Child penalties in earnings. By the firm's net income



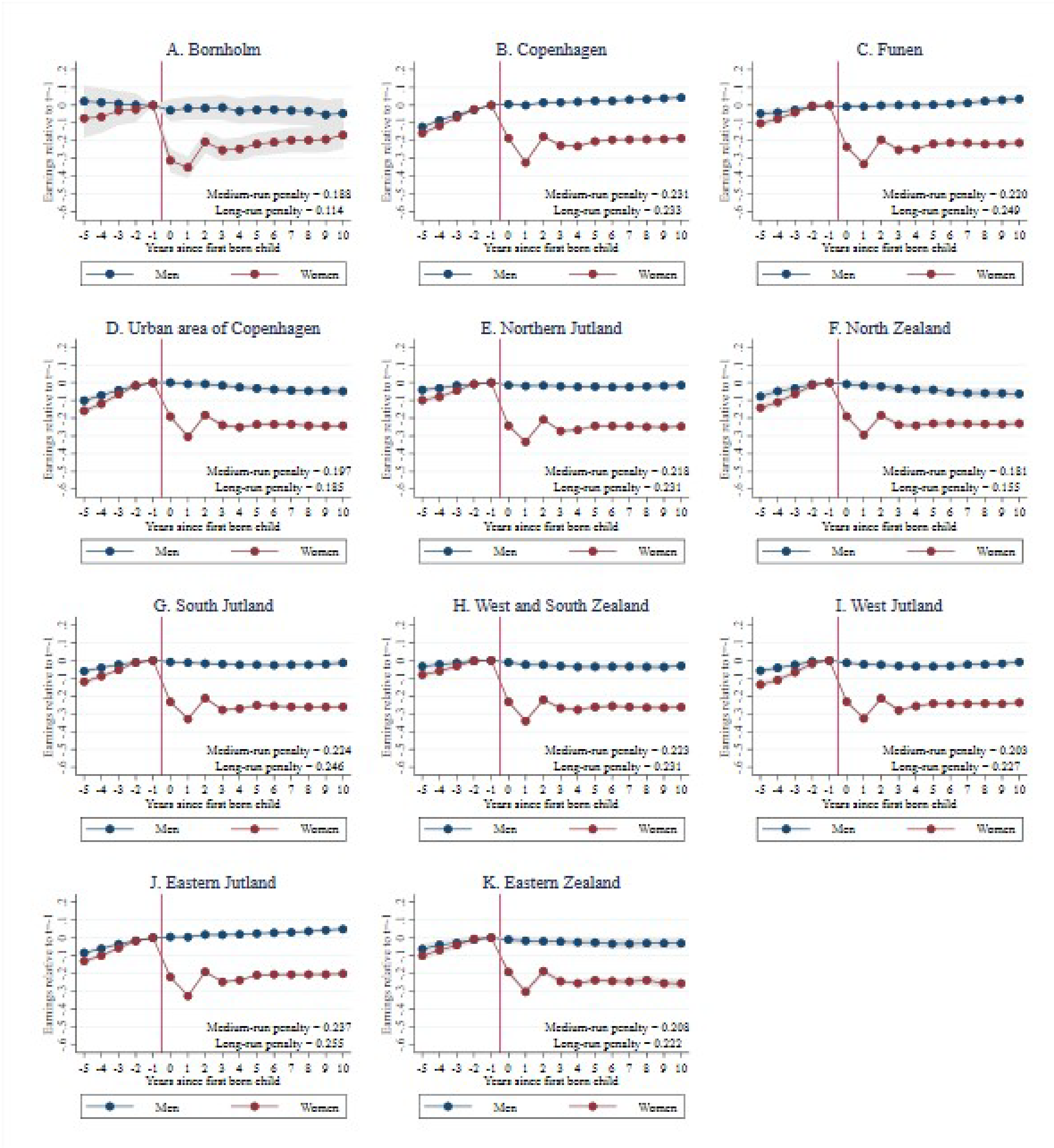
Notes: The figure shows the impact of children on earnings conditional on the size of the firm's net income at t = −1. We divide our sample into two subsamples based on net income. We define high net income as net income above the median and low net income as net income below the median. In addition to the sample restrictions in Table 2, we condition on available accounting information. The vertical line indicates the birth of the first child and the reported medium-run and long-run child penalties are measured in t = 5 and t = 10, respectively by Equation 2. Similar to Kleven et al. (2019b) we construct 99% confidence bands using robust estimates of the standard errors.

Figure A.2.2. Child penalties in earnings. By above/below median ROA



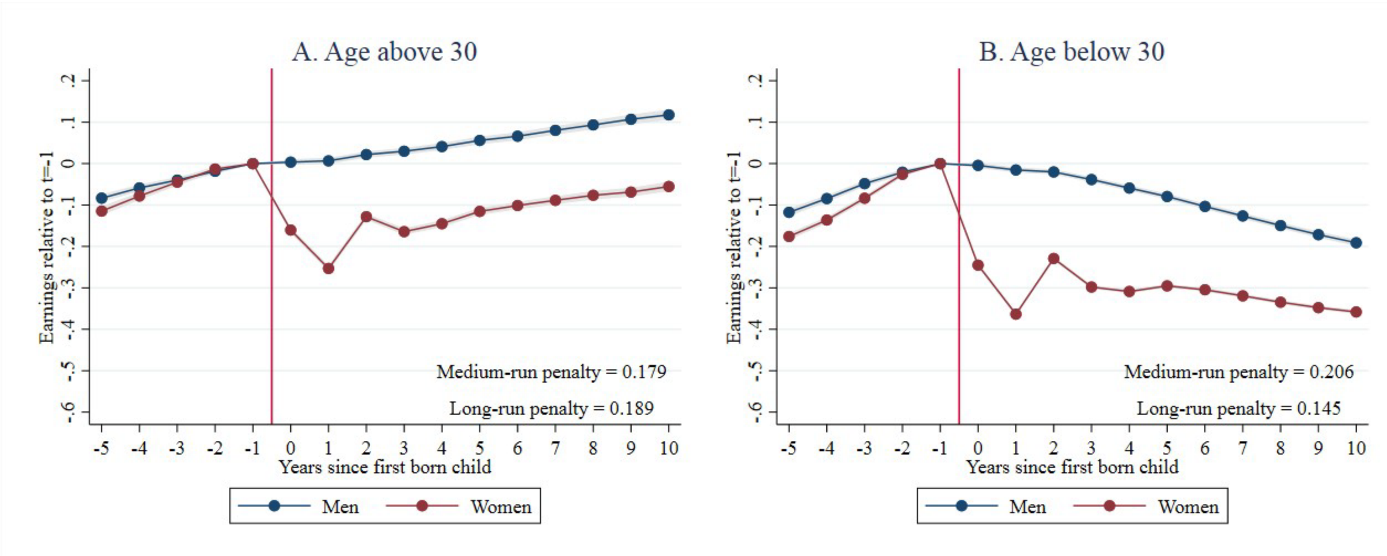
Notes: The figure shows the impact of children on earnings conditional on the firm being above/below median ROA at $t = -1$. We divide our sample into two subsamples based on ROA. We define high return on assets as ROA above the median and low return on assets as ROA below the median. In addition to the sample restrictions in Table 2, we condition on available accounting information. The vertical line indicates the birth of the first child and the reported medium-run and long-run child penalties are measured in $t = 5$ and $t = 10$, respectively by Equation 2. Similar to Kleven et al. (2019b) we construct 99% confidence bands using robust estimates of the standard errors.

Figure A.2.3: Child penalties in earnings. By area of residence



Notes: The figure shows the impact of children on earnings conditional on area of residence at t = 1. In addition to the sample restrictions in Table 1, we condition on available area of residence information. The vertical line indicates the birth of the first child. And the reported medium-run and long-run child penalties are measured in t = 5 and t = 10, respectively by Equation 2. Similar to Kleven et al. (2019b) we construct 99% confidence bands using robust estimates of the standard errors.

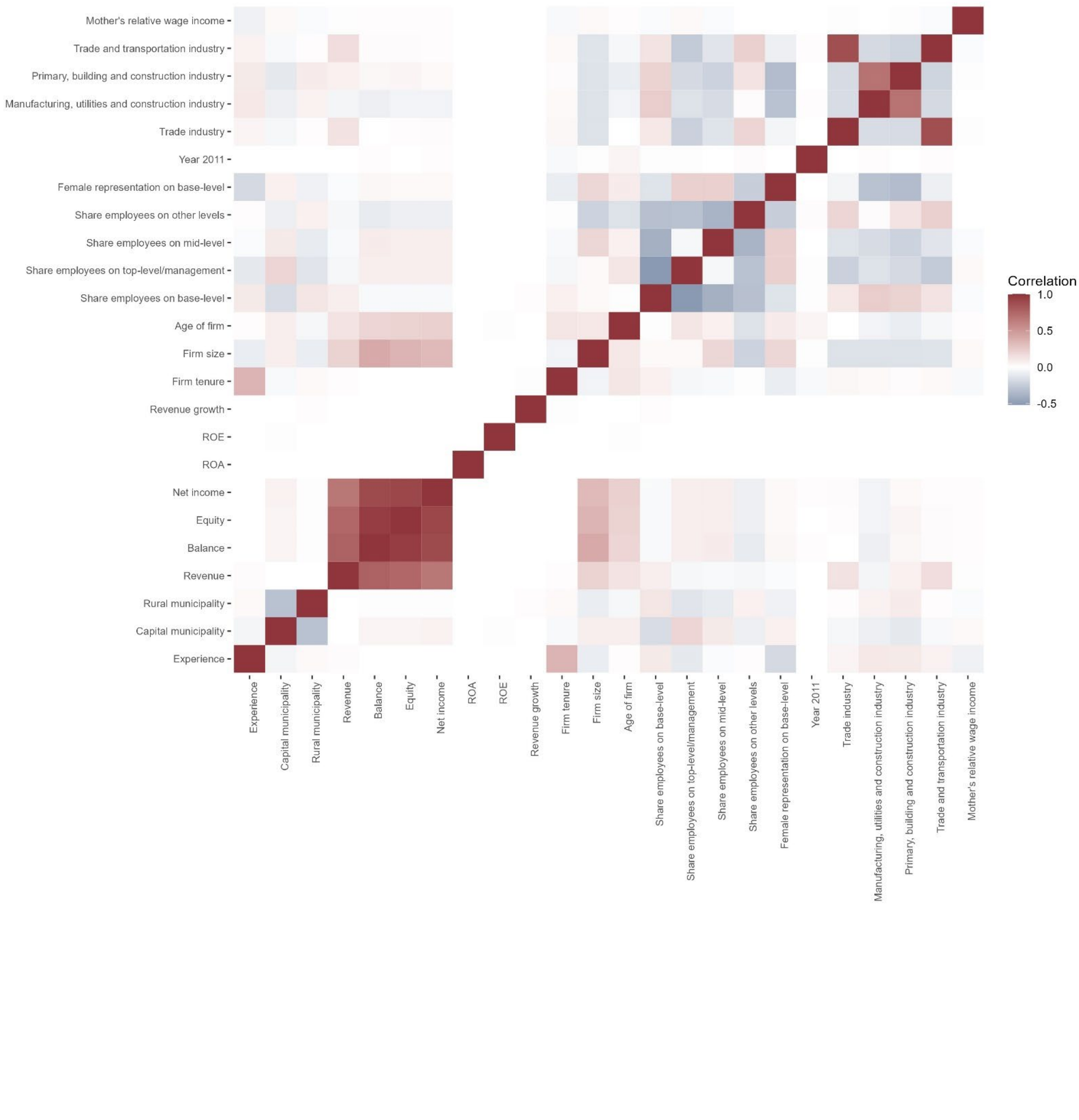
Figure A.2.4: Child penalties in earnings conditional on age



Notes: The figure shows the impact of children on earnings conditional on age at t = 0. i.e.. we divide our sample into two subsamples based on age of the parents. The sample is subject to the sample restrictions in Table 2. The vertical line indicates the birth of the first child. And the reported medium-run and long-run child penalties are measured in t = 5 and t = 10, respectively by Equation 2. Similar to Kleven et al. (2019b) we construct 99% confidence bands using robust estimates of the standard errors.

A3. Feature correlations

Figure A.3.1. Correlation between selected variables in the covariate space



A4. Heat Maps

Figure A.4.1: CATE distribution by firm characteristics & firm financial variables

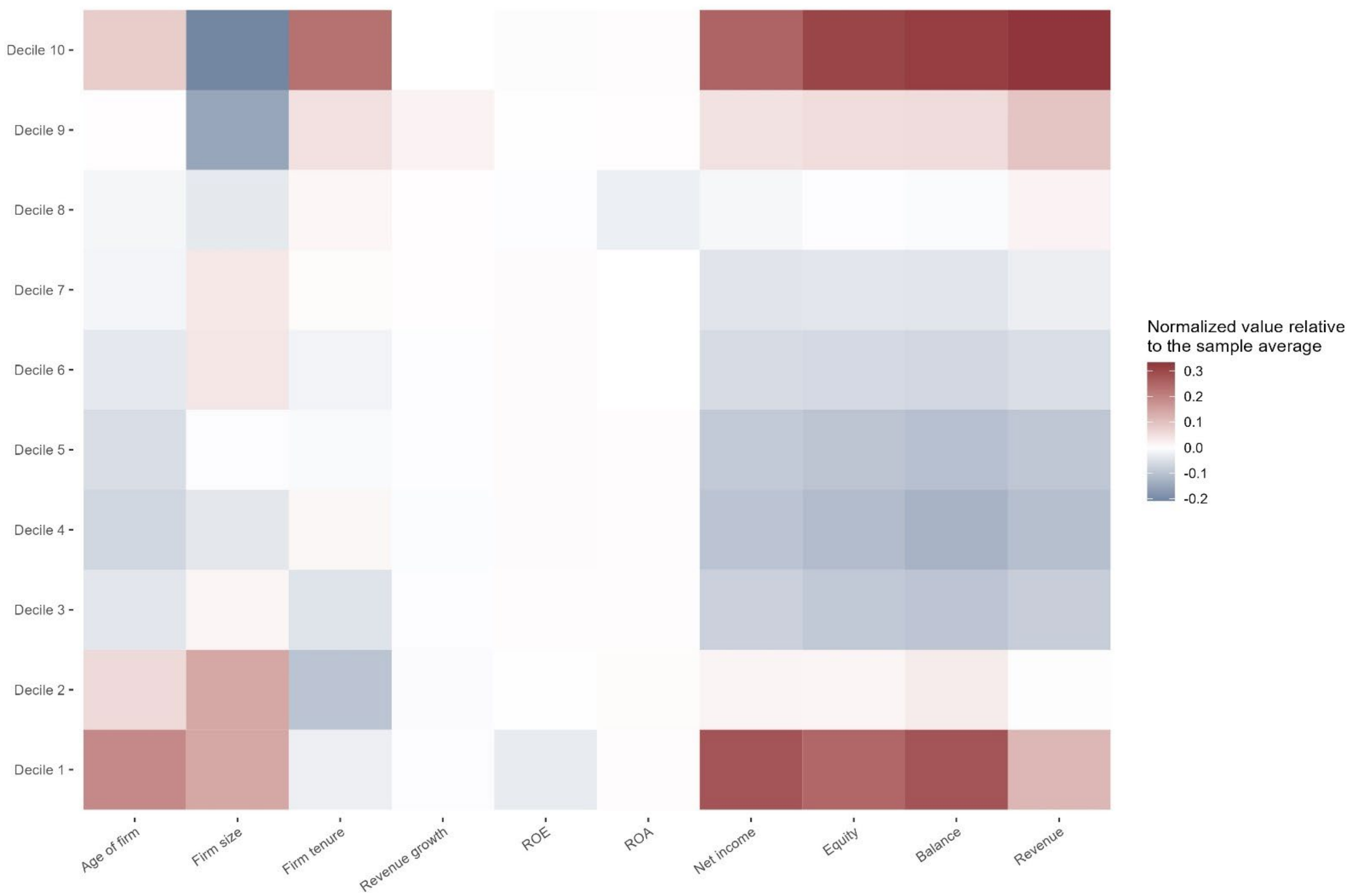


Figure A.4.2: CATE distribution by municipality and industry groupings

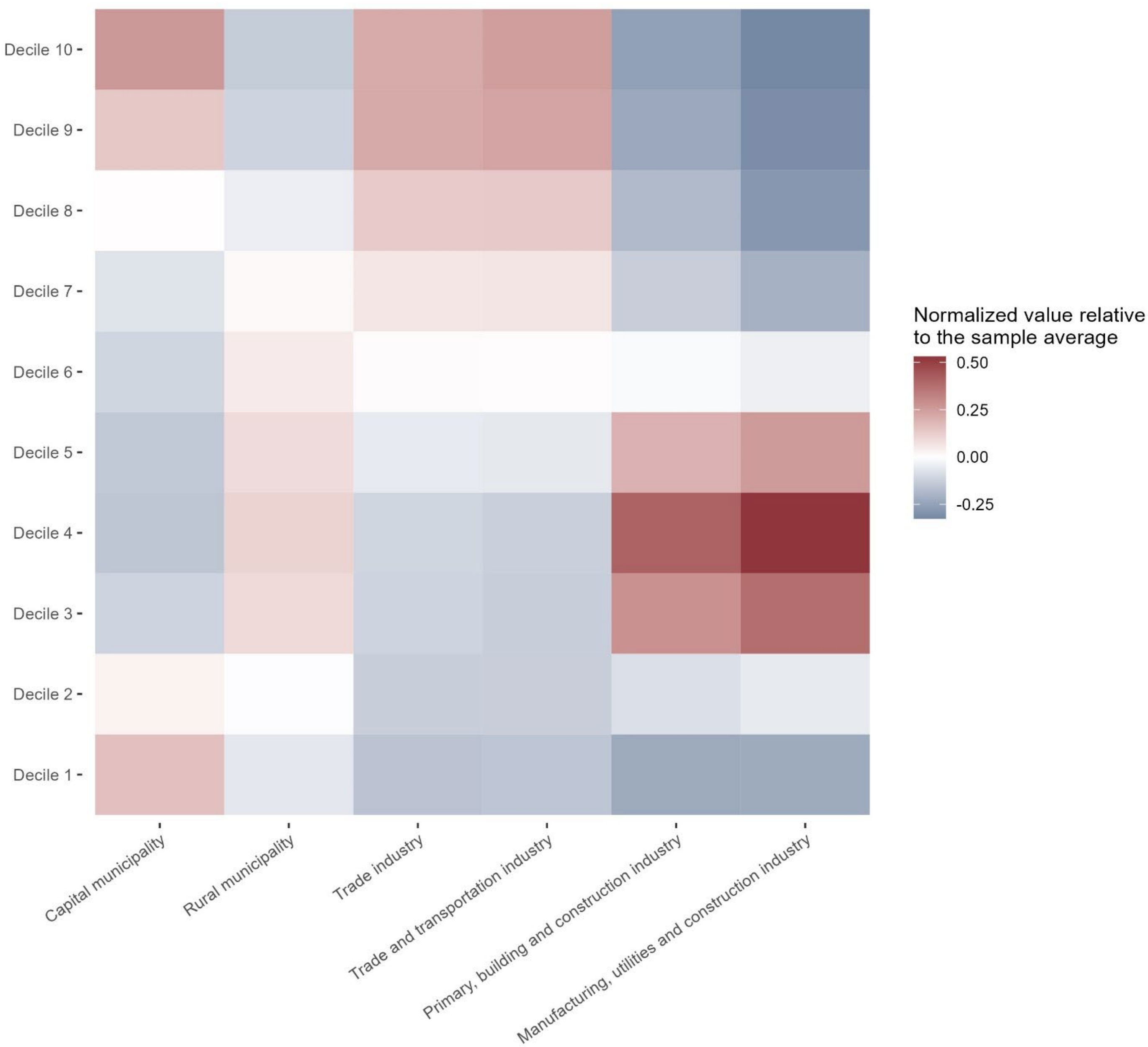
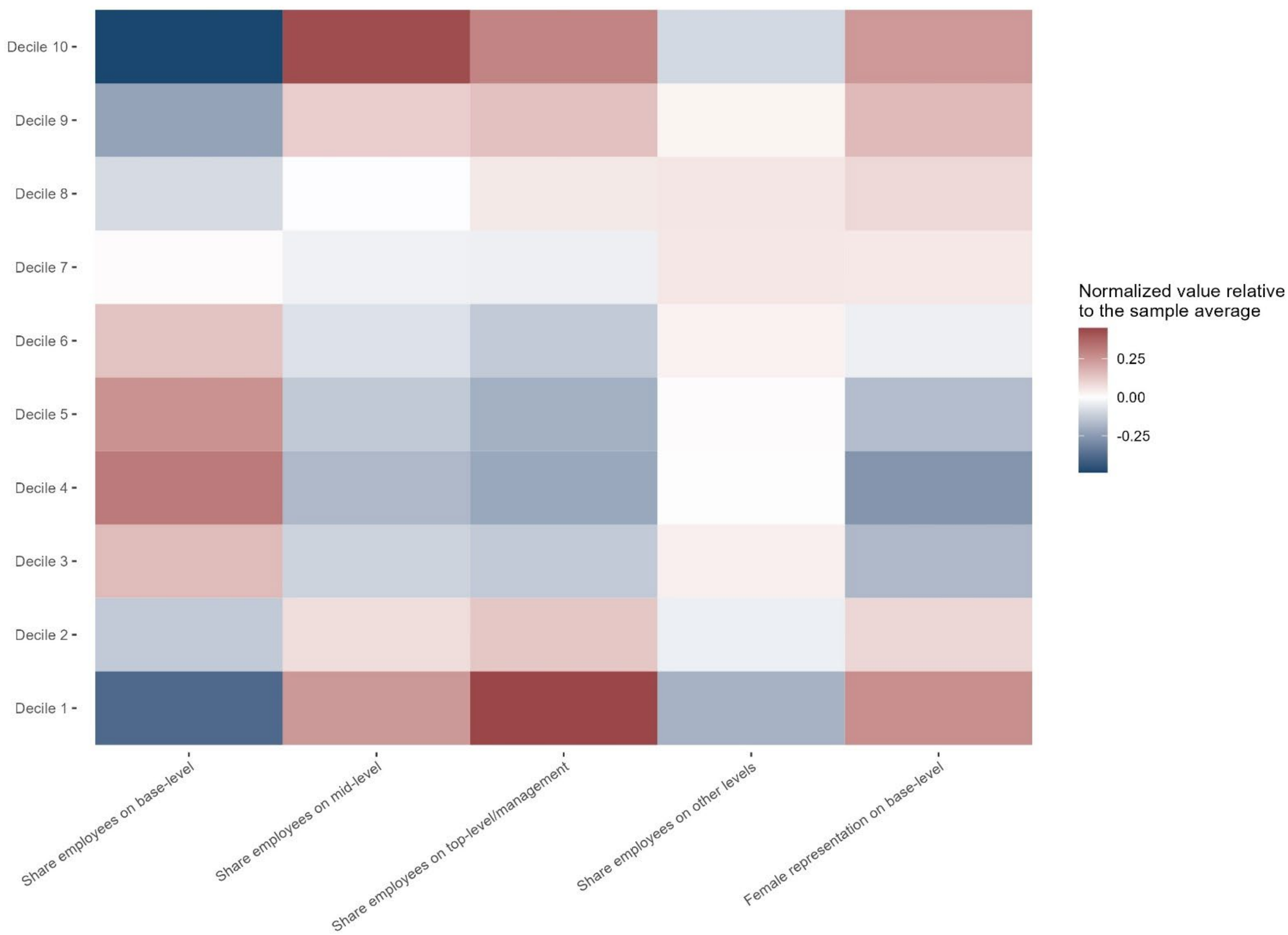
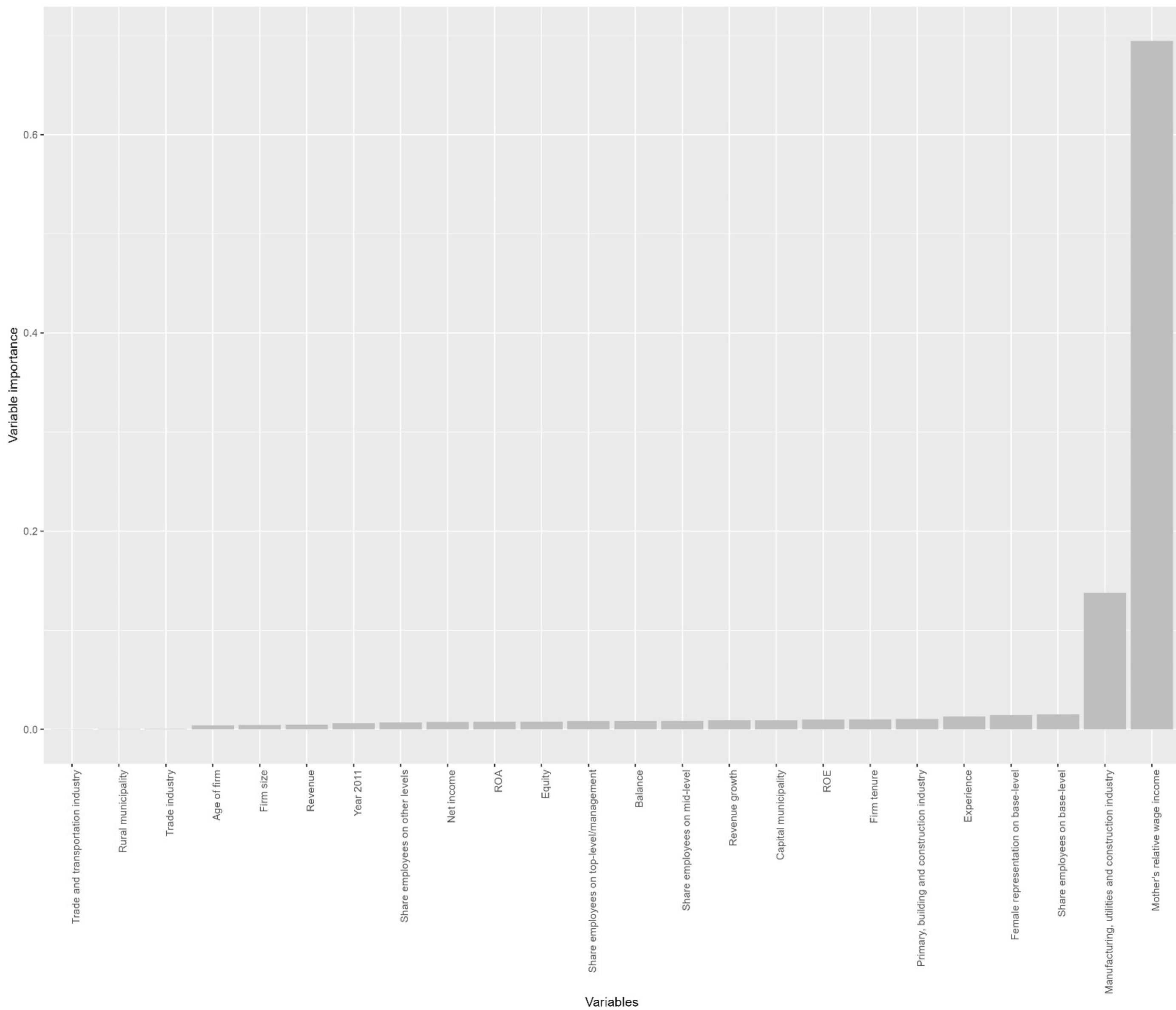


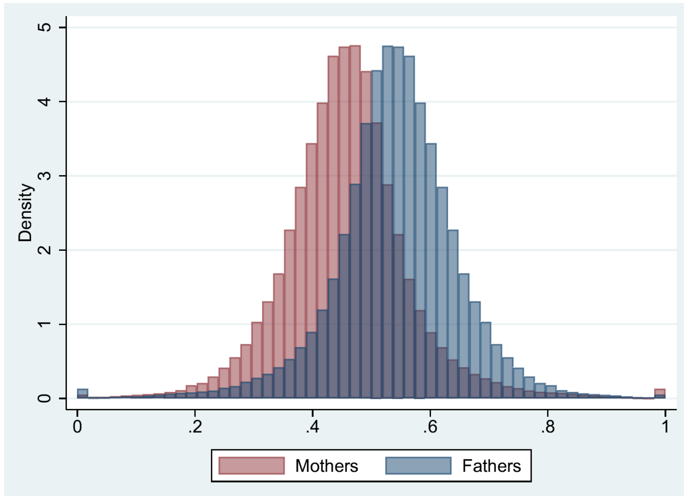
Figure A.4.3: CATE distribution by share of employees & female representation at various levels



A5. Variable importance – all features



A6. Distribution of relative wage income



A7. Parameter settings, pilot and causal forest

Pilot forest

Number of trees	2000
Sample weights	NONE
Clusters	NONE
Fraction of sample used to grow trees	0,5
Minimum node side	5
Number variables tried for each split	31
Honesty	YES
Honest fraction	0,5
Pruning of empty leaves	YES
Maximum imbalance of a split (alpha)	0,05
Imbalance penalty	0

Causal forest

Number of trees	2000
Sample weights	NONE
Clusters	NONE
Fraction of sample used to grow trees	0,5
Minimum node side	5
Number variables tried for each split	25
Honesty	YES
Honest fraction	0,5
Pruning of empty leaves	YES
Maximum imbalance of a split (alpha)	0,05
Imbalance penalty	0

See also: Causal forest — causal_forest • grf (grf-labs.github.io)