Quantifying the heterogeneous link between housework and labor market outcomes through Meta-learners

Master's Thesis

Presented to the
Department of Economics at the
Rheinische Friedrich-Wilhelms-Universität Bonn

In Partial Fulfillment of the Requirements for the Degree of
Master of Science (M.Sc.)

Supervisor: Prof. Dr. Hans-Martin von Gaudecker

Submitted in September 2024 by

Purti Sadhwani

Matriculation Number: 3500185

Contents

List of Figures

List of Tables

1 Introduction

Housework and its implications for labor market outcomes have been subjects of significant interest within the fields of economics, sociology, and gender studies. Defined as unpaid domestic labor, housework includes tasks such as cleaning, cooking, running errands, childcare, and other household management activities (Kan (2014), Shelton and John (1996)). In the past, research has mostly focused on the more restricted category of housework, consisting physical activities like cleaning, laundry, running for errands and cooking. Few studies have also incorporated some other components of household labor- child-care, emotional labor such as providing encouragement or advice, and mental labor such as planning or household management (Coltrane (2000)). These activities, although crucial for the functioning of households, are often undervalued in economic analyses (Bridgman (2016) and L. Stratton (2020)).

Understanding the link between housework and labor market outcomes is crucial for several reasons. Housework can significantly influence personal income by affecting the time and effort available for paid work, understanding the trade-off between housework and paid work and, housework and leisure can potentially help individuals make better choices regarding time and effort allocation. Becker (1981), in his well-known contribution *A Theory of the Allocation of Time* argued that "consumers maximize their utility by choosing commodities that are produced with market goods and time by a consumer facing both budget and time constraint". He described a model with a fixed amount of energy to be allocated amongst different activities. And therefore, housework activities being tiring, reduce the amount of effort available for paid work, potentially affecting the productivity and wages negatively. A similar model focusing on timing and flexibility of housework was presented by Bonke, N. Gupta, and Smith (2005). He notes that "Timing and flexibility of housework turn out to be more important than the level of housework, and women, particularly at the high end of the conditional wage distribution, who time their housework immediately before or after market work or engage in home tasks that require contiguous blocks of time are significantly penalized in terms of lower wages."

Other theories focus more explicitly on labour market structure and job characteristics. Employees with significant housework responsibilities may choose jobs that offer flexible hours or less demanding working conditions, which might be associated with lower wages due to a negative compensating differential (Hersch and Stratton (1997a), and Hersch (2009)), or pay less because monopsonistic employers take account of workers' preferences for these jobs (Sigle-Rushton and Waldfogel (2007)). The relationship between housework and personal income has been a critical area of inquiry within the field of Labor Economics and research indicates that the

time and effort devoted to housework can have substantial effects on personal income, primarily by reducing the time available for paid work and, consequently, limiting opportunities for career advancement and income growth.

Moreover, the effects of housework on personal growth, such as skill development and career advancement, are also significant. Time spent on housework can reduce the time available for further education, professional development, and networking opportunities, which are critical for career progression (Bertrand, Kamenica, and Pan (2013)). These channels illustrate how housework can lead to wage disparities, not just between different genders, but across individuals with varying household responsibilities, influencing their overall labor market outcomes. Thus, examining the impact of housework on these dimensions can provide valuable insights into the broader implications of unpaid domestic labor on long-term economic outcomes.

The literature on the relationship between housework and income suggests a primary negative relationship with some mixed effects within a few subgroups. A study by Fendel (2021) on how this relationship varies between immigrants and non-immigrants in Germany found a negative effect on income across all the subgroups studied, it further showed significant negative effects of housework on wages for both migrant and native-born women compared to that of men. Bryan and Sevilla-Sanz (2010) found the wages of full-time employees decrease by about 0.25% per hour of weekly housework, implying an extra ten hours of housework per week would lower wages by 2.5%. They further found the effect to be constant within occupations for full-time workers. The negative housework-income relationship persists irrespective of having children in a household although they found some evidence of housework having a larger effect on the wages of married mothers.

These findings suggest that the amount of housework may play a more significant role in influencing wages than the specific type or timing of housework. As (Bryan and Sevilla-Sanz, 2010, p. 206) noted,"The similarity of the housework penalty across sub-groups which are characterized by different types and timing of housework could be an indication that the amount of housework matters more than the type or timing."

Another study by Keith and Malone (2005) found that housework significantly reduces wages for young and middle-aged married women, with each additional hour of housework lowering wages by 0.1–0.4%. This effect aligns with the demands of childcare during these life stages. However, no consistent wage impact is observed for older women or men. Housework also contributes to 3–10% of the explained gender wage gap, particularly affecting younger workers. And numerous studies have established that housework significantly impacts wages,

with notable gender-specific effects. Research consistently shows that women, and particularly married women and those with children, face substantial wage penalties due to housework, as compared to men (Coverman (1983), Firestone and Shelton (1988), Hersch (1991), Hersch and Stratton (1994, 1997c), Hundley (2000, 2001), Noonan (2001), L. S. Stratton (2001), Hersch and Stratton (2002), Shirley and Wallace (2004), Keith and Malone (2005), and Hersch (2009)). While some studies, such as those by Hersch and Stratton (1997b), have found no support for the hypothesis that marriage increases men's wages due to household specialization, others suggest that the wage effects of housework differ by marital status, with 14% of the gender pay wage gap explained by including housework in wage equations Hersch and Stratton (1994). This negative relationship between housework and wages is also evident in studies conducted in Australia McAllister (1990), Canada Phipps, Burton, and Lethbridge (2001), and Denmark Bonke, N. Gupta, and Smith (2005).

Many studies in this area have focused primarily on women, with less attention paid to how housework links with other subgroups' labor market outcomes. Additionally, much of the existing literature assumes that the relationship between housework and labor market outcomes is uniform across all individuals, without considering potential heterogeneity in this link (Gupta (2006)). This thesis aims to fill this gap by exploring how the effects of housework vary across different demographic groups, including by gender, age, education level, and household size.

The shift in family dynamics has spurred extensive research focused on understanding not only the factors that influence the division of household labor but also the broader implications for families and individuals. Scholars have particularly focused on the relationship between the time spouses dedicate to housework and the earnings from paid employment (Becker (1981), Budig and England (2001), Jacobs and Gerson (2004), England (2005), and Hochschild and Machung (2012)).

The relative resources hypothesis suggests that within couples, the spouse with greater resources, such as income or education, can negotiate out of housework responsibilities. This dynamic is particularly evident in couples with relatively equal earnings, leading to a more equitable division of routine housework, where husbands tend to increase their participation, and wives do less. However, when wives out-earn their husbands, traditional gender norms often resurface, according to the gender display and gender deviance neutralization perspectives. In these scenarios, wives may increase their housework to conform to societal expectations of femininity, while husbands reduce their involvement in domestic tasks to assert traditional masculinity. Autonomy theory further complements this discussion by suggesting that higher

absolute earnings, particularly for women, provide the financial means to outsource housework, thereby reducing their direct involvement (Carlson and Lynch (2017)).

This dynamic involving bargaining among spouses and also potentially in a larger household among more members can skew household decision-making processes. It is interesting to study if housework has a negative income effect, in case of a hiring decision for the work; if the additional income from no housework could be used for hiring. Furthermore, by quantifying the economic value of housework, households could make more equitable decisions about the division of labor, ultimately leading to better outcomes for all members (Bittman et al. (2003)).

Labor supply decisions stem from a structural model of work-leisure trade-off and that has been well-documented in labor economics, however, the role of housework in this equation is less understood (Gimenez-Nadal and Sevilla (2012)) therefore a study of how housework fits in the equation could potentially be an important factor determining optimal labor supply. Moreover, these decisions are not constant over time, life-cycle effects as mentioned by Keith and Malone (2005) for example childbearing, family changes, layoffs, job switches, changing priorities and time investments could be worthwhile exploring when housework can have long term impacts on labor market outcomes.

Understanding these trade-offs is crucial for making optimal work choices at different stages of life, depending on personal goals. For example, the impact of housework on wages may vary by education level, age, health, marital status, and family size. Research has found that the effects of housework on wages are more pronounced for less-educated individuals and those with larger families, suggesting that these factors should be considered when making decisions about education, family planning, and career development (S. M. Bianchi et al. (2000), Treas and Drobnič (2010), and S. Bianchi et al. (2012))

The insights gained from studying the link between housework and labor market outcomes could also help individuals who have been out of the labor force for extended periods, such as full-time homemakers, to re-enter the workforce. By attaching a monetary value to housework, these individuals could better demonstrate their skills and contributions, making it easier for them to find employment (Folbre (2006)).

Lastly, social norms and gender stereotypes also play a significant role in shaping the division of housework within households. Women, in particular, are often expected to take on a larger share of housework, which can contribute to wage differentials between men and women (Hook (2006)). Measuring housework could potentially help entangle the role of social norms pointing towards a little more of gender equity in the workforce.

The following section of the paper will discuss the methodological framework that will be used later for the empirical analysis. I would also elaborate upon the theoretical assumptions for the causal analysis. We then proceed with Section 3 to understand the institutional setup, data and the variables used for the analysis. Section 4 consists of the simulation study setup done in order to find the suitable methodology in the context of our empirical study. Section 5.2 proceeds on to the real-data application where I describe the assumptions in the context specific to the analysis, followed by the analysis and results. Towards the end, we follow through limitations (Section 6) of the study and lastly Conclusions in Section 7.

2 Methodological Framework

2.1 Overview

Following the research focus from the previous section which is to understand the effect of housework on labor market outcomes across different individuals and consequently subgroups, I employ Conditional Average Treatment Effect (CATE) estimation as the primary methodology. CATE is particularly well-suited for this analysis because it allows for the investigation of treatment effects across different subgroups within the population, rather than assuming a homogeneous treatment effect across all individuals. This is crucial in understanding how different factors, such as gender, age, or household structure, may interact with the treatment effect of housework on labor market outcomes.

CATE estimates the average treatment effect conditional on certain covariates, which means it can capture how the treatment effect varies across different subgroups defined by these covariates. This is particularly important when there is reason to believe that the treatment effect is not uniform across the population. For instance, smaller subgroups may exhibit different treatment effects, which would be obscured if only the overall Average Treatment Effect (ATE) were considered (Wager and Athey (2018)). By focusing on CATE, we can then look at the heterogeneity and try to link characteristics (covariates) that are drivers for different treatment effects. Additionally, we can also explore potential interactions between the treatment and various covariates, which could lead to more nuanced insights into the relationship between housework and labor market outcomes (Jacob (2021)).

A common method to estimate CATE involves calculating two conditional mean functions—one for the treated observations and another for the control group. For each observation,

the predicted outcomes under both treatment and control are determined by applying these functions. The CATE is then derived by taking the difference between these two predicted outcomes.

2.2 Fundamental Assumptions

I will start by introducing the potential outcome framework: Each observation has two potential outcomes, Y_1 and Y_0, of which we only observe one. Specifically, we observe Y_1 if the individual was treated and Y_0 if the individual was not treated. This is denoted by the binary treatment indicator $D \in \{0, 1\}$. The observed covariates are denoted as $X \in \mathbb{R}^p$. To interpret the estimated parameter as indicative of a causal relationship, the following assumptions must be met; see, for instance, Rubin (1980).

1. Conditional independence (Exogeneity or conditional unconfoundedness):

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i \mid X_i.$$

2. Stable Unit Treatment Value Assumption (SUTVA):

$$Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i}).$$

3. Common support Assumption (Overlap condition):

$$\forall x \in \text{supp}(X_i), \quad 0 < P(D_i = 1 \mid X_i = x) < 1,$$

$$P(D_i = 1 \mid X_i = x) \equiv e(x).$$

4. Exogeneity covariates:

$$X_i \perp D_i.$$

Assumptions 1 and 4 are very similar and correlated; they state that the treatment assignment is independent of the two potential outcomes conditional on the covariates and that the covariates are not affected by the treatment. Assumption 2 ensures that there is no interference, no spillover effects, and no unobserved differences between observations in both the groups. Assumption 3 asserts that no subpopulation defined by $X_i = x$ is exclusively in the treatment or control group; thus, the probability of receiving treatment must remain within bounds, neither zero nor one (Jacob (2021)).

The propensity score is defined as the probability of receiving the treatment given the covariates, denoted as:

$$e(x) = P(D_i = 1 \mid X_i = x).$$

This score represents the likelihood that an individual with covariates $X_i = x$ is assigned to the treatment group.

The conditional expectation of the outcome for the treatment or control group is defined as:

$$\mu_d(x) = \mathbb{E}[Y_i \mid X_i = x, D_i = d] \quad \text{with } D \in \{0, 1\}.$$

Our parameter of interest is the Conditional Average Treatment Effect (CATE), denoted by $\tau(x)$, which is formally defined as:

$$\tau(x) = \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = x] = \mu_1(x) - \mu_0(x).$$

The following shows how the two conditional mean functions can represent the two potential outcomes and hence, by taking the difference, lead to the CATE:

$$\begin{aligned}
\tau(x) &= \mu_1(x) - \mu_0(x) \\
&= \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x] \\
&= \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x] \\
&= \mathbb{E}[Y_{1i} \mid X_i = x] - \mathbb{E}[Y_{0i} \mid X_i = x] \\
&= \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = x].
\end{aligned}$$

And here is what Average treatment effect will be ATE $:= \mathbb{E}[Y(1) - Y(0)]$. This clearly shows the distinction between ATE and CATE. And from the latter we can estimate the treatment effects for various subgroups as is the aim of the paper.

In order to estimate CATE effectively, I use a meta-learning technique. Meta-learners are a class of algorithms designed to estimate CATE by leveraging existing machine learning models. These techniques are particularly useful when dealing with high-dimensional data or when the relationship between the covariates and the treatment effect is complex and potentially nonlinear. Traditional linear models, while useful, often require significant feature engineering and may struggle with overfitting, especially when dealing with numerous covariates or complex interactions between them (Jacob (2021)).

"Meta-algorithms decompose estimating the CATE into several subregression problems that can be solved with any regression or supervised learning method. The most common metaalgorithm for estimating heterogeneous treatment effects takes two steps. First, it uses so-called base learners to estimate the conditional expectations of the outcomes separately for

units under control and those under treatment. Second, it takes the difference between these estimates" as explained by Künzel et al. (2019).

This approach has been analyzed when the base learners are linear-regression (Foster (2013)), Random Forests (RF) (Breiman (2001)) , Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch (2010)), XGBoost (Chen and Guestrin (2016)), Generalized Additive Model (GAM) (Hastie and Tibshirani (1986)), Neural Network (NN) (Hopfield (1982)), Model-Based recursive partitioning (MOB) (Achim Zeileis and Hornik (2008) and Seibold, Zeileis, and Hothorn (2016)), and Super Learner (SL) (Laan, Polley, and Hubbard (2007)).

The choice of meta-learner depends on the data structure and the assumptions one is willing to make. These methods allow us to flexibly model interactions between the treatment and covariates, which is essential for capturing the heterogeneity of treatment effects across different subgroups in the data.

2.3 Understanding the Meta-learners

In the context of estimating heterogeneous treatment effects, I have used the following three meta-learners to find the best fit for my analysis:

2.3.1 S-Learner

The S-learner approach estimates a single joint function for both the treatment and control groups. Specifically, the conditional outcome function $\mu(x, w) := \mathbb{E}[Y^{obs} \mid X = x, D = d]$, where $D \in \{0, 1\}$, is estimated. The conditional average treatment effect (CATE) between treatment 1 and control 0, can then be computed as:

$$\hat{\tau}_S(x) \;=\; \hat{\mu}(x, 1) - \hat{\mu}(x, 0).$$

The S-learner uses the entire dataset to estimate the function $\mu(\cdot)$, considering the treatment D as one of the covariates. However, this can lead to issues, particularly in high-dimensional settings, where the treatment variable may be underweighted or neglected. In my simulations, this algorithm performs poorly, and I have not used it for the real data application, but it may do well in other settings. The results of the simulation study using S-learner can be found in the Appendix.

2.3.2 T-Learner

In contrast, the T-learner approach estimates separate conditional outcome functions for each group. It operates by first estimating the:

Control response function $\mu_0(x) = \mathbb{E}[Y(0) \mid X = x]$, using the observations in the control group,

Treatment response function $\mu_1(x) = \mathbb{E}[Y(1) \mid X = x]$, using the observations in the treatment group.

Both using different models or supervised learning algorithms. The treatment effect $\hat{\tau}(x)$ is then the difference between these two estimates:

$$\hat{\tau}(x) \;=\; \hat{\mu}_1(x) - \hat{\mu}_0(x).$$

While the T-learner is straightforward and leverages the separate estimation of treatment effects, it can lose efficiency because each model is estimated independently, potentially leading to a loss of information shared across treatment groups. This is also used for the Simulation study in the Appendix but not for the real data analysis because the following X-learner had the lowest Expected mean squared error.

2.3.3 X-Learner

The X-learner, as proposed by Künzel et al. (2019), addresses the data efficiency issues inherent in the T-learner by focusing directly on the treatment effects rather than the response surfaces. This estimates a treatment effect separately for the control and the treatment group and therefore it might be especially helpful in situations where the proportion of the two groups is highly imbalanced. The procedure for the X-learner in a binary treatment setting can be described in the following steps:

(1) **Estimate Response Functions:** Estimate the conditional response functions for both control and treatment groups using any supervised learning or regression algorithm. Specifically, we estimate:

$$\mu_0(x) \;=\; \mathbb{E}[Y(0) \;\mid\; X \;=\; x],$$

$$\mu_1(x) \;=\; \mathbb{E}[Y(1) \;\mid\; X \;=\; x],$$

using any supervised learning or regression algorithm and denote the estimators as $\hat{\mu}_0(\cdot)$ and $\hat{\mu}_1(\cdot)$, respectively. The algorithms used here are what we refer to as base learners of first stage.

(2) **Impute Treatment Effects:** Impute the treatment effects for individuals in the treated group using the control outcome estimator, and similarly impute the treatment effects for individuals in the control group using the treatment outcome estimator. For each individual i, these imputed treatment effects are defined as:

$$\tilde{D}_i^1 \;:=\; Y_i^1 - \hat{\mu}_0(X_i^1),$$

$$\tilde{D}_i^0 \;:=\; \hat{\mu}_1(X_i^0) - Y_i^0.$$

here the estimator for the controls is subtracted from the observed treated outcomes, and, similarly, the observed control outcomes are subtracted from estimated treatment outcomes to obtain the imputed treatment effects. Y_i^0 and Y_i^1 are the ith observed outcome of the control and the treated group, respectively. X_i^0 , X_i^1 are the corresponding covariate vectors.

(3) **Fit Models on Imputed Effects:** Using the imputed treatment effects $\tilde{D}_i^1$ and $\tilde{D}_i^0$ as response variables, fit models to predict $\hat{\tau}_1(x)$ for the treatment group and $\hat{\tau}_0(x)$ for the control group:

$$\hat{\tau}_1(x) \;=\; \mathbb{E}[\tilde{D}^1 \;\mid\; X^1 \;=\; x], \quad \hat{\tau}_0(x) \;=\; \mathbb{E}[\tilde{D}^0 \;\mid\; X^0 \;=\; x].$$

The supervised learning algorithms used here are referred as base learners of the second stage.

(4) **Combine Treatment Effect Models:** Finally, combine the two treatment effect estimators $\hat{\tau}_1(x)$ and $\hat{\tau}_0(x)$ to estimate the Conditional Average Treatment Effect (CATE) as:

$$\hat{\tau}(x) \;=\; g(x)\hat{\tau}_0(x) + (1 - g(x))\,\hat{\tau}_1(x),$$

where $g(x)$ is a weighting function, often chosen as the propensity score.

The X-learner can use information from the control group to derive better estimators for the treatment group and vice-versa

As noted by Künzel et al. (2019), the X-learner allows structural information about the CATE to make efficient use of an unbalanced design. However, a limitation of the X-learner is that it is based on a squared loss function, making it suitable primarily for continuous outcomes and not for other types of outcomes.

Finally, because of an unbalanced division between treatment and control groups and a potentially complex relationship between the covariates and treatment effects, X-learner provided me with the least error and therefore was selected as the method to be used for the analysis.

3 Data

3.1 Institutional Setup

This research draws on data from the German Socio-Economic Panel (SOEP), a longitudinal survey conducted annually that covers households and individuals living in Germany (Wagner, Frick, and Schupp (2007)). The SOEP, established in 1984, provides comprehensive data across various life domains such as work, health, time use, and education, with an average annual sample size of approximately 35,000 individuals. For this analysis, data from the years 2011 to 2020 are considered, with 2020 being the latest year with available information. The eight-year period was chosen to observe long-term trends, making 2012 the treatment year to evaluate the link over time. The study aims to measure the link between housework and labor market outcomes from 2013-2020 with a specific focus on short medium and long term results in 2013, 2016 and 2020 respectively.

Average number of housework hours per day are derived from the question: "What is a typical day like for you? How many hours do you spend on the following activities on a typical weekday, Saturday, and Sunday?" specifically focusing on three aspects, *Errands* i.e. shopping, trips to government agencies, etc., *Chores* i.e. housework such as washing, cooking, cleaning, and *Repair* i.e. performing building, flat/apartment/house repairs, car repairs, gardening work. Approximately 14.47% of those in the sample who report a positive number of housework hours indicate that they spend 1 hour per day on housework. 24.64% of the respondents report spending 2 hours per day on housework, while the remaining 60.89% report spending 3 or more hours per day on housework. A binary variable is created from this data, a method chosen based on several considerations outlined in Section 5.2.

Firstly, I filtered individuals during the pre-treatment period to identify those who engaged in less than 2.5 hours of housework per day. Based on this, the treatment group is defined as individuals who spend more than 2.5 hours per day on housework during the treatment period. The selection of 2.5 hours as the threshold is not arbitrary; it is grounded in an observed breakpoint in the data. Specifically, the data indicates that housework hours typically increase up to this point, after which the frequency of individuals spending more than 2.5 hours for

housework begins to decline, as illustrated in Figure 1. This threshold effectively captures the transition from moderate to high levels of housework involvement.

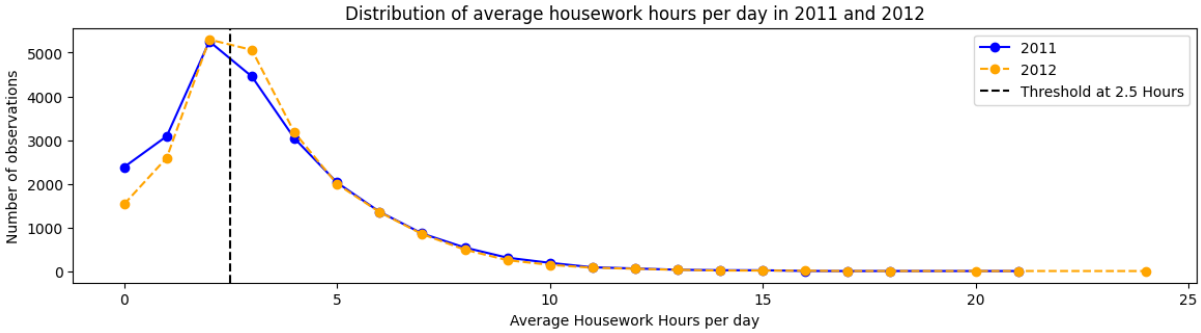


Figure 1. Distribution of average housework hours per day in 2011 and 2012

Choosing a starting point of zero housework hours would have resulted in an insignificant control group since housework is a near-universal activity, with most individuals contributing at least some time regularly. By establishing 2.5 hours as the critical threshold, this study aims to distinguish between individuals who engage in moderate versus extensive housework, thereby allowing for a more precise analysis of the effect of housework on labor market outcomes. This approach ensures that the treatment and control groups are meaningfully differentiated, with an aim to make the analysis robust and findings more insightful.

The analysis focuses on a cohort of individuals aged 18 to 66, representing the prime working-age population. Observations are drawn from multiple time points, spanning from t=-1 to t=8. Here t=-1 refers to the pre-treatment year 2011 and t=0 denotes the treatment year 2012. In 2011, individuals performing less than 2.5 hours of housework per day were selected and assigned the treatment defined as working more than 2.5 hours in 2012. Thus the treatment group consists of those who increased their housework to more than 2.5 hours per day in 2012, while those who maintained less than 2.5 hours per day constitute the control group. And the labor market outcomes for these individuals is measured from 2013-2020 (t=1 to t=8)

The effects of housework on labor market outcomes are measured over three distinct periods. In the short-term (t=1), effects on income are measured, as other variables were not collected in this year. In the medium-term (t=4), effects are measured on income, job prospects, and feelings of personal advancement, with all variables being self-perceived and rank-based. The timeline extends to t=8, enabling a comprehensive analysis of the long-term impacts of housework on labor market outcomes. This structured approach enables a detailed examination of how housework affects labor market outcomes over different time horizons, providing

valuable insights into the interplay between domestic responsibilities and professional life. Figure 2 illustrates the timeline of the analysis. Finally, we observe with 10,282 individuals in the pre-treatment period who are assigned housework as treatment in 2012.

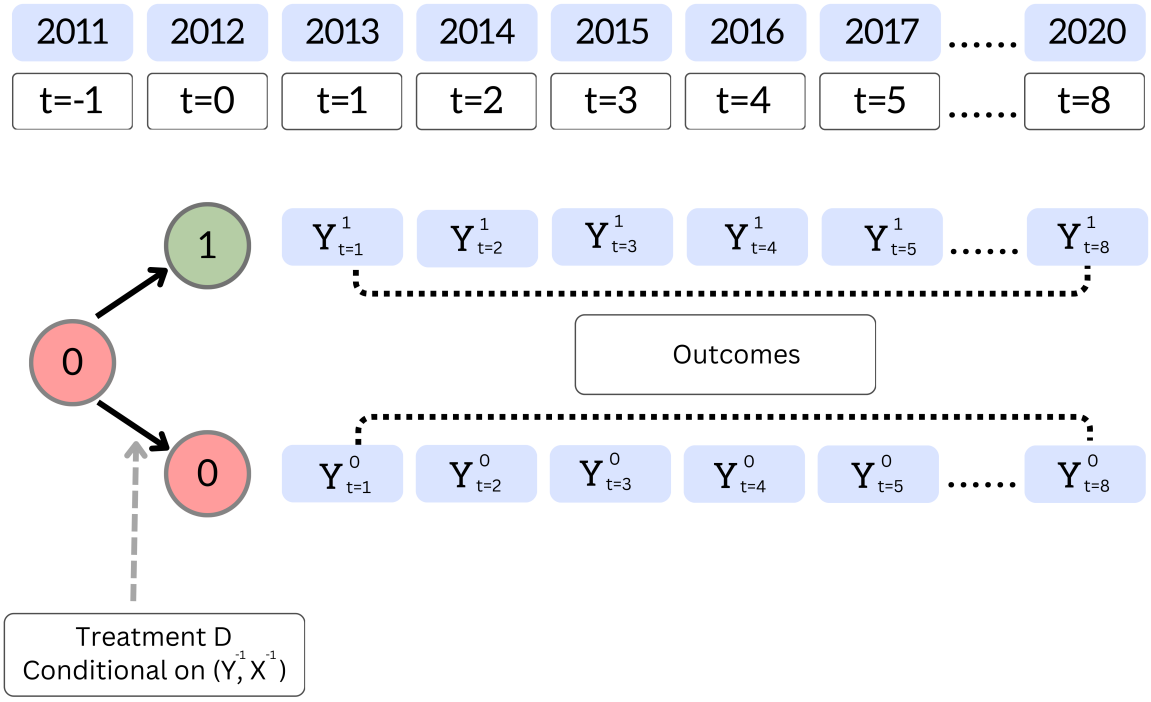


Figure 2. Timeline for the analysis

Figure 1 shows the distribution of average household work per day on the entire population for baseline and treatment years, 2011 and 2012, respectively. I also extended this analysis for different subgroups and found some interesting patterns but the overall picture strongly indicates that the threshold of 2.5 hours is pervasive across the different population subgroups in both 2011 and 2012. These graphs are presented and explained in the Appendix

Figure 3 shows the patterns in average housework hours per day across different age cohorts from 2011 to 2021, along with the corresponding standard errors. We can see a clear difference in the average hours worked across different age groups, although since the age groups 18-29 and 60-66 have higher standard errors, the estimates should be interpreted with caution.

Figure 4 shows the patterns of housework hours per day based on years of education. This is an interesting graph as it shows clear differences in the amount of time spent on housework for individuals with less, medium and high years of education. And this strengthens the reason to estimate conditional average treatment effects for the analysis.

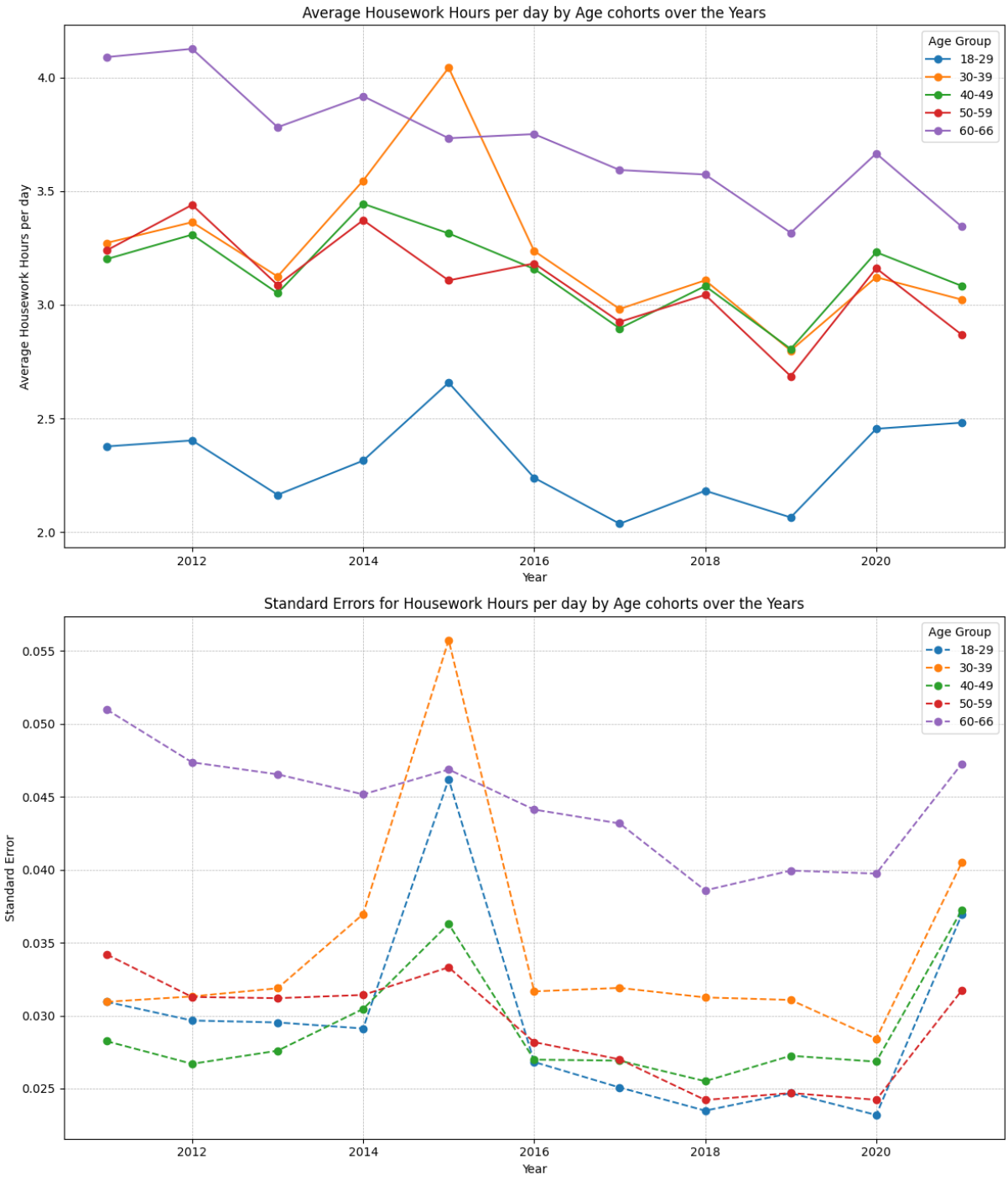


Figure 3. Average Housework Hours and Corresponding Standard Errors by Age Cohort

Figure 5 shows the patterns of housework hours per day based on gender, it is clear that housework hours of women stay consistently greater than that of men over all the years with little differences in the standard errors.

Finally, Figure 6 shows the trajectory of average housework hours based on household size and this graph too shows differences in household sizes ranging from 1-3 members and 8-12 members. But at the same time, the standard errors of the latter are quite high and therefore the patterns can be indicative but not definitive.

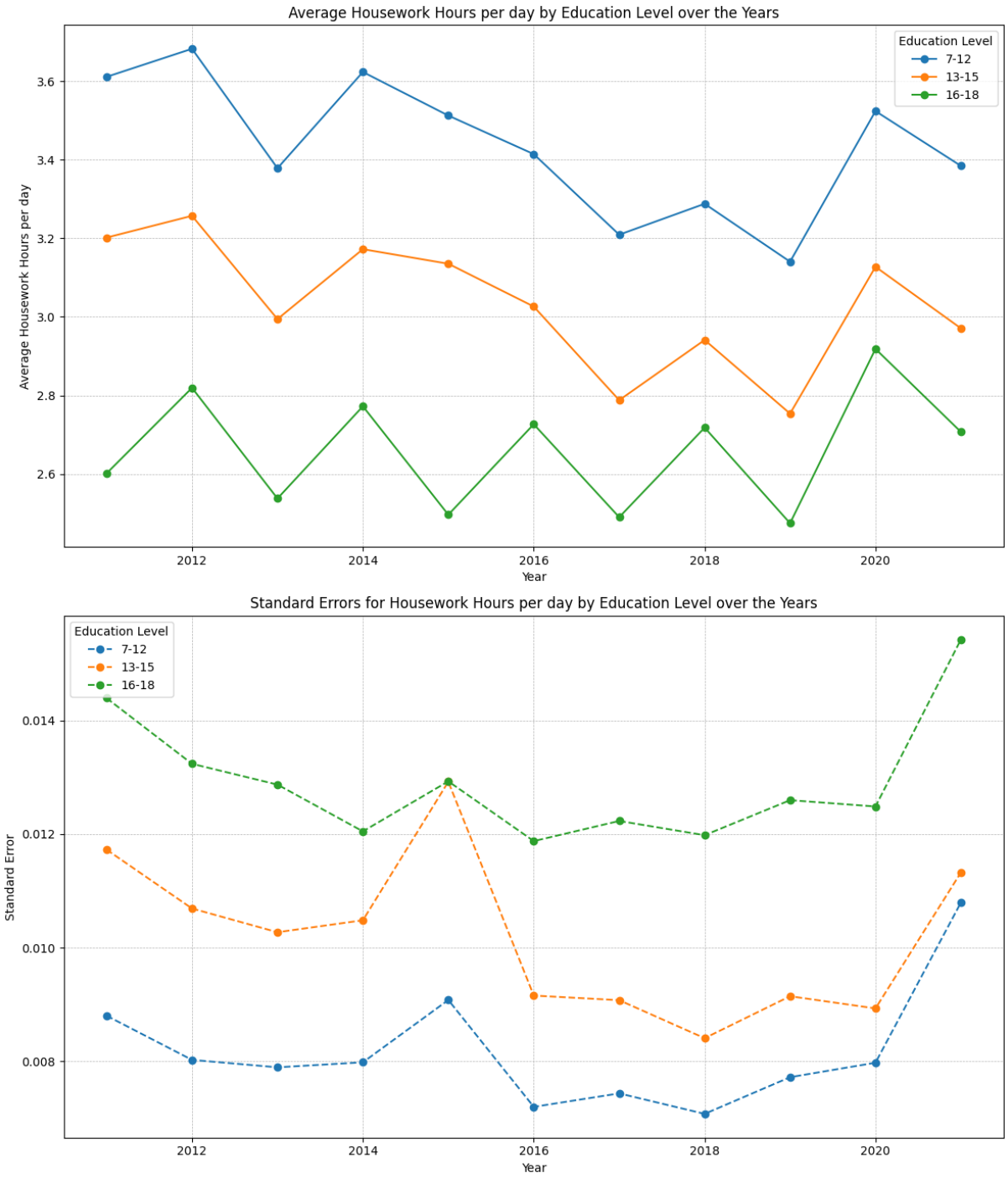


Figure 4. Average Housework Hours and Corresponding Standard Errors by Education level

3.2 Control Variables

In analyzing the relationship between housework and labor market outcomes, it is essential to account for several key covariates—age, education, household size, living status, and gender—that significantly influence both domestic responsibilities and labor market dynamics. These factors are critical for understanding the diverse ways in which housework impacts em-

Figure 5. Average Housework Hours and Corresponding Standard Errors by gender

ployment, earnings, and career progression, as they interact with and potentially modify the effects of housework.

Age is a crucial factor because it often correlates with both housework responsibilities and employment. Many aspects of housework are likely to vary over the life cycle, for example an individuals' fertility decisions are closely related with age. A study by Keith and Malone (2005) found that housework hours significantly reduce wages for young and middle-aged women, with each additional hour lowering wages by 0.1% to 0.4%. This effect diminishes for older women, who, despite doing more housework, show no consistent wage impact. For men, housework hours do not significantly affect wages across any age group. The wage penalty for women appears linked to child care responsibilities, which are more intense for younger

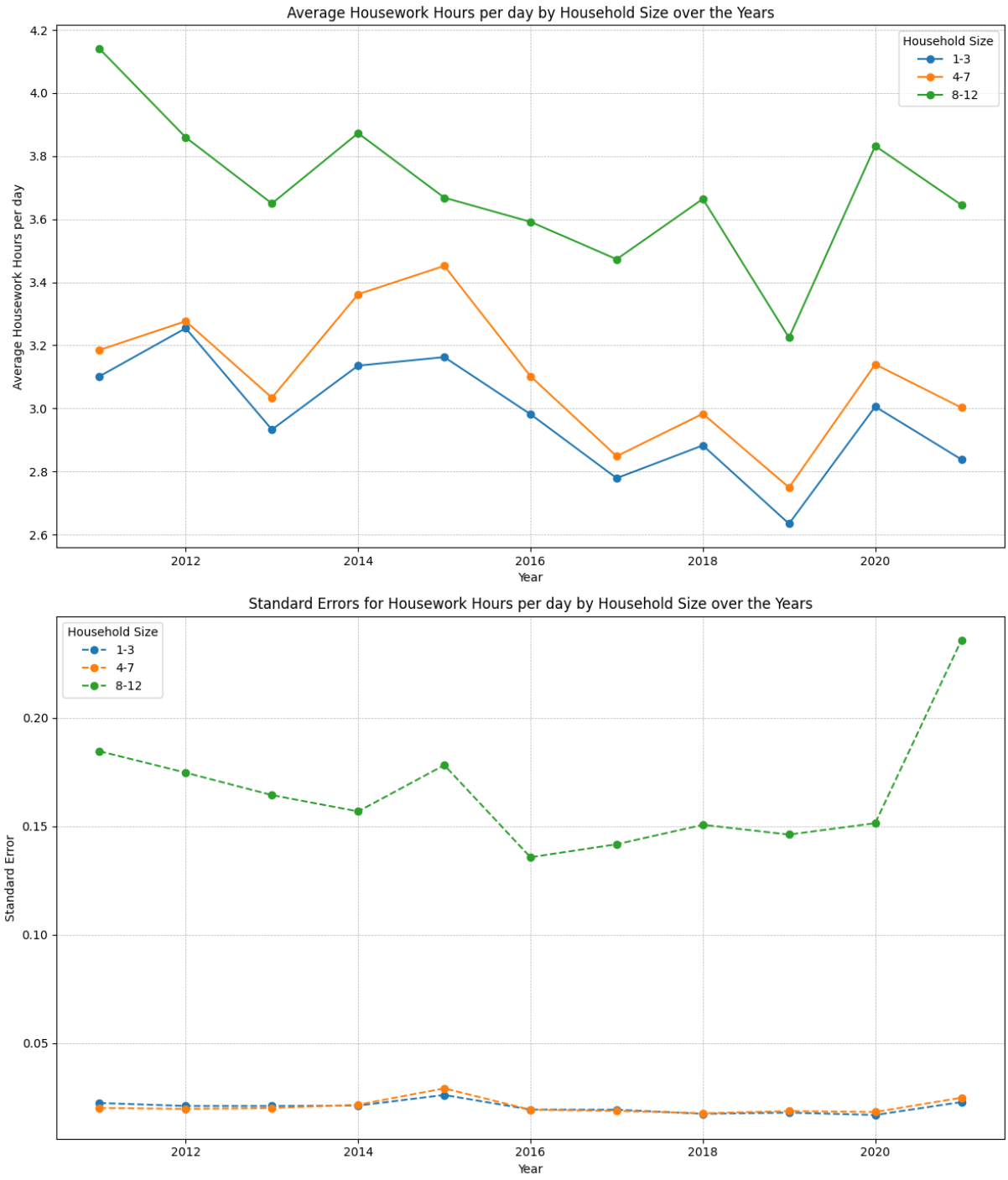


Figure 6. Average Housework Hours and Corresponding Standard Errors by Household size

women. Housework time contributes 3-10% to the gender wage gap, particularly for younger workers, where it explains an additional 1-3 percentage points. Unobserved individual factors and societal changes might influence these dynamics, potentially leading to more equitable sharing of housework among younger couples.

S. M. Bianchi et al. (2000) Relative to younger persons (those aged 25 to 34, the omitted category in the regressions), all older age groups do significantly more housework. Men aged

55 to 64 average almost 5 more hours per week than men aged 25 to 34. For women, housework hours are marginally higher after age 35 and appear to rise again after age 45 and then level off. Further a review of different studies by De Lange et al. (2021) shows a strong negative relationship between age and wages, particularly for older workers, due to reduced job mobility, lower employability perceptions, and less favorable labor market conditions. Factors like declining health, job tenure, and limited skill development contribute to this decline. And similar negative relationship with older age and income is also estimated by Neumark and Johnson (1997), Blau and Kahn (2000), and Lahey (2008).

Education strongly influences labor market outcomes and also affects how housework is distributed within households. Higher educational attainment typically leads to better employment opportunities, higher earnings, and greater career stability, which may enable individuals to outsource housework or negotiate its distribution more effectively. Additionally, education can shape attitudes towards gender roles, impacting how housework is divided between partners. By including education in the analysis, I can explore whether and how the trade-offs between domestic responsibilities and career progression vary across educational levels. S. M. Bianchi et al. (2000) found educational differentials to be relatively small in the multivariate models, with college graduate men doing over an hour more and college graduate women over an hour less than those with a high school education or less. Hersch and Stratton (1997c) found that more educated women face a larger wage penalty for housework due to higher opportunity costs.

Sullivan (2011) suggested that education may mitigate the negative impact of housework on wages, as educated individuals often have better access to resources that reduce this effect. Moreover, higher education significantly boosts income and labor market outcomes. Card (1999) found that each additional year of schooling increases wages by 7-10%. Oreopoulos and Petronijevic (2013) highlighted that the return on higher education is substantial, with college graduates earning much more than non-graduates. Autor, Kearney, and Katz (2008) noted that the rising demand for skilled labor has widened the wage gap between those with higher education and those without, contributing to increased income inequality.

Household size directly impacts the amount of housework required. Larger households generally demand more time for domestic tasks, potentially increasing the burden on individuals who also have work commitments. The household composition, including the presence of children or elderly dependents, further complicates this dynamic. It is reasonable to assume that the total amount of housework required to maintain a family would depend on the size of the family and the members' ages. The housework responsibilities vary over the life cycle

of children as well with higher workload in infancy and early childhood and lower after the higher education. Therefore, including household size as a covariate allows for a more precise assessment of how housework affects labor market outcomes by accounting for varying domestic demands.

Additionally, because of less data on the marital status of the individuals in the sample, I consider household size as a proxy, categorizing the number of household members into groups will likely solve the problem of understanding if an individual lives alone, with a partner and also if with a larger family(Keith and Malone (2005)). Estimates from Bryan and Sevilla-Sanz (2010) provide strong evidence of differential impacts of housework across marital status for women. Moreover, larger household sizes generally reduce labor supply and earnings, particularly for women, due to increased household responsibilities and childcare demands. Studies by Angrist and Evans (1998), Becker (1981), and Leibowitz (1974) all highlight how having more children or a larger household can negatively impact labor market participation and income.

Gender is perhaps the most critical covariate, given the well-documented gender disparities in the distribution of housework. Women typically perform more housework than men, which can limit their time and energy for career pursuits, potentially resulting in lower earnings and fewer opportunities for advancement. Gender is a common theme among all the above covariates mentioned since it interacts with all these factors and potentially more of them in complex ways. Hersch and Stratton (1994) found that women spend about 20 hours per week on housework compared to men's 7 hours, contributing significantly to the gender wage gap, increasing it by about 14 percentage points. Studies like Bertrand, Kamenica, and Pan (2013) show that social norms further exacerbate this penalty for high-earning women. Age also matters, with younger women facing a 0.1–0.4% wage reduction per additional hour of housework Keith and Malone (2005). The inflexibility of housework schedules leads to even greater wage penalties for women Bonke, N. D. Gupta, and Smith (2010).

Lastly, I include pre-treatment outcome variables, specifically previous labor force status, to account for individual differences in employment history that may influence current labor market outcomes and gross income from the baseline year 2011. These adjustments ensure a more comprehensive and precise analysis. The variable used for gross income is defined in the section below, and the labor force status is based on the annual question on current employment status, combined with additional information on activities of non-working individuals. The variable provides a differentiation between "working" and "non-working" across all waves. The

two categories are divided into more sub-categories and for the analysis in this study, I have used these two primary categories to treat this as a binary variable.

3.3 Outcome variables

In my analysis, I used gross income as the primary outcome variable due to its continuous nature and consistent availability across all years in the dataset, which allowed for robust conclusions about the relationship between housework and income. The variable is named as Current gross labor income in euros and is taken from subfile "generated variables". Additionally, I included actual weekly working hours, defined as actual average working hours (including overtime) reported by individuals, rather than contractual hours. The variable is named as Actual weekly work time and the data are obtained by asking respondents how many hours they work on average per week.

4 Simulation study

For the simulation study, I conducted a series of simulations with varying Data Generating Processes (DGPs) to assess the performance of different machine learning models in estimating treatment effects. Specifically, we explored three primary meta-learning frameworks: TLearner, SLearner, and XLearner. These frameworks were implemented using both Random Forest (RF) and Gradient Boosting (GB) models. The objective was to identify the scenarios where each learner and model combination performs optimally, providing insights into the strengths and weaknesses of each approach.

4.1 Data Generating Process (DGP)

The simulated dataset consisted of 10,000 observations, with covariates designed to mimic real-world factors influencing treatment effects. The key covariates included: *Age Group (Cohort)*, a categorical variable with five levels representing age ranges (1-25, 26-35, 36-45, 46-55, 56-66); *Gender*, a binary variable indicating gender (0 for female, 1 for male); *Household Size*, an integer variable ranging from 1 to 12; *Education Level*, a continuous variable representing years of education, normally distributed around 16 years with a standard deviation of 2 years; *Baseline Income*, a continuous variable calculated as a function of the covariates, representing initial income before any treatment effect; *Labor Force Status*, a binary variable indicating participation in the labor force (1 for participation, 0 otherwise); *Treatment Indicator (W)*, a binary variable indicating whether an observation is in the treatment group (1) or control

group (0); and *Observed Income*, the final income observed after applying the treatment effect, depending on the treatment assignment.

4.2 Metrics Used for Evaluation

Across all simulations, I evaluated the models using the following metrics:

- *Bias*: The average difference between estimated and true treatment effects, indicating the accuracy of the models.
- *Mean Squared Error (MSE)*: This metric measures the average squared difference between estimated and true treatment effects, reflecting both variance and bias in the models.
- *Root Mean Squared Error (RMSE)*: The square root of MSE, providing an interpretable error metric in the same units as the treatment effect.
- *Mean Absolute Error (MAE)*: This metric averages the absolute differences between estimated and true treatment effects, emphasizing the impact of outliers.
- *R-squared (R^2)*: The proportion of variance in the true treatment effect explained by the model, indicating the overall fit.
- *Explained Variance*: This metric indicates how well the model captures the variance in treatment effects, providing insights into the model's robustness.

First part of the simulation study focused on identifying a suitable meta-learner by applying the same data-generating process across three meta-learners—S-learner, T-learner, and X-learner—paired with two base learners: Random Forest and Gradient Boosting regressors. As detailed in Appendix C, the X-learner consistently delivered superior results across various combinations, thus emerging as the optimal choice for subsequent analyses.

In the second phase, the goal was to determine the most effective base learner for the analysis. This phase involved running a series of simulations that included variations such as introducing linear and non-linear terms for both the baseline outcome and treatment effects, implementing purely linear terms for the treatment effect, applying a balanced 50:50 treatment assignment split, increasing the complexity by incorporating additional non-linear terms in the treatment effect, and adjusting noise levels. I tailored the original data-generating process to create distinct DGPs for each scenario. The X-learner, combined with both Random Forest and Gradient Boosting base learners was used to assess performance across these simulations.

Among all simulations, the X-learner with Gradient Boosting consistently yielded the most optimal results. Subsequently, I modified the DGP by altering the linearity, treatment assignment

split, and noise levels, while maintaining the X-learner as the meta-learner and experimenting with both Random Forests and Gradient Boosting as base learners. Across these simulations, Gradient Boosting consistently produced the lowest Root mean squared errors a higher (R^2) values and explained variance. Although bias fluctuated slightly across the simulations, the consistency in the other metrics, i.e. MSE, RMSE,R^2 when using Gradient Boosting, made it the final choice for the empirical analysis. Please find detailed results in Appendix C.

5 Empirical Analysis

5.1 Assumptions

There are 3 key assumptions for my analysis, Conditional independence assumption, unconfoundedness and SUTVA, Conditional Independence assumption being the identifying assumption for the analysis. The CIA (as mentioned in Section 2) posits that, conditional on a set of observed covariates, the assignment to treatment (in this case, performing 2.5 or more hours of housework) is independent of the potential outcomes. Formally, this means that once you control for a comprehensive set of relevant variables, the decision to engage in more intensive housework is effectively random, and any differences in labor market outcomes between those who do and do not perform high levels of housework can be attributed to the treatment itself.

As mentioned in Section 3, we start with the pre-treatment period t=-1 and restrict the analysis to a subsample of individuals with D_1=2.5, we then observe the housework hours in the year t=0 and then we measure the effects on labor market outcomes from t=1 to t=8. Each individual has two potential outcomes per period, Y_t^1 and Y_t^0, where the superscripts denote potential outcomes with or without performing 2.5 or more hours of housework in period 1. The causal effect of performing 2.5 or more hours of housework in period 1 on labor market outcomes in period t is $Y_t^1 - Y_t^0$. This individual treatment effect is a well-defined parameter but impossible to determine for the researcher as only the factual, not the counterfactual, outcome is observed. The observational rule is:

$$Y_t = D_1 Y_t^1 + (1 - D_1) Y_t^0$$

where D_1 is an indicator that equals 1 if the individual performs 2.5 or more hours of housework in period 1, and 0 otherwise, irrespective of any changes in D_t over time.

Ideally, I would like to randomly assign individuals to different levels of housework in $t = 1$, follow them over several years, and observe their performance and outcomes in the labor market compared to those who are not assigned to perform 2.5 or more hours of housework. This kind of an expperiment would help evaluate the average causal effect of performing 2.5 or more hours of housework in $t = 1$, irrespective of how long the person actually continues this level of housework. However, we are dealing with observational data, and housework is most naturally a voluntary decision. Individuals weigh costs and benefits based on the opportunity, willingness, and ability to perform housework.

While it is not my aim to fully model the decision to perform housework, I proceed with controlling for all variables that affect both treatment and outcomes, leaving the decision to engage in 2.5 or more hours of housework a random event (conditional on controls)(Schmitz and Westphal (2017)). In other words, in order to identify the link between housework and labor market outcomes, we make a conditional independence assumption:

$$Y_t^1, Y_t^0 \perp D_1 \mid X_0, Y_0 \quad \forall t > 0$$

In Section 3 I have detailed the variables we account for in X in order to justify this assumption. We exploit detailed information on individuals' household size, years of education, gender, age and, most importantly, pre-treatment outcomes Y_0, which will capture the labor force status, gross income and gross weekly working hours in the baseline period (2011). The identifying assumption here is that, conditional on all covariates and past outcomes, the observed treatment is as good as random.

Additionally, note that the treatment status is only defined in $t = 0$ which is 2012 and not affected by later housework engagement, as this might be endogenously affected by future realizations of the outcome or control variables

On the basis of the definition of the common supprt assumption (Section 2), for all values of the covariates X_i, the probability of assignment lies between 0 and 1. In this analysis, since housework is performed by all subgroups, as demonstrated in the graphs in Section 3, it is reasonable to conclude that this assumption is satisfied before proceeding with the analysis.

Finally, the covariates are observed in the baseline year 2011 so as to satisfy the exogeneity assumption which implies that the covariates X_i are not influenced by the treatment D_i and are determined prior to the treatment assignment. Since I observe the age, gender, household size and education in years in the pre-treatment year, for the observed covariates, we proceed with the exogeneity condition satisfied.

As mentioned in Section 2 the parameter we are primarily interested in is the Conditional Average Treatment Effect (CATE), which measures the expected effect of performing 2.5 or more hours of housework on labor market outcomes, conditional on the defined set of covariates:

$$\text{CATE}(x) \;=\; E(Y_t^1 - Y_t^0 \;\mid\; X \;=\; x)$$

This is the difference in expected outcomes between those who perform 2.5 or more hours of housework and those who do less, given the covariates $X = x$ (age, gender, household size and education in years). The difference estimated will be for gross income and gross weekly working hours.

5.2 Analysis

The objective of our analysis is to find the heterogeneous link between average housework hours and labor market outcomes, specifically income, in the short, medium, and long term.

Treatment here is defined as a binary variable for spending more than 2.5 housework hours per day on average . We are using housework hours as a treatment variable because continuous variables, especially those based on self-reported data like hours of housework, can be prone to small inaccuracies that blur the effects when used directly. Binarization reduces the impact of these inaccuracies, leading to more reliable and robust results.

With a small dataset, dichotomizing helps avoid overfitting by simplifying the model and providing more stable estimates. Treating hours of housework as a continuous variable can result in uneven distribution, with clusters around certain hours (e.g., 4-6 hours) and sparse observations at others (e.g., 1, 9, 11 hours). This unevenness can complicate analysis by introducing noise or distortion. Binarizing the data addresses this by grouping hours into broader categories, ensuring balanced representation and reducing the impact of sparsely populated categories on the results. Additionally, the data identification strategy, particularly the dichotomization and timeline, is drawn with the help of Schmitz and Westphal (2017), who dealt with a similar case involving the impact of informal work hours. Moreover, my chosen methodology requires a binary treatment, further justifying this approach

And lastly, while non-linearity can occur after dichotomization, simplifying the treatment variable can mitigate complex non-linear relationships, making the effects more interpretable,

and Figure 1 shows that the distribution naturally divides into distinct categories, with 2.5 hours being the threshold.

I am implementing X-learner as the meta-learner and Gradient Boosting as the base learner in my analysis. My decision for these algorithms is based on both the literature (Künzel et al. (2019)) and the simulation studies explained in Section 4.

This analysis investigates the effect of engaging in more than 2.5 hours of housework per day (the treatment) on labor market outcomes over the years 2013 to 2020. The analysis leverages Conditional Average Treatment Effect (CATE) estimates, which have been computed using the aforementioned base-learner and meta-learner for outcome modeling and Logistic Regression for propensity score estimation. As emphasized in the Methodological framework (Section 2) section of the paper, CATE was chosen as the estimation strategy in order to capture heterogeneous effects of housework on labor market outcomes. I aim to understand how the estimates would differ across different subgroups based on age, gender, education level, and number of household members.

The various parts of the analysis include estimating the treatment effects on gross monthly income and actual weekly hours of work. With the covariates age, gender, education level, and household size in the baseline year 2011, the treatment is defined as working more than 2.5 hours per day on average in the year 2012. The longitudinal analysis measures the impact of this treatment on the aforementioned outcomes from years 2013 to 2020. For this estimation, the same individuals are followed throughout the eight years, resulting in some loss of sample size for the estimates. In order to preserve as much data as possible, I have also estimated the treatment effects separately throughout all the years and have discussed the results in the following sections. I have further extracted the CATE estimates for various subgroups to analyze how they differ based on these subgroups, for instance, how the CATE estimates will differ across different age groups, genders, education levels, and household sizes. I have lastly estimated CATE estimates separately for men and women as well, i.e. CATE (women) given the other three covariates age, education, and household size, and likewise CATE (men). With this, I aim to isolate and understand the gender-specific effects of housework on the outcomes. This approach is added basically to make the analysis more robust and to observe the difference in CATE estimates (if any) based specifically on gender. The analysis aims to also observe the potential difference in outcomes specifically across genders, and therefore with this extra mile, I aim to understand more clearly the effects of housework on men and women, further breaking down their interactions with the other covariates.

Regarding the handling of missing values, I compared the results of dropping missing values with K-Nearest Neighbors (KNN) imputation. Although both methods have the potential to introduce bias—dropping can reduce the sample size, and imputation can introduce inaccuracies—dropping missing values was determined to be the more reliable approach for this analysis. This decision was made to avoid the risk of imputation introducing additional variability or distortion into the data, thereby ensuring that the results are more directly reflective of the observed data.

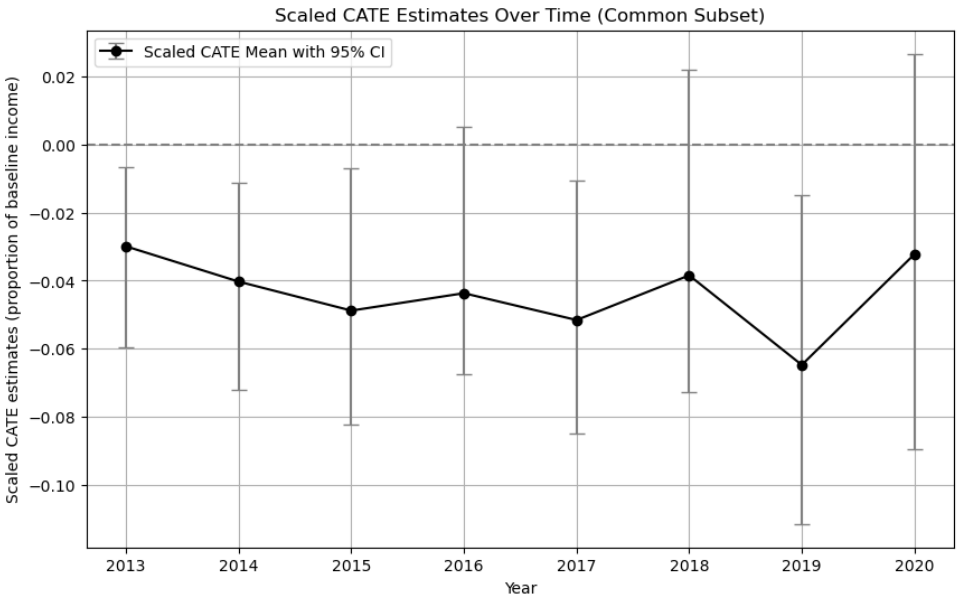


Figure 7. CATE Estimates for Gross Income (Longitudinal data)

For figure 7, the results were obtained by following a consistent subset of 4,133 individuals from 2013 to 2020, with 1,380 of them in the treatment group who met the housework threshold. I had started with 8,826 individuals in 2011 who were treated in 2012, and then we see that around 50% of the sample is followed up until 2020. The CATE estimates represent the percentage change in income from the baseline income in 2011, allowing us to compare the impact over time on a consistent basis. By normalizing the CATE estimates in this way, I aim for an intuitive comparison of treatment effects across different years. These are the mean CATE estimates over the years. The figure also includes 95% confidence interval bars.

Across the years, these scaled CATE estimates remain consistently negative, suggesting that those who spent more time on housework experienced a reduction in income compared to the baseline. The magnitude of these effects varied across years, with the most substantial negative impacts observed in 2015, 2017, and 2019, where the CATE estimates dipped significantly. In 2013, the negative impact on income is approximately 3% lower than the baseline year. This effect worsens slightly in 2014, with a mean CATE of -4%. By 2015, the negative impact

reaches -4.9%, suggesting that individuals performing additional housework during this period experienced nearly a 5% reduction in income compared their income in 2011. The year 2017 sees a significant negative impact at -5.2%, indicating a consistent reduction in income due to the treatment. In 2018 and 2019, the mean treatment effects are -3.8% and -6.5%; 2020, however, sees a 3.2% lower mean income than in 2011.

The statistical significance of the negative effects on income is consistent across the years, with low p-values confirming the robustness of the observed reductions. The effects are statistically significant from 2013 through 2015, with 2016 and 2017 also showing significant negative impacts, despite a slight recovery in 2016. In 2018, the statistical significance weakens somewhat, indicating potential variability, but the negative effect intensifies again in 2019, reaffirming its significance. By 2020, while the negative impact persists, the reduction in statistical significance suggests a decrease in the magnitude of the effect. Overall, the analysis reveals a persistent and statistically significant negative effect of performing additional housework on income across the years, with some fluctuations in the magnitude and significance of the impact.

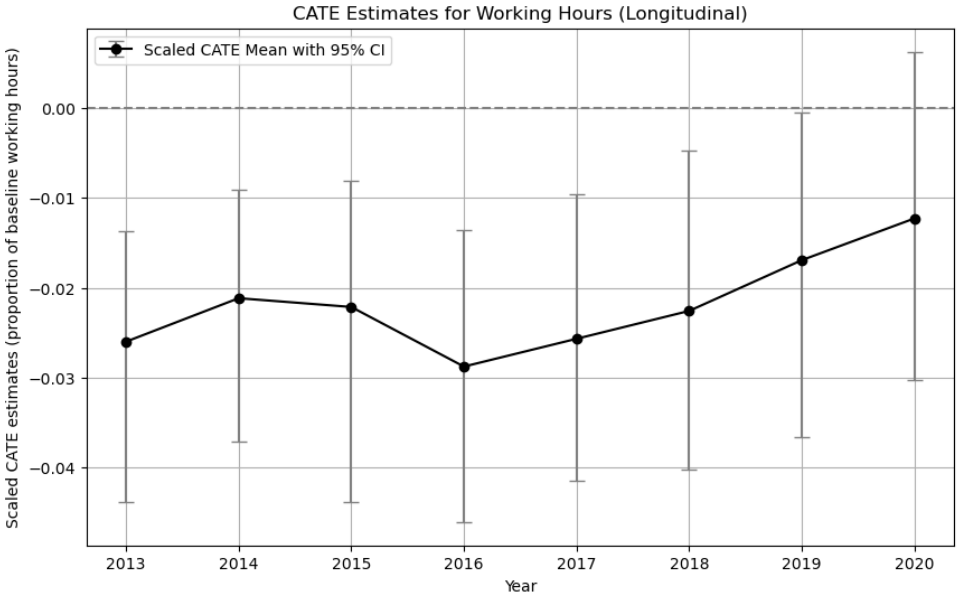


Figure 8. CATE Estimates for Working hours(Longitudinal data)

Figure 8 illustrates the treatment effects on weekly working hours from 2013 to 2020, using 2011 as the baseline. In 2013, working hours decreased by approximately 2.6% compared to the baseline. This reduction lessened slightly to -2.1% in 2014 and remained around -2.2% in 2015. The largest decline occurred in 2016, with a reduction of -2.9%. The years 2017 and 2018 showed a decrease in the negative impact, with mean treatment effects of -2.6% and -2.3%, respectively. By 2019 and 2020, the impact further lessened to -1.7% and -1.2%, indicating a gradual recovery in working hours.

The statistical significance of the negative effects on working hours is evident across most years, with p-values confirming the robustness of the observed reductions. The effects are statistically significant from 2013 through 2018, with 2016 and 2017 showing particularly strong negative impacts. In 2019, the statistical significance weakens, indicating potential variability, but the negative effect remains evident. By 2020, while the negative impact persists, the reduction in statistical significance suggests a decrease in the magnitude of the effect. Overall, the analysis reveals a persistent and statistically significant negative effect of performing additional housework on working hours across the years, with some fluctuations in the magnitude and significance of the impact.

I will now present the year-specific analysis of the CATE estimates, this offers insights into the immediate impact of housework on income by evaluating each year independently, rather than tracking the same individuals over multiple years. This method allows for a larger sample size for each year, as it does not limit the analysis to only those individuals who are consistently present from 2013 to 2020. By analyzing each year separately, we capture a broader snapshot of the population at different points in time, providing a more comprehensive understanding of how housework influences income across various years, without the constraints of longitudinal tracking.

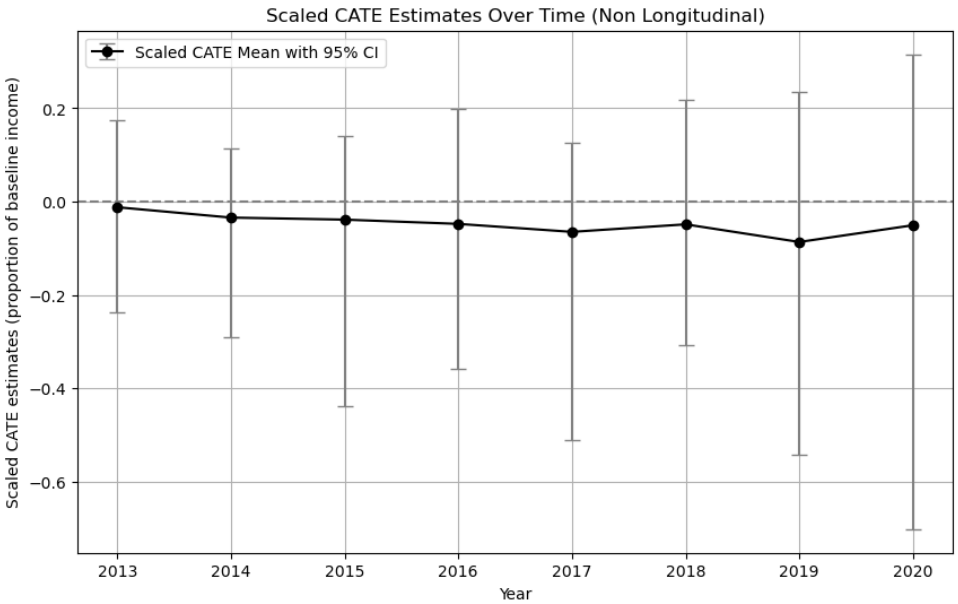


Figure 9. CATE Estimates for Income (Not-longitudinal)

In Figure 9, the scaled CATE estimates for income show a persistent negative impact of housework, with varying intensity over time. The effect starts modestly at -1.2% in 2013, increases to -3.4% in 2014, and peaks at -6.5% in 2017. It then fluctuates, reaching -8.6% in

2019 before receding to -5.1% in 2020. These results indicate a consistent negative impact on income, varying in magnitude across years. The confidence intervals suggest some uncertainty in 2013 and 2014, but from 2015 onwards, they reflect more robust and consistently negative effects. The year-specific CATE estimates differ significantly from those based on longitudinal data, since the confidence intervals here include zero, the effects might not be statistically significant.

The significance of the CATE estimates varies over time. In 2013, 2014, and 2015, the effects are not statistically significant, with p-values of 0.69, 0.25, and 0.20, respectively. A significant effect is found in 2017 (p = 0.03), but in 2018 and 2020, the p-values around 0.10 suggest the effects are again not significant, highlighting variability in the impact over time.

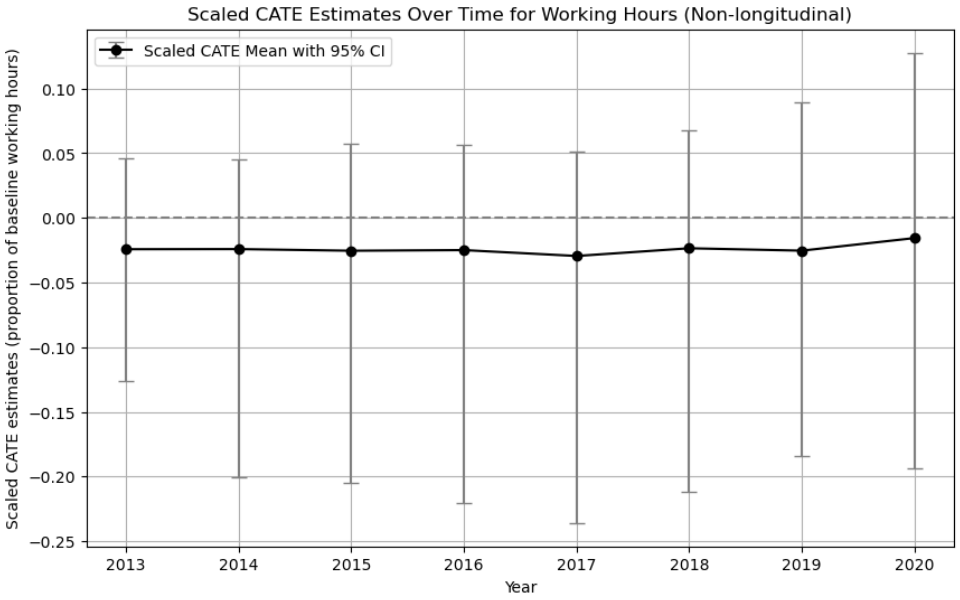


Figure 10. CATE Estimates for Working hours (Not-longitudinal)

Figure 10 above shows the treatment effects on the variable: weekly working hours. This analysis considers the non-longitudinal data for individuals from 2013 to 2020, with 2011 as the baseline year. In 2013, the negative impact on working hours is approximately 2.4% lower than the baseline year remaining relatively stable in 2014 and 2015, with mean CATE estimates of -2.4% and -2.5%, respectively. The most significant negative impact is observed in 2017. This suggests that individuals performing additional housework during this period experienced a notable reduction in working hours compared to 2011. The subsequent years, 2018 and 2019, maintain a similar pattern with mean treatment effects of -2.4% and -2.5%, while 2020 shows a negative impact of -1.6%, lesser than all years before. The confidence intervals indicate that the effect cannot be certainly random.

he effects are statistically significant from 2013 to 2019, with p-values below 0.05 and strongest in 2017 ($p < 0.001$). In 2020, the p-value rises to 0.06, indicating the effect is no longer significant, though the negative impact persists

Table 1. Mean CATE and Standard Errors by Education Group and Year

Education Group	**7-10**		**11-12**		**13-15**		**16-18**	
Year	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**
2013	-0.0218	0.0115	-0.0240	0.0043	-0.0178	0.0051	-0.0224	0.0045
2014	-0.0668	0.0205	-0.0514	0.0045	-0.0606	0.0098	-0.0463	0.0064
2015	-0.0342	0.0099	-0.0565	0.0066	-0.0565	0.0107	-0.0449	0.0078
2016	-0.0639	0.0294	-0.0486	0.0060	-0.0489	0.0086	-0.0423	0.0070
2017	-0.0575	0.0130	-0.0575	0.0055	-0.0544	0.0075	-0.0626	0.0079
2018	-0.0384	0.0137	-0.0378	0.0050	-0.0192	0.0059	-0.0339	0.0069
2019	-0.0452	0.0158	-0.0502	0.0037	-0.0339	0.0092	-0.1078	0.0094
Sample Sizes	161		1261		593		802	

Table 1, shows the mean CATE Estimates and the standard errors by the years of education, revealing a consistent negative impact of the treatment across all education groups from 2013 to 2019. The impact is observed to be more severe for individuals with lower education levels, particularly in the group with 7-10 years of education, where the negative effects are most pronounced in 2014 (-6.68%) and 2016 (-6.39%). In contrast, the group with 16-18 years of education experiences a sharp decline in 2019, reaching -10.78%, indicating a significant negative treatment effect. The standard errors across the groups suggest varying degrees of significance, with some estimates (e.g., 11-12 years of education in 2019) showing high precision, while others, particularly in smaller sample sizes (e.g., 7-10 years), exhibit larger standard errors, indicating less certainty in those estimates.

Table 2. Mean CATE and Standard Errors by Age Group and Year

Age-groups	**18-25**		**26-35**		**36-45**		**46-55**		**56-66**	
Years	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**
2013	-0.0189	0.0089	-0.0197	0.0048	-0.0305	0.0045	-0.0367	0.0047	-0.0330	0.0101
2014	-0.0338	0.0070	-0.0388	0.0113	-0.0456	0.0060	-0.0360	0.0063	-0.0444	0.0113
2015	-0.0586	0.0221	-0.0630	0.0118	-0.0457	0.0066	-0.0499	0.0071	-0.0329	0.0125
2016	-0.0338	0.0126	-0.0549	0.0111	-0.0383	0.0066	-0.0564	0.0082	-0.0138	0.0076
2017	-0.0317	0.0140	-0.0586	0.0105	-0.0617	0.0055	-0.0403	0.0058	-0.0486	0.0086
2018	-0.0373	0.0164	-0.0446	0.0092	-0.0348	0.0072	-0.0378	0.0066	-0.0365	0.0133
2019	-0.0486	0.0126	-0.0769	0.0115	-0.0556	0.0074	-0.0691	0.0082	-0.0735	0.0251
Sample Sizes	115		469		902		834		170	

Table 2 is a similar table for the subgroup age. The estimates indicate a consistent negative impact across all age groups, with the severity of the effect increasing over time for most age groups. The 56-66 age group shows the most significant decline in income by 2019, with a -7.35% decrease compared to baseline income. The relatively small standard errors across most years suggest that these findings are statistically significant, particularly in the later years where the negative impact is more pronounced. The consistency of the negative impact across

age groups highlights the broad and persistent adverse effect of housework on labor market outcomes, especially as individuals age.

Table 3. Mean CATE and Standard Errors by Household Group and Year

Household size	**1**		**2**		**3-4**		**5-12**	
Years	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**
2013	-0.0204	0.0060	-0.0319	0.0053	-0.0277	0.0034	-0.0414	0.0088
2014	-0.0294	0.0124	-0.0440	0.0070	-0.0339	0.0055	-0.0641	0.0102
2015	-0.0286	0.0082	-0.0517	0.0086	-0.0539	0.0067	-0.0547	0.0099
2016	-0.0310	0.0087	-0.0483	0.0085	-0.0420	0.0065	-0.0628	0.0115
2017	-0.0654	0.0137	-0.0386	0.0048	-0.0466	0.0047	-0.0835	0.0128
2018	-0.0445	0.0129	-0.0444	0.0078	-0.0356	0.0055	-0.0262	0.0114
2019	-0.0664	0.0151	-0.0672	0.0097	-0.0636	0.0061	-0.0642	0.0142
Sample Sizes	310		699		1136		345	

Table 3 shows the Mean CATE estimates and standard errors across household sizes, which are categorized into groups. I wanted to account for the estimates for individuals living alone versus those living with a partner potentially versus those with larger family than of two. The results consistently show negative CATE values, indicating a decline in income due to treatment across all household sizes and years. Larger households (5-12 members) generally experience more significant income reductions.The standard errors are relatively small, suggesting that these estimates are statistically reliable. However, the consistency in the trend of more considerable negative effects for larger households across years is notable, reflecting the economic pressures these households might face due to treatment.

Table 4. Mean CATE and Standard Errors by Gender and Year

Year	**Men**		**Women**	
Years	**Mean CATE**	**Std Error**	**Mean CATE**	**Std Error**
2013	-0.0314	0.0031	-0.0271	0.0046
2014	-0.0432	0.0050	-0.0352	0.0056
2015	-0.0517	0.0051	-0.0462	0.0075
2016	-0.0426	0.0053	-0.0496	0.0073
2017	-0.0557	0.0049	-0.0442	0.0044
2018	-0.0376	0.0052	-0.0390	0.0062
2019	-0.0679	0.0062	-0.0607	0.0072
2020	-0.0392	0.0075	-0.0197	0.0068
Sample Sizes	1608		882	

The results from table 4, indicate that both men and women consistently experience a negative impact from the treatment across all years, with the magnitude varying. The negative effects are slightly more pronounced for men. For women, the most significant negative impact occurs in 2019, with a CATE of -6.07%. The standard errors are generally small, indicating that these estimates are statistically significant. The larger sample size for males compared to

females may explain the slightly smaller standard errors observed for the male group, providing more precise estimates. Overall, the results suggest that both genders are adversely affected by the treatment, though males may experience a slightly larger negative impact.

In Appendix A I have shown the subgroups' data in figures with the help of faceted plots with error bars for each subgroup. The error bars correspond to 95% confidence intervals for the CATE estimates for each subgroup and year. Among all the subgroups, with household groups one can clearly see the differences in CATE estimates from 2013-2016, the trends are mostly consistent over the years with lowest CATE estimates for single households gradually decreasing with the increase in household size.

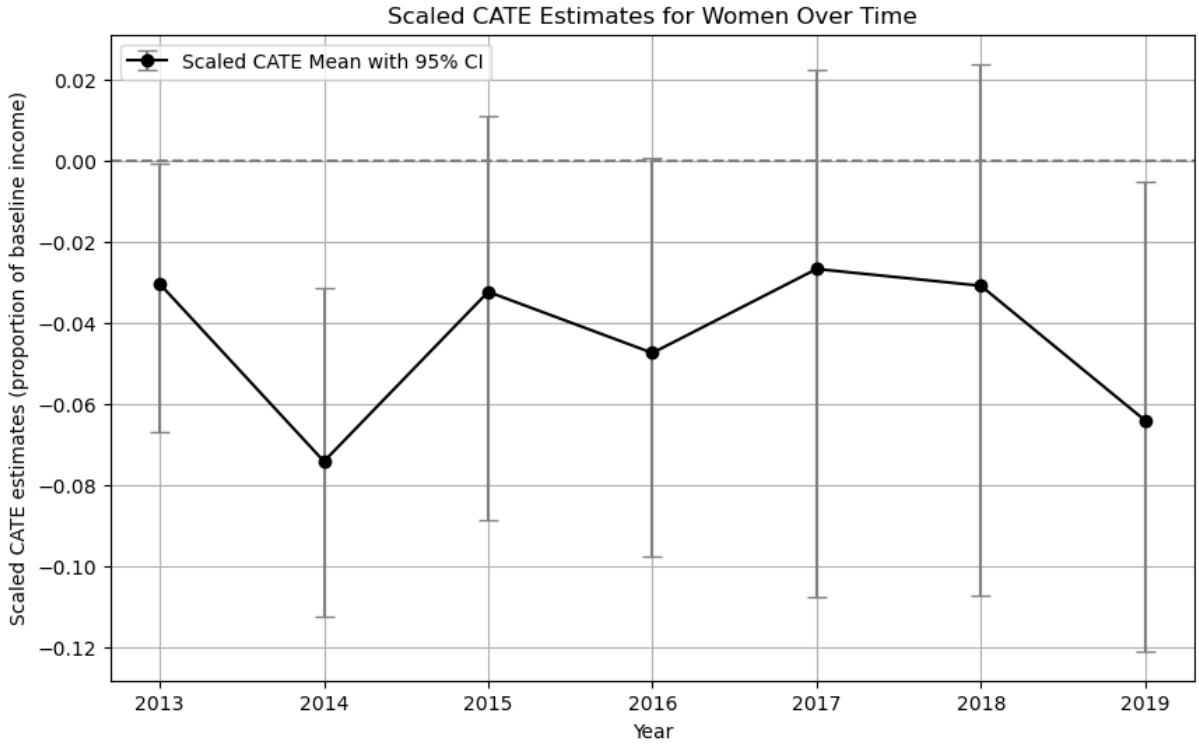


Figure 11. CATE estimation results separately for women [Longitudinal]

Figure 11 shows the scaled CATE estimates for women over the years from 2013 to 2019 indicate that the effect of the treatment on income relative to baseline levels is generally negative. The year 2014 shows the most substantial negative impact, indicating that the treatment resulted in a 7.4% reduction in income relative to the baseline on average. Other years, such as 2015 and 2016, also show negative effects, though these are less pronounced than in 2014, with mean scaled CATEs of around -3.2% and -4.7%, respectively. The impact becomes slightly less negative in the later years. However, by 2019, the effect intensifies again to around -6.4%.

The statistical significance of these estimates varies across the years. The t-values and p-values indicate that the effects in 2014 and 2019 are statistically significant, suggesting that the negative impact observed in these years is unlikely due to random chance. On the other hand, the effects in other years are not statistically significant, meaning we cannot confidently assert that the observed negative impacts in these years are different from zero. This suggests that while there is evidence of a significant negative impact in certain years, the overall trend shows varying levels of impact and statistical certainty.

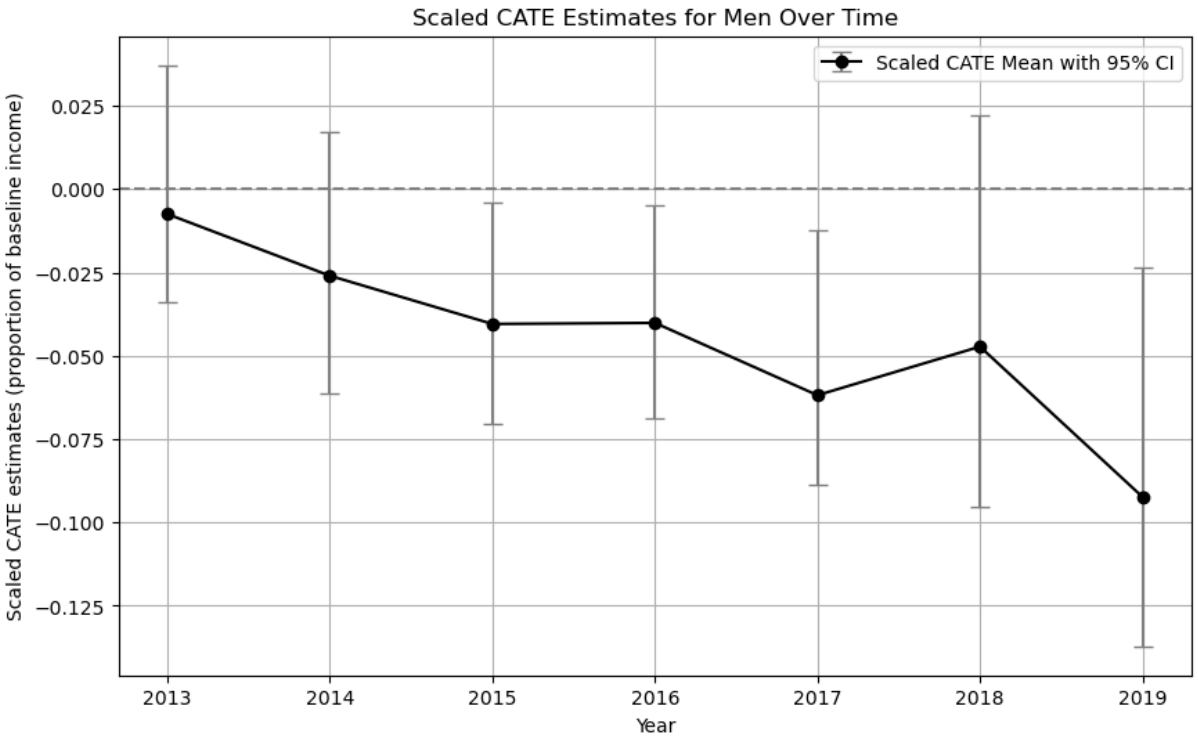
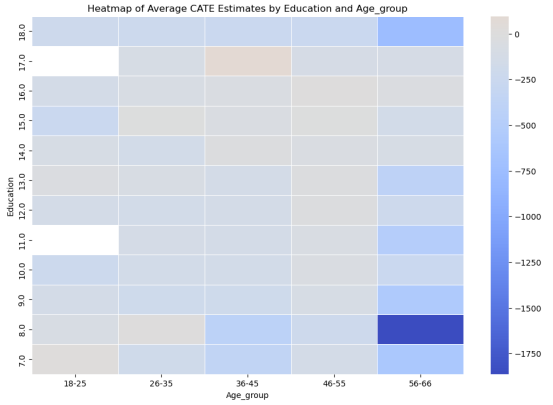


Figure 12. CATE estimation results separately for men [Longitudinal]

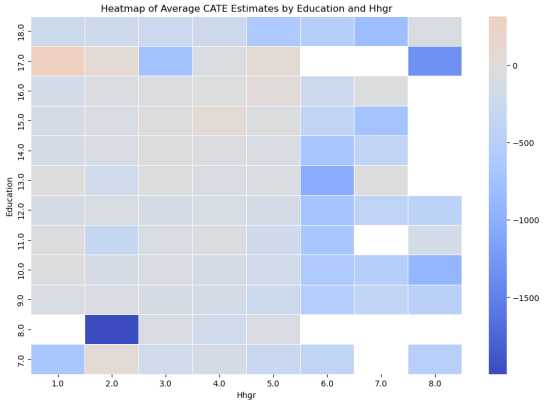
Figure 12 shows the scaled CATE estimates for men over the years from 2013 to 2019 reveal a consistent trend of negative impacts from the treatment on income relative to baseline levels. In 2013, the effect starts relatively small with a mean scaled CATE of approximately -0.7%, suggesting a minor reduction in income. However, this negative impact becomes more pronounced over time, with a marked increase in 2017 where the mean scaled CATE reaches around -6.2%. The most substantial negative effect is observed in 2019 indicating the treatment resulted in a 9.2% reduction in income relative to the baseline for men in that year.

The statistical significance of these estimates varies across the years. The p-values show that the effects in 2015, 2016, 2017, and 2019 are statistically significant, indicating that the observed negative impacts in these years are unlikely to be due to random chance. Conversely,

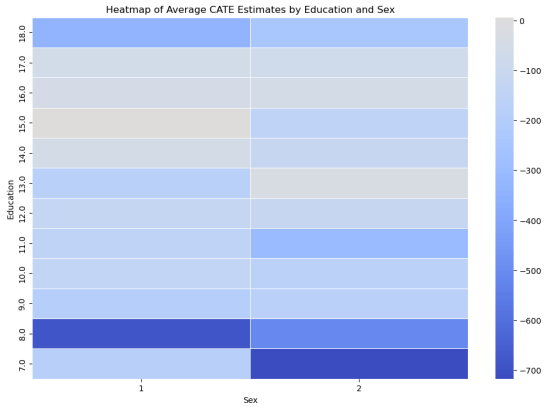
the effects in 2013, 2014, and 2018 are not statistically significant. The variability in statistical significance of the effects over years generally suggest a weak relationship between gender and treatment effects.



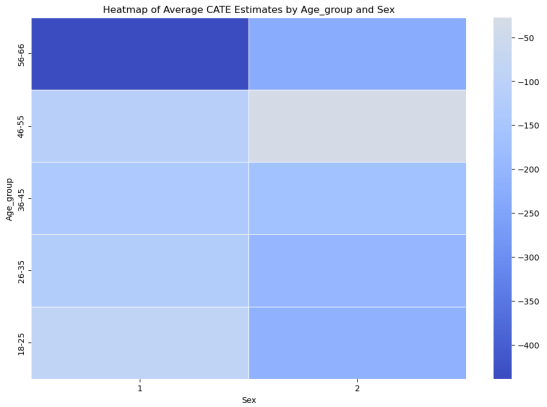
(a) Heatmap of average CATE estimates by years of Education and age groups combined



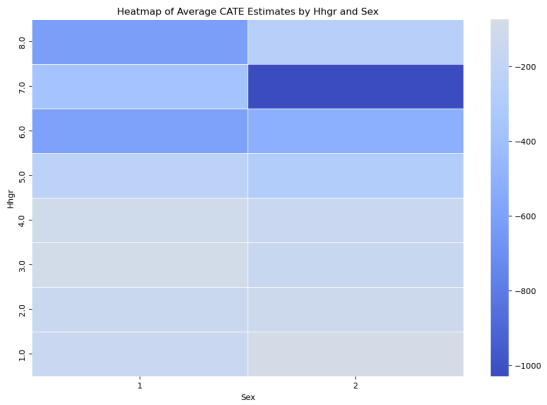
(b) Heatmap of average CATE estimates by years of Education and household size combined



(c) Heatmap of average CATE estimates by years of Education and gender combined



(d) Heatmap of average CATE estimates by age groups and gender combined



(e) Heatmap of average CATE estimates by household size and gender combined

Figure 13. Heatmaps for average CATE estimates' interaction among subgroups

Lastly, I have generated heatmaps in order to analyze the link between the different subgroups, more specifically to see how the different subgroups interact among each other. The analysis comprises the short-term and medium-term link with income. We can see in the

heatmaps with age groups and education levels that the CATE estimates are highly negative for higher age cohorts and lower education levels combined. The heatmaps with the interactions among household size and education levels show a relatively stronger negative link between more housework and lower income in one-person households. The heatmap with interactions between gender and education shows a relatively stronger negative link for women compared to men, and for both genders with lower education levels. The heatmap with the interactions between gender and age groups clearly shows a stronger negative link for women compared to men, with an exception for higher age groups where the effects are stronger for men. And lastly, the ones visualizing the interactions between gender and household sizes in both the years show a relatively stronger negative link for women with higher household sizes. Note that the tables with mean CATE estimates, standard errors and sample sizes are shown in Appendix B.

6 Limitations

Despite the strengths and robustness of the analysis, several limitations should be acknowledged, which may impact the interpretation and generalizability of the findings.

Although the X-Learner methodology used in this study is designed for estimating causal effects, it is important to clarify that the analysis presented here should not be interpreted as strictly causal. The methodology operates under the assumption that the treatment variable is exogenous, meaning it is not influenced by other variables that simultaneously affect the outcome. However, in the context of this analysis, housework cannot be considered entirely exogenous due to several factors that intrinsically link it with labor market outcomes.

Housework, as a treatment variable, is influenced by a multitude of factors that are deeply intertwined with an individual's socioeconomic environment and personal circumstances. Literature suggests that housework is not merely a voluntary activity chosen independently by individuals but is rather shaped by external factors such as gender norms, cultural expectations, economic necessity, and household composition. For instance, studies have shown that women, regardless of their employment status, tend to spend more time on housework than men, driven by traditional gender roles that prescribe domestic responsibilities primarily to women (S. Bianchi et al. (2012), Lachance-Grzela and Bouchard (2010)). This gendered division of labor is often perpetuated by cultural norms that persist across different societies and socioeconomic strata, influencing how housework is distributed among household members (Coltrane (2000)).

Moreover, household dynamics, including the number of dependents, spousal employment status, and the presence of extended family, can significantly affect the distribution of housework. For example, research has found that individuals in households with young children or elderly dependents are likely to engage in more housework due to caregiving responsibilities (Hook (2006)). Similarly, in dual-earner households, the allocation of housework often depends on the relative earning power of each partner, with higher earners potentially bargaining their way out of more domestic tasks (Killewald (2011)). These dynamics suggest that housework is not an isolated variable but one that is deeply connected to the socioeconomic and cultural fabric of an individual's life.

These complexities make it challenging to treat housework as an exogenous variable in this analysis. The observed association between housework and labor market outcomes could be confounded by unobserved variables such as intrinsic motivation, time preferences, or unmeasured aspects of socioeconomic status. For instance, individuals with higher intrinsic motivation or stronger time management skills might be better at balancing housework and paid work, potentially skewing the results if these factors are not adequately controlled for in the model. Furthermore, socioeconomic factors like access to paid domestic help, which is often unmeasured, could lead to differential impacts of housework on labor market outcomes.

Given these considerations, the use of the X-Learner methodology in this study, while methodologically sound for capturing heterogeneous treatment effects, does not fully address the endogeneity of the housework variable. Consequently, the analysis presented here should be interpreted with caution. The results highlight significant associations or links between housework and labor market outcomes, but these should not be mistaken for definitive causal effects. The findings are more accurately described as indicative of a strong correlation, reflecting the intertwined nature of domestic responsibilities and economic activities, rather than a direct cause-and-effect relationship.

Secondly, the decision to use X-Learner combined with Gradient Boosting, while optimal for continuous outcome variables like income and working hours, limited the analysis by precluding the use of categorical outcome variables. Meta-learners like the X-Learner are typically more suited to continuous variables, and while extensions to categorical outcomes exist, they often require more complex adaptations or alternative modeling strategies that were beyond the scope of this study. This limitation restricts the analysis to certain types of outcomes, potentially overlooking important categorical outcomes such as employment status (employed vs. unemployed), type of employment (full-time vs. part-time), or job satisfaction. More variables

such as perception of future job prospects, chances of career advancement although observed and collected were not used as they were rank based which again makes them categorical and therefore challenging to include in the analysis with the aforementioned methodology.

Lastly, the analysis further faced data limitations that led to the exclusion of potentially important covariates such as health status and marital status. The limited availability of consistent data on health, in particular, posed a challenge, as health is a crucial factor that can influence both the ability to engage in housework and labor market outcomes. Marital status was also excluded due to data constraints; however, household size was included as a proxy to account for some aspects of family structure and its impact on housework responsibilities. Nonetheless, the absence of direct data on marital status may mean that certain nuances in the relationship between housework and labor market outcomes, particularly those related to spousal dynamics, are not fully captured.

In summary, while the analysis provides valuable insights into the relationship between housework and labor market outcomes, it is crucial to interpret the findings within the context of these limitations. Future research could address these limitations by incorporating more comprehensive datasets, exploring alternative modeling approaches for categorical outcomes, and considering methods that better account for the endogenous nature of housework.

7 Conclusion

Based on the results in the above section, we discuss the main conclusions and interpretations, also in consideration of the hypothesis of the paper. The analysis provides comprehensive insights into the effects of housework on labor market outcomes, particularly income and working hours, across various demographic and socioeconomic groups from 2013 to 2020. By employing Conditional Average Treatment Effect (CATE) estimates through an X-Learner methodology, the results confirm a negative impact of housework on these outcomes, with most of the results being statistically significant as well.

The CATE estimates show a significant negative impact in the short run through the long run, which is to say the average income falls from 2013 to 2019 relative to that of the baseline year 2011. This corresponds to a significant link between more housework and lower gross income, especially until 2017. We see some fluctuations in 2018 and 2019 and a decreased negative effect in 2020, which means the link between more housework and less income weakens in 2020.*The link therefore weakens in the long term*. Furthermore, the CATE estimates show a similar pattern for working hours over the years, showing a negative and significant effect of

more housework on working hours in the short run through the medium run, i.e., until 2016. In the later years, the effect seems to worsen, and the working hours increase continuously until 2020, still negative in that year.

The analysis of Conditional Average Treatment Effect (CATE) estimates, disaggregated by gender, reveals a persistent negative impact of the treatment on income for both men and women across the years 2013 to 2019. However, the magnitude and statistical significance of these effects vary over time and between genders. The results indicate that the negative impact of housework on income is statistically significant in more years for men than for women. Specifically, for men, the effects are statistically significant in 2015, 2016, 2017, and 2019, suggesting a more consistent and robust negative impact over time. In contrast, for women, statistically significant negative effects are observed in two separate years. This implies that while both genders experience adverse effects from housework on income, these effects are more consistently significant for men across the years analyzed.

Moreover, with subgroup analysis, we can also see the distribution of CATE estimates across age groups, education levels, household sizes, and gender. Based on the analysis of Table 1 through Table 4, it is evident that the treatment under study consistently yields negative effects across various demographic and socioeconomic groups, albeit with varying degrees of severity. The impact of the treatment appears to be particularly pronounced among certain age groups (26-35 and 56-66) and those with lower levels of education.

The analysis further indicates that household size plays a role in the magnitude of the treatment's effects, with larger households generally experiencing more significant negative impacts. This trend is particularly evident over most of the years in the analysis, the largest households facing the most substantial declines. This could be due to the increased financial and caregiving burdens in larger households, which exacerbate the economic pressures during adverse conditions. The positive effect observed in 2016 for the largest households might be an anomaly or could suggest a temporary economic relief or policy intervention that somehow benefited this group.

Furthermore, while the differences in impact between genders are relatively small, males consistently experience slightly more negative effects than females, particularly by 2019. The negative impacts on both genders remain fairly consistent over time, suggesting that while there is a small but persistent difference in how males and females are affected, these differences do not fluctuate significantly from year to year.

It is crucial to interpret these results with caution. The analysis presented is not causal and merely identifies associations between the treatment and various outcomes across different demographic groups. The findings should be seen as indicative of potential trends and correlations rather than definitive evidence of cause and effect.

And lastly, from the heatmaps that show the interactions among the various subgroups, some persistent results are evident from the short run through the medium run. Overall, the negative effects of housework are more pronounced among individuals with lower education levels, particularly when these individuals are also older or part of smaller households. Women generally experience stronger negative impacts compared to men, especially when they belong to lower education levels or larger households. However, in higher age groups, men tend to face more severe negative effects. These patterns suggest that lower education, advanced age, and larger household sizes amplify the negative link between housework and income, with women being more adversely affected across most subgroups.

In this study, although housework is typically unpaid, our results clearly demonstrate its notable relationship with the time and effort available for paid work, subsequently affecting income. The observed link between housework and income emphasizes the importance of studying this relationship to better understand the trade-offs individuals make when allocating time between housework, paid work, and leisure. Housework, while not financially compensated, demands time that could otherwise be invested in paid employment, leading to potential income losses. This trade-off is crucial for individuals to understand and make informed choices regarding their time and effort allocation in various activities, including housework.

Moreover, understanding the trade-off between housework and leisure is equally important. Housework can sometimes serve as a form of self-improvement or leisure, blurring the lines between productive and non-productive time. Therefore, studying this aspect can provide insights into how individuals balance their time across different activities to maximize both economic and non-economic benefits. “This balance is particularly important in modern society, where time has become a scarce resource, and the decision to allocate it to different tasks can significantly influence one’s overall quality of life”(S. M. Bianchi et al. (2000)).

The dynamics of housework are particularly relevant in dual-income households, where responsibilities can be shared but at the same time, can also be influenced by the bargaining power among partners. This division of labor can be shaped by the income and work effort each partner contributes to the household. Our analysis shows that single households experience a stronger negative link between housework and income, while households with two or three

members exhibit less pronounced negative effects. This finding underscores the complexity of household labor dynamics, and with a better value attached to housework for different subgroups, there could potentially be a better time and effort allocation in a household as well.

Measuring the impact of housework on income is also valuable because it helps quantify the monetary value of housework, as mentioned in Section 1. Understanding this value allows individuals to make more informed decisions about their time investments. For instance, if the income loss due to housework is substantial, it might be more economically viable to hire external help for household chores, effectively converting potential income loss into a financial gain that could be allocated toward hiring services (Folbre (2006)). However, to determine the optimal amount of external help, one needs to understand the precise income-housework relationship, which this study aims to shed light on.

Finally, attaching a monetary value to housework has broader economic implications. It could lead to the inclusion of housewives' work in national income and production measures, thereby recognizing their contribution to the economy. This recognition would bring those engaged in unpaid housework into the economic chain, highlighting their role in sustaining the economy even without direct financial compensation (Folbre (2006)).

In conclusion, this study underscores the significance of housework in shaping labor market outcomes and highlights the need for a more comprehensive approach to evaluating time use and its economic implications. The findings could contribute to the academic understanding of household labor dynamics but also have practical implications for individuals seeking to optimize their time and policymakers aiming to recognize and value all forms of labor.

Building on the findings of this study, future research could focus on understanding the long-term impacts of housework on career progression, promotional chances, mental and, physical health. Different methods and treatments could be explored to analyze this relationship more deeply. It would be innovative to consider approaches where housework could be treated as exogenous or perhaps even rewarded, to better evaluate causal impacts. My motivation for pursuing this topic stems from growing up in a 'developing nation,' where housework responsibilities are often unevenly distributed, significantly affecting labor market outcomes for certain sub-populations. The results of this study highlight a small yet distinct impact, which may partly reflect the differences in the dataset drawn from a 'developed nation.' Future research should aim to uncover these nuances further, potentially leading to more equitable economic policies globally.

Appendix A Additional Graphs

Referencing Section 3 here, I am adding some graphs here to illustrate that the threshold of 2.5 hours as the distinction between treatment and control groups.

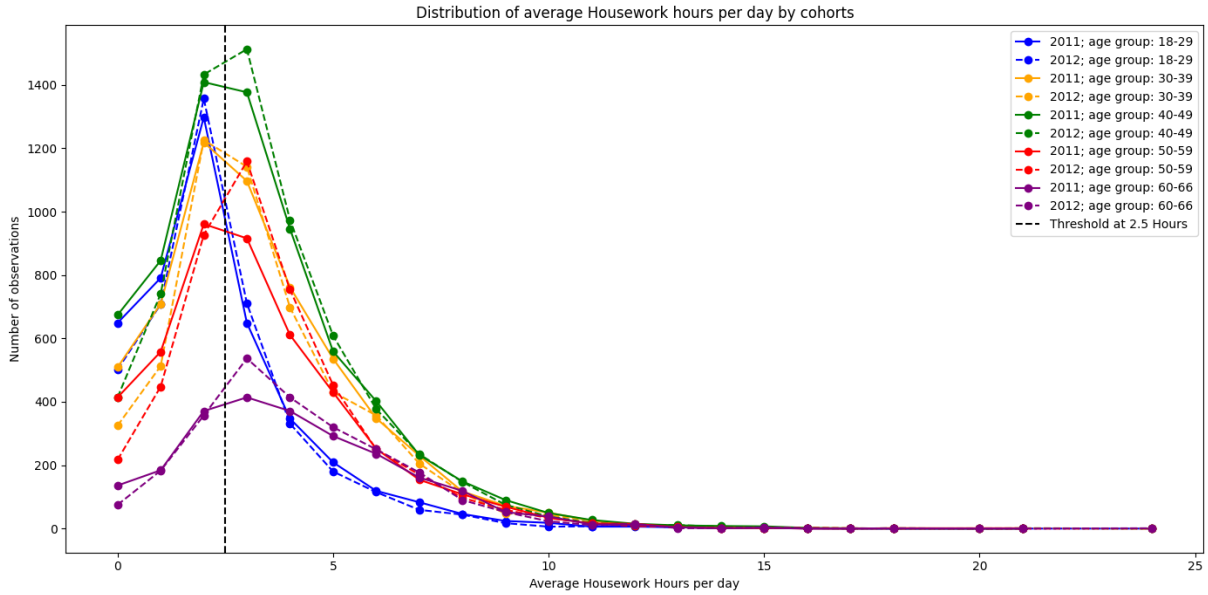


Figure A.1. Distribution of hosuework hours across the population by age-cohorts

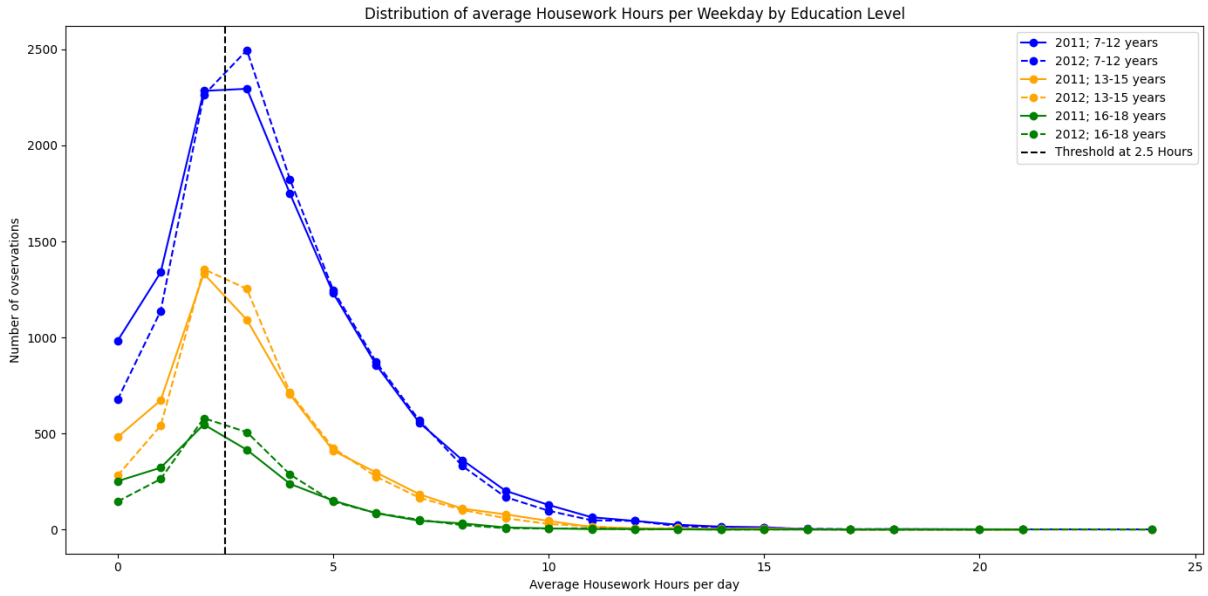


Figure A.2. Distribution of hosuework hours across the population by Education in years

Figure Figure A.1 shows the frequency of individuals with different levels of housework hours segregating on the basis of age cohorts, and the data used is for 2011 and 2012 for the entire population. We can clealy see 2.5 comes out to be a reasonable threshold. Similarly, Figure A.2 shows the frequency across different groups based on years of education, Figure A.3 and Figure A.4 show similarly across the sexes and household sizes respectively.

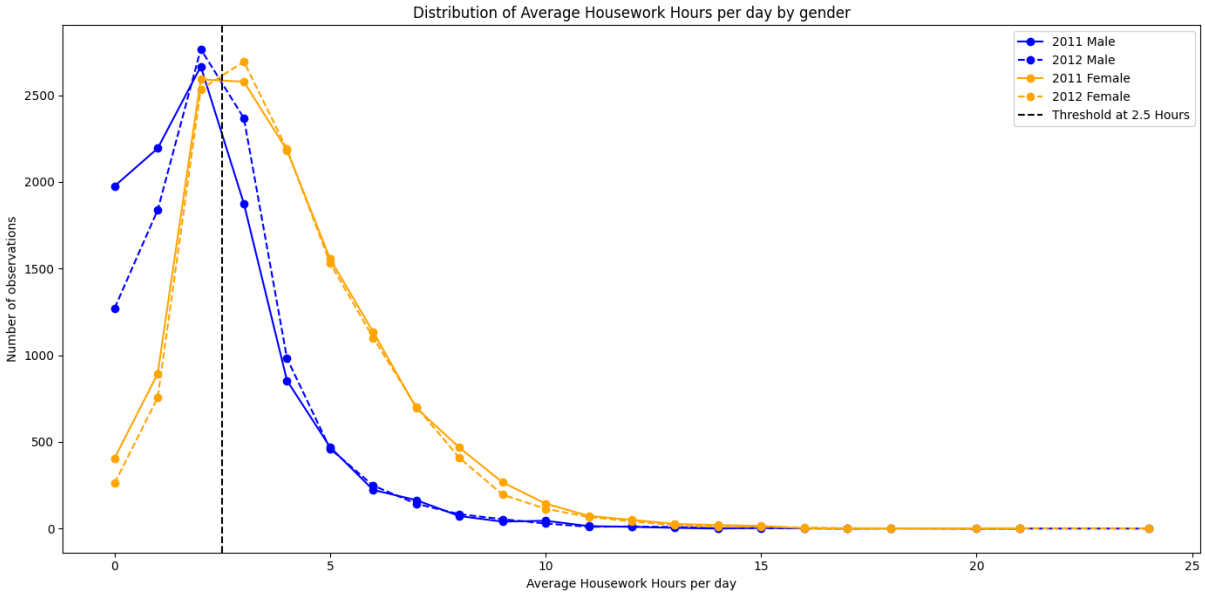


Figure A.3. Distribution of hosuework hours across the population by gender

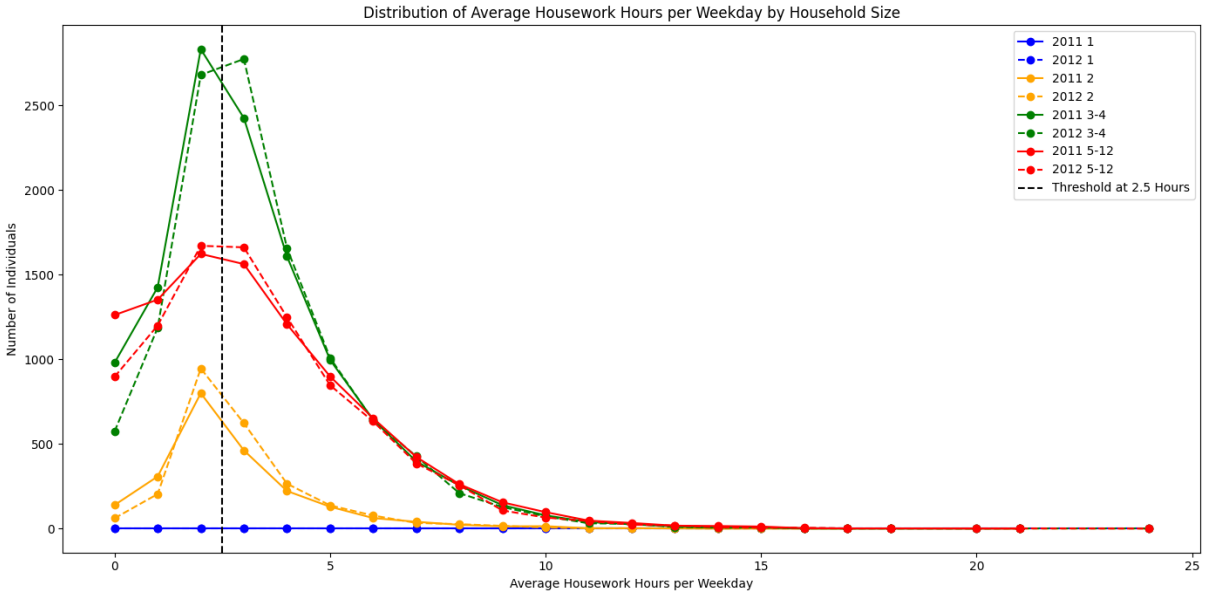


Figure A.4. Distribution of hosuework hours across the population by household size

Secondly, in Section 5.2's Subgroup analysis over the years, I have used tables to show mean CATE estimates and standard errors. Here I am adding the graphs for all the years seprately with the confidence intervals as well.

The graphs show the same trends as those in the tables. I created them to understand the patterns visually and also visualise the Confidence intervals.

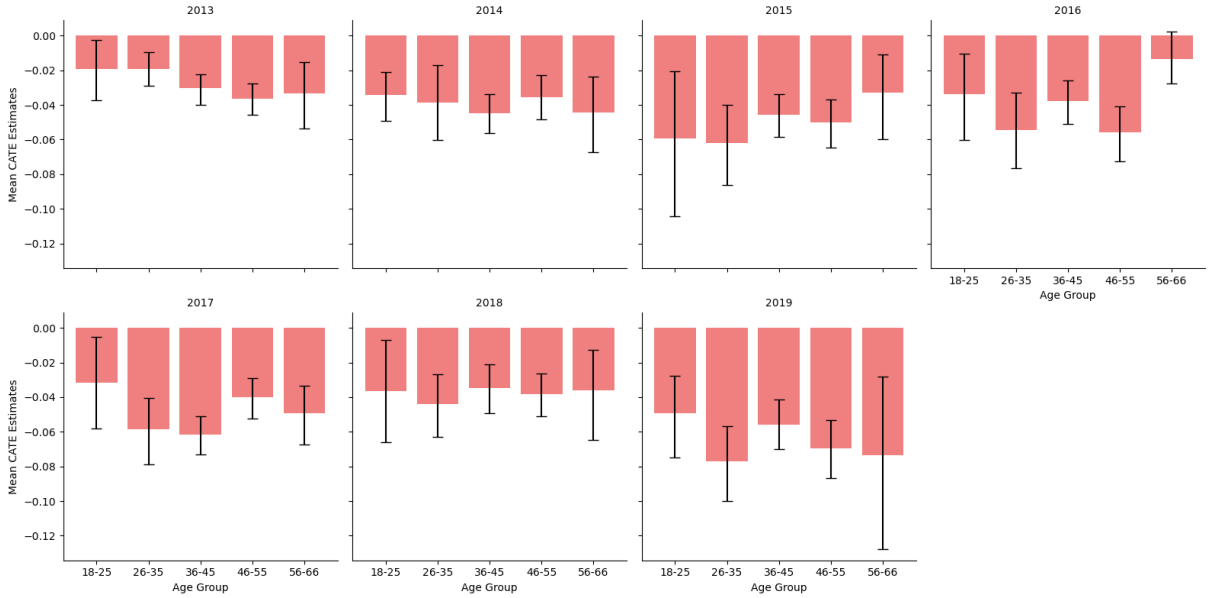


Figure A.5. Distribution of CATE estimates over the years by age-cohorts

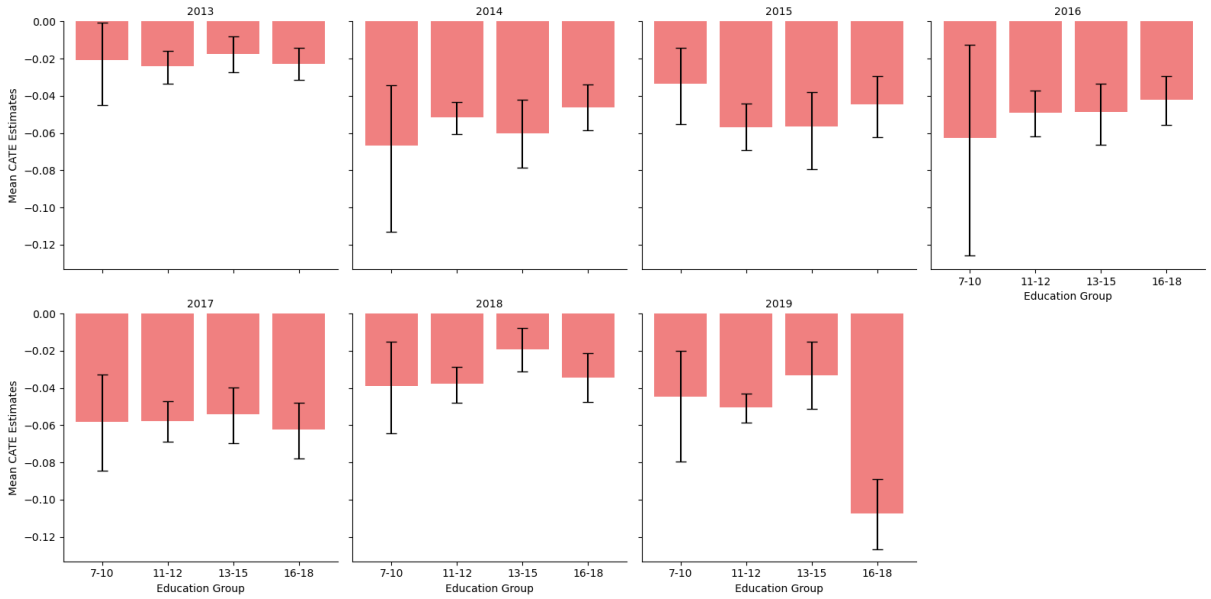


Figure A.6. Distribution of CATE estimates over the years by education in years

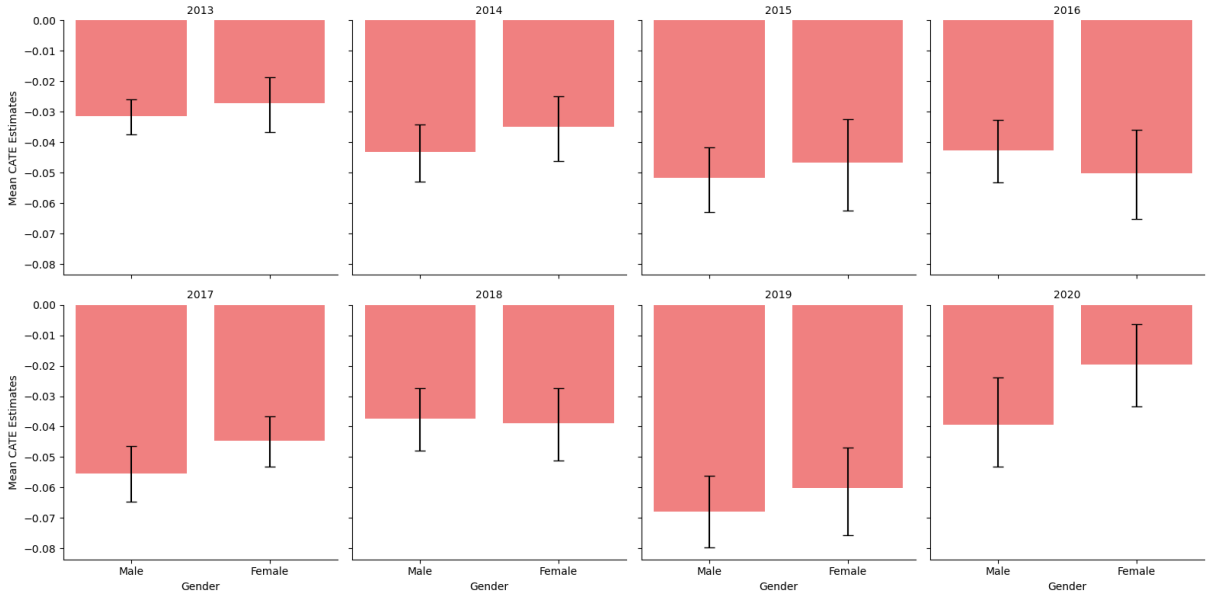


Figure A.7. Distribution of CATE estimates over the years by gender

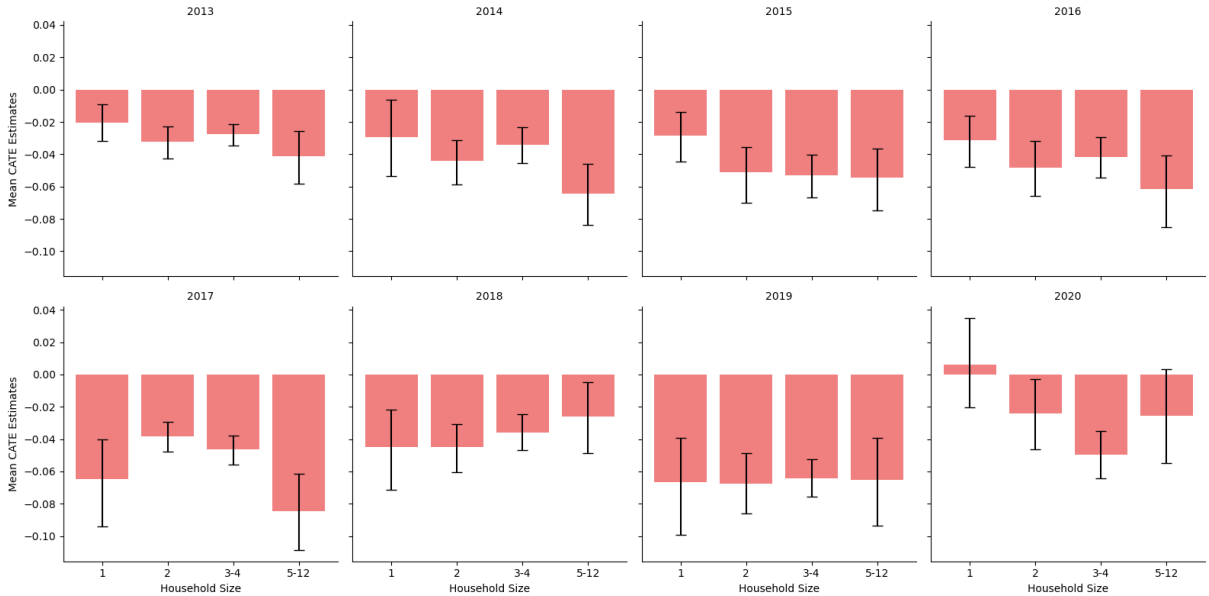


Figure A.8. Distribution of CATE estimates over the years by hosuehold size

Appendix B Additional tables

Tables B.1, B.2, B.3, B.4 and, B.5 corresponnd to the heatmaps presented in Section 5.2, showing the sample sizes and standard errors along with mean CATE estimates for the combined subgroups.

Table B.1. Summary Statistics for the interaction results among subgroups- Education and Age cohorts

Age-groups	18-25			26-35			36-45			46-55			56-66		
Education in years	Mean	Std Error	Sample Size	Mean	Std Error	Sample Size	Mean	Std Error	Sample Size	Mean	Std Error	Sample Size	Mean	Std Error	Sample Size
7-10	-0.0377	0.0203	37	-0.0549	0.0116	41	-0.0661	0.0051	67	-0.0313	0.0109	55	-0.2390	0.0952	17
11-12	-0.0690	0.0162	72	-0.0445	0.0045	110	-0.0477	0.0046	264	-0.0164	0.0081	286	-0.0925	0.0301	102
13-15	-0.0301	0.0042	126	-0.0443	0.0025	401	-0.0347	0.0045	666	-0.0111	0.0062	481	-0.0665	0.0222	130
16-18	-0.0500	0.0105	13	-0.0173	0.0129	125	-0.0153	0.0075	220	-0.0049	0.0122	186	-0.0324	0.0249	69

Table B.2. Summary Statistics for the interaction results among subgroups- Education and Household size

Household size	2			3-4			5-12		
Education	Mean	Std Error	Sample Size	Mean	Std Error	Sample Size	Mean	Std Error	Sample Size
7-10	-0.0493	0.0230	22	-0.0502	0.0234	78	-0.0758	0.0074	117
11-12	-0.0100	0.0059	80	-0.0390	0.0089	403	-0.0572	0.0064	351
13-15	-0.0298	0.0075	207	-0.0280	0.0037	972	-0.0404	0.0057	625
16-18	-0.0325	0.0126	82	-0.0168	0.0090	312	-0.0064	0.0095	219

Table B.3. Summary Statistics for the interaction results among subgroups- Education and gender

Gender	Men			Women		
Education	Mean	Std Error	Sample Size	Mean	Std Error	Sample Size
7-10	-0.0643	0.0119	166	-0.0625	0.0134	51
11-12	-0.0408	0.0057	649	-0.0547	0.0113	185
13-15	-0.0334	0.0046	1090	-0.0311	0.0022	714
16-18	-0.0065	0.0090	384	-0.0298	0.0049	229

Table B.4. Summary Statistics for the interaction results among subgroups- Household size and gender

Gender	Men			Women		
Household size	Mean	Std Error	Sample Size	Mean	Std Error	Sample Size
2	-0.044996	0.010242	280	-0.022168	0.005030	201
3-4	-0.033683	0.006176	1210	-0.041067	0.005430	941
5-12	-0.058201	0.006616	1352	-0.060643	0.004957	305
1	NaN	NaN	0	NaN	NaN	0

Table B.5. Summary Statistics for the interaction results among subgroups- Age grouos and gender

Gender	**Men**			**Women**		
Age Group	**Mean**	**Std Error**	**Sample Size**	**Mean**	**Std Error**	**Sample Size**
18-25	-0.026776	0.009563	137	-0.064200	0.005928	117
26-35	-0.036932	0.004235	523	-0.059237	0.003101	308
36-45	-0.041578	0.006247	1047	-0.049243	0.003138	476
46-55	-0.033838	0.007270	856	-0.008193	0.007083	414
56-66	-0.131045	0.026397	279	-0.068244	0.031013	132

Appendix C Results from Simulation study

Table C.1. Average Bias across different Meta-learners and Base learners

Learner	**Random Forest Regressor**	**Gradient Boosting Regressor**
S-learner	41.5339	26.7253
T-learner	-3.5925	-4.1316
X-learner	-0.7905	0.5306

Table C.2. Average MSE across different Meta-learners and Base learners

Learner	**Random Forest Regressor**	**Gradient Boosting Regressor**
S-learner	1982648.61	1804728.86
T-learner	33215.63	11209.21
X-learner	15536.80	10156.92

Table C.3. Average RMSE across different Meta-learners and Base learners

Learner	**Random Forest Regressor**	**Gradient Boosting Regressor**
S-learner	1408.07	1343.40
T-learner	182.25	105.87
X-learner	124.65	100.78

Table C.4. Average MAE across different Meta-learners and Base learners

Learner	**Random Forest Regressor**	**Gradient Boosting Regressor**
S-learner	1012.27	971.73
T-learner	141.11	78.47
X-learner	99.58	80.34

Table C.5. R-squared across different Meta-learners and Base learners

Learner	Random Forest Regressor	Gradient Boosting Regressor
S-learner	0.0329	0.1197
T-learner	0.1865	0.7253
X-learner	0.6168	0.7493

Table C.6. Explained Variance across different Meta-learners and Base learners

Learner	Random Forest Regressor	Gradient Boosting Regressor
S-learner	0.0338	0.1200
T-learner	0.1951	0.7294
X-learner	0.6184	0.7505

Table C.7. Simulation results with X-learner across RF and GB regressors

Index	Statistic measured	Random Forest Regressor	Gradient Boosting Regressor
Simulation 1	Average Bias	-0.7905	0.5306
	Average MSE	15536.7981	10156.9249
	Average RMSE	124.6467	100.7816
	Average MAE	99.5787	80.3387
	R-squared	0.6168	0.7493
	Explained Variance	0.6184	0.7505
Simulation 2	Average Bias	1.6575	0.1209
	Average MSE	2165399.0497	2075126.0040
	Average RMSE	1471.5295	1440.5298
	Average MAE	1170.1319	1074.0084
	R-squared	0.0468	0.1320
	Explained Variance	0.0512	0.1375
Simulation 3	Average Bias	-0.7905	0.7858
	Average MSE	15536.7981	9975.2680
	Average RMSE	124.6467	99.8763
	Average MAE	99.5787	79.4866
	R-squared	0.6168	0.7538
	Explained Variance	0.6184	0.7551
Simulation 4	Average Bias	0.7900	0.3033
	Average MSE	11822.6325	9718.6998
	Average RMSE	108.7319	98.5835
	Average MAE	86.3927	77.5155
	R-squared	0.7082	0.7601
	Explained Variance	0.7101	0.7618

References

Achim Zeileis, Torsten Hothorn, and Kurt Hornik. 2008. "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics* 17 (2): 492–514. https://doi.org/10.1198/106186008X319331. [8]

Angrist, Joshua D., and William N. Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review* 88 (3): 450–77. Accessed August 31, 2024. http://www.jstor.org/stable/116844. [19]

Autor, David, Melissa Kearney, and Lawrence Katz. 2008. "Trends in U.S. Wage Inequality: Revising the Revisionists." *Review of Economics and Statistics* 90: 300–323. https://doi.org/10.1162/rest.90.2.300. [18]

Becker, Gary S. 1981. *A Treatise on the Family*. Cambridge, MA: Harvard University Press. [1, 3, 19]

Bertrand, Marianne, Emir Kamenica, and Jessica Pan. 2013. "Gender Identity and Relative Income Within Households." *Quarterly Journal of Economics* 130. https://doi.org/10.2139/ssrn.2216750. [2, 19]

Bianchi, Suzanne, Liana Sayer, Melissa Milkie, and John Robinson. 2012. "Housework: Who Did, Does or Will Do It, and How Much Does It Matter?" *Social Forces - SOC FORCES* 91: 55–63. https://doi.org/10.2307/41683183. [4, 35]

Bianchi, Suzanne M., Melissa A. Milkie, Liana C. Sayer, and John P. Robinson. 2000. "Is Anyone Doing the Housework? Trends in the Gender Division of Household Labor*." *Social Forces* 79 (1): 191–228. https://doi.org/10.1093/sf/79.1.191. [4, 17, 18, 39]

Bittman, Michael, Paula England, Liana Sayer, Nancy Folbre, and George Matheson. 2003. "When Does Gender Trump Money? Bargaining and Time in Household Work." *American Journal of Sociology* 109 (1): 186–214. Accessed August 28, 2024. http://www.jstor.org/stable/10.1086/378341. [4]

Blau, Francine D., and Lawrence M. Kahn. 2000. "Gender Differences in Pay." *Journal of Economic Perspectives* 14 (4): 75–99. https://doi.org/10.1257/jep.14.4.75. [18]

Bonke, Jens, Nabanita Datta Gupta, and Nina Smith. 2005. "The timing and flexibility of housework and men and women's wages" [in English]. In *The economics of time use,* edited by Daniel S. Hamermesh and Gerard A. Pfann, 43–77. Contributions to Economic Analysis. Pergamon Press. Bidraget er opstillet i ringbind på Tidsskriftlæsesalen - Forskning / Handelshøjskolen i Århus. [1, 3]

Bonke, Jens, Nabanita Datta Gupta, and Nina Smith. 2010. "Timing and Flexibility of Housework and Men and Women's Wages." *Journal of Family and Economic Issues* 31 (3): 318–39. https://doi.org/10.1007/s10834-010-9201-5. [19]

Breiman, Leo. 2001. *Random forests*. 45: 5–32. 1. Springer. [8]

Bridgman, Benjamin. 2016. "Home productivity." *Journal of Economic Dynamics and Control* 71: 60–76. https://doi.org/https://doi.org/10.1016/j.jedc.2016.08.003. [1]

Bryan, Mark L., and Almudena Sevilla-Sanz. 2010. "Does Housework Lower Wages? Evidence for Britain." *Oxford Economic Papers* 62 (2): 187–210. https://doi.org/10.1093/oep/gpq011. [2, 19]

Budig, Michelle J., and Paula England. 2001. "The Wage Penalty for Motherhood." *American Sociological Review* 66 (2): 204–25. [3]

Card, David. 1999. "Chapter 30 - The Causal Effect of Education on Earnings," edited by Orley C. Ashenfelter and David Card, 3: 1801–63. Handbook of Labor Economics. Elsevier. https://doi.org/https://doi.org/10.1016/S1573-4463(99)03011-4. [18]

Carlson, Daniel L., and Jamie L. Lynch. 2017. "Purchases, Penalties, and Power: The Relationship Between Earnings and Housework." *Journal of Marriage and Family* 79 (1): 199–224. Accessed August 28, 2024. http://www.jstor.org/stable/26646153. [4]

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–94. [8]

Chipman, Hugh A, Edward I George, and Robert E McCulloch. 2010. "BART: Bayesian additive regression trees." *Annals of Applied Statistics,* 266–98. [8]

Coltrane, Scott. 2000. "Research on Household Labor: Modeling and Measuring the Social Embeddedness of Routine Family Work." *Journal of Marriage and Family* 62 (4): 1208–33. https://doi.org/https://doi.org/10.1111/j.1741-3737.2000.01208.x. [1, 36]

Coverman, Shelley. 1983. "Gender, Domestic Labor Time, and Wage Inequality." *American Sociological Review* 48 (5): 623–37. Accessed August 28, 2024. http://www.jstor.org/stable/2094923. [3]

De Lange, Annet, Beatrice van der Heijden, Tinka Van Vuuren, Trude Furunes, Christiane de Lange, and Josje Dikkers. 2021. "Employable as We Age? A Systematic Review of Relationships Between Age Conceptualizations and Employability." *Frontiers in Psychology* 11. https://doi.org/10.3389/fpsyg.2020.605684. [18]

England, Paula. 2005. "Emerging Theories of Care Work." *Annual Review of Sociology* 31: 381–99. [3]

Fendel, Tanja. 2021. "The Effect of Housework on Wages: A Study of Migrants and Native-Born Individuals in Germany." *Journal of Family and Economic Issues* 42: 1–16. https://doi.org/10.1007/s10834-020-09733-5. [2]

Firestone, Juanita, and Beth Anne Shelton. 1988. "An Estimation of the Effects of Women's Work on Available Leisure Time." *Journal of Family Issues* 9 (4): 478–95. https://doi.org/10.1177/019251388009004004. [3]

Folbre, Nancy. 2006. "Measuring Care: Gender, Empowerment, and the Care Economy." *Journal of Human Development* 7 (2): 183–99. https://doi.org/10.1080/14649880600768512. [4, 40]

Foster, Jared. 2013. "Subgroup Identification and Variable Selection from Randomized Clinical Trial Data." [8]

Gimenez-Nadal, Jose Ignacio, and Almudena Sevilla. 2012. "Trends in time allocation: A cross-country analysis." *European Economic Review* 56 (6): 1338–59. https://doi.org/https://doi.org/10.1016/j.euroecorev.2012.02.011. [4]

Gupta, Sanjiv. 2006. "Her Money, Her Time: Women's Earnings and their Housework Hours." *Social Science Research* 35: 975–99. https://doi.org/10.1016/j.ssresearch.2005.07.003. [3]

Hastie, Trevor J, and Robert J Tibshirani. 1986. "Generalized additive models." *Statistical Science* 1 (3): 297–310. [8]

Hersch, Joni. 1991. "The Impact of Nonmarket Work on Market Wages." *American Economic Review* 81 (2): 157–60. Accessed August 29, 2024. http://www.jstor.org/stable/2006845. [3]

Hersch, Joni. 2009. "Home production and wages: Evidence from the American Time Use Survey." *Review of Economics of the Household* 7: 159–78. https://doi.org/10.1007/s11150-009-9051-z. [1, 3]

Hersch, Joni, and Leslie S. Stratton. 1994. "Housework, Wages, and the Division of Housework Time for Employed Spouses." *American Economic Review* 84 (2): 120–25. http://www.jstor.org/stable/2117814. [3, 19]

Hersch, Joni, and Leslie S. Stratton. 1997a. "Housework, Fixed Effects, and Wages of Married Workers." *Journal of Human Resources* 32 (2): 285–307. Accessed August 28, 2024. http://www.jstor.org/stable/146216. [1]

Hersch, Joni, and Leslie S. Stratton. 1997b. "Housework, Fixed Effects, and Wages of Married Workers." *Journal of Human Resources* 32 (2): 285–307. Accessed August 29, 2024. http://www.jstor.org/stable/146216. [3]

Hersch, Joni, and Leslie S. Stratton. 1997c. "Housework, Wages, and the Division of Housework Time for Employed Spouses." *American Economic Review* 87 (2): 87–92. https://www.jstor.org/stable/2950902. [3, 18]

Hersch, Joni, and Leslie S. Stratton. 2002. "Housework and Wages." *Journal of Human Resources* 37 (1): 217–29. Accessed August 29, 2024. http://www.jstor.org/stable/3069609. [3]

Hochschild, A., and A. Machung. 2012. *The Second Shift: Working Families and the Revolution at Home*. Penguin Publishing Group. https://books.google.de/books?id=St_6kWcPJS8C. [3]

Hook, Jennifer L. 2006. "Care in Context: Men's Unpaid Work in 20 Countries, 1965–2003." *American Sociological Review* 71 (4): 639–60. https://doi.org/10.1177/000312240607100406. [4, 36]

Hopfield, John. 1982. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." *Proceedings of the National Academy of Sciences of the United States of America* 79: 2554–8. https://doi.org/10.1073/pnas.79.8.2554. [8]

Hundley, Gregory. 2000. "Male/Female Earnings Differences In Self-Employment: The Effects Of Marriage, Children, And The Household Division Of Labor." *ILR Review* 54. https://doi.org/10.2307/2696034. [3]

Hundley, Gregory. 2001. "Why Women Earn Less than Men in Self-Employment." *Journal of Labor Research* 22: 817–29. https://doi.org/10.1007/s12122-001-1054-3. [3]

Jacob, Daniel. 2021. "CATE meets ML: Conditional average treatment effect and machine learning." IRTG 1792 Discussion Papers 2021-005. Humboldt University of Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series". https://ideas.repec.org/p/zbw/irtgdp/2021005.html. [5–7]

Jacobs, Jerry A., and Kathleen Gerson. 2004. *The Time Divide: Work, Family, and Gender Inequality*. Cambridge, MA: Harvard University Press. [3]

Kan, Man Yee. 2014. "Housework Participation Measurement." In *Encyclopedia of Quality of Life and Well-Being Research,* edited by Alex C. Michalos, 2967–71. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_3988. [1]

Keith, Kristen, and Paula Malone. 2005. "Housework and the Wages of Young, Middle-Aged, and Older Workers." *Contemporary Economic Policy* 23: 224–41. https://doi.org/10.1093/cep/byi017. [2–4, 16, 19]

Killewald, Alexandra. 2011. "Opting Out and Buying Out: Wives Earnings and Housework Time." *Journal of Marriage and Family* 73 (2): 459–71. https://doi.org/https://doi.org/10.1111/j.1741-3737.2010.00818.x. [36]

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. "Metalearners for estimating heterogeneous treatment effects using machine learning." *Proceedings of the National Academy of Sciences* 116 (10): 4156–65. https://doi.org/10.1073/pnas.1804597116. [8–10, 25]

Laan, Mark J van der, Eric C Polley, and Alan E Hubbard. 2007. "Super learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1). [8]

Lachance-Grzela, Mylene, and Genevieve Bouchard. 2010. "Why Do Women Do the Lions Share of Housework? A Decade of Research." *Sex Roles* 63: 767–80. https://doi.org/10.1007/s11199-010-9797-z. [35]

Lahey, Joanna N. 2008. "Age, Women, and Hiring: An Experimental Study." *Journal of Human Resources* 43 (1): 30–56. Accessed August 31, 2024. http://www.jstor.org/stable/40057338. [18]

Leibowitz, Arleen. 1974. "Home Investments in Children." *Journal of Political Economy* 82 (2, Part 2): S111–S131. https://doi.org/10.1086/260295. [19]

McAllister, IAN. 1990. "Gender and the Household Division of Labor: Employment and Earnings Variations in Australia." *Work and Occupations* 17 (1): 79–99. https://doi.org/10.1177/0730888490017001004. [3]

Neumark, David, and Richard Johnson. 1997. "Age Discrimination, Job Separations, and Employment Status of Older Workers: Evidence from Self-Reports." *Journal of Human Resources* 32: 779–811. https://doi.org/10.2307/146428. [18]

Noonan, Mary C. 2001. "The Impact of Domestic Work on Men's and Women's Wages." *Journal of Marriage and Family* 63 (4): 1134–45. https://doi.org/10.1111/j.1741-3737.2001.01134.x. [3]

Oreopoulos, Philip, and Uros Petronijevic. 2013. "Making College Worth It: A Review of the Returns to Higher Education." *Future of Children* 23. https://doi.org/10.2307/23409488. [18]

Phipps, Shelley, Peter Burton, and Lynn Lethbridge. 2001. "In and out of the labour market: long-term income consequences of child-related interruptions to women's paid work." *Canadian Journal of Economics* 34 (2): 411–29. https://ideas.repec.org/a/cje/issued/v34y2001i2p411-429.html. [3]

Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *Journal of the American Statistical Association* 75 (371): 591–93. Accessed August 30, 2024. http://www.jstor.org/stable/2287653. [6]

Schmitz, Hendrik, and Matthias Westphal. 2017. "Informal care and long-term labor market outcomes." *Journal of Health Economics* 56: 1–18. https://doi.org/https://doi.org/10.1016/j.jhealeco.2017.09.002. [23, 24]

Seibold, Heidi, Achim Zeileis, and Torsten Hothorn. 2016. "Model-Based Recursive Partitioning for Subgroup Analyses." *International Journal of Biostatistics* 12. https://doi.org/10.1515/ijb-2015-0032. [8]

Shelton, Beth Anne, and Daphne John. 1996. "The Division of Household Labor." *Annual Review of Sociology* 22 (Volume 22, 1996): 299–322. https://doi.org/https://doi.org/10.1146/annurev.soc.22.1.299. [1]

Shirley, Carla, and Michael Wallace. 2004. "Domestic Work, Family Characteristics, and Earnings: Reexamining Gender and Class Differences." *Sociological Quarterly* 45 (4): 663–90. Accessed August 29, 2024. http://www.jstor.org/stable/4121205. [3]

Sigle-Rushton, Wendy, and Jane Waldfogel. 2007. "Motherhood and women's earnings in Anglo-American, Continental European, and Nordic Countries." *Feminist Economics* 13 (2): 55–91. https://doi.org/10.1080/13545700601184849. [1]

Stratton, Leslie. 2020. "The determinants of housework time." *IZA World of Labor,* no. 133, https://doi.org/10.15185/izawol.133.v2. [1]

Stratton, Leslie S. 2001. "Why Does More Housework Lower Women's Wages? Testing Hypotheses Involving Job Effort and Hours Flexibility." *Social Science Quarterly* 82 (1): 67–76. https://doi.org/10.1111/0038-4941.00007. [3]

Sullivan, Oriel. 2011. "An End to Gender Display Through the Performance of Housework? A Review and Reassessment of the Quantitative Literature Using Insights From the Qualitative Literature." *Journal of Family Theory and Review* 3: 1–13. https://doi.org/10.1111/j.1756-2589.2010.00074.x. [18]

Treas, Judith, and Sonja Drobnič, eds. 2010. *Dividing the Domestic: Men, Women, and Household Work in Cross-National Perspective.* Edited by Judith Treas and Sonja Drobnič. 280. Stanford, CA: Stanford University Press. http://www.sup.org/books/title/?id=16339. [4]

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42. https://doi.org/10.1080/01621459.2017.1319839. [5]

Wagner, Gert G., Joachim R. Frick, and Jürgen Schupp. 2007. "The German Socio-Economic Panel Study (SOEP): Scope, Evolution and Enhancements." SOEPpapers on Multidisciplinary Panel Data Research 1. DIW Berlin, The German Socio-Economic Panel (SOEP). https://ideas.repec.org/p/diw/diwsop/diw_sp1.html. [11]

Selbstständigkeitserklärung

Ich versichere hiermit, dass ich die vorstehende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass die vorgelegte Arbeit noch an keiner anderen Hochschule zur Prüfung vorgelegt wurde und dass sie weder ganz noch in Teilen bereits veröffentlicht wurde. Wörtliche Zitate und Stellen, die anderen Werken dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall kenntlich gemacht.

2. September 2024

Purti Sadhwani