I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13613

Two-Stage Least Squares Random Forests with an Application to Angrist and Evans (1998)

Martin Biewen
Philipp Kugler

AUGUST 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13613

Two-Stage Least Squares Random Forests with an Application to Angrist and Evans (1998)

Martin Biewen
University of Tübingen and IZA

Philipp Kugler
IAW T˙ubingen

AUGUST 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Two-Stage Least Squares Random Forests with an Application to Angrist and Evans (1998)[1]

We develop the case of two-stage least squares estimation (2SLS) in the general framework of Athey et al. (Generalized Random Forests, Annals of Statistics, Vol. 47, 2019) and provide a software implementation for R and C++. We use the method to revisit the classic application of instrumental variables in Angrist and Evans (Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size, American Economic Review, Vol. 88, 1998). The two-stage least squares random forest allows one to investigate local heterogenous effects that cannot be investigated using ordinary 2SLS.

JEL Classification: C26, C55, J22, J13, C14

Keywords: machine learning, generalized random forests, fertility, instrumental variable estimation

Corresponding author:
Martin Biewen
School of Business and Economics
University of Tübingen
Mohlstr. 36
72074 Tübingen
Germany

E-mail: martin.biewen@uni-tuebingen.de

[1] We thank Stefan Wager, Helmut Farbmacher, Michael Lechner, Tony Strittmatter, Nicolaj Mühlbach, Karsten Schweikert as well as participants of the THE Christmas Workshop Hohenheim and the Causal Machine Learning Workshop St. Gallen 2020 for valuable comments and suggestions.

1 Introduction

Random forests (Breimann, 2001) are a successful and increasingly popular method for fitting flexible regression models based on statistical learning. The method consists in successively splitting a given sample into heterogenous subgroups (yielding regression trees), and on repeating this procedure on random variations of the data (leading to random forests). Athey et al. (2019) have generalized the concept of random forests to a general class of estimation methods that solve conditional moment conditions. The applications presented in Athey et al. (2019) include the estimation of conditional average partial effects under exogeneity and conditional instrumental variable estimation based on the classic one-instrument formula (Wald's formula, e.g., Angrist and Pischke, 2008). Unfortunately, this formula does not easily extend to the case with multiple instruments, which is the case often encountered by practitioners.

This paper has two goals. The first one is to develop a conditional instrumental variable estimator based on multiple instruments in the general framework introduced by Athey et al. (2019), and to work out the expressions for estimation, sample splitting and variance estimation needed for implementation in software. This contributes to completing the toolbox of machine learning techniques for classical econometric problems and to verifying the validity of Athey et al. (2019)'s general framework. We also address the problem of tuning an instrumental variables forest which, to our best knowledge, has not been considered in the literature before. Finally, we provide an implementation in R and C++, extending previous codes contributed by Athey et al. (2019). Our second goal is to use this estimator to revisit a classic application of instrumental variables by Angrist and Evans (1998), who used sibling-sex composition instruments in order to investigate the effect of family size on parental labor supply. Including coarse group categories in their 2SLS regressions, they also provided a basic analysis of heterogeneity of these effects across characteristics such as mother's education or husband's earnings. We revisit this question using instrumental variables random forests, which allow one to plot detailed maps of heterogenous effects across multiple dimensions which is not possible using standard regression techniques. This yields deeper insights into the nature of heterogeneity in these effects, going beyond the analysis in Angrist and Evans (1998).

The rest of the paper is structured as follows. Section 2 describes the extension of instrumental variables forests to the case with multiple instruments (two-stage least squares random forests).

Section 3 presents our empirical application. Section 4 concludes.

2 Two-Stage Least Squares (2SLS) random forests

2.1 Generalized random forests

Athey et al. (2019) develop a general framework for building random forests for the estimation of local (i.e. conditional) effects $\theta(x)$ that are the solutions to moment conditions

$$\mathbb{E}\left[\psi_{\theta(x),\nu(x)}(O_i)|X_i = x\right] = 0 \text{ for all } x \in \mathcal{X}, \tag{1}$$

where $\nu(x)$ are nuisance parameters and $O_i, i = 1, \ldots n$ are i.i.d. sample data.

The generalized random forests estimates $\hat{\theta}(x), \hat{\nu}(x)$ are obtained as

$$\left(\hat{\theta}(x), \hat{\nu}(x)\right) \in \arg\min_{\theta,\nu} \left\| \sum_{i=1}^{n} \alpha_i(x)\psi_{\theta,\nu}(O_i) \right\|_2, \tag{2}$$

with

$$\alpha_{bi}(x) = \frac{1(\{X_i \in L_b(x)\})}{|L_b(x)|},\ \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x), \tag{3}$$

where $L_b(x)$ is the set of observations falling into the same leaf as the test point x in tree b. The weights $\alpha_i(x)$ count how often observation X_i was in the same leaf as x across all fitted trees $b = 1, \ldots, B$. They thus determine the relevance of different observations i for fitting $\hat{\theta}(x)$ in the estimating equation (2) (i.e. a local weight).

As described in Athey et al. (2019), the tree-building algorithm proceeds by producing successive splits that maximize heterogeneity. Let $P \subset \mathcal{X}$ be a parent node which is to be split into two children $C_1, C_2 \subset \mathcal{X}$. Athey et al., (2019) show that this can be done by first generating pseudo-outcomes

$$\rho_i = -\xi' A_P^{-1} \psi_{\hat{\theta}_P,\hat{\nu}_P}(O_i), \tag{4}$$

where

$$A_p = \frac{1}{|\{i : X_i \in P\}|} \sum_{i:X_i \in P} \nabla\psi_{\hat{\theta}_P,\hat{\nu}_P}(O_i), \tag{5}$$

and ξ is a vector that picks out the θ coordinate from the (θ, ν) vector. The parameters $\hat{\theta}_P, \hat{\nu}_P$ are the estimators solving the empirical estimation equation in the parent node, i.e.

$$\left(\hat{\theta}_P, \hat{\nu}_P\right) \in \arg\min_{\theta,\nu} \left\| \sum_{i:X_i \in P} \psi_{\theta,\nu}(O_i) \right\|_2 . \tag{6}$$

The splitting is then done on the pseudo-outcomes, i.e. P is split into C_1, C_2 by maximizing

$$\tilde{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{|\{i : X_i \in C_j\}|} \left(\sum_{i:X_i \in C_j} \rho_i \right)^2 . \tag{7}$$

In order to achieve consistency, Athey et al. (2019) use in addition a subsample splitting technique ('honesty'), which divides subsamples of the data in order to grow trees on one part of the data and to solve the estimating equation (2) on another part of the data.

For statistical inference, Athey et al. (2019) show that the variance of $\hat{\theta}(x)$ can be consistently estimated as

$$\hat{\sigma}_n^2(x) = \xi' \hat{V}_n(x)^{-1} \hat{H}_n(x) \hat{V}_n(x)^{-1'} \xi \tag{8}$$

with

$$H_n(x) = var\left(\sum_{i=1}^{n} \alpha_i(x) \psi_{\theta,\nu}(O_i) \right) \tag{9}$$

and $V_n(x)$ a consistent estimator of

$$V(x) = \frac{\partial}{\partial(\theta,\nu)} \mathbb{E}\left(\psi_{\theta,\nu}(O_i) | X_i = x\right) \Big|_{\theta(x),\nu(x)} . \tag{10}$$

2.2 Application to instrumental variables estimation

We now describe the application of this framework to instrumental variables estimation. As one of their applications, Athey et al. (2019) consider the structural model

$$Y_i = \mu(X_i) + \tau(X_i) W_i + \epsilon_i, \tag{11}$$

where Y_i is the outcome, and W_i is a treatment variable that is potentially correlated with ϵ_i. In order to estimate $\tau(X_i)$, they consider the case of a scalar instrumental variable Z_i assumed to be independent of ϵ_i conditional on X_i. This yields the moment condition $\mathbb{E}(Z_i(Y_i - \mu(X_i) - \tau(X_i)W_i)|X_i = x) = 0$, which, along with the moment condition defining the intercept $\mu(X_i)$, defines the function $\psi(\cdot)_{\tau(x),\mu(x)}$ in (2) (Athey et al., 2019, p. 1172).

If there is more than one instrumental variable (i.e. if Z_i is a $M \times 1$ vector), one might be tempted to use the (vector) moment condition $\mathbb{E}(Z_i(Y_i - \mu(X_i) - \tau(X_i)W_i)|X_i = x) = 0$, but this is not possible because the framework described in Athey et al. (2019) is tailored to the just-identified case with as many moment conditions as estimated parameters (this is evident from A_P^{-1} in (4) and $\hat{V}_n(x)^{-1}$ in (8), showing that the function $\psi(\cdot)$ has as many arguments as it has dimensions). In order to arrive at a just-identified representation, we use a local variant of the two-stage least squares estimator (2SLS) to which we apply the technique of stacking moment conditions to form two-step estimation procedures (e.g., Wooldridge, 2010, p. 529).

This yields the representation

$$\psi_{\substack{\tau(x),\mu(x),\\ \gamma_1(x),\gamma_0(x)}}(Y_i, W_i, Z_i) = \begin{pmatrix} \widetilde{W}_i(Y_i - \mu(x) - \tau(x)\widetilde{W}_i) \\ Y_i - \mu(x) - \tau(x)\widetilde{W}_i \\ Z_i\left(W_i - \gamma_0(x) - Z_i'\gamma_1(x)\right) \\ W_i - \gamma_0(x) - Z_i'\gamma_1(x) \end{pmatrix}, \tag{12}$$

where $\widetilde{W}_i$ is the abbreviation of $\gamma_0(x) + Z_i'\gamma_1(x)$. Defining $W_i^c = (1\ W_i)'$, $Z_i^c = (1\ Z_i)'$, $\widehat{W}_i = \hat{\gamma}_0(x) + Z_i'\hat{\gamma}_1(x)$ and $\widehat{W}_i^c = (1\ \widehat{W}_i)'$, the $M + 3$ resulting moment conditions in (2) can be solved analytically yielding

$$\begin{aligned} \begin{pmatrix} \widehat{\gamma}_0(x) \\ \widehat{\gamma}_1(x) \end{pmatrix} &= \left(\sum_{i=1}^{n} \alpha_i(x) Z_i^c Z_i^{c'}\right)^{-1} \left(\sum_{i=1}^{n} \alpha_i(x) Z_i^c W_i\right) \\ \begin{pmatrix} \hat{\mu}(x) \\ \hat{\tau}(x) \end{pmatrix} &= \left(\sum_{i=1}^{n} \alpha_i(x) \widehat{W}_i^c \widehat{W}_i^{c'}\right)^{-1} \left(\sum_{i=1}^{n} \alpha_i(x) \widehat{W}_i^c Y_i\right) \\ &= \left(\left(\sum_{i=1}^{n} \alpha_i(x) Z_i^c W_i^{c'}\right)' \left(\sum_{i=1}^{n} \alpha_i(x) Z_i^c Z_i^{c'}\right)^{-1} \left(\sum_{i=1}^{n} \alpha_i(x) Z_i^c W_i^{c'}\right)\right)^{-1} \\ &\quad \left(\sum_{i=1}^{n} \alpha_i(x) Z_i^c W_i^{c'}\right)' \left(\sum_{i=1}^{n} \alpha_i(x) Z_i^c Z_i^{c'}\right)^{-1} \left(\sum_{i=1}^{n} \alpha_i(x) Z_i^c Y_i\right). \end{aligned} \tag{13}$$

Note that the score function $\psi(\tau, \mu, \gamma_1, \gamma_0)$ as defined in (12) is the negative gradient of a convex function $.5\,\psi(\tau, \mu, \gamma_1, \gamma_0)'\psi(\tau, \mu, \gamma_1, \gamma_0)$ which is required for consistency of the random forest (assumption 6 in Athey et al., 2019). The score function is very well-behaved and satisfies all other regularity assumptions listed in Athey et al. (2019).

The partial derivatives of the score function that are needed to compute the matrix A_P and the pseudo-outcomes ρ_i are given by

$$\nabla\psi_{\substack{\tau(x),\mu(x),\\ \gamma_1(x),\gamma_0(x)}}(Y_i, W_i, Z_i) = \begin{pmatrix} \widetilde{W}_{Pi}\widetilde{W}_{Pi} & \widetilde{W}_{Pi} & -Z_i'\left(Y_i - \mu_P(x) - \tau_P(x)\widetilde{W}_{Pi}\right) + \tau_P(x)Z_i'\widetilde{W}_{Pi} & & & & -\left(Y_i - \mu_P(x) - \tau_P(x)\widetilde{W}_{Pi}\right) + \tau_P(x)\widetilde{W}_{Pi} \\ \widetilde{W}_{Pi} & 1 & \tau_P(x)Z_i' & & & & \tau_P(x) \\ 0 & 0 & Z_{i1}Z_{i1} & Z_{i1}Z_{i2} & ..... & Z_{i1}Z_{iM} & Z_{i1} \\ . & . & Z_{i2}Z_{i1} & Z_{i2}Z_{i2} & ..... & Z_{i2}Z_{iM} & Z_{i2} \\ . & . & ..... & ..... & ..... & ..... & . \\ . & . & Z_{iM}Z_{i1} & Z_{iM}Z_{i2} & ..... & Z_{iM}Z_{iM} & Z_{iM} \\ 0 & 0 & Z_{i1} & Z_{i2} & ..... & Z_{iM} & 1 \end{pmatrix}$$

where $\tau_P(x)$ and $\mu_P(x)$ denote estimates in the parent node and $\widetilde{W}_{Pi} = \gamma_{P0}(x) + Z_i'\gamma_{P1}(x)$ (for simplicity, we have changed the sign of $\psi(\cdot)$ before taking the derivative).

Defining $\gamma(x) = (\gamma_0(x), \gamma_1(x))'$, the matrix $V(x)$ needed for variance estimation is given by

$$V(x) = \begin{pmatrix} \mathbb{E}\left[\widetilde{W}_i\widetilde{W}_i|X=x\right] & \mathbb{E}\left[\widetilde{W}_i|X=x\right] & \mathbb{E}\left[-Z_i'\left(Y_i - \mu(x) - \tau(x)\widetilde{W}_i\right) + \tau(x)Z_i'\widetilde{W}_i|X=x\right] & & & \mathbb{E}\left[-\left(Y_i - \mu(x) - \tau(x)\widetilde{W}_i\right) + \tau(x)\widetilde{W}_i|X=x\right] \\ \mathbb{E}\left[\widetilde{W}_i|X=x\right] & 1 & \mathbb{E}\left[\tau(x)Z_i'|X=x\right] & & & \mathbb{E}\left[\tau(x)|X=x\right] \\ 0 & 0 & \mathbb{E}\left[Z_{i1}Z_{i1}|X=x\right] & ..... & \mathbb{E}\left[Z_{i1}Z_{iM}|X=x\right] & \mathbb{E}\left[Z_{i1}|X=x\right] \\ . & . & \mathbb{E}\left[Z_{i2}Z_{i1}|X=x\right] & ..... & \mathbb{E}\left[Z_{i2}Z_{iM}|X=x\right] & \mathbb{E}\left[Z_{i2}|X=x\right] \\ . & . & ..... & ..... & ..... & . \\ . & . & \mathbb{E}\left[Z_{iM}Z_{i1}|X=x\right] & ..... & \mathbb{E}\left[Z_{iM}Z_{iM}|X=x\right] & \mathbb{E}\left[Z_{iM}|X=x\right] \\ 0 & 0 & \mathbb{E}\left[Z_{i1}|X=x\right] & ..... & \mathbb{E}\left[Z_{iM}|X=x\right] & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \gamma(x)'\mathbb{E}[Z_i^c Z_i^{c'}|X]\gamma(x) & \gamma(x)'\mathbb{E}[Z_i^c|X] & -\mathbb{E}[Y_i Z_i'|X] + \mu(x)\mathbb{E}[Z_i'|X] + 2\tau(x)\gamma(x)'\mathbb{E}[Z_i^c Z_i'|X] & & & -\mathbb{E}[Y_i|X] + \mu(x) + 2\tau(x)\gamma(x)'\mathbb{E}[Z_i^c|X] \\ \gamma(x)'\mathbb{E}[Z_i^{c'}|X] & 1 & \tau(x)\mathbb{E}[Z_i'|X] & & & \tau(x) \\ & & & & & \mathbb{E}\left[Z_{i1}|X=x\right] \\ 0 & 0 & \mathbb{E}\left[Z_{i1}Z_{i1}|X=x\right] & ..... & \mathbb{E}\left[Z_{i1}Z_{iM}|X=x\right] & \mathbb{E}\left[Z_{i2}|X=x\right] \\ . & . & \mathbb{E}\left[Z_{i2}Z_{i1}|X=x\right] & ..... & \mathbb{E}\left[Z_{i2}Z_{iM}|X=x\right] & . \\ . & . & ..... & ..... & ..... & \mathbb{E}\left[Z_{iM}|X=x\right] \\ . & . & \mathbb{E}\left[Z_{iM}Z_{i1}|X=x\right] & ..... & \mathbb{E}\left[Z_{iM}Z_{iM}|X=x\right] & 1 \\ 0 & 0 & \mathbb{E}\left[Z_{i1}|X=x\right] & ..... & \mathbb{E}\left[Z_{iM}|X=x\right] & \end{pmatrix}$$

whose entries we estimate as in Athey et al. (2019) by the honest regression forests produced by the underlying estimation problem.

As described in Athey et al. (2019), the matrix $H_n(x)$ can be estimated by a bootstrap of little bag technique (Sexton and Laake, 2009), for which the overall number of trees $b = 1, \ldots, B$ is partitioned into $g = 1, \ldots G$ bags, where all the $l = B/G$ trees in one bag are estimated on the same half sample. Let tree b be the dth tree in bag g (i.e. tree gd), and denote $\Psi_b = \Psi_{gd} = \sum_{i=1}^n \alpha_{bi}\psi(O_i)$ (i.e. using only data from tree $b = gd$), an estimate $\hat{H}_n(x)$ is then obtained as the solution to

$$
\begin{aligned}
&\frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{l}\sum_{d=1}^{l}\Psi_{gd}\right)\left(\frac{1}{l}\sum_{d=1}^{l}\Psi_{gd}\right)' \\
&= \hat{H}_n(x) + \frac{1}{l-1}\frac{1}{G}\sum_{g=1}^{G}\left[\frac{1}{l}\sum_{d=1}^{l}\left(\Psi_{gd}-\frac{1}{l}\sum_{d=1}^{l}\Psi_{gd}\right)\left(\Psi_{gd}-\frac{1}{l}\sum_{d=1}^{l}\Psi_{gd}\right)'\right].
\end{aligned}
\tag{14}
$$

We follow Athey et al. (2019) who recommend to carry out all of the above computations not for the original outcomes $\{Y_i, W_i, Z_i\}$, but for conditionally centered outcomes $\{Y_i^*, W_i^*, Z_i^*\}$. Let $y(x) = E(Y_i|X = x), w(x) = E(W_i|X = x)$ and $z(x) = E(Z_i|X = x)$, then $Y_i^* = Y_i - \hat{y}^{(-i)}(X_i)$, $W_i^* = W_i - \hat{w}^{(-i)}(X_i)$ and $Z_i^* = Z_i - \hat{z}^{(-i)}(X_i)$, where $\hat{y}^{(-i)}, \hat{w}^{(-i)}$ and $\hat{z}^{(-i)}$ are estimated using separate regression forests not using information based on observation i.

2.3 Tuning

Growing random forests requires the choice of basic tuning parameters such as the minimal node size, the subsample fraction, the number of variables used for splitting and parameters that control the imbalance of splits. The optimal choice of such tuning parameters is an open research topic. A common practical approach is to minimize a suitable loss function based on out-of-bag predictions. If W_i in (11) was exogenous, then a possible loss function would be the so-called R-learner

$$R = \frac{1}{n}\sum_{i=1}^{n}\Big(\left[Y_i - \hat{y}^{(-i)}(X_i)\right] - \hat{\tau}(X_i)\left[W_i - \hat{w}^{(-i)}(X_i)\right]\Big)^2 \tag{15}$$

(Nie and Wager, 2019).

This would lead to a spurious fit, however, because W_i is endogenous. The endogeneity of W_i makes the identification of a suitable loss function more difficult than in the case of unconfoundedness. For parameter tuning, we therefore resort to an idea based on Chernozhukov and Hansen (2008) who argue that the reduced form of an instrumental variables problem is a representation that is always valid and informative about the relationship studied.[2] We, therefore, define our loss function as

$$L = \sum_{i=1}^{n}\Big(\left[Y_i - \hat{y}^{(-i)}(X_i)\right] - \hat{\rho}(X_i)'\left[Z_i - \hat{z}^{(-i)}(X_i)\right]\Big)^2 \tag{16}$$

[2] We thank Stefan Wager for pointing this out to us.

with

$$\hat{\rho}(X_i) = \Big(\sum_{i=1}^{n} \alpha_i(x)(Z_i - \hat{z}^{(-i)}(X_i))(Z_i - \hat{z}^{(-i)}(X_i))'\Big)^{-1} \sum_{i=1}^{n} \alpha_i(x)(Z_i - \hat{z}^{(-i)}(X_i))(Y_i - \hat{y}^{(-i)}(X_i))'. \tag{17}$$

3 Revisiting Angrist and Evans (1998) using two-stage least squares random forests

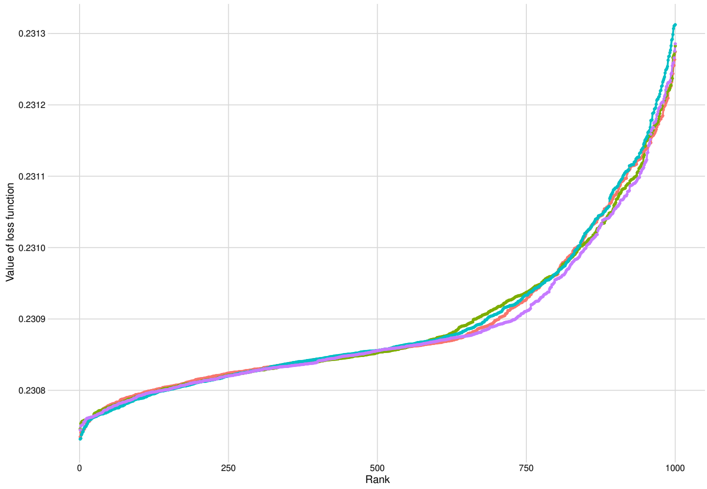
We apply the above estimator to the estimation of the effect of the number of children on the labor supply of married women as described in more detail in Angrist and Evans (1998). Based on over 250.000 observations of married women aged between 21 and 35 years from the 1980 US Census, we follow as closely as possible the specifications in Angrist and Evans (1998), but use the alternative method of two-stage least squares random forests. The main difference between the 2SLS regression models used in Angrist and Evans (1998) and the random forests used here is that the latter allow us to estimate local effects, i.e. the effect of additional children on female labor supply for narrow subgroups of women defined by their observed characteristics.

The variables used to measure female labor supply ($= Y_i$) are either $Worked\ for\ pay$ (indicating whether the woman reported to work for pay in the given year) or $Weeks\ worked$ (representing the number of weeks worked in the given year). The treatment variable ($= W_i$) is whether the woman had more than two children, which is instrumented by the two instrumental variables $Two\ boys$ ($= Z_1$) and $Two\ girls$ ($= Z_2$) indicating whether the first two children were either boys or girls. As argued by Angrist and Evans (1998), these instruments are credibly random but influence the likelihood of having more than two children.[3] Following the construction of the instrument, the estimation sample is restricted to women who had at least two children. The covariates of the analysis ($= X_i$) are dummies for race ($= Black,\ Hispanic,\ Other\ race$), schooling of the woman in years ($= Mother's\ schooling$), her age ($= Age$), age at first birth ($= Age\ at\ first\ birth$), whether the first child was a boy ($= Boy\ 1st$) and father's income ($= Father's\ income$).

[3] Our setup requires that these instruments are valid *conditional on observed covariates* X_i. Given the nature of the instruments, this is apriori plausible. In addition, Farbmacher et al. (2020) have shown that the validity of these instruments cannot be rejected even conditional on observable characteristics.

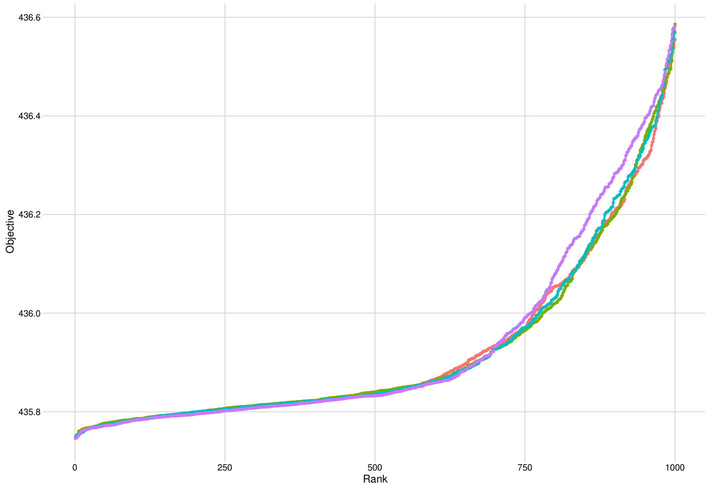
Before we present our empirical results, figures 1 and 2 show more details on our tuning procedure.

Figure 1 – Worked for pay: ordered values of loss function for different candidate tuning parameters



Tuning parameters optimized: minimal node size, subsample fraction, number of splitting variables, split balance parameter α, imbalance penalty. The different colours show the results for four separate Kriging runs (Roustant et al., 2012). Four each Kriging run, 200 random points from the space of tuning parameters are drawn. These are complemented by Kriging interpolations to generate 1000 points of the loss function surface.

Figure 2 – Weeks worked: ordered values of loss function for different candidate parameters



Tuning parameters optimized: minimal node size, subsample fraction, number of splitting variables, split balance parameter α, imbalance penalty. The different colours show the results for four separate Kriging runs (Roustant et al., 2012). Four each Kriging run, 200 random points from the space of tuning parameters are drawn. These are complemented by Kriging interpolations to generate 1000 points of the loss function surface.

Following the implementation in Athey et al. (2019), we use a Kriging procedure (Roustant et al., 2012) to approximate the loss function surface and then choose the tuning parameters that correspond to the minimal value on the approximated loss function surface. For this purpose, we draw 200 random points from the space of tuning parameters and complement them by Kriging interpolations to generate 1000 points of the loss function surface. We carry out this procedure four times in order to minimize the risk of unrepresentative random draws. As in Athey et al. (2019)'s implementation, these computations use forests with a smaller number of trees than our final forests to save computation time. We optimize the following tuning parameters: minimal node size, subsample fraction, number of splitting variables, split balance parameter α, imbalance penalty.[4]

Figures 1 and 2 show that the minimal values and the shape of the loss functions are very similar across the four Kriging runs, making us confident that they are representative examples of the loss function surface. The final minimizing values for the tuning parameters are given in table 1. These were obtained as the smallest minimum out of the four Kriging runs. In general, our random forest results were quite robust to changes of the tuning parameters in a neighborhood of the loss function minimizing values, and only moderately sensitive to larger deviations from them.

Table 1 – Loss function minimizing tuning parameters

	Worked for pay	*Weeks worked*
Minimal node size	1066	601
Subsample fraction	0.101144	.1299106
# Splitting variables	6	4
Balance parameter α	2.831253e-03	.0118385
Imbalance penalty	3.672790e-01	1.0620941
Minimal loss function	0.2307317	435.7457

We now present our 2SLS random forest results. Figures 3 to 14 show the estimated treatment effects of having more than two children on whether the woman reported to work for pay and for the number of weeks worked per year along different covariate dimensions. Covariates not shown in a graph were set to median values. All random forests are based on 100,000 trees.

Figure 3 plots treatment effects of having more than two children on working for pay along the different values of the observed variables father's income and mother's education. The estimates along the dimension of father's income correspond quite well to the ones in Angrist and Evans for fathers income in the bottom third (-.122), the middle third (-.165) and the upper third (-.078)

[4] For the definition of these parameters, see the software implementation of Athey et al. (2019). When splitting a parent node, the size of each child node must not be smaller than α times the size of the parent node. The imbalance parameter penalizes size differences between children of a parent node.

Figure 3 – Treatment effects of more than two children on worked for pay along the dimensions father's income and mother's education (2SLS random forest)

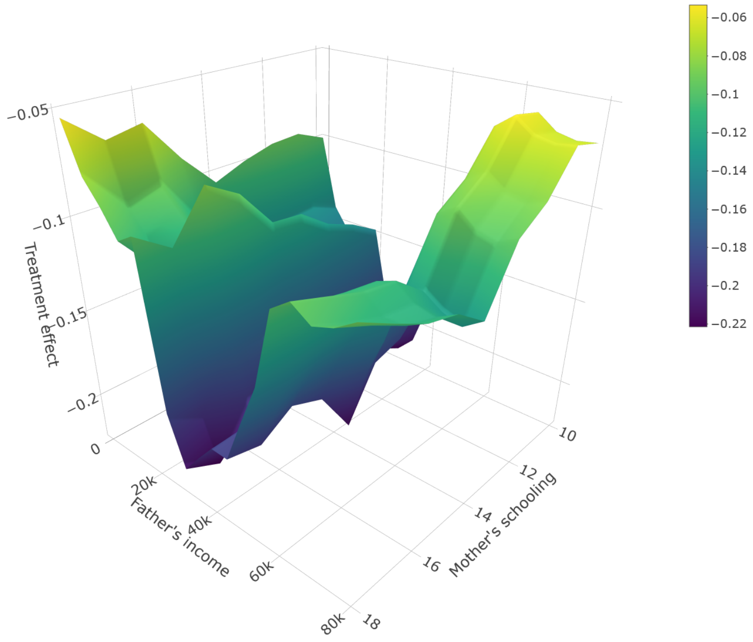
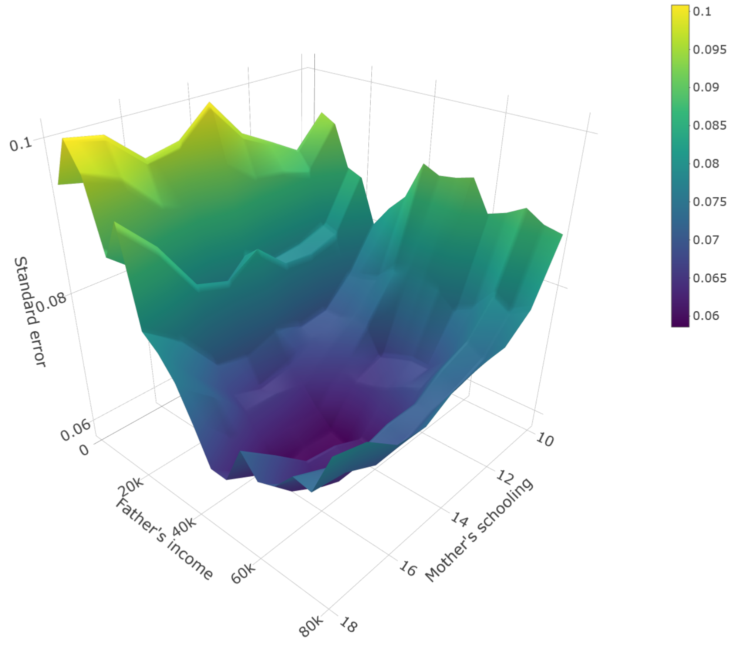


Figure 4 – Standard errors worked for pay along the dimensions father's income vs. mother's education (2SLS random forest)



(Angrist and Evans, 1998, p. 468, table 9, panel A, column 5) if one considers mothers with a high school degree (12 years of education). The estimates for different values of mother's education in Angrist and Evans (1998) (table 9, panel B, less than high school: -.121, high school: -.147, more than high school: -.082) are also similar to the corresponding averaged values in figure 3, but the simple categorization into three groups misses the complex interaction effects uncovered by figure 3: for mother's with low husband's income, there is a strictly positive education gradient (higher education leads to a lower loss in labor supply due to children), while the effects become V-shaped for mother's with higher husband's income. If one looks more specifically at the effects of mother's education for mothers whose husband's income is in the middle third (the median is around 36,000 dollars), then the labor supply effects become even more negative (-.2 or lower). This is also the case in Angrist and Evans, although the estimates there tend to become rather imprecise when smaller subgroups are being considered (Angrist and Evans, 1998, table 9, panel C).

Figure 4 shows that most of the effects in figure 3 are reasonably precisely estimated, with most estimated standard errors ranging between .06 and .1.[5] The plot also nicely reflects the density of observations along the two dimensions (low standard errors in the center and rising standard errors towards areas with few observations).[6]

Figure 5 displays the corresponding estimates for weeks worked per year. Again, the estimates in Angrist and Evans for father's income (table 9, panel A, bottom third: -7.55 weeks, middle third: -7.11 weeks, top third: -3.17 weeks) are quite similar to the ones in the graph for mothers with a high school degree (12 years of education). In the direction of mother's education, the Angrist and Evans' estimates (less than high school: -7.12, high school: -6.42, more than high school: -2.93) correspond well to the estimates shown in the graph for a range of father's income between 40,000 and 60,000 dollars, but the random forest estimates suggest that the effects depend a lot on the exact value of father's income. For median father's income (around 36,000 dollars) and low values of mother's education, labor supply effects again get more negative (-10 weeks per year), both in the graph and in Angrist and Evans (1998, table 9, panel C, last column).

The patterns revealed by figure 5 make much sense: for women with high husband's income, the

[5] These standard errors were computed according to equations (10) and (14) using a bag size of 200.

[6] The axes typically cover around 90 percent of the sample observations in each dimension (each axis approximately ranges from the 5th to the 95th percentile of the corresponding variable).

Figure 5 – Treatment effects of more than two children on weeks worked along the dimensions father's income and mother's education (2SLS random forest)

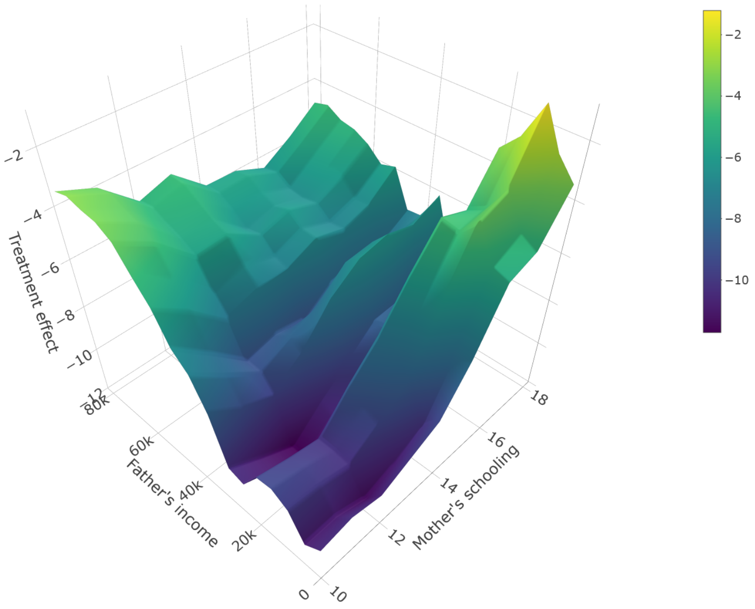
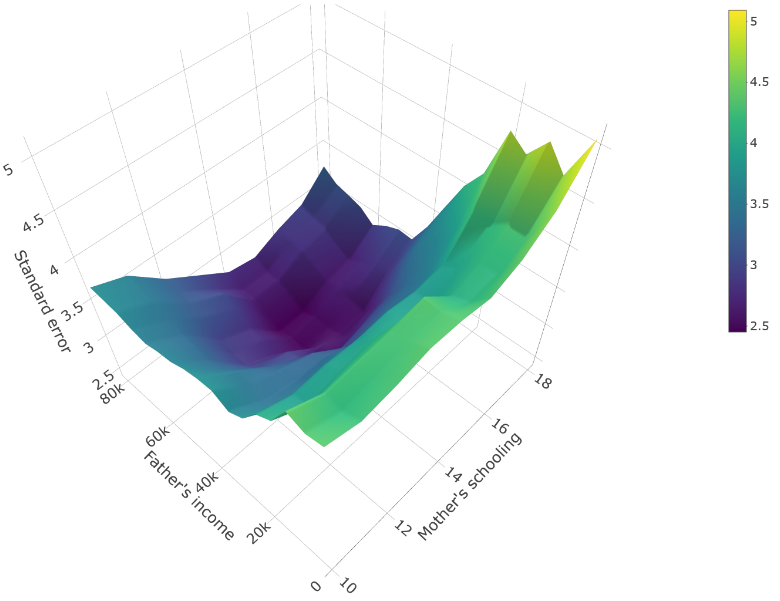


Figure 6 – Standard errors weeks worked along the dimensions father's income vs. mother's education (2SLS random forest)



loss in labor supply is small and not very sensitive to own education, while highly educated women in poorer households face high opportunity costs (in terms of foregone household income) if they do not participate in the labor market. The corresponding standard errors shown in figure 6 suggest that most effects are reasonably precisely estimated, although there are always areas in which this is not the case (areas with sparse density, e.g., few husbands have earnings close to zero).

Summing up our comparison of the instrumental variables random forests with the estimates in Angrist and Evans (1998) based on including basic group categories into 2SLS regressions (e.g. low/middle/high father's income, or low/middle/high education), we find that the general magnitude of the effects as well as basic qualitative patterns generally coincide well, but that the random forest shows in a much more detailed way, and simultaneously in more than one dimension, the exact geometry of effect heterogeneity.

A strength of the random forest methodology is that it models interaction effects in a fully unrestricted and automatic way. Figure 7 presents another example of such interaction effects, plotting the effects of having more than two children on the number of weeks worked per year across the dimensions father's income and mother's age. While the labor supply effect for low to medium values of father's income is U-shaped (young and old mothers do not reduce labor supply as much as middle aged mothers), it more and more transforms into a monotonically falling pattern for higher values of father's income. A possible explanation is that women in low income households face the necessity to return to the labor market once the children have reached a certain age, while those in high income households do not. Figure 9 shows the corresponding graph for the extensive margin (i.e. whether the woman worked for pay). The overall pattern is similar, but the interaction effect is less strong.

Another example for heterogenous effects with interactions is given in figure 11, showing the labor supply effects at the extensive margin along the dimensions of mother's age and mother's education. For older mothers, higher schooling is related to a lower loss in labor supply in the presence of children, while for younger mothers, the gradient is inversely U-shaped. Again, an explanation may be that mothers with high levels of education have a large incentive to return to work once their children have reached a certain age. Figure 13 suggests that this interaction effect does not apply to the number of weeks worked per year, where the effects of higher schooling is the same at all ages (i.e. more schooling leads to a smaller loss in labor supply due to children).

Figure 7 – Treatment effects of more than two children on weeks worked along the dimensions father's income and mother's age (2SLS random forest)

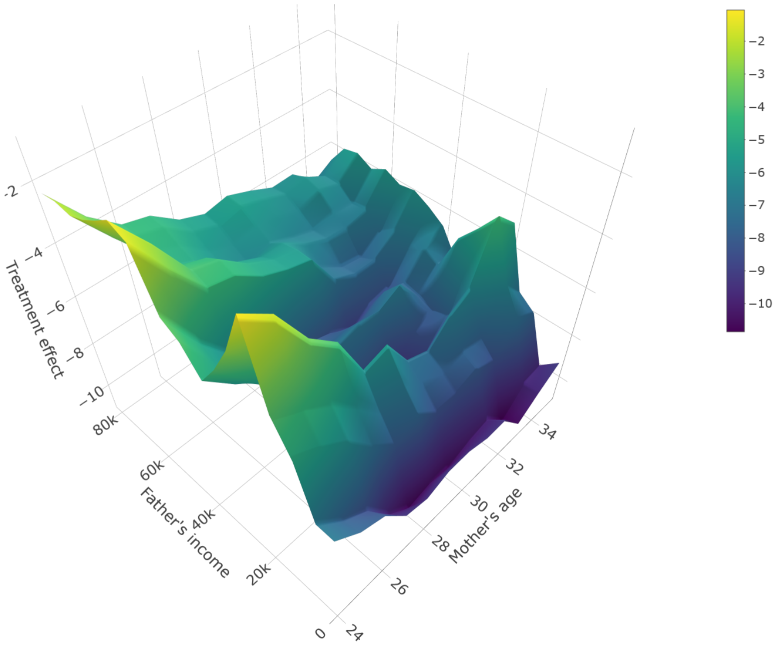


Figure 8 – Standard errors weeks worked along the dimensions father's income vs. mother's age (2SLS random forest)

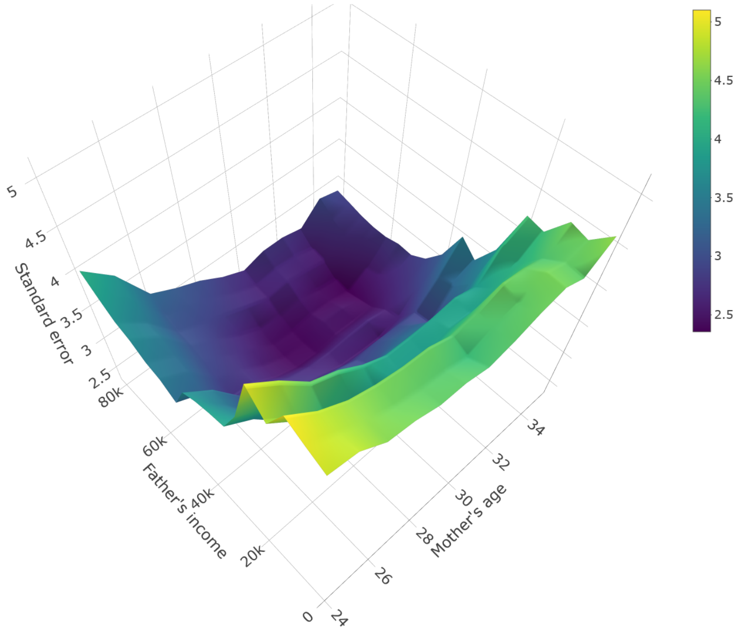


Figure 9 – Treatment effects of more than two children on worked for pay along the dimensions father's income and mother's age (2SLS random forest)

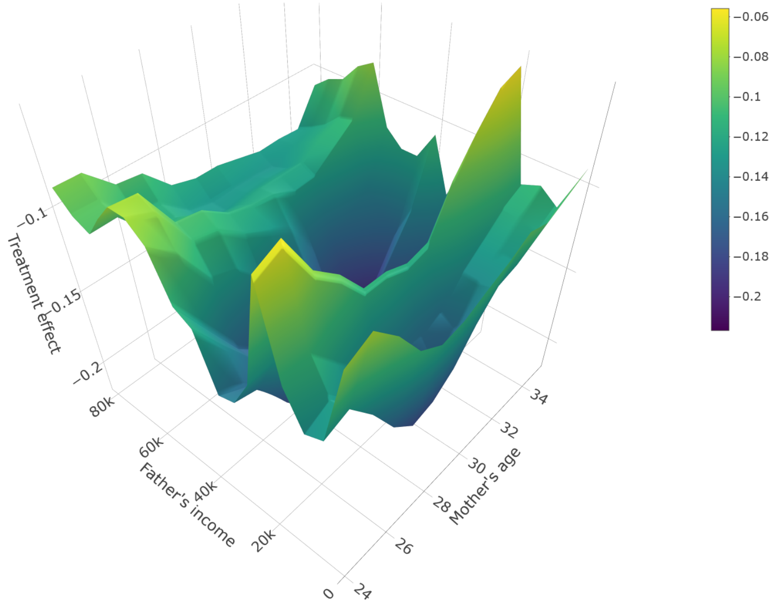


Figure 10 – Standard errors worked for pay along the dimensions father's income vs. mother's age (2SLS random forest)

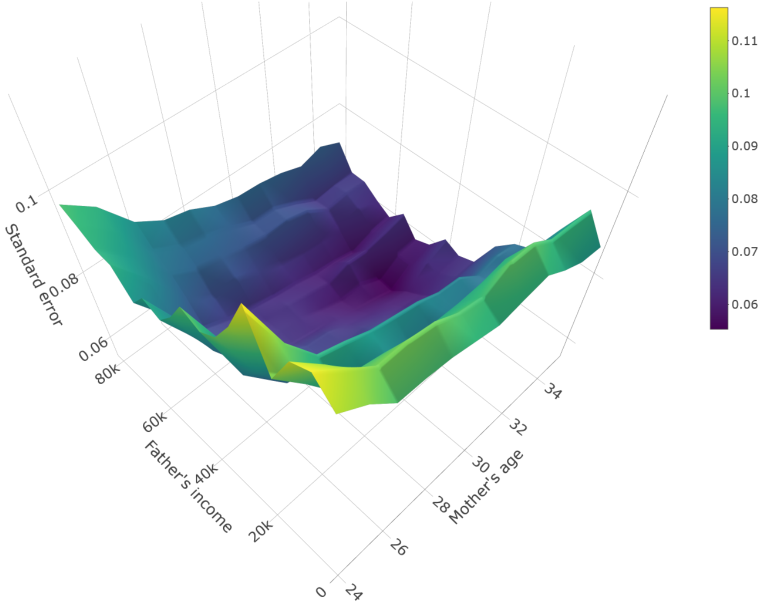


Figure 11 – Treatment effects of more than two children on worked for pay along the dimensions mother's education and mother's age (2SLS random forest)

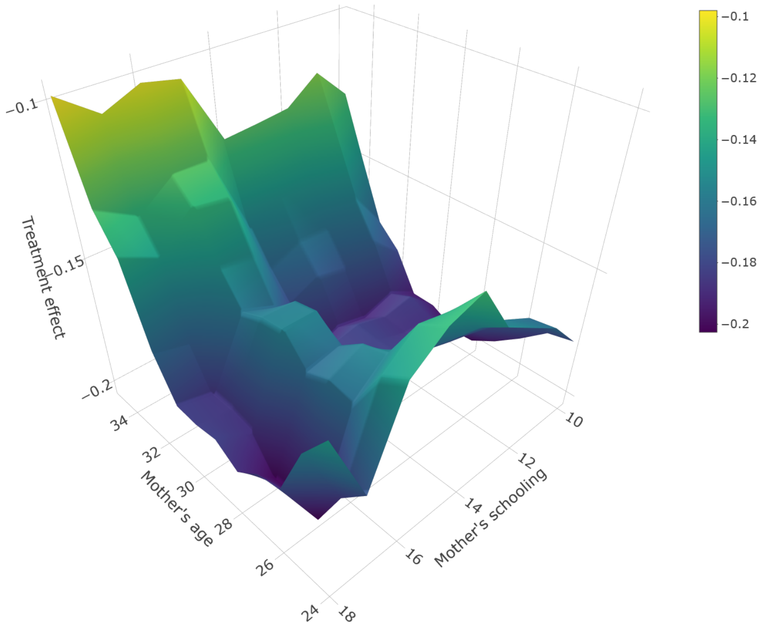


Figure 12 – Standard errors worked for pay along the dimensions mother's education vs. mother's age (2SLS random forest)

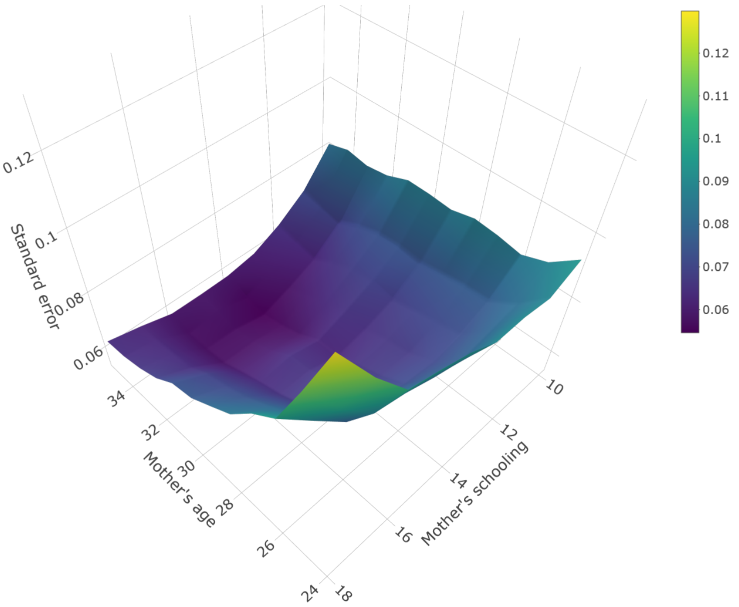


Figure 13 – Treatment effects of more than two children on weeks worked along the dimensions mother's education and mother's age (2SLS random forest)

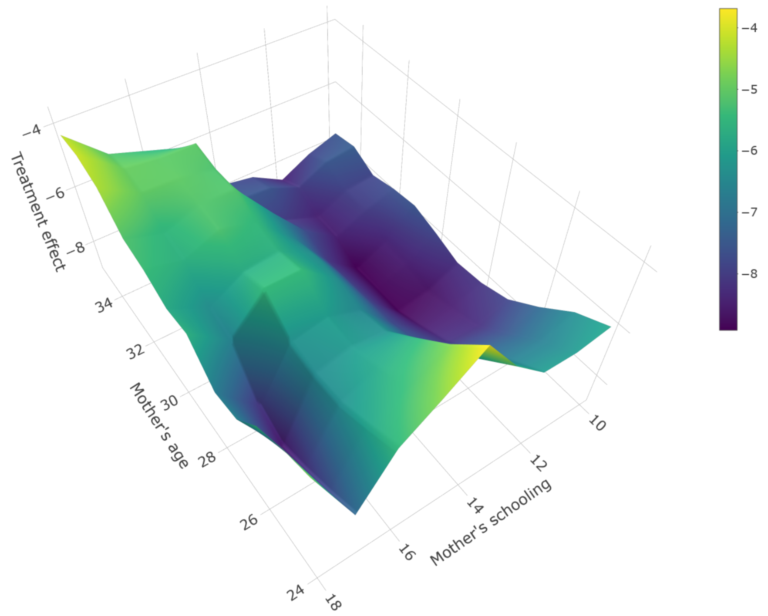
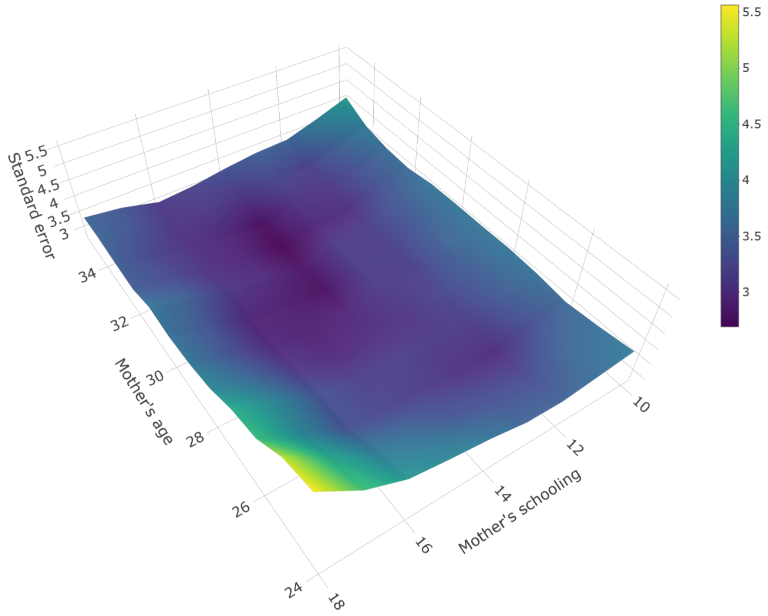
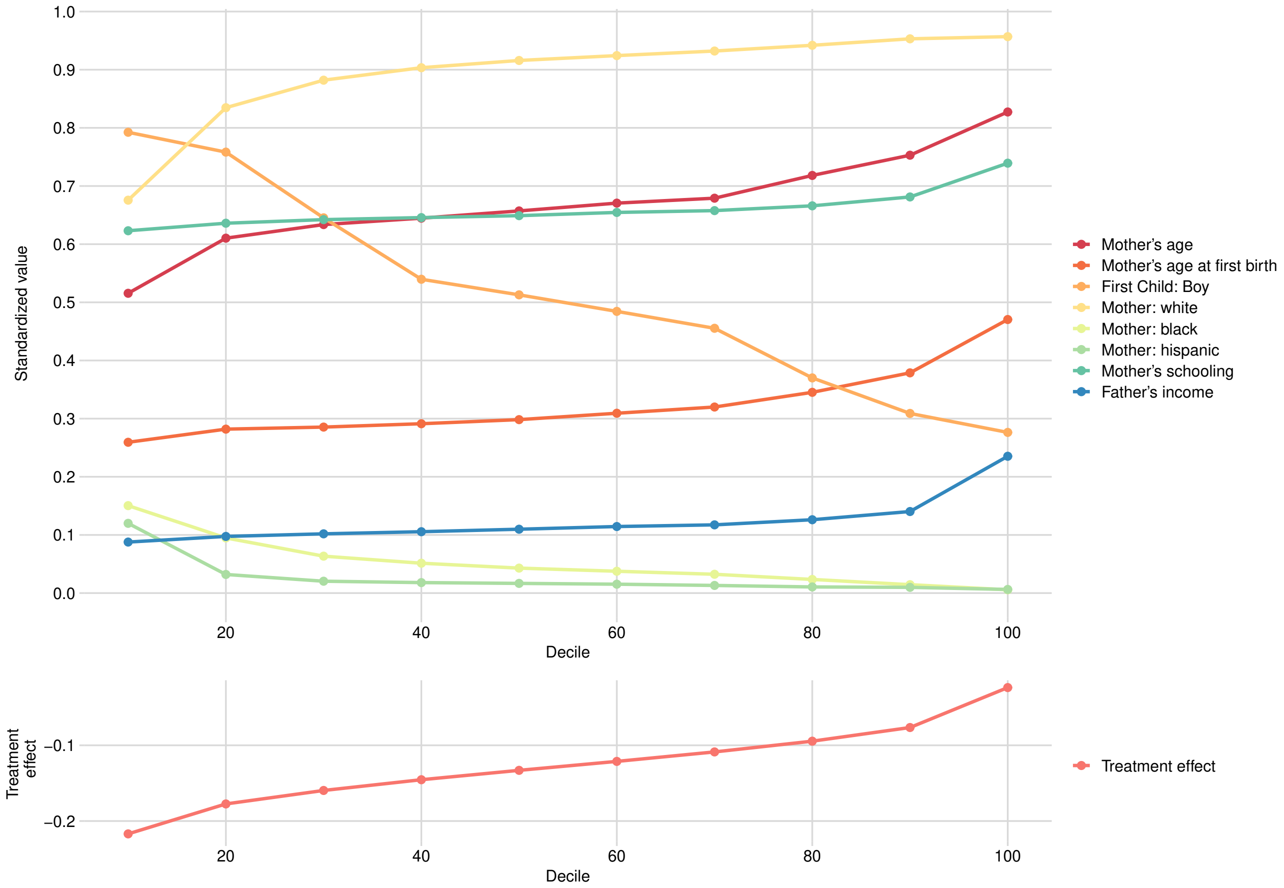


Figure 14 – Standard errors weeks worked along the dimensions mother's education income vs. mother's age (2SLS random forest)



We now present a summary of the effect heterogeneity as detected by our two-stage least squares random forests.[7] Figures 15 and 16 display the mean values of each covariate at different points of the treatment effect distribution. To fit all results on one scale, we standardize the covariates by dividing them by the difference of their maximal and minimal values (for dummy variables, this yields the fraction of cases at a particular point in the distribution).

Figure 15 – Worked for pay: average values of covariates across different points of the treatment effect distribution



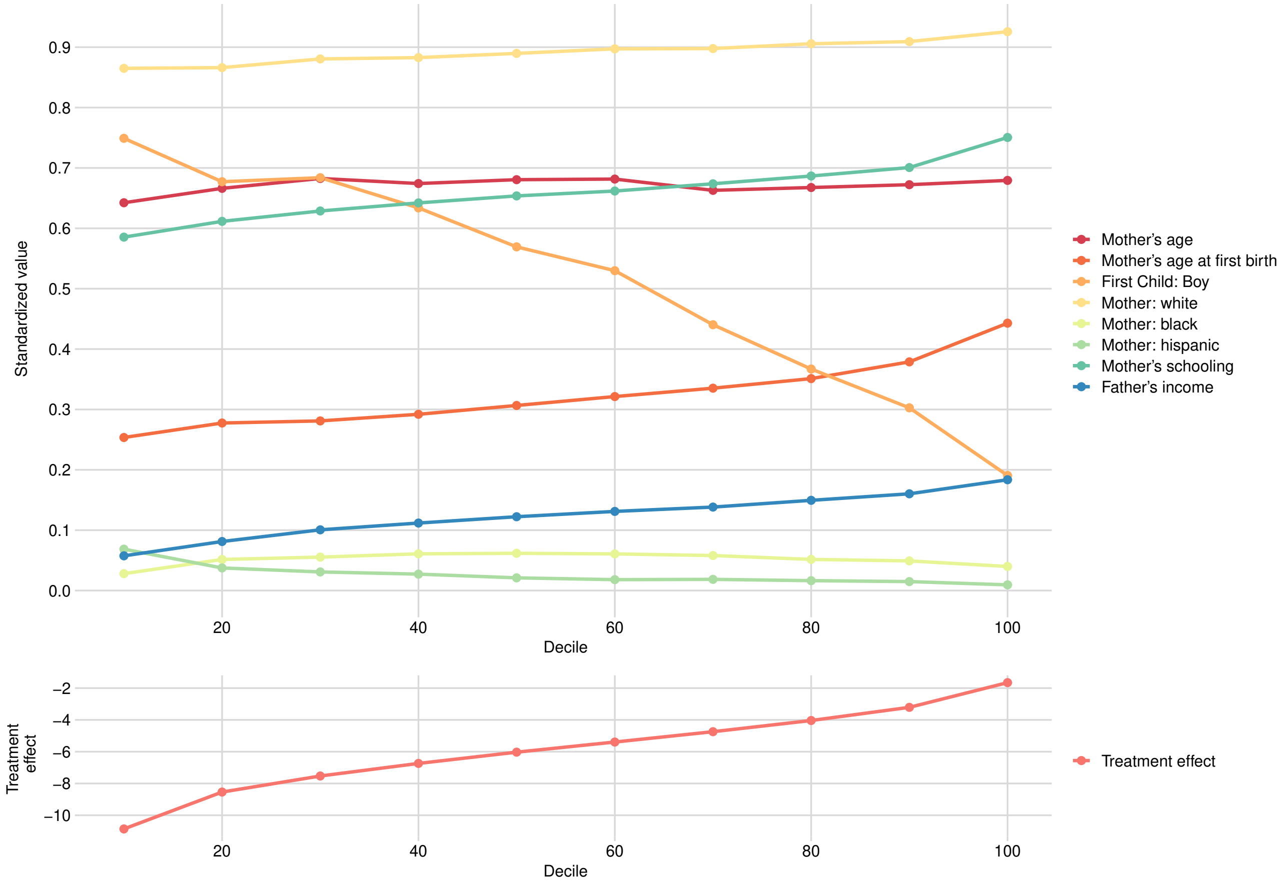
The y-axis shows the mean standardized values of the given covariate for a given decile of the distribution of treatment effects. The standardized values are obtained by dividing the value of each covariate by the difference between the maximal and the minimal value. In the case of dummy variables, this shows the fraction of cases.

The two graphs provide interesting insights into the structure of labor supply effects of children across different covariate dimensions. For example, it turns out that ethnic minorities and young mothers are very much overrepresented among large reductions in participation due to children

[7] To our best knowledge, this way of summarizing effect heterogeneity was first proposed by Athey et al. (2020).

(see lower deciles in figure 15). On the other hand, older women, women with more years of education, women who were older at the time of their first birth as well as women with high husband's earnings are much overrepresented among the cases with relatively mild reductions in labor supply due to children (see upper deciles in figure 15).

Figure 16 – Weeks worked: average values of covariates across different points of the treatment effect distribution



The y-axis shows the mean standardized values of the given covariate for a given decile of the distribution of treatment effects. The standardized values are obtained by dividing the value of each covariate by the difference between the maximal and the minimal value. In the case of dummy variables, this shows the fraction of cases.

The most striking pattern in figure 15 is that of the variable indicating whether the first child was a boy. It turns out that extremely negative labor supply reactions are very tightly related to having a boy as the first child, while relatively mild labor market reactions are tightly related to not having a boy as first child. At first glance, this appears to be consistent with the findings in Ichino et al. (2014) who find that women with first-born boys are less likely to work and work fewer hours in variety of countries. However, they argue that a likely channel for this is reduced marital stability after a first-born girl. This channel does not apply to our application (we only consider married women), suggesting an independent effect of having a first-born boy on later labor supply.[8]

Figure 16 shows the corresponding results for the number of weeks worked per year. The general patterns are very similar to those in figure 15, but the differences are more gradual across percentiles. This is not suprising given that the outcome modeled is a continuous variable (weeks worked). By contrast, the patterns are more nonlinear for the binary case of working vs. not working (mostly horizontal for the central deciles and strongly changing towards the boundaries, see figure 15). Also note that the median effects shown in the lower panels of figures 15 and 16 are close to the estimated two-stage least squares coefficients reported by Angrist and Evans (1998) for the whole sample (-.113 for worked for pay, and -5.15 for weeks worked, Angrist and Evans, 1998, Table 7).

4 Conclusion

This paper develops the case of two-stage least squares random forests (2SLS random forests) based on the general framework of Athey et al. (2019). Our application to Angrist and Evans (1998) demonstrates the usefulness of the method for evaluating effect heterogeneity along multiple dimensions, leading to a richer description of effect differences across observable characteristics and their interactions compared to conventional methods for average effects.

[8] See Ichino et al. (2014) for more discussion. A possibility is sample selection bias, i.e. women still married may differ in unobserved ways from those not married.

5 References

Angrist, J.D., J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

Angrist, J.D., W.N. Evans (1998): Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *American Economic Review*, Vol. 88, pp. 450-477.

Athey, S., J. Tibshirani, S. Wager (2019): Generalized Random Forests. *Annals of Statistics*, Vol. 47, pp. 1148-1178.

Athey, S., R. Friedberg, N. Mühlbach, H. Steimer, S. Wager (2020): Estimating Heterogenous Treatment Effects of an Early Retirement Reform, unpublished manuscript, Stanford University.

Breimann, L. (2001): Random forests, *Machine Learning*, Vol. 45, pp. 123-140.

Chernozhukov, V., C. Hansen (2008): The reduced form: A simple approach to inference with weak instruments, *Economics Letters*, Vol. 100, pp. 68-71.

Farbmacher, H., R. Guber, S. Klaassen (2020): Instrument Validity Tests with Causal Forests, *MEA Discussion Paper 13-2020*, Munich Center for the Economics of Aging.

Ichino, A., E.-A. Lindström, E. Viviano (2014): Hidden consequences of a first-born boy for mothers, *Economics Letters*, Vol. 123, pp. 274-278.

Nie, X., S. Wager (2019): Quasi-Oracle Estimation of Heterogeneous Treatment Effects, unpublished manuscript, Stanford University, https://arxiv.org/abs/1712.04912

Roustant, O., D. Ginsbourger, Y. Deville (2012): DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization, *Journal of Statistical Software*, Vol. 51, pp. 1-55.

Sexton, J., P. Laake (2009): Standard errors for bagged and random forest estimation. *Computational Statistics and Data Analysis*, Vol. 53, pp. 801-811.

Wooldridge, J.M. (2010): *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge/MA.