A boosted multistate model of partnership trajectories in Germany

Angela Carollo[1], Guillermo Briseño Sanchez[2], Valeria Ferraretto[3], and Nicole Hiekel[4]

[1]Max Planck Institute for Demographic Research, Laboratory of Fertility and Well-being
[2]Methods for Big Data, Scientific Computing Center, Karlsruhe Institute of Technology
[3]University of Trento, Department of Sociology and Social Research
[4]Max Planck Institute for Demographic Research, Research Group Gender Inequalities and Fertility

Abstract

Romantic partnerships trajectories over time can be analyzed by means of multistate models. Individuals move between states such as dating, cohabiting, marriage and union dissolution in non-random ways, usually determined by some observable and non-observable characteristics. To disentangle the heterogeneity in these transitions, researchers usually fit regression models, choosing sets of predictors based on theoretical considerations, previous findings, and available data. In order to exploit the full potential of rich survey data, such as the German Family Panel (pairfam), we suggest to combine theoretical considerations with data-driven approaches of variable selection, such as statistical boosting, to identify the best set of predictors for each transition in a multistate model. The result is an interpretable statistical model for each transition, in which the covariates' effect where selected and estimated automatically by the boosting algorithm. In this study, we combine statistical boosting algorithms with multistate models to study partnership trajectories in Germany and to identify the best predictors of transitions between states in a relationship. We discuss advantages and disadvantages of the proposed approach, and we detail future directions of research.

Keywords: Multistate models; Statistical Boosting; Partnership trajectories

1 Introduction

Partnership trajectories involve a series of transitions between states like dating, cohabiting, marrying or separating. Individuals move over time through these several states in a systematic way, and while successfully advancing the partnership increases the level of commitment, some partnerships will eventually end in union dissolution. Understanding the complexity of these trajectories, and what differentiates individuals who break up their relationships from those who successfully advance the partnership are relevant questions in contemporary family demography research. Increasing variation in whether and when to find a partner and decisions about whether, when, and how many children to have, are all relevant drivers of low fertility in many European countries.

An appropriate model to analyse partnership trajectories is a multistate model where individuals in a romantic relationship move across different states of the partnership over time, as measure in

relationship duration or age. To investigate individual heterogeneity in transitions between these states, a suitable set of covariates is usually identified from the relevant literature and the available data. Usually, studies tend to focus on specific classes of predictors to avoid over-fitting and collinearity issues, such as demographic variables, socio-economic factors or variables measuring relationships' quality.

Building a multistate model for such a complex process is not an easy task. On the one hand, the analyst aims to select a model that is simple enough to provide interpretable and general results, on the other hand, by doing so, some important predictors of the transitions of interest may be overlooked. Additional choices need to be made such as the choice of the time scale(s) to be used, for example a clock-forward if the time scale progresses forward with each transition, or a clock-reset approach if the time scale is reset to zero every time a new state is visited. As regards covariates' effect, the analyst can consider linear or non-linear effects of covariates, if any interaction between predictors should be considered, and whether some of the covariates share the same effect on more than one transition, referred to as *cross-transition-type* effects.

Statistical boosting combines the predictive power of machine learning approaches with interpretable statistical modelling techniques. It also provides a robust solution to multicollinearity issues (Mayr and Hofner 2018). Hence, statistical boosting might help identifying which factors determine how people navigate their relationship trajectories, what makes them likely (not likely) to experience a transition and when. At the same time, boosting will provide a more flexible way, than classical regression approaches, to deal with some of these factors being intertwined in complex ways. Here, we apply statistical boosting approach to multistate model to identify the best predictors of transitions between states of a partnership of young individuals in Germany.

2 Method

2.1 A multistate model of partnerships trajectories

In this paper We model partnership trajectories of young women and men living in Germany. This section will introduce our application, and using the example, we formally review the basic concepts of multistate models.

We specify the multistate model in Figure 1.

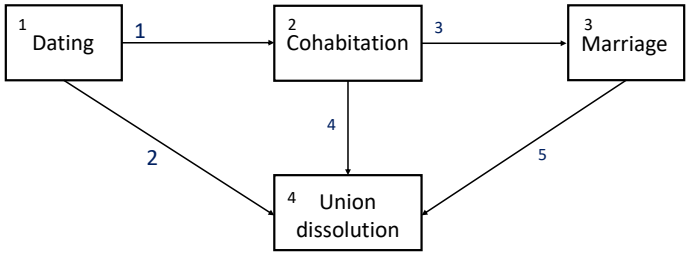

Figure 1: Multistate model of partnership trajectories.

Multistate models describe successive transitions between states over time, and are used to study the effect of predictors on transition rates. A multistate model usually comprises of one initial

state, a set of intermediate (or transient) states, and one or more absorbing states. Individuals can move back and forth between the transient states, but once they reached one of the absorbing states, they leave the risk set (Putter et al. 2007). An example of a multistate model for romantic relationships is presented in Figure 1, where the initial (but also intermediate) state is *Dating*, the other intermediate states are *Cohabitation* and *Marriage*, and the absorbing state is *Union dissolution*. Additionally, a set of possible transitions between states should be specified, and the are usually represented by means of directional arrows connecting the states. In the example presented in this paper, we specify transitions from Dating to Cohabitation, from Cohabitation to Marriage and from each of the intermediate states to the absorbing state of Union Dissolution. The model is non-reversible, as each relationship is only allowed to move forward, and it can end either by entering the state Union Dissolution or as censored observation, at the end of the observation period. Multiple relationships per individual are allowed, and individuals can enter the model in any of the states Dating, Cohabitation and Marriage.

Formally, we consider a multistate model with K states, Q possible transitions, such that $q = k \rightarrow l$ indicates a transition from state k to state l, with $k \neq l$, and where t is the time scale. $Z(t)$ indicates the state occupied at time t. The transition rates give the instantaneous risk of transition from state k to state l at time t:

$$\lambda_{kl}(t) = \lim_{\Delta t \to 0} \frac{P(Z(t + \Delta t) = l | Z(t) = k)}{\Delta t} \tag{1}$$

assuming continuous functions of time. The same transition rate can be indicated as $\lambda_q(t)$, and we will follow this notation hereafter.

Consider now a set of, possibly time-dependent, covariates of interest, indicated with X. A regression model for the transition rates in a multistate model can be formulated as:

$$\lambda_q(t) = \lambda_{0,q}(t) \exp(\eta)\,, \tag{2}$$

where $\lambda_{0,q}(t)$ is the baseline rate and η is a linear predictor, expressed as a (linear) combination of the covariates and their effects. The baseline hazard is estimated non-parametrically and the linear predictor is estimated through minimization of the negative log-stratified-partial likelihood w.r.t. η. Equation (2) indicates a Cox-type proportional hazards model for the transition rates.

We distinguish between *global* covariates, sharing the same effect on all the transitions in the model, and *transition-specific* covariates, having a different effect for each transition. Additionally, some covariates may share the same effect for specific combination of transitions rates, but different effects for other transitions, and we indicate them as *cross-transitions-type* covariates.

Formally, $x_{p.q,i} = x_{p,i} \cdot I_{trans_i = q}$ indicates the transition-specific value of covariate x_p for individual i and transition q. In case only transition-specific covariates are included in the linear predictor η in an additive way, for individual i:

$$\eta_i = \sum_{p=1}^{P} \left(\sum_{q=1}^{Q} f_{x_{p.q}}(x_{p.q,i}) \right)\,, \tag{3}$$

where $f_{x_{p.q}}$ is a function of the transition-specific covariate x_p for transition q. The best fitting model for transition λ_q can be any combination of global, cross-transition-type and transition-specific covariates.

Putter et al. (2007) discuss strategies to select the best fitting model, that is selecting covariates with transition-specific effect, and those with global effect, and to identify which transitions may be

modelled proportionally. The task of selecting such model is feasible when the number of possible predictors, and the number of transitions in the model, are small. However, it quickly becomes computationally unfeasible when both quantities are of moderate to large size. For each predictor, we must choose how it enters the model across all possible transitions: excluding it completely, as a global covariate, or as a transition-specific covariate in a subset of transitions (excluding for simplicity the case of cross-transitions effects).

For example, in a multistate model with $Q = 5$ possible transitions and $P = 10$ potential predictors, excluding cross-transitions effects, and assuming that all the transitions are modelled with a separate baseline hazard, the total number of possible model specifications is 33^{10} (generally calculated as: $(2^Q + 1)^P$). Additionally, some covariates might have a non-linear effect on the hazard rates, and some other might interact with the levels of another predictor.

While theoretical reasoning can guide researchers in the task of selecting a suitable set of predictors, this approach may lead to overlooking important predictors, or to opt for a too simple, or not simple enough, model. Data-driven approaches, can be helpful in model building tasks, and help navigating the complexity of rich data sources, and specification of effects.

In the next section, the statistical boosting approach will be introduced. Additionally, we show how the approach works in the context of continuous time multistate models, and we discuss strategies to select the optimal stopping iteration.

2.2 Statistical boosting

Boosting originated in machine learning and has been extended towards estimating statistical models (Bühlmann and Hothorn 2007), hence the name *Statistical Boosting*. Boosting can be seen as an alternative to the maximum likelihood algorithm to fit a regression model (Mayr and Hofner 2018) by also providing a data-driven variable selection approach.

In a nutshell: the algorithm iteratively minimizes an objective function by conducting small updates to the model. Every small update is comprised of a single covariate effect which is deemed as *best-fitting* within each iteration. As a results of this updating mechanism, some variables will be left-out of the linear predictor, resulting in automatic/data-driven variable selection.

Statistical boosting requires the specification of a *loss function* that measures the differences between the real data and the predicted values from the model. If the negative log-likelihood function is used, then the algorithm mimics maximum likelihood estimation. In boosting terminology, the covariate effects are referred to as *base-learners*. Typically the base-learners are based on a single regressor, specific to the type of covariate effect. These can be linear functions, non-linear effects (for example P-splines) or spatial effects. Additionally, interactions between several variables can also be specified as base-learners.

Technically, boosting minimizes the *empirical risk* $\omega_n = \frac{1}{n}\sum_{i=1}^{n}\omega(\boldsymbol{y}_i;\boldsymbol{\eta}_i)$ iteratively. Here, $\omega(\cdot)$ is a *loss function* of interest. At every iteration, the algorithm fits each of the pre-specified base-learners individually to the negative gradient of the loss function w.r.t. to the linear predictor,

$$-\partial\omega(\boldsymbol{y}_i;\boldsymbol{\eta}_i)/\partial\eta_i \tag{4}$$

. Only the base-learner which leads to highest decrease in the empirical risk is updated. The model is then updated by a small factor, that is, the selected estimated effect is shrunken by a shrinkage factor $\sim 90\%$.

The algorithm is run for a pre-specified number of iterations, denoted by m_{stop}. This acts as the main tuning parameter, similarly to how the λ penalty parameter acts in a LASSO model. By

conducting *early stopping*, i.e. using $m_{stop}^{opt} < m_{stop}$ iterations, some base-learners will effectively be left-out of the model due to not being selected in any of the iterations. This results in data-driven variable selection as well as shrinkage (towards 0) of covariate effects.

The stopping parameters m_{stop} is the only parameters that needs to be tuned in order to prevent over-fitting and achieve a sparse model. Usually, m_{stop} is optimally selected via cross-validation techniques, or out-of-sample evaluation.

Statistical boosting for multistate models was introduced by Reulen and Kneib (2016). They show that the negative log-stratified partial likelihood is a valuable choice as loss function for the gradient boosting algorithm. In a boosted multistate model, cross-transitions-type effects, as well as non-linear effects, are selected automatically in a data-driven way, effectively solving the model-selection issues outlined in the previous section.

We indicate with $x_{p.q,i} = x_{p,i} \cdot I_{trans_i=q}$ the transition-specific value of covariate x_p for individual i and transition q. In case only transition-specific covariates are included in the linear predictor η in an additive way, for individual i:

$$\eta_i = \sum_{p=1}^{P} \left(\sum_{q=1}^{Q} f_{x_{p.q}}(x_{p.q,i}) \right) , \tag{5}$$

where $f_{x_{p.q}}$ is a function of the transition-specific covariate x_p for transition q. The boosting algorithm fits η at the same time as it selects the best $f_{x_{p.q}}$.

Figure 2 schematically illustrates the boosting algorithm that is outlined in the following. A high-dimensional covariate space, comprised of all possible specifications of the base-learners $f_{x_{1,q}}(x_{1,q}), \ldots, f_{x_{p,q}}(x_{p,q})$ enters the boosting algorithm. The iterative updating runs until the optimal stopping iteration m_{stop}^{opt}. Only a subset of base-learners is selected and their estimated effect returned, for example $\hat{f}_{x_{2,1}}(x_{2,1}), \hat{f}_{x_{2,2}}(x_{2,2}), \hat{f}_{x_{6,5}}(x_{6,5})$

3 Data

We use data from waves 1-13 of the German family panel (pairfam) (Brüderl et al. 2020; Huinink et al. 2011), which is a longitudinal survey providing detailed information on romantic relationships of German individuals sampled from four different cohorts. At each wave, respondents and their partner are asked a set of questions concerning their relationship, their values, their satisfaction and fertility plans, as well as updated information on their relationship status and socio-economic status.

Such rich data provides the perfect opportunity to study predictors of transitions between partnership states and investigate heterogeneity in these transitions. In order to do so, we fit a boosted multistate model with 307 base-learners of transitions-specific covariates and let the algorithm select the set of best predictors for each transition.

We select the following covariates: age at union formation, attained age, sex, region of residence, cohort, ethnicity, migration background, religion, emplyoment status, educational attainment, net income, household size, number of years of education, number of kids, number of kids of partner, number of relationships, and family values and views.

Our time scale is time in relationship (time since start dating). We observe a total of 10,577 individuals and 9,675 observed transitions.

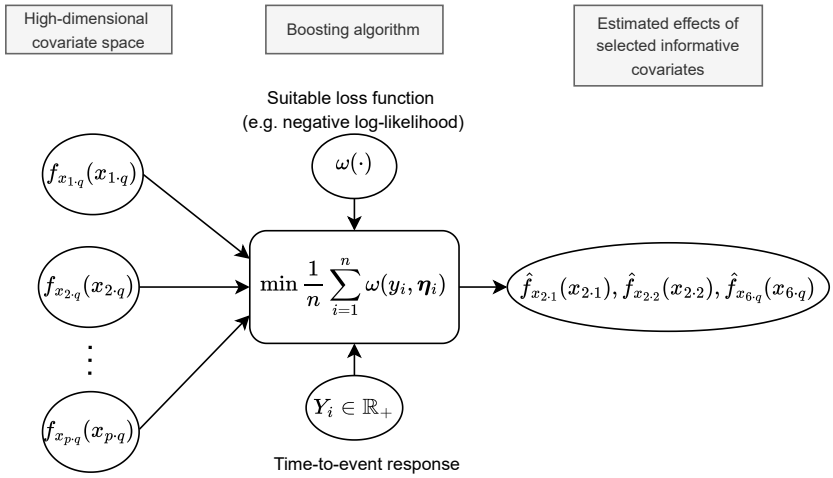

Figure 2: Schematic illustration of the boosting algorithm.

4 Results

In our initial model, we allow each covariate to have transition-specific effects, and so our initial model comprises of 282 base-learners. We stop the algorithm at $m_{stop}^{opt} = 5,920$ iterations, and we observe that 96 transition-specific base-learners have been selected, effectively reducing the number of effects by almost 2/3.

The biggest number of predictors is selected for the transitions dating to union dissolution, dating to cohabitation and cohabitation to marriage. The smallest number of predictors is selected for the transition from dating to marriage directly (that is also the less common one). All selected coefficients are presented in Table 1.

Individuals who are older when starting dating their partner, have lower risks of dissolution both in dating and cohabiting unions. The algorithm also selects current age, independently of age at union, for the transition from dating to union dissolution. Women have higher rates for the transition from dating to cohabitation, and lower rates for the transition from dating to dissolution, compared to men. Individuals living in East Germany have higher rates of transitioning from dating to cohabitation and from marriage to union dissolution, and lower rates of marrying if in a cohabiting union. Both ethnicity and migration background are selected as predictors of some transitions. For example, compared to German individuals, individuals with a Turkish background or another non-German ethnicity have lower rates of transitioning from dating to cohabitation, and higher rates of moving from cohabitation to marriage. Quite interestingly, first generation migrants move directly from dating to marriage with higher rates than individuals with no migration background. This could signal be a signal for trailing partners (it might be easier for a spouse to enter the country

legally after having married the partner who already emigrated to Germany). Second generation migrants have lower transitions from dating to cohabitation, and higher rates of moving out of the cohabitation state in either marriage or dissolution.

Religion affiliation explains some of the transitions. For example, non-religious individuals have higher rates of union dissolution, and lower rates of moving from cohabitation to marriage. Contrarily, individuals who belong to religions other than Catholic or Protestant move to marriage with very high rates either from dating or from cohabiting relationships.

More broadly, employment and education seem to be more predictive of the transitions between partnership states than income, the latter being left-out of the model. Individuals who are employed full-time have higher rates for further institutionalize the relationship and lower rates of dissolution. Unemployed individuals have lower rates of moving from cohabitation to marriage and very high rates of marriage dissolution. Interestingly, both educational level and years of completed education are selected as predictors of the transition from dating to cohabitation, indicating that having at least some level of education (with respect to being in school) is associated to higher rates of moving in with a partner. Having higher education (that is tertiary) is predictive of the transition from cohabitation to marriage, and is associated with lower rates of dissolution for both cohabiting and married couples. The presence of children, of both the anchor and the partner, but mostly partner's children, is predictive of further institutionalization of the relationship and lower dissolution rates.

It is of particular interest to focus on the estimated effects of the *values* variable on each of the transitions. These are presented in Figure 3. Here, transitions are represented on the bottom-axis, the 5 different values variable are represented on the vertical-axis, and the levels of agreement to each of the statements are represented on the top-axis. The color of the cells indicates the sign and strength of the selected effects; red colors represents negative coefficients (lower hazards) while blue colors represents positive coefficients (higher hazards). White cells indicate that the corresponding combination of value-agreement-transition has not been selected.

For example, individuals who disagree with the idea that couples should get married if they live together, have higher rates of union dissolution, while agreeing or completely agreeing with this statement is associated with higher rates of moving from cohabitation to marriage, not surprisingly, and lower rates of moving to union dissolution. Individuals with more traditional views about gender equality, for example disagreeing with the statement that men should participate in housework as much as women, have much higher rate of dissolution of both dating and cohabiting relationship. At the same time, though, completely agreeing that women should be more concerned with family than career is associated with lower rates of union dissolution of dating relationships. Not surprisingly, completely disagreeing with the statement "Marriage is a lifelong union that should not be broken" is associated with very high rates of marriage dissolution, but lower rates of moving from dating to cohabitation and from cohabitation to marriage. At the other end of the spectrum, completely agreeing with this statement is predictive of higher rates of transitioning from cohabitation to marriage and lower rates of dissolution of either dating unions or marriages. Finally, individuals that completely agree that couples should marry at the latest after a child is born have much higher rates of moving directly from dating to marriage, which could be maybe explained by couples who have a child during their dating stage.

In the next steps of our analysis, we will consider different cross-transition-type effects, as well as global effects, non-linear effect of continuous variables, possibly time-varying effects and interactions effects. We will also refine the initial choices of the base-learners and the tuning of the model.

As a trade-off for the intrinsic data-driven variable selection mechanism, statistical boosting lacks "out-of-the-box" uncertainty quantification measures such as standard errors of the estimated

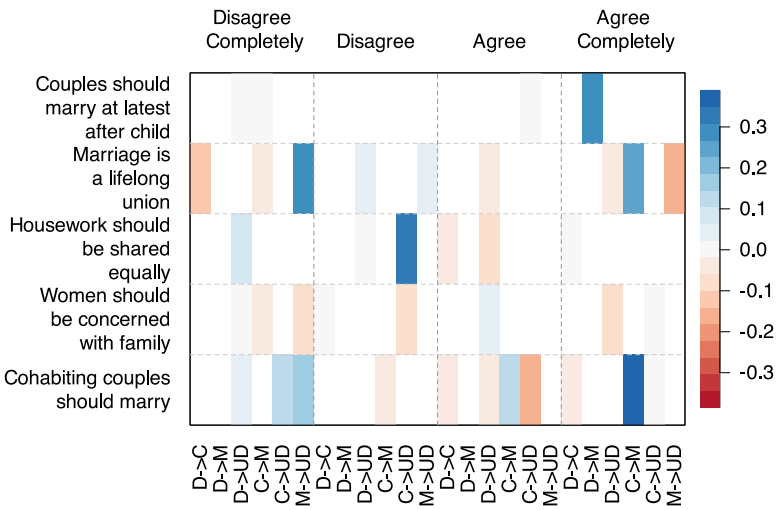

Figure 3: Estimated effects of the *values* covariates.

coefficients. Therefore, we plan to estimate the standard errors via non-parametric bootstrap, which would also correct for the extra heterogeneity introduced by observing multiple relationships per individual.

Data-driven variable selection and regularization of effect estimates resulting from boosting with early stopping are particularly suitable for exploratory data analyses or prediction modeling. Therefore, statistical boosting can serve as a valuable instrument to generate insights about a process, by automatically selecting relevant variables without requiring prior knowledge of their importance. One possibility for future work is to conduct data-driven variable selection via boosting, followed by estimation using classical techniques such as maximum likelihood or Bayesian inference using the subset of selected variables.

In conclusion, these preliminary results show that statistical boosting provides an attractive approach to study heterogeneity in partnerships trajectories over time and it is ultimately resulting in a better understanding of the complex ways in which couples navigate their relationship life course.

References

Brüderl, J., Drobnič, S., Hank, K., Neyer, F. J., Walper, S., Alt, P., Bozoyan, C., Finn, C., Frister, R., Garrett, M., Avilés, T. G., Greischel, H., Gröpler, N., Hajek, K., Herzig, M., Huyer-May, B., Lenke, R., Minkus, L., Peter, T., Reim, J., Schmiedeberg, C., Schütze, P., Schumann, N., Thönnissen, C., Wetzel, M., and Wilhelm, B. (2020). The German Family Panel (pairfam). GESIS Data Archive, Cologne. ZA5678 Data file Version 11.0.0.

Bühlmann, P. and Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, 22(4).

Huinink, J., Brüderl, J., Nauck, B., Walper, S., Castiglioni, L., and Feldhaus, M. (2011). Panel Analysis of Intimate Relationships and Family Dynamics (pairfam): Conceptual framework and design. *Zeitschrift für Familienforschung - Journal of Family Research*, 23:77–101.

Mayr, A. and Hofner, B. (2018). Boosting for statistical modelling-A non-technical introduction. *Statistical Modelling*, 18(3-4):365–384.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

Reulen, H. and Kneib, T. (2016). Boosting multi-state models. *Lifetime Data Analysis*, 22(2):241–262.

Included covariates (base-learners)	Dat.→ Coh.	Dat.→ Mar.	Dat.→ Diss.	Coh.→ Mar.	Coh.→ Diss.	Mar.→ Diss.
Age at union*			-0.005		-0.013	
Age*			-0.005			
Women (ref. Men)	0.008		-0.161			
East Germany (ref. West)	0.131			-0.111		0.113
Cohort (ref. 1991–93)						
1981–83						
1971–73						
2001–2003						
Ethnicity (ref. German)						
Ethnic-German immigrant	0.016		-0.008			
Half-German						
Turkish background	-0.210		0.176	0.156		
Other non-German ethnicity	-0.200			0.067		
Migration background (ref. no migration background)						
First generation		0.479				
Second generation	-0.022			0.048	0.061	
Religion (ref. Catholic)						
Not religious	0.029		0.049	-0.072		
Protestant	0.124					
Other religions		1.733		0.650		
Undetermined			-0.067			
Employment (ref. In Education)						
Full-time employed	0.427		-0.110	0.291	-0.110	
Part-time and other forms of employment	0.068				0.005	
Self-employed						
Out of labor force					0.221	
Unemployed	0.005			-0.025		0.421
Education (ref. Enrolled)						
Low education	0.327		0.058	-0.023		0.233
Medium education	0.232		0.016	0.098		
High education				0.255	-0.295	-0.206
Net income*						
Household size*	-0.065					
Number of years of education*	0.062		-0.020			
Number of kids*			-0.056			-0.066
Number of partner's kids*	0.193		-0.068	0.048	-0.079	
Number of relationships (including current)*						
You should get married if you permanently live with your partner (ref. Neither agree/nor disagree + I don't know)						
Completely disagree			0.025	-0.387	0.125	0.157
Disagree				-0.044		
Agree	-0.020		-0.049	0.131	-0.174	
Completely agree	-0.021			0.358	-0.014	
Women should be more concerned about family than about career (ref. Neither agree/nor disagree + I don't know)						
Completely disagree			0.004		-0.059	
Disagree	-0.009				-0.070	
Agree			0.036			
Completely agree			-0.096		0.020	
Men should participate in housework to the same extent as women (ref. Neither agree/nor disagree + I don't know)						
Completely disagree			0.083			
Disagree			-0.003		0.309	
Agree	-0.051		-0.060			
Completely agree	0.020					
Marriage is a lifelong union that should not be broken (ref. Neither agree/nor disagree + I don't know)						
Completely disagree	-0.112			-0.037		0.269
Disagree			0.023			0.038
Agree			-0.032			
Completely agree			-0.045	0.251		-0.179
Couples should marry at the latest after a child is born (ref. Neither agree/nor disagree + I don't know)						
Completely disagree						
Disagree			-0.018	-0.017		
Agree				-0.008		
Completely agree		0.277				

Table 1: Estimated coefficients for included covariates (base-learners). Fitted model with $m_{stop}^{opt} = 5,920$ and $\nu = 0.2$. $*$ indicates numerical variables. All other variables are categorical.