Author/s:
Jacobi, L;Wagner, H;Fruehwirth-Schnatter, S

Title:
Bayesian Treatment Effects Models with Variable Selection for Panel Outcomes with an Application to Earnings Effects of Maternity Leave

Date:
2016-07

Citation:
Jacobi, L., Wagner, H. & Fruehwirth-Schnatter, S. (2016). Bayesian Treatment Effects Models with Variable Selection for Panel Outcomes with an Application to Earnings Effects of Maternity Leave. Journal of Econometrics, 193 (1), pp.234-250. https://doi.org/10.1016/j.jeconom.2016.01.005.

Persistent Link:
https://hdl.handle.net/11343/122211

Bayesian Treatment Effects Models with Variable Selection for Panel Outcomes with an Application to Earnings Effects of Maternity Leave[1]

Liana Jacobi
Department of Economics, University of Melbourne, Level 4 FBE Building, 111 Barry, 3010 Parkville, Victoria, Australia

Helga Wagner
Department of Applied Statistics, Johannes Kepler University Linz, Altenbergerstr.69, 4040 Linz, Austria

Sylvia Frühwirth-Schnatter[2]
Department of Finance, Accounting and Statistics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

December 22, 2015

JEL classification: C11, C31, C33, C38, C52, J31, J13, J16

keywords: treatment effects models, switching regression model, shared factor model, factor analysis, spike and slab priors, variable selection, Markov Chain Monte Carlo method, earnings effects, maternity leave

Abstract

Child birth leads to a break in a woman's employment history and is considered one reason for the relatively poor labor market outcomes observed for women compared to men. However, the time spent at home after child birth varies significantly across mothers and is likely driven by observed and, more importantly, unobserved factors that also affect labor market outcomes directly. In this paper we propose

[1]We are grateful to J.J. Heckman, Remi Piatek, Siddharta Chib and seminar participants at the University of Melbourne, Monash University, University of Innsbruck, as well as at the ESOBE 2012 meeting, the 5th Rimini Bayesian Workshop, and the Melbourne Bayesian Econometrics Workshop 2013 for helpful comments and suggestions. We gratefully acknowledge funding from the Austrian Science Fund (FWF): S10309-G16 and University of Melbourne.

[2]Corresponding author. email: Sylvia.Fruehwirth-Schnatter@wu.ac.at

two alternative Bayesian treatment effect modeling and inferential frameworks for panel outcomes to estimate dynamic earnings effects of a long maternity leave on mothers' earnings in the years following the return to the labor market. The frameworks differ in their modeling of the endogeneity of the treatment and the panel structure of the earnings, with the first framework based on the modeling tradition of the Roy switching regression model, and the second based on the shared factor approach. We show how stochastic variable selection can be implemented within both frameworks and can be used, for example, to test for the heterogeneity of the treatment effects. Our analysis is based on a large sample of mothers from the Austrian Social Security Register (ASSD) and exploits a recent change in the maternity leave policy to help identify the causal earnings effects. We find substantial negative earnings effects from long leave over a 5 years period after mothers' return to the labor market, with the earnings gap between short and long leave mothers steadily narrowing over time.

1 Introduction

In this paper we introduce two modeling and inferential frameworks within the Bayesian paradigm to estimate the causal effect of an endogeneous binary treatment on panel outcomes and implement Bayesian variable selection. We apply the methods to analyze the dynamic causal earnings effects of a long leave after childbirth for mothers returning to the labour market.

The estimation of treatment effects has become a focus of many econometric papers, in particular the identification and estimation of the effect of an endogenous treatment variable on some outcome of interest. Several approaches have been popular to identify causal treatment effects in such settings, in particular instrumental variable approaches, the LATE estimator and joint modeling approaches (Lee, 2005; Heckman, Ichimura, and Todd, 1998; Heckman and Navarro-Lozano, 2004). Bayesian inferential methods to treatment effect estimation are commonly based on some flexible joint modeling approach, often in the spirit of Roy's switching regression model (Roy, 1951; Lee, 1978) and have addressed a range of issues such as panel outcomes and heterogeneity in treatment across subjects (Koop and Poirier, 1997; Chib and Hamilton, 2000; Munkin and Trivedi, 2003; Chib, 2007; Chib and Jacobi, 2007; Li and Tobias, 2011).

Building on this literature, we consider two modeling frameworks within the Bayesian paradigm to estimate the causal effect of an endogeneous binary treatment on panel outcomes. Both models are formulated within the potential outcome framework following the standard approach in the treatment literature. The first framework is formulated in the tradition of Bayesian treatment effects models in terms of a joint modeling framework for the treatment and the potential outcomes based on the Roy switching regression model (Roy, 1951; Lee, 1978) to capture the endogeneity of the treatment, and does not require the specification of the unidentified joint distribution of the two potential outcome sequences. Within this framework, we discuss two alternatives to model the dependence across the panel outcomes. The second framework employs the more recent factor approach to model the endogeneity of the treatment as well as the panel structure of the earnings following Aakvik, Heckman, and Vytlacil (2005); Carneiro, Hansen, and Heckman (2003); Heckman, Lopes, and Piatek (2014). Both frameworks contain flexible parametric mean formulations of the potential outcomes to

capture possible interactions between observed covariates with the treatment level, allowing for different effects of the treatment across subjects and, importantly, different time dynamics in the two treatment groups.

As an additional innovative and useful feature of these frameworks we introduce Bayesian variable selection in the context of treatment effects models, which has been implemented in many Bayesian papers in the context of "non-treatment" models (for example George and McCulloch (1993, 1997); Geweke (1996); Ley and Steel (2009); Frühwirth-Schnatter and Wagner (2010)). This feature together with a suitable specification of the model will enable us to determine which covariates should be included in the model and to test for the existence of common and level-specific effects of the treatment as well as covariates.

Our empirical analysis provides the first investigation of the dynamics of a range of treatment effects of a mother's yearly earnings after her return to the labour market as a results of her decision regarding the length of leave after childbirth. A number of empirical studies have investigated the effects of maternity leave policies, a key determinant in the length of leave taken, on labour market outcomes and found mixed results as reported in Lalive, Schlosser, Steinhauer, and Zweimüller (2014). However, the amount of time mothers spend at home before returning to the labor market varies considerably, even among mothers covered by the same maternity leave policy. A mother's decision when to return to the labor market depends also on a range of additional factors and is likely driven by observed and, more importantly, unobserved factors that also affect labor market outcomes directly.

We consider a set of model specifications formulated within the two inferential frameworks for treatment effect models with variable selection for panel data introduced in the paper to obtain estimates of short and medium run earnings effects of short versus long leave based on the average and marginal treatment effects, as well as the treatment effect on the treated and untreated. We exploit a recent exogenous change in the parental leave policy in Austria to help identify causal dynamic earnings effects of the endogenously determined length of leave. Our empirical analysis is based on a large sample of mothers from a unique administrative data set with global coverage from the Austrian Social Security Register.

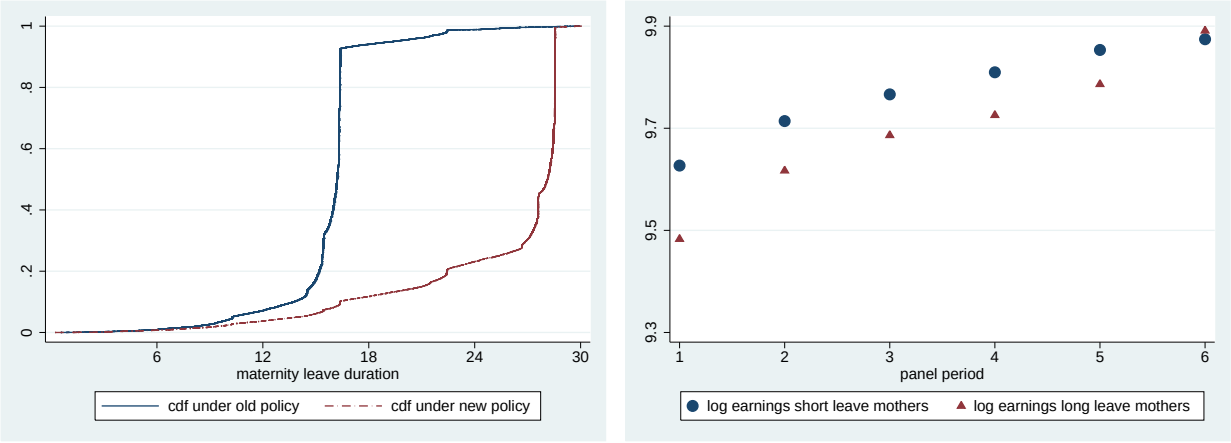
The remainder of the paper is organized as follows. In Section 2 we provide some background about the maternity policy change in Austria and provide a context for the model discussions. Section 3 describes our modeling frameworks and in Section 4 we discuss Bayesian inference including variable selection. Section 5 discusses the definition and estimation of various treatment effects within our modeling framework. Section 6 contains the empirical analysis and Section 7 concludes the paper.

2 Background: Parental Leave Policy in Austria

In Austria, the earliest time mothers can return to work is two months after the birth of the child which is the end of the standard mother protection period. The parental leave policy starts after the end of this period. In Austria, the parental leave policy has two components: job protection and the payment of parental leave benefits. Since July 1990, the job protection and leave benefits periods were extended from previously 12 months since the birth of the child to 24 months. The length of the benefits payment period has undergone several changes more recently. A reduction to 18 months in July 1996 has been followed by an extension of the leave period to up to 30 months, 6 months beyond the job protection period, in July 2000.

For our analysis we consider the 2000 policy change. The extension of the benefits period by

one year and beyond the job protection period in July 2000 induced a substantial proportion of mothers to delay return to work leading to a considerable exogenous variation in time mothers spent at home (Lalive et al., 2014). Panel (a) in Figure 1 shows the empirical cdfs of the duration of leave after child birth by policy regime based on a sample of mothers who gave birth in a two year window before and after the 2000 policy change. The graph is based on



(a) Cdf of Maternity Leave by Policy

(b) Average Log Earnings by Leave

Figure 1: Empirical cdf of the duration of leave after the child birth and average log earnings for mothers with short and long leave by panel period

a sample of mothers taken from the Austrian Social Security Register (ASSD) which contains the complete individual employment histories for the universe of Austrian employees since 1972, including information on number of births and maternity and parental leave spells. The mothers could not predict the policy change as it was made public on August 7, 2001 with an effective date of January 1st 2002. Further, to ensure equal treatment of mothers who were on leave August 7, 2001 and gave birth after July 1st, 2000 they could extend the job protected leave to 2 years and parental leave payments to 30 months. As we can see from panel (a), mothers start to return to work after the end of the mother protection period, within each group the majority of mothers return in the months leading up to the end of the benefit period under the relevant policy scheme. Under the old policy regime a large proportion of mothers return just before month 18, while under the new policy regime most mothers return just before month 30. This strong response of mothers to the length of the benefit period has been observed and studied in Lalive and Zweimüller (2009); Lalive et al. (2014) who also find negative short-term consequences for wages from the new policy. Their model on job search offers an explanation for the higher relative importance via strong effects on the reservation wage of mothers. Interestingly, change of the parental leave duration has no effect on the number of children, but some effect on the timing of subsequent births (Lalive and Zweimüller, 2009). However, since the policy change was only announced in August 2001 but applied to all mothers who gave birth from July 2000, we can define our sample to minimize any effects of mothers who delay the birth of the 2nd or 3rd child under the new more generous policy and consider the fertility decision of a mother as exogenous in our empirical analysis.

For our analysis we therefore consider two groups of mothers based on their leave, those with a maternity leave up to 18 months (short leave) and those with a maternity leave beyond 18

month (long leave). This paper focuses on the identification of the effect of a long versus a short maternity leave, the binary treatment, on the subsequent earnings of mothers following their return to the labor market. As discussed, there are several potential reasons to believe that mothers with a longer maternity leave receive lower earnings at their return to the labor market such as a higher loss of human capital and loss of good job matches. Panel (b) in Figure 1 shows the average log yearly earnings for mothers in both leave groups for six consecutive panel periods (years) following their return to the labor market. The graph suggests that mothers with longer leave start out with substantially lower earnings in their first full year in the labor market than mothers with a short leave, and continue to earn less in the 5 years after their return before the gap closes.

However, we have to be careful with the interpretation of the differences in terms of earnings effects as we do not account for the endogenous choice of the maternity leave state that may affect the earnings across the two groups. While a mother's choice of length of maternity leave is heavily affected by the length of the benefit period under the policy regime in place, their choice also depends on a range of unobserved factors related to later yearly earnings such as availability and attitudes to child care and personal investment in child rearing after their return to the labor market. The latter might have strong affects on hours worked after a mother's return and thus yearly earnings. Further, while the graph compares mothers across the two treatment groups in the same panel period, the data points across the two treatment groups also contain calender year effects.

3 Alternative Treatment Model Specifications for Panel Outcomes

In this section we discuss two alternative Bayesian treatment modeling frameworks to isolate the effect of long maternity leave on panel earnings. Both are phrased within the potential outcome framework commonly used in the treatment literature, allowing for heterogeneous panel treatment effects on earnings, and specify joint regression type models for selection into treatment and the panel outcomes. The differences in the modeling frameworks are with respect to modeling second order moments such as variance and dependence needed to capture the key features of the data. While the first framework is based on the switching regression approach used previously, the second framework employs a factor model approach to modeling of treatment effect data.

In the following we first formally describe the data structure motivated by our application in terms of useful notation, before introducing the mean structure for the selection into treatment and the potential outcome sequences common across the two modeling frameworks. We then specify switching regression and shared factor models as alternative approaches to capture the dependence across the outcomes and the dependence between the treatment and discuss their implications for the observed data models and treatment effects.

3.1 Data Structure and Notation

Consider the following setting based on the empirical example with a sample of $i = 1, ..., n$ mothers that gave birth between July 1998 and June 2002. Following from the previous discussion we define the exogenous binary policy variable z_i, our instrumental variable, for

each mother as

$$z_i = \begin{cases} 0, & \text{if child born before July 2000,} \\ 1, & \text{if child born after June 2000.} \end{cases}$$

As discussed before, mothers with $z_i = 0$ receive maternity leave benefits up to 18 months, while mothers with $z_i = 1$ receive benefits up to 30 months. Job protection ends at 24 months under both policy regimes.

We let the variable m_i denote the number of months a mother spends on maternity leave before returning to the labor market and define the endogeneous binary maternity leave treatment variable x_i for short (0) or long (1) leave as

$$x_i = \begin{cases} 0, & \text{if } m_i \leq 18, \\ 1, & \text{if } m_i > 18. \end{cases}$$

Each mother has a vector $\mathbf{v}_i$ of baseline characteristics at treatment, such as demographic characteristics and earnings before maternity leave, that affect selection into the treatment in addition to the policy regime in place at the time.

For each mother we further observe a vector of labor market outcomes $\mathbf{y}_i = \{y_{i1}, y_{i2}, ..., y_{iT_i}\}$, here measured as log earnings. The time subscript refers to the period (year) since return to the labor market where T_i denotes the number of consecutive panel periods (years) for which we observe the mother i in the labor market after her return. We also observe a matrix of demographic and job related variables that affect earnings, $\mathbf{W}_i = \{\mathbf{w}_{i1}, \mathbf{w}_{i2}, ..., \mathbf{w}_{iT_i}\}$ with some elements varying over time.

For each mother we define the two potential outcome sequences under the two possible treatments as $\mathbf{y}_{0i}$ and $\mathbf{y}_{1i}$, with $\mathbf{y}_{0i} = \{y_{0,i1}, y_{0,i2}, ..., y_{0,iT_i}\}$ and $\mathbf{y}_{1i} = \{y_{1,i1}, y_{1,i2}, ..., y_{1,iT_i}\}$ referring, respectively, to the log potential earnings under short ($x_i = 0$) and under long leave ($x_i = 1$). Depending on the realized treatment x_i, we observe only one of these potential outcome sequences, i.e.:

$$\mathbf{y}_i = \mathbf{y}_{0i}(1 - x_i) + \mathbf{y}_{1i}x_i.$$

Due to this data restriction, the correlation between the potential outcomes and thus individual level treatment effects $(\mathbf{y}_{1i} - \mathbf{y}_{0i}|\mathbf{W}_i)$ cannot be identified from the data without additional identification assumptions that can never be verified by the data. Hence the treatment effect literature focuses on the estimation of the average causal treatment effect (ATE)

$$\text{ATE}(\mathbf{W}) = E[\mathbf{y}_{1i}|\mathbf{W}] - E[\mathbf{y}_{0i}|\mathbf{W}],$$

for a particular matrix $\mathbf{W}$ of demographic and job related variables. Further average treatment effects for different sub-populations, such as the average causal treatment effects on the treated and untreated, can be identified and estimated in this setting as discussed in Section 5.

3.2 Modeling the Mean Structures of the Endogenous Leave Treatment and Potential Earnings Sequences

As noted above, the two modelling approaches differ only with respect to modelling second order moments such as dependencies and variances structures. In both approaches, selection into the endogenous treatment x_i is specified as a standard probit model via a normal latent variable as $x_i = I\{x_i^* > 0\}$ with

$$x_i^* = \mathbf{v}_i'\boldsymbol{\alpha}_1 + z_i\alpha_2 + \eta_i\,, \qquad \eta_i \sim \mathcal{N}\left(0, \sigma_x^2\right), \tag{3.1}$$

where z_i is the instrument based on the policy regime defined above. Subsequently, we will refer to x_i^* as latent utility. For the sake of a simpler notation and clearer description of the model specifications and features we assume a balanced panel with $T_i = T$ yearly observations for each mother in the following discussion.

To capture the heterogeneous effect of long maternity leave on log earnings we specify the basic observation models for the two potential outcome vectors $\mathbf{y}_{0i}$ and $\mathbf{y}_{1i}$ as

$$\mathbf{y}_{0i} = \mathbf{1}_T\mu + \mathbf{W}_i\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_{0i}, \tag{3.2}$$

$$\mathbf{y}_{1i} = \mathbf{1}_T(\mu + \kappa) + \mathbf{W}_i(\boldsymbol{\gamma} + \boldsymbol{\theta}) + \boldsymbol{\varepsilon}_{1i}, \tag{3.3}$$

where $\mathbf{y}_{0i}$ and $\mathbf{y}_{1i}$ refer to the potential log earnings of a mother under a short leave and long leave, respectively, and $\mathbf{1}_T$ refers to a vector of ones. This general formulation assumes that all covariate effects can vary with the treatment. We have a common and heterogeneous effects of the treatment captured by the coefficients κ and $\boldsymbol{\theta}$, respectively. This specification is also useful when we introduce variable selection into the modeling framework to test for the presence of the common and heterogenous treatment effects.

3.3 Modeling the Dependence and Variance Structures

Further assumptions concerning the unobserved errors η_i in the selection equation (3.1) and the composite errors $\boldsymbol{\varepsilon}_{ji}$ in the two potential outcome equations (3.2) and (3.3) are necessary for identification.

A key feature and modeling challenge is the dependence between the treatment and the outcomes. Mothers choose the length of the maternity leave based on a range of considerations, including their job market prospects and unobserved factors related to subsequent earnings, implying that treatment is endogenous. This endogeneity can be captured by specifying a joint distribution for the error terms in the treatment and potential outcome equations $(\boldsymbol{\varepsilon}_{0i}, \boldsymbol{\varepsilon}_{1i}, \eta_i)$ as

$$\begin{pmatrix} \boldsymbol{\varepsilon}_{0i} \\ \boldsymbol{\varepsilon}_{1i} \\ \eta_i \end{pmatrix} \sim \mathcal{N}_{2T+1}\left(\mathbf{0}, \begin{pmatrix} \Omega_0 & \Omega_{01} & \boldsymbol{\omega}_0 \\ \Omega_{01} & \Omega_1 & \boldsymbol{\omega}_1 \\ \boldsymbol{\omega}_0' & \boldsymbol{\omega}_1' & \sigma_x^2 \end{pmatrix}\right). \tag{3.4}$$

Marginally, the composite errors $\boldsymbol{\varepsilon}_{ji}$ follow a normal distribution, $\boldsymbol{\varepsilon}_{ji} \sim \mathcal{N}_T(0, \Omega_j)$, hence the covariance matrices Ω_j capture dependence between subsequent outcomes under treatment j.

As we never observe both sequences of potential outcomes for one individual, the covariance matrix Ω_{01} of the potential outcome sequences cannot be identified directly from the observed data. While it is possible to identify bounds on Ω_{01} based on the positive-definiteness restriction and the remaining components of the covariance matrix of the three error terms in the simple cross-section case as shown in Koop and Poirier (1997), such an undertaking appears to be infeasible for our larger (unbalanced) panel data case.

We will consider two modelling approaches, which differ in the concrete specification of the structure of the joint covariance matrix $\mathsf{Cov}(\boldsymbol{\varepsilon}_{0i}, \boldsymbol{\varepsilon}_{1i}, \eta_i)$. As shown in Chib (2007), identification of Ω_{01} can be avoided by specifying only the $(T+1)$-variate distribution $(\boldsymbol{\varepsilon}_{ji}, \eta_i)$ for each subject for both treatments. This is our first modelling approach which is based on the switching regression model, also referred to as the Roy model. In the second model, the shared factor model, we will assume that all three errors $(\boldsymbol{\varepsilon}_{0i}, \boldsymbol{\varepsilon}_{1i}, \eta_i)$ share a common latent factor to which all correlation between these terms can be attributed.

3.3.1 Switching Regression Models

In the switching regression model (SR) we have to specify only the structure of $\boldsymbol{\omega}_j$ and Ω_j for both treatments. To do so, we consider two variants of the SR model. In a first variant, the SRI model, dependence between subsequent outcomes is captured by introducing an individual- and treatment-specific random intercept $b_{ji} \sim \mathcal{N}(0, D_j)$ and specifying, both for $j = 0, 1$,

$$\boldsymbol{\varepsilon}_{ji} = \mathbf{1}_T b_{ji} + \boldsymbol{\epsilon}_{ji}, \tag{3.5}$$

where b_{ji} is assumed to be independent of the errors $\boldsymbol{\epsilon}_{ji}$, $\boldsymbol{\epsilon}_{ji} \sim \mathcal{N}_T(0, \Sigma_j)$ and Σ_j is a diagonal matrix with elements $\boldsymbol{\sigma}_j^2 = (\sigma_{j,1}^2, \ldots, \sigma_{j,T}^2)$. Marginalizing over the random intercept b_{ji}, the shocks $\boldsymbol{\varepsilon}_{ji}$ in the potential outcome equations have a multivariate normal distribution with covariance matrix Ω_j defined as

$$\mathsf{Cov}(\boldsymbol{\varepsilon}_{ji}) = \Omega_j = \Sigma_j + D_j \mathbf{1}_T \mathbf{1}_T' = \begin{pmatrix} \sigma_{j,1}^2 + D_j & D_j & \ldots & D_j \\ D_j & \sigma_{j,2}^2 + D_j & \ldots & D_j \\ \vdots & \vdots & \ddots & \vdots \\ D_j & D_j & \ldots & \sigma_{j,T}^2 \end{pmatrix}. \tag{3.6}$$

The compound symmetry structure of the covariance matrix implies that the covariance between outcomes at different points in time remains constant.

As this assumption is potentially too restrictive for the data at hand, we consider as a more flexible alternative a latent factor structure of Ω_j. To specify the switching regression model with a latent factor (SRF) we introduce individual- and treatment-specific latent factors $\tilde{b}_{ji} \sim N(0, 1)$ and vectors of time-varying factor loadings $\boldsymbol{\lambda}_j = (\lambda_{j,1}, \ldots, \lambda_{j,T})$, to explain the covariance structure in $\boldsymbol{\varepsilon}_{ji}$ as

$$\boldsymbol{\varepsilon}_{ji} = \boldsymbol{\lambda}_j \tilde{b}_{ji} + \boldsymbol{\epsilon}_{ji}, \tag{3.7}$$

where $\tilde{b}_{ji}$ is assumed to be independent of the idiosyncratic errors $\boldsymbol{\epsilon}_{ji}$. As above, the errors $\boldsymbol{\epsilon}_{ji}$ follow a normal distribution, $\boldsymbol{\epsilon}_{ji} \sim \mathcal{N}_T(0, \Sigma_j)$, with Σ_j being a diagonal matrix with elements $\boldsymbol{\sigma}_j^2 = (\sigma_{j,1}^2, \ldots, \sigma_{j,T}^2)$. Marginalizing over the latent factors yields the following covariance matrix of the potential outcomes,

$$\mathsf{Cov}(\boldsymbol{\varepsilon}_{ji}) = \Omega_j = \Sigma_j + \boldsymbol{\lambda}_j \boldsymbol{\lambda}_j' = \begin{pmatrix} \sigma_{j,1}^2 + \lambda_{j,1}^2 & \lambda_{j,1}\lambda_{j,2} & \ldots & \lambda_{j,1}\lambda_{j,T} \\ \lambda_{j,1}\lambda_{j,2} & \sigma_{j,2}^2 + \lambda_{j,2}^2 & \ldots & \lambda_{j,2}\lambda_{j,T} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{j,1}\lambda_{j,T} & \lambda_{j,2}\lambda_{j,T} & \ldots & \sigma_{j,T}^2 + \lambda_{j,T}^2 \end{pmatrix}, \tag{3.8}$$

which obviously has a more flexible structure than in the SRI model. The compound symmetry structure of the SRI model arises as that special case where the factor loadings are constant, i.e. $\boldsymbol{\lambda}_j = \sqrt{D_j}\mathbf{1}_T$.

As shown by Chib (2007) it is sufficient to specify only the joint distribution of the errors in the selection model and the outcome sequence under treatment j if interest lies in the identification of the average treatment effects. This means that only the two $(T+1)$-variate distributions of the errors $(\boldsymbol{\varepsilon}_{ji}, \eta_i)$ have to be specified for both treatments $j = 0$ and $j = 1$. We follow here Chib and Jacobi (2008) and model the dependence structure of $(\boldsymbol{\varepsilon}_{ji}, \eta_i)$ indirectly,

by specifying both for the SRI model (3.5) as well as the SRF model (3.7), the $(T+1)$-variate joint distribution of η_i and the idiosyncratic errors $\boldsymbol{\epsilon}_{ji}$ in both outcome equations as:

$$\begin{pmatrix} \boldsymbol{\epsilon}_{ji} \\ \eta_i \end{pmatrix} \sim \mathcal{N}_{T+1}\left(\mathbf{0}, \begin{pmatrix} \Sigma_j & \boldsymbol{\omega}_j \\ \boldsymbol{\omega}_j' & 1 \end{pmatrix}\right). \tag{3.9}$$

The restriction $\mathsf{V}(\eta_i) = \sigma_x^2 = 1$ is based on the standard identification argument for probit models. This covariance structure implies that $\boldsymbol{\omega}_j = \Sigma_j^{1/2}\boldsymbol{\rho}_j$, with correlation $\boldsymbol{\rho}_j = \mathsf{Cor}(\boldsymbol{\epsilon}_{ji}, \eta_i)$. No further structure is assumed for the correlations $\boldsymbol{\rho}_j = (\rho_{j,1}, \ldots, \rho_{j,T})$, however restrictions can arise from the assumption that $(\boldsymbol{\epsilon}_{ji}, \eta_i)$ has a proper $(T+1)$-variate normal distribution and hence the covariance matrix in (3.9) has to be positive definite.

In the following we will denote the models as SRI and SRF if we want to emphasize the different covariance structures and as SR otherwise. We would like to emphasize here some general specifics of both models. First, as indicated by the subscript j, the latent variables b_{ji} or $\tilde{b}_{ji}$ are specific to each potential outcome vector. Second, the error $\boldsymbol{\varepsilon}_{ji}$ of the potential outcome model is decomposed into the contribution of the latent variables, which captures only dependence within the potential outcome sequences and the idiosyncratic error $\boldsymbol{\epsilon}_{ji}$ which captures dependence between potential outcomes and latent utility.

3.3.2 Shared Factor Model

An alternative approach to model the panel dependence between outcomes and the treatment is the factor approach (Aakvik et al., 2005; Carneiro et al., 2003; Heckman et al., 2014). We employ this approach to specify a joint, shared factor model (SF) for the panel outcomes and the treatment selection by introducing a latent factor that captures both the panel correlation within the potential outcome sequences, as well as the dependence between the outcomes and selection into treatment.

We specify the errors in equations (3.1) to (3.3) to depend on a shared latent factor $f_i \sim N(0,1)$:

$$\eta_i = \lambda_x f_i + \upsilon_i, \quad \upsilon_i \sim \mathcal{N}(0,1), \tag{3.10}$$

$$\boldsymbol{\varepsilon}_{0i} = \boldsymbol{\lambda}_0 f_i + \boldsymbol{\epsilon}_{0i}, \quad \boldsymbol{\epsilon}_{0i} \sim \mathcal{N}_T(0, \Sigma_0), \tag{3.11}$$

$$\boldsymbol{\varepsilon}_{1i} = \boldsymbol{\lambda}_1 f_i + \boldsymbol{\epsilon}_{1i}, \quad \boldsymbol{\epsilon}_{1i} \sim \mathcal{N}_T(0, \Sigma_1). \tag{3.12}$$

As in Subsection 3.3.1 we assume that the idiosyncratic errors $\boldsymbol{\epsilon}_{ji}$ follow a normal distribution, $\boldsymbol{\epsilon}_{ji} \sim \mathcal{N}_T(0, \Sigma_j)$, with Σ_j being a diagonal matrix with elements $\boldsymbol{\sigma}_j^2 = (\sigma_{j,1}^2, \ldots, \sigma_{j,T}^2)$, and that the factor f_i is independent of the idiosyncratic errors $\boldsymbol{\epsilon}_{ji}$ as well as υ_i.

To identify all parameters in the model, we fix the variance of the latent factor and the variance of the pure error υ_i in the selection model at 1, but do not restrict any of the factor loadings. An implication of this specification is that $\mathsf{V}(\eta_i) = \sigma_x^2 = 1 + \lambda_x^2$.

Marginalizing over the latent factor f_i, the joint distribution of the error terms is given by the $(2T+1)$-variate normal distribution

$$\begin{pmatrix} \boldsymbol{\varepsilon}_{0i} \\ \boldsymbol{\varepsilon}_{1i} \\ \eta_i \end{pmatrix} \sim \mathcal{N}_{2T+1}\left(\mathbf{0}, \begin{pmatrix} \Sigma_0 + \boldsymbol{\lambda}_0\boldsymbol{\lambda}_0' & \boldsymbol{\lambda}_0\boldsymbol{\lambda}_1' & \boldsymbol{\omega}_0 \\ \boldsymbol{\lambda}_1\boldsymbol{\lambda}_0' & \Sigma_1 + \boldsymbol{\lambda}_1\boldsymbol{\lambda}_1' & \boldsymbol{\omega}_1 \\ \boldsymbol{\omega}_0' & \boldsymbol{\omega}_1' & \sigma_x^2 \end{pmatrix}\right), \tag{3.13}$$

where $\boldsymbol{\omega}_j = \lambda_x \boldsymbol{\lambda}_j$. Note that, though the signs of factor f_i and factor loadings $\lambda_x, \boldsymbol{\lambda}_j$ are not fully identified since the likelihood for $\lambda_x f_i$ is the same as the likelihood for $(-\lambda_x)(-f_i)$, and also the likelihoods for $\boldsymbol{\lambda}_j f_i$ and $(-\boldsymbol{\lambda}_j)(-f_i)$ are equal, the signs of all elements in the joint covariance matrix in (3.13) are identified.

The covariance of the errors in each potential outcome equation is given as $\mathsf{Cov}(\boldsymbol{\varepsilon}_{ji}) = \Omega_j = \Sigma_j + \boldsymbol{\lambda}_j \boldsymbol{\lambda}_j'$ which is identical to that of the SRF model given in equation (3.8). Hence, in both models, the factor loadings $\boldsymbol{\lambda}_j$ capture correlation in $\boldsymbol{\varepsilon}_{ji}$, caused e.g. by unobserved characteristics that effect earning outcomes.

Whereas SR models do not make any assumptions about the joint distributions of the potential outcome sequences, an implication of the shared factor structure in the treatment and outcome models is implicit modelling of the dependence between potential outcome sequences, with the covariance given by (3.13) as $\mathsf{Cov}(\mathbf{y}_{0i}, \mathbf{y}_{1i}) = \Omega_{01} = \boldsymbol{\lambda}_0 \boldsymbol{\lambda}_1'$.

Finally, the shared factor model implies that the dependence between the treatment and the outcomes resulting from the endogeneity of the treatment is captured by $\mathsf{Cov}(\mathbf{y}_{ji}, x_i^*) = \boldsymbol{\omega}_j = \lambda_x \boldsymbol{\lambda}_j$. Hence, both for the SR as well as the SF model, the correlation between x_i^* and $y_{j,it}$ is given by

$$\mathsf{Cor}(y_{j,it}, x_i^*) = \frac{\omega_{j,t}}{\sigma_x \sqrt{\Omega_{j,tt}}}, \tag{3.14}$$

where $\sigma_x^2 = 1$ for the SR model. For the SF model, $\omega_{j,t} = \lambda_{j,t} \lambda_x$ is entirely determined by the other model parameters, whereas for SR models $\omega_{j,t} = \sigma_{j,t} \rho_{j,t}$ depends on the correlation parameter $\rho_{j,t}$ which is independent of the other model parameters. Hence, SR models allow for a more flexible structure in the dependence between the treatment and the outcomes and as a result may provide a better fit to the data at hand.

3.4 Observed data models

A specific feature of treatment effects models is their specification in terms of unobserved variables: The potential outcome sequence $\mathbf{y}_{0i}$ is observed only for $x_i = 0$ corresponding to $x_i^* < 0$, whereas for $x_i = 1$ corresponding to $x_i^* > 0$ only $\mathbf{y}_{1i}$ is observed. Moreover the latent utility x_i^* is never observed. We investigate now the implications of the models for the observed data.

To simplify notation, we will use $\mu(\mathbf{y}_{ji})$ to denote the conditional expectation of $\mathbf{y}_{ji}$ for $j = 0, 1$ given the covariates $\mathbf{W}_i$, i.e. $\mu(\mathbf{y}_{0i}) = \mathbf{1}_T \mu + \mathbf{W}_i \boldsymbol{\gamma}$ and $\mu(\mathbf{y}_{1i}) = \mathbf{1}_T(\mu + \kappa) + \mathbf{W}_i(\boldsymbol{\gamma} + \boldsymbol{\theta})$. In addition, we will denote by $\mu(x_i^*) = \mathbf{v}_i' \boldsymbol{\alpha}_1 + z_i \alpha_2$ the conditional expectation of x_i^* given the covariates $(\mathbf{v}_i, z_i)$.

In both modelling approaches the joint distributions of the potential earnings and the latent utility, i.e. $p(\mathbf{y}_{0i}, x_i^*)$ and $p(\mathbf{y}_{1i}, x_i^*)$, given $\mathbf{W}_i, \mathbf{v}_i, z_i$, are $(T+1)$-variate normal distributions, given as

$$(\mathbf{y}_{ji}, x_i^*) \sim \mathcal{N}_{T+1}\left(\begin{pmatrix} \mu(\mathbf{y}_{ji}) \\ \mu(x_i^*) \end{pmatrix}, \begin{pmatrix} \Omega_j & \boldsymbol{\omega}_j \\ \boldsymbol{\omega}_j' & \sigma_x^2 \end{pmatrix} \right), \tag{3.15}$$

where Ω_j is defined, respectively, in (3.6) and (3.8) for the two SR models and below (3.13) for the SF model.

The observed treatment x_i restricts the range of the latent utility to either $I_0 = (-\infty, 0]$ or $I_1 = (0, \infty)$, so that even if the latent utilities were available, $(\mathbf{y}_{ji}, x_i^*)$ were not observed

in the full support $\Re^{T+1}$ but only in the restricted support $\Re^T \times I_0$ and $\Re^T \times I_1$, respectively, for $j = 0$ and $j = 1$. As the distribution of the latent utility x_i^* conditional on the potential outcome $\mathbf{y}_{ji}$ is $x_i^*|\mathbf{y}_{ji} \sim \mathcal{N}\left(\tilde{\mu}_{ji}, \tilde{\sigma}_j^2\right)$ with moments

$$\tilde{\mu}_{ji} = \mu(x_i^*) + \boldsymbol{\omega}_j \Omega_j^{-1}(\mathbf{y}_{ji} - \mu(\mathbf{y}_{ji})), \qquad \tilde{\sigma}_j^2 = \sigma_x^2 - \boldsymbol{\omega}_j' \Omega_j^{-1} \boldsymbol{\omega}_j,$$

the joint distribution of the observed earnings sequence $\mathbf{y}_{x_i,i}$ and the treatment x_i is given by:

$$p(\mathbf{y}_{0i}, x_i = 0) = p(\mathbf{y}_{0i}) \int_{-\infty}^{0} p(x_i^*|\mathbf{y}_{0i})\, dx_i^* \;=\; p(\mathbf{y}_{0i})(1 - \Phi(\tilde{\mu}_{0i}/\tilde{\sigma}_0)),$$
$$p(\mathbf{y}_{1i}, x_i = 1) = p(\mathbf{y}_{1i}) \int_{0}^{\infty} p(x_i^*|\mathbf{y}_{1i})\, dx_i^* \;=\; p(\mathbf{y}_{1i})\Phi(\tilde{\mu}_{1i}/\tilde{\sigma}_1),$$

where $\Phi(z)$ denotes the cdf of the standard normal distribution, and $p(\mathbf{y}_{ji})$ is equal to the marginal distribution of $\mathbf{y}_{ji}$, given by $\mathcal{N}_T(\mu(\mathbf{y}_{ji}), \Omega_j)$. Which of these two joint distributions has generated the data depends on the realized treatment x_i, so that the data $(\mathbf{y}_i, x_i)$ observed for subject i come from

$$p(\mathbf{y}_i, x_i) = p(\mathbf{y}_{0i}, x_i = 0)\, I(x_i = 0) \;+\; p(\mathbf{y}_{1i}, x_i = 1)\, I(x_i = 1). \tag{3.16}$$

Based on the marginal distribution $p(x_i)$, derived from the probit specification (3.1), the conditional distribution of the observed outcome sequence $\mathbf{y}_i$ given treatment x_i follows as

$$p(\mathbf{y}_i|x_i) = \frac{p(\mathbf{y}_i, x_i)}{p(x_i)} = \begin{cases} p(\mathbf{y}_{0i}) \frac{1 - \Phi(\tilde{\mu}_{0i}/\tilde{\sigma}_0)}{1 - \Phi(\mu(x_i^*)/\sigma_x)}, & x_i = 0, \\ p(\mathbf{y}_{1i}) \frac{\Phi(\tilde{\mu}_{1i}/\tilde{\sigma}_1)}{\Phi(\mu(x_i^*)/\sigma_x)}, & x_i = 1, \end{cases} \tag{3.17}$$

which obviously is not a multivariate normal distribution as conditional on x_i, $\tilde{\mu}_{x_i,i}$ is a function of the observed outcome $\mathbf{y}_i$.

4 Bayesian Inference

For all models introduced in Section 3, a fully Bayesian inference is applied to estimate the unknown model parameters. We discuss the choice of prior distributions for all model parameters and specify a set of standard priors in Subsection 4.1, which is extended in Subsection 4.2 to a more flexible set of prior distributions to implement variable selection in the context of both modeling frameworks. Subsection 4.3 discusses posterior inference by means of Markov chain Monte Carlo (MCMC) methods under both sets of prior specifications given panel data as in our specific application.

To simplify the subsequent discussion, we introduce a more compact notation and write the structural mean of the selection equation (3.1) as $\mu(x_i^*) = \mathbf{Z}_i \boldsymbol{\alpha}$, where $\mathbf{Z}_i = (\mathbf{v}_i, z_i)$ denotes the $1 \times d_\alpha$ covariate vector for subject i in the selection model and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \alpha_2)$ is the corresponding vector of d_α regression coefficients. The structural mean of outcome under treatment j (equations (3.2) and (3.3)) is denoted by $\mu(\mathbf{y}_{ji}) = \mathbf{W}_{ji}\boldsymbol{\beta}$, where

$$\mathbf{W}_{ji} = \begin{cases} \begin{pmatrix} \mathbf{1}_T & \mathbf{W}_i & \mathbf{0}_T & \mathbf{O} \end{pmatrix}, & \text{for } j = 0, \\ \begin{pmatrix} \mathbf{1}_T & \mathbf{W}_i & \mathbf{1}_T & \mathbf{W}_i \end{pmatrix}, & \text{for } j = 1, \end{cases}$$

denotes the $T \times d_\beta$ covariate matrix for $\mathbf{y}_{ji}$ in the outcome model and $\boldsymbol{\beta} = (\mu, \boldsymbol{\gamma}, \kappa, \boldsymbol{\theta})$ is the corresponding vector of d_β regression coefficients.

4.1 Priors

Bayesian model specification is completed by assigning prior distributions to all model parameters. For each model $\mathcal{M} \in \{SRI, SRF, SF\}$ we use a prior $p^{\mathcal{M}}(\Theta^{\mathcal{M}})$, where $\Theta^{\mathcal{M}}$ denotes the collection of all parameters in model $\mathcal{M}$. Our priors are of the following structure

$$\begin{aligned} p^{SRI}(\Theta^{SRI}) &= p(\boldsymbol{\beta})p(\boldsymbol{\alpha}) \prod_{j=0}^{1} p^{SR}(\boldsymbol{\sigma}_j)p^{SR}(\boldsymbol{\rho}_j)p^{SRI}(D_j), \\ p^{SRF}(\Theta^{SRF}) &= p(\boldsymbol{\beta})p(\boldsymbol{\alpha}) \prod_{j=0}^{1} p^{SR}(\boldsymbol{\sigma}_j)p^{SR}(\boldsymbol{\rho}_j)p^{SRF}(\boldsymbol{\lambda}_j), \\ p^{SF}(\Theta^{SF}) &= p(\boldsymbol{\beta})p(\boldsymbol{\alpha})\, p^{SF}(\lambda_x) \prod_{j=0}^{1} p^{SF}(\boldsymbol{\sigma}_j)p^{SF}(\boldsymbol{\lambda}_j). \end{aligned}$$

In the shared factor model the property that x_i^* and $\mathbf{y}_{ji}$ are independent conditioning on the latent factor f_i allows to specify conditionally conjugate priors for all parameters, namely independent inverse Gamma-priors for the idiosyncratic variances, i.e. $\sigma_{j,t}^2 \sim \mathcal{G}^{-1}(s_{0,jt}, S_{0,jt})$, and normal prior for the factor loadings, i.e. $\lambda_x \sim \mathcal{N}(l_{x0}, L_{x0})$ and $\boldsymbol{\lambda}_j \sim \mathcal{N}_T(\mathbf{l}_{j0}, \mathbf{L}_{j0})$.

In the switching regression models, prior specification for idiosyncratic variances and correlations is more involved. From the lower triangular Cholesky factor $\mathbf{G}_j$ of $\mathsf{Cov}(\boldsymbol{\epsilon}_{ji}, \eta_i)$, defined in (3.9), which is given as

$$\mathbf{G}_j = \begin{pmatrix} \Sigma_j^{1/2} & \mathbf{0}_T \\ \boldsymbol{\rho}_j' & (1 - \sum_{t=1}^{T} \rho_{j,t}^2)^{1/2} \end{pmatrix},$$

it is obvious that positive-definiteness of $\mathbf{G}_j$, and hence $\mathsf{Cov}(\boldsymbol{\epsilon}_{ji}, \eta_i)$, is guaranteed iff $(1 - \sum_{t=1}^{T} \rho_{j,t}^2) > 0$. Following Chib and Jacobi (2007), we assign to $\boldsymbol{\rho}_j$ a T-variate Normal prior $\mathcal{N}_T(\mathbf{r}_0, \mathbf{R}_0)$ truncated to the region yielding a positive definite Cholesky factor $\mathbf{G}_j$ and to $\ln \boldsymbol{\sigma}_j$ a T-variate Normal distribution $\mathcal{N}_T(\mathbf{c}_{0j}, \mathbf{C}_{0j})$. Finally, we specify conditionally conjugate priors for the random intercept variances, $D_j \sim \mathcal{G}^{-1}(d_{j0}, D_{j0})$, in the SRI model as well as for the factor loadings, $\boldsymbol{\lambda}_j \sim \mathcal{N}_T(\mathbf{l}_{j0}, \mathbf{L}_{j0})$, in the SRF model.[3]

For all three models, priors for the regression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have to be specified. As usual, standard normal distributions priors $\boldsymbol{\alpha} \sim \mathcal{N}_{d_\alpha}(\mathbf{a}_0, \mathbf{A}_0)$ and $\boldsymbol{\beta} \sim \mathcal{N}_{d_\beta}(\mathbf{b}_0, \mathbf{B}_0)$ can be applied. Alternatively, to incorporate variable selection we use spike and slab prior distributions, which will be described in Subsection 4.2.

4.2 Variable Selection with spike and slab priors

We now introduce variable selection to decide which covariates should be included both in the probit model for selection into treatment as well in the potential outcome models. In the latter, variable selection will also enable us to test for the presence of heterogeneous treatment effects captured by $\boldsymbol{\theta}$, as we have specified the models very generally to allow for covariate effects which differ by treatment, but this general model might also be overspecified.

In a Bayesian approach, selection of relevant regressors can be performed by specifying spike and slab prior distributions for the corresponding regression effects. These prior distributions are mixtures of two components, a spike at zero to shrink small coefficients to zero and a fairly uninformative slab component, to leave large effects (nearly) unshrunk. Spike and slab prior distribution have been widely used in different variants in regression type models, e.g.

[3] Alternatively, for the SRI model we could use its representation as a restricted SRF (see Subsection 3.3.1) and specify normal priors on the factor loading $\sqrt{D_j}$.

Mitchell and Beauchamp (1988); George and McCulloch (1997); Geweke (1996); Ishwaran and Rao (2005) but also for more general model selection problems (see e.g. Frühwirth-Schnatter and Tüchler (2008); Frühwirth-Schnatter and Wagner (2010)).

We will use spike and slab prior distributions for the regression effects $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in all models. To this aim, we introduce for each regression effect a binary indicator which takes the value 1, if the effect is assigned to the slab component, and zero otherwise. We assume conditional independence of the regression effects assigned to the slab component, but other choices are also possible.

Conditional on a vector of binary indicators $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{d_\alpha})$, the prior for $\boldsymbol{\alpha}$ is specified as

$$p(\boldsymbol{\alpha}|\boldsymbol{\nu}) = \prod_{j:\nu_j=1} p_{\text{slab}}(\alpha_j) \prod_{j:\nu_j=0} p_{\text{spike}}(\alpha_j). \tag{4.1}$$

The indicators $\boldsymbol{\nu}$ are assumed independent apriori with $p(\nu_j = 1) = \pi_\alpha$ and π_α is assigned a uniform hyper-prior $\pi_\alpha \sim \mathcal{U}[0,1]$. We use a Dirac spike, i.e. a point mass at zero, $p_{\text{spike}}(\alpha_j) = \Delta_0(\alpha_j)$ and a normal slab $p_{\text{slab}}(\alpha_j) = p(\alpha_j|\mathcal{N}(0, A_0))$. Similarly, we define a further vector of binary indicators $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{d_\beta})$ and specify the prior for $\boldsymbol{\beta}$ as

$$p(\boldsymbol{\beta}|\boldsymbol{\delta}) = \prod_{j:\delta_j=1} p(\beta_j|\mathcal{N}(0, B_0)) \prod_{j:\delta_j=0} \Delta_0(\beta_j), \tag{4.2}$$

where the components of $\boldsymbol{\delta}$ are independent apriori with $p(\delta_j = 1) = \pi_\beta$ and $\pi_\beta \sim \mathcal{U}[0,1]$.

4.3 Model Estimation

We subsume now in $\Theta^{\mathcal{M}}$ all parameters of model $\mathcal{M}$, including the indicators $\boldsymbol{\nu}$ and $\boldsymbol{\delta}$ and the hyper-parameters π_α and π_β, if variable selection is performed. The goal is posterior inference for $\Theta^{\mathcal{M}}$ based on the respective posterior distribution, which is proportional to likelihood times prior,

$$p(\Theta^{\mathcal{M}}|\mathbf{x}, \mathbf{y}) \propto p^{\mathcal{M}}(\Theta^{\mathcal{M}}) \prod_{i=1}^{n} p(\mathbf{y}_i, x_i|\Theta^{\mathcal{M}}). \tag{4.3}$$

The prior distributions $p^{\mathcal{M}}(\Theta^{\mathcal{M}})$ for all models are defined in the previous sections and the likelihood contributions for the two models are given in equation (3.16).

Given the different model specifications, we employ different MCMC sampling schemes for the switching regression and the shared factor models. However, for both models we augment the parameter space to improve the tractability of the posterior distribution. As standard, the MCMC implementation relies on the latent utility specification of the probit model (Albert and Chib, 1993) and the likelihood contribution of observed outcome and treatment intake, augmented by the latent utility, is given as

$$p(x_i^*, \mathbf{y}_i, x_i) = p(x_i^*, \mathbf{y}_{x_i,i}|x_i)p(x_i) \qquad x_i \in \{0,1\}.$$

We emphasize here that though $p(x_i^*, \mathbf{y}_{x_i,i}|x_i)$ is the likelihood of a $(T+1)$-dimensional truncated normal distribution,

$$p(x_i^*, \mathbf{y}_{x_i,i}|x_i) = \begin{cases} p(x_i^*, \mathbf{y}_{0i})/P(x_i^* \leq 0), & \text{if } x_i = 0, \\ p(x_i^*, \mathbf{y}_{1i})/P(x_i^* > 0), & \text{if } x_i = 1, \end{cases}$$

the total likelihood contribution of subject i is identical with the likelihood contribution from the joint $(T+1)$-variate normal distribution of $(x_i^*, \mathbf{y}_{x_i,i})$.

To facilitate simulating from the posterior (4.3) by standard MCMC methods we augment the parameter space not only with the latent utilities $\{x_i^*\}$ but also with the latent factors $\mathbf{f} = (f_1, \ldots, f_n)$ in the SF model and the latent variables $\mathbf{b}$ in the SR model, where $\mathbf{b} = (b_{x_1,1}, \ldots, b_{x_n,n})$ is equal to the collection of random intercepts in the SRI model and $\mathbf{b} = (\tilde{b}_{x_1,1}, \ldots, \tilde{b}_{x_n,n})$ is equal to the latent factors in the SRF model.

4.3.1 Posterior Inference for the Switching Regression Model

The basis for posterior inference is the augmented posterior distribution

$$p(\Theta^{SR}, \mathbf{x}^*, \mathbf{b}|\mathbf{y}, \mathbf{x}) \propto p^{SR}(\Theta^{SR})\, p(\mathbf{b}) \prod_{i=1}^{n} p(x_i^*, \mathbf{y}_i, x_i|\Theta^{SR}, \mathbf{b}), \tag{4.4}$$

where $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$. For data augmentation with the latent utilities x_i^*, these are sampled from their full conditional distribution. From the joint error distribution given in equation (3.9) we derive

$$x_i^*|x_i = j, \mathbf{y}_{ji}, \mathbf{b}, \Theta^{SR} \sim \mathcal{N}\left(\mathbf{Z}_i\boldsymbol{\alpha} + \boldsymbol{\omega}_j'\Sigma_j^{-1}\boldsymbol{\epsilon}_{ji}, 1 - \boldsymbol{\omega}_j'\Sigma_j^{-1}\boldsymbol{\omega}_j\right), \tag{4.5}$$

which is truncated to the interval I_{x_i} due to conditioning x_i, and $\boldsymbol{\epsilon}_{ji} = \mathbf{y}_{ji} - \mathbf{W}_{ji}\boldsymbol{\beta} - \mathbf{1}_T b_{ji}$ in the SRI and $\boldsymbol{\epsilon}_{ji} = \mathbf{y}_{ji} - \mathbf{W}_{ji}\boldsymbol{\beta} - \boldsymbol{\lambda}_j\tilde{b}_{ji}$ in the SRF model.

We sample the indicators $(\boldsymbol{\nu}, \boldsymbol{\delta})$, the regression effects $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and the random intercepts (or factors) $\mathbf{b}$ in one block. Marginalizing over the random intercepts (factors), $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the vector of regression effects in the multivariate linear normal regression model for $(\mathbf{y}_{x_i,i}, x_i^*)$ given in (3.15), where variable selection is a standard step. Conditional on the rest of the parameters, also sampling of random intercepts (in the SRI model) or latent factors (in the SRF model) is a standard Gibbs draw. As discussed in Subsection 3.3.2, for the SRF model all elements of the covariance matrix Ω_j are fully identified though the signs of the factor loadings and the factors $\mathbf{b}$ are not separately identified. To guarantee sampling from the whole range of the posterior this non-identifiability is taken into account in posterior sampling by performing a random sign-switch of $\mathbf{b}$ and $\boldsymbol{\lambda}_j$.

Finally, the full conditionals of the remaining parameters $\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_1, \boldsymbol{\rho}_0, \boldsymbol{\rho}_1$, given as

$$p(\boldsymbol{\sigma}_j|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\rho}_j, \mathbf{y}, \mathbf{x}^*, \mathbf{x}) \propto p^{SR}(\boldsymbol{\sigma}_j) \prod_{i:x_i=j} p(\mathbf{y}_{ji}, x_i^*|\boldsymbol{\sigma}_j, \boldsymbol{\rho}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, x_i), \tag{4.6}$$

$$p(\boldsymbol{\rho}_j|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\sigma}_j, \mathbf{y}, \mathbf{x}^*, \mathbf{x}) \propto p^{SR}(\boldsymbol{\rho}_j) \prod_{i:x_i=j} p(\mathbf{y}_{ji}, x_i^*|\boldsymbol{\rho}_j, \boldsymbol{\sigma}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, x_i), \tag{4.7}$$

are not of closed form and hence these parameters are sampled from their full conditionals using a Metropolis-Hastings (MH)-step. Full details of all sampling steps are given in Appendix A.1.

4.3.2 Posterior Inference for the Shared Factor Model

The basis for the posterior inference is the augmented posterior distribution

$$p(\Theta^{SF}, \mathbf{x}^*, \mathbf{f}|\mathbf{y}, \mathbf{x}) \propto p^{SF}(\Theta^{SF})\, p(\mathbf{f})\, p(\mathbf{x}^*, \mathbf{y}, \mathbf{x}|\Theta^{SF}, \mathbf{f}). \tag{4.8}$$

Conditional on the latent factor $\mathbf{f}$, the likelihood of the joint model is given as

$$p(\mathbf{y}, \mathbf{x}, \mathbf{x}^* | \Theta^{SF}, \mathbf{f}) = p(\mathbf{y} | \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \mathbf{f})\, p(\mathbf{x}, \mathbf{x}^* | \boldsymbol{\alpha}, \lambda_x, \mathbf{f}),$$

which is the product of the likelihood contributions from two standard normal regression models, where all parameters can be drawn directly from their full conditional posterior distributions under the conjugate priors introduced in Subsection 4.1. From the specification of the error terms in equations (3.10) to (3.12), conditional on the factor loadings, the other parameters and the observed data, the latent factors f_i have a normal posterior distribution $\mathcal{N}(f_{n,i}, F_{n,i})$ with the posterior moments depending on $x_i = j$:

$$F_{n,i} = 1/(1 + \lambda_x^2 + \sum_{t=1}^{T} \frac{\lambda_{j,t}^2}{\sigma_{j,t}^2}), \quad f_{n,i} = F_{n,i}(\lambda_x, \frac{\lambda_{j,1}}{\sigma_{j,1}}, \ldots, \frac{\lambda_{j,T}}{\sigma_{j,T}}) \begin{pmatrix} x_i^* - \mathbf{Z}_i \boldsymbol{\alpha} \\ \mathbf{y}_{ji} - \mathbf{W}_{ji} \boldsymbol{\beta} \end{pmatrix}. \tag{4.9}$$

The non-identifiability of the signs of the latent factors and the factor loadings (discussed in Section 3.3.2) is taken into account in the sampling scheme via a random sign-switch.

The detailed steps of the sampler are provided in Appendix A.2. The sampling scheme does not rely on imputation of the unobserved potential outcome, although this would be possible based on the joint distribution of both potential outcome vectors. However, while we observed that such a sampler provides essentially the same results, it is slower and less efficient with respect to autocorrelation of the draws.

4.4 Estimation Performance

We have performed a small simulation study with all three models (SRI, SRF, SF) fit to data generated by each of the three models to test the performance of the MCMC samplers and to explore consequences of mis-specifications of the covariance structure of the panel outcomes and the dependence between outcome and latent utility. The detailed results are reported in Appendix B. We find that the proposed MCMC methods perform well with posterior estimates recovering the true parameter values well if inference is based on the correct model. Given the simpler sampling scheme for the SF model the computation is much faster than for the SR models (by a factor of roughly 20) with the SRF model being the most computational intensive model to fit. The inefficiency factors are generally higher in the SR models where MH-steps, usually associated with higher autocorrelation in the chain, are required.

We observe some bias in parameter estimates, including the regression effects in the potential outcome models and the parameters for the dependence structure, when the incorrect model is fit. An interesting point that emerges is that despite the more flexible correlation structure in the SR models, the positive definiteness condition on $\mathsf{Cov}(y_{ij}, x_i^*)$ restricts the range of the estimated possible correlations yielding biased estimates for the SF data. Overall, therefore inference based on the incorrect model introduces some bias also in the average panel treatment effect estimates. An exception is the SRF model, which given its more flexible dependence structure than the SRI model, performs well for data generated from the SRI model. While we fit all three models in our empirical analysis, based on these results our discussion focuses on the SRF and SF models.

5 Analysing Treatment Effects

5.1 Deriving Treatment Effects

Under our flexible specifications of the potential outcome models the difference in the potential outcome sequences is given by the $T \times 1$ vector

$$\Delta_i = \mathbf{y}_{1i} - \mathbf{y}_{0i} = \mathbf{1}_T \kappa + \mathbf{W}_i \boldsymbol{\theta} + (\varepsilon_{1i} - \varepsilon_{0i}).$$

In both modeling frameworks we can identify a sequence of causal average treatment effects (ATE) based on the differences in the means of the potential outcome sequences as:

$$ATE(\mathbf{W}) = E(\Delta_i|\mathbf{W}) = \mu(\mathbf{y}_1|\mathbf{W}) - \mu(\mathbf{y}_0|\mathbf{W}) = \mathbf{1}_T \kappa + \mathbf{W}\boldsymbol{\theta}, \tag{5.1}$$

where, for $j = 0, 1$, $\mu(\mathbf{y}_j|\mathbf{W}) = E(\mathbf{y}_j|\mathbf{W})$ denotes the conditional expectation of $\mathbf{y}_j$, given covariates $\mathbf{W}$.[4]

Hence, κ captures a common (homogeneous) treatment effect and $\mathbf{W}\boldsymbol{\theta}$ is a heterogeneous treatment effect depending on subjects' demographic characteristics such as whether a mother is a blue or a white collar worker. These two types may face very different penalties from time spent on leave and different dynamic patterns of the treatment effects over the panel periods. In Subsection 5.2 we discuss in detail how we can estimate these effects and take the empirical distribution of the demographic factors in $\mathbf{W}$ in the computation of the ATE into account.

While the average treatment effect provides an estimate of the expected gain or loss from maternity leave of a "typical" mother from the sample controlling for any selection bias, it is likely to either overstate or understate the gains or losses of a long maternity leave for mothers who chose the short leave (untreated) or the long leave (treated) as they will differ in terms of their observed and unobserved characteristics. We therefore also estimate the average treatment effects on the treated (TT) and untreated (TU), that are defined as differences in expectations of the potential outcomes conditional on $x = 1$ and $x = 0$ respectively:

$$\begin{aligned} TT(\mathbf{W}, \mathbf{Z}) &= E(\mathbf{y}_1 - \mathbf{y}_0|\mathbf{W}, \mathbf{Z}, x = 1) = E(\mathbf{y}_1|\mathbf{W}, \mathbf{Z}, x = 1) - E(\mathbf{y}_0|\mathbf{W}, \mathbf{Z}, x = 1), \\ TU(\mathbf{W}, \mathbf{Z}) &= E(\mathbf{y}_1 - \mathbf{y}_0|\mathbf{W}, \mathbf{Z}, x = 0) = E(\mathbf{y}_1|\mathbf{W}, \mathbf{Z}, x = 0) - E(\mathbf{y}_0|\mathbf{W}, \mathbf{Z}, x = 0), \end{aligned}$$

that also depend on the covariates $\mathbf{Z}$ in the selection model.

Note that as TT and TU are expected values, the joint distribution of $x, \mathbf{y}_0, \mathbf{y}_1$ is not required as $E(\mathbf{y}_0|\mathbf{W}, \mathbf{Z}, x = 1)$ and $E(\mathbf{y}_0|\mathbf{W}, \mathbf{Z}, x = 0)$ can be computed from the joint distribution of $(x, \mathbf{y}_0)$ and correspondingly $E(\mathbf{y}_1|\mathbf{W}, \mathbf{Z}, x = 1)$ and $E(\mathbf{y}_1|\mathbf{W}, \mathbf{Z}, x = 0)$ from the joint distribution of $(x, \mathbf{y}_1)$. We can thus identify the effects under both panel treatment models from the results derived in Subsection 3.4. From (3.15), we obtain for $j = 0, 1$:

$$E(\mathbf{y}_j|\mathbf{W}, \mathbf{Z}, x^*) = \mu(\mathbf{y}_j|\mathbf{W}) + \frac{\boldsymbol{\omega}_j}{\sigma_x^2}(x^* - \mu(x^*|\mathbf{Z})),$$

where $\mu(x^*|\mathbf{Z}) = E(x^*|\mathbf{Z}) = \mathbf{Z}\boldsymbol{\alpha}$ is the expectation of x^*, given covariates $\mathbf{Z}$ in the selection model. Hence, the expected difference between the potential earnings is given by

$$E(\mathbf{y}_1|\mathbf{W}, \mathbf{Z}, x^*) - E(\mathbf{y}_0|\mathbf{W}, \mathbf{Z}, x^*) = ATE(\mathbf{W}) + (\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)\left(\frac{x^* - \mathbf{Z}\boldsymbol{\alpha}}{\sigma_x^2}\right). \tag{5.2}$$

[4]Note that we used in the previous sections the simpler notation $\mu(\mathbf{y}_j)$ and $\mu(x^*)$ which did not indicate explicitly the dependence on the covariates.

Integrating (5.2) over $x^* > 0$ (treated) and $x^* < 0$ (untreated), yields the treatment effect on the treated (TT) and untreated (TU) as

$$\begin{aligned} TT(\mathbf{W}, \mathbf{Z}) &= \mathbf{1}_T \kappa + \mathbf{W}\boldsymbol{\theta} + (\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0) u_{TT}(\mathbf{Z}), \\ TU(\mathbf{W}, \mathbf{Z}) &= \mathbf{1}_T \kappa + \mathbf{W}\boldsymbol{\theta} - (\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0) u_{TU}(\mathbf{Z}). \end{aligned} \tag{5.3}$$

Since the first two terms are equal to $ATE(\mathbf{W})$, the last term captures the modification of the ATE for the treated and untreated, respectively, based on selection on unobservables. The factor $(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)$ varies across the SR and SF models as a result of the different parameterizations of the dependence structure. Finally, the last term in (5.3), reflecting the values of the unobservables, is identical across the models, with

$$u_{TT}(\mathbf{Z}) = \frac{\phi(-\mathbf{Z}\boldsymbol{\alpha}/\sigma_x)}{\sigma_x \Phi(\mathbf{Z}\boldsymbol{\alpha}/\sigma_x)}, \qquad u_{TU}(\mathbf{Z}) = \frac{\phi(-\mathbf{Z}\boldsymbol{\alpha}/\sigma_x)}{\sigma_x (1 - \Phi(\mathbf{Z}\boldsymbol{\alpha}/\sigma_x))}, \tag{5.4}$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution. Note that $TT(\mathbf{W}, \mathbf{Z})$ and $TU(\mathbf{W}, \mathbf{Z})$ are evaluated at different sets of covariates $\mathbf{W}$ and $\mathbf{Z}$, depending on their distribution in the respective treatment groups (see also estimation of TT and TU in Subsection 5.2).[5]

Hence, differences between the TT (TU) and ATE effects will be driven by selection on observables as well as the presence of correlation and the difference in the realized errors. We would expect mothers to choose the length of the leave based on some information regarding the expected penalty from the length of leave or their attitudes towards work versus child care efforts. For example, mothers in jobs with high career prospects might choose a short leave to avoid extensive loss of human capital that would have strong negative effects if they were to take a long leave. As a result we would expect the TU and TT effects to be quite different from the ATE, with a negative shift from selection on unobservables for the untreated.

Another treatment effect that has been discussed in the recent literature (Heckman and Vytlacil, 1999, 2005; Heckman, Tobias, and Vytlacil, 2001) is the marginal treatment effect introduced in Björklund and Moffitt (1987). The marginal treatment effect is defined for a certain value of the observable characteristics and unobservables in the latent utility model. In our empirical analysis we focus on a common interpretation of the MTE of a mother that is indifferent between short and long treatment, i.e. a mother with a latent utility of zero ($x^* = 0$):

$$MTE(\mathbf{W}, \mathbf{Z}) = E(\mathbf{y}_1 - \mathbf{y}_0 | \mathbf{W}, \mathbf{Z}, x^* = 0).$$

As $x^* = 0$ implies that $\mu(x^*|\mathbf{Z}) = -\eta$ the marginal treatment effect can be written either as function of $\mu(x^*|\mathbf{Z})$ or of η:

$$MTE(\mathbf{W}, \mathbf{Z}) = ATE(\mathbf{W}) - (\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0) \frac{\mu(x^*|\mathbf{Z})}{\sigma_x^2} = ATE(\mathbf{W}) + (\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0) \frac{\eta}{\sigma_x^2}. \tag{5.5}$$

In the MTE, we adjust the ATE to account for the role played by selection on unobservables based on a particular value of the error term in the latent utility chosen to represent a mother at the margin. The sign of the adjustment term will depend on the correlations and the value of η. The formula is a special case of the formula derived for the TT/TU effects. In general the concept of the MTE can be used to describe the standard treatment effects parameters such as ATE, TT and TU as discussed in Heckman and Vytlacil (2005).

[5] Note that in the shared factor model, TT and TU could also be derived as in Heckman et al. (2014) from the distribution of the difference in the potential outcome sequences of subjects, which is based on the joint distribution of the errors $(\varepsilon_{1i}, \varepsilon_{0i})$.

5.2 Treatment Effects Estimation

Following Heckman et al. (2014), inference about the four treatment effects ATE, TT, TU, and MTE is based on their respective posterior distributions (in the three models), which accounts for the uncertainty in parameter estimation.

For our case study, the set of individual covariates will be expanded both by calendar effects and a set of dummy variables accounting for the panel period t which runs from $t = 1, \ldots, 6$, see equation (6.1). For such panel outcomes, we are typically interested in the evolvement of treatment effects over time. To capture dynamics in the treatment effect that are not attributable to changes in covariates, subsequently treatment effects will be studied under the assumption that the demographic and job related covariates of a mother remain constant over all panel periods, i.e. $\mathbf{w}_t \equiv \mathbf{w}$. The corresponding row vector of expanded regressors in panel period t will be denoted by $\tilde{\mathbf{w}}_t$.

For fixed covariates $\mathbf{w}$ and $\mathbf{Z}$, with corresponding regressor $\tilde{\mathbf{w}}_t$, the posterior distribution of all treatment effects is easily derived from the M draws of the parameters $(\kappa^{(m)}, \boldsymbol{\theta}^{(m)}, \boldsymbol{\omega}_1^{(m)}, \boldsymbol{\omega}_0^{(m)}, \boldsymbol{\alpha}^{(m)}, (\sigma_x^2)^{(m)})$, obtained from the MCMC algorithm of the corresponding model (see Appendix A).[6] Consider, e.g. the marginal treatment effect $MTE(t|\mathbf{w}, \mathbf{Z})$ at panel time t which will be computed in our case study for fixed covariates $\mathbf{Z}$ and $\mathbf{w}$ over $t = 1, \ldots, 6$. Based on (5.5), we easily obtain posterior draws of the marginal treatment effect at panel time t as:

$$MTE(t|\mathbf{w}, \mathbf{Z})^{(m)} = \kappa^{(m)} + \tilde{\mathbf{w}}_t \boldsymbol{\theta}^{(m)} - (\omega_{1,t}^{(m)} - \omega_{0,t}^{(m)}) \frac{\mathbf{Z}\boldsymbol{\alpha}^{(m)}}{(\sigma_x^2)^{(m)}}, \tag{5.6}$$

which can be used to estimate the posterior expectation of $MTE(t|\mathbf{w}, \mathbf{Z})$ through the sample average as $\widehat{MTE}(t|\mathbf{w}, \mathbf{Z}) = \frac{1}{M} \sum_{m=1}^{M} MTE(t|\mathbf{w}, \mathbf{Z})^{(m)}$ for all panel periods $t = 1, \ldots, 6$, see Figure 5.

For all other treatment effects we are interested in the evolvement of the in-sample treatment effect over time under the assumption that the demographic and job related covariates of each mother remain constant over all panel periods. Hence, for each mother i observed in the panel, the demographic and job related covariates $\mathbf{w}_{i1}$ observed at panel period $t = 1$ will be used to define the expanded regressors, including panel dummies, etc., which will be denoted by $\tilde{\mathbf{w}}_{it}$. In-sample treatment effects are then obtained by integration with respect to the empirical distribution of the covariates $\tilde{\mathbf{w}}_{it}$ and $\mathbf{Z}_i$ in the panel.

Consider, e.g. the average treatment effect defined in (5.1). For each posterior draw, the corresponding in-sample ATE at panel time t is obtained as the average over the distribution of covariate $\tilde{\mathbf{w}}_{it}$ for all n mothers in the panel, i.e.

$$ATE(t)^{(m)} = \frac{1}{n} \sum_{i=1}^{n} \left(\kappa^{(m)} + \tilde{\mathbf{w}}_{it}\, \boldsymbol{\theta}^{(m)} \right) = \kappa^{(m)} + \tilde{\mathbf{w}}_t^n \boldsymbol{\theta}^{(m)}, \qquad \tilde{\mathbf{w}}_t^n = \frac{1}{n} \left(\sum_{i=1}^{n} \tilde{\mathbf{w}}_{it} \right). \tag{5.7}$$

The posterior draws $ATE(t)^{(m)}$ can be used not only to estimate $\widehat{ATE}(t) = \frac{1}{M} \sum_{m=1}^{M} ATE(t)^{(m)}$ for all panel periods $t = 1, \ldots, 6$ as in Table 5 and Figure 3, but also to explore the uncertainty with respect to inference about $ATE(t)$. This can be done by estimating the posterior standard deviations through the standard deviation of the posterior draws $ATE(t)^{(m)}$, see Table 5, or by visualizing the spread of the posterior distribution of $ATE(t)$ at panel period t through 95% credibility intervals as in Figure 3.

[6] Note that $(\sigma_x^2)^{(m)} \equiv 1$ for the SR model.

Based on the sampled coefficients in MCMC iteration m, we also obtain samples of the expected log potential earnings $y_0(t)$ and $y_1(t)$ for all panel periods t which are given by:

$$y_0(t)^{(m)} = \left(\frac{1}{n} \sum_{i=1}^{n} \mu^{(m)} + \tilde{\mathbf{w}}_{it} \boldsymbol{\gamma}^{(m)} \right) = \mu^{(m)} + \tilde{\mathbf{w}}_t^n \boldsymbol{\gamma}^{(m)},$$
$$y_1(t)^{(m)} = \mu^{(m)} + \kappa^{(m)} + \tilde{\mathbf{w}}_t^n \left(\boldsymbol{\gamma}^{(m)} + \boldsymbol{\theta}^{(m)} \right) = y_0(t)^{(m)} + ATE(t)^{(m)},$$

where $\tilde{\mathbf{w}}_t^n$ is defined as in (5.7). As above, these posterior draws can be used to estimate the posterior expectations of both potential outcomes $\hat{y}_j(t) = \frac{1}{M} \sum_{m=1}^{M} y_j(t)^{(m)}$, $j = 0, 1$ over all panel periods $t = 1, \ldots, 6$, see e.g. Figure 3.

To derive in-sample treatment effects on treated and untreated we integrate, respectively, over the empirical distribution of covariates in the respective subsample of mothers that were treated and untreated. The corresponding sample sizes n_1 and n_0 are equal to the number of treated and untreated persons in panel period $t = 1$. For our case study, we have data for $n_1 = 16,939$ women to derive TT and for $n_0 = 14,115$ to derive TU.

For each MCMC draw, we use (5.3) to derive the in-sample treatment effect on treated and untreated at panel period t, by averaging with respect to the empirical distribution of the covariates $\tilde{\mathbf{w}}_{it}$ and $\mathbf{Z}_i$ in the respective subsample:

$$TT(t)^{(m)} = \kappa^{(m)} + \left(\frac{1}{n_1} \sum_{i:x_i=1} \tilde{\mathbf{w}}_{it} \right) \boldsymbol{\theta}^{(m)} + (\omega_{1,t}^{(m)} - \omega_{0,t}^{(m)}) \left(\frac{1}{n_1} \sum_{i:x_i=1} u_{TT}(\mathbf{Z})^{(m)} \right),$$
$$TU(t)^{(m)} = \kappa^{(m)} \left(\frac{1}{n_0} \sum_{i:x_i=0} \tilde{\mathbf{w}}_{it} \right) \boldsymbol{\theta}^{(m)} - (\omega_{1,t}^{(m)} - \omega_{0,t}^{(m)}) \left(\frac{1}{n_0} \sum_{i:x_i=0} u_{TU}(\mathbf{Z})^{(m)} \right),$$

where $u_{TT}(\mathbf{Z})^{(m)}$ and $u_{TU}(\mathbf{Z})^{(m)}$ are given by (5.4), with regression coefficient $\boldsymbol{\alpha}^{(m)}$.

As above, the posterior draws $TT(t)^{(m)}$ can be used to estimate $\widehat{TT}(t) = \frac{1}{M} \sum_{m=1}^{M} TT(t)^{(m)}$ over all panel periods $t = 1, \ldots, 6$, but also to determine 95% credibility intervals for $TT(t)$, see Figure 4. The same is true of the posterior draws $TU(t)^{(m)}$.

6 Application

As mothers spent time at home to care for the newborn, this break in the employment history can lead to lower human capital of mothers, for example due to the depreciation of general and firm-specific skills during absences from the labor market and lost rents associated with good job matches. According to empirical work this lower human capital, in particular, years out of the labor force, and unobserved heterogeneity are significant contributors to the widely observed motherhood wage penalty, a key driver of the gender earnings gap (Anderson, Binder, and Krause, 2002, 2003; Budig and England, 2001; Waldfogel, 1998a,b; Lundberg and Rose, 2000). A considerable literature has investigated the effect of parental leave policies on labour market outcomes of women as these play a key role in the decision about the length of leave (financial benefits, job protection) and can affect mothers' attachment to the labour market positively or negatively. Overall parental leave policies vary widely across countries and over time and the studies have yielded mixed results with some pointing to negative effects on employment and wages and other studies finding no effects. A recent study on parental leave policy changes in Austria (Lalive et al., 2014) finds no effect on earnings from the 2000 extension

of the parental leave benefit period from 18 to 30 months from a cross-section analysis on (cumulative) earnings of mothers 5 years after childbirth.

However, while parental leave policies are a key determinant in a mother's decision regarding the length of leave taken after childbirth, a range of other factors, both observed and unobserved, will affect her decision. Differences in unobserved ability, career plans and other factors may affect how mothers respond to a policy. In our empirical analysis we therefore focus on the mother's endogenous leave decision and investigate its dynamic labour market consequences, exploiting the policy change in Austria in 2000 described earlier. Given the pronounced response of mothers to the benefit period as shown in Figure 1 (a) we have defined a binary leave variable in terms of short leave (18 months of less) and long leave (over 18 months) in Section 3 that is based on the final length chosen by the mother as reported in the data. We employ the flexible Bayesian panel treatment modelling framework presented in the previous sections to estimate the earnings effect of short versus long leave on a mother's yearly earnings in the 5 years following her return to the labour market, taking into account the endogeneity of the leave decision, dependence across a mother's earnings over time and allowing for different dynamic patterns across the treatment groups. We identify a vector of causal dynamic earnings effects of long leave (as described in Section 5) to investigate the presence and time dynamics of a potential earnings gap from long leave.

6.1 Data and Sample

The data for our analysis come from two data sets, one is the Austrian Social Security Data Base (ASSD), which is an administrative data set of the universe of Austrian employees and provides detailed information on employment spells and maternity leave spells, as well as demographic information on mothers and information on employers (Zweimüller, Winter-Ebmer, Lalive, Kuhn, Wuellrich, Ruf, and Büchi, 2009). The second is a data set collected as basis for wage taxes. These data are well suited for our empirical analysis. In addition to the global coverage, they contain precise information on earnings, whether and how long a mother took maternity leave and whether she returned to the same employer. A weakness of these data are the limited demographic information and lack of household information, such as partner's earnings and household wealth, and health information that will affect a mother's leave decision. These factors will be captured by the error term in the selection equation and some may be correlated with unobserved factors in the potential earnings which is then captured by the correlations and will not bias our results. While it would be preferable to include such variables and learn about their effect on the length of maternity leave, the main purpose of the analysis is to account for the endogeneity of the mother's leave decision to obtain causal estimates of the length of leave on earnings.

For our analysis we focus on women that gave birth within a 4 years period around the change in the parental leave policy in July 2000 and consider those who gave birth between July 1998 and June 2002. This period was chosen to (i) create a sample with a balanced window of mothers before and after the policy change, and (ii) to ensure that we can observe a reasonable number of periods for each mother after her return to the labor market for our panel analysis. For women who have more than one child between July 1998 and June 2002 we will consider the last child birth in the period. Our analysis is restricted to mothers that were employed in the private sector in the year before child birth to ensure eligibility for the standard maternity leave policy regimes in place at the time. It also allows us to compute a

baseline earnings variable to account for the earnings level before the birth of the child (first child if more than one in the considered period).

Our analysis focuses on mothers who returned to the labour market at the end of the maternity leave. A considerable proportion of mothers have no strong attachment to the labour market and did not return to work within a few years of the end of the maternity leave. In order to be able to identify dynamic earnings effects we focus on mothers returning to the labour market for whom we have at least 4 consecutive panel observations (63,800). To create a comparable sample of mothers with a strong attachement to the labour market we focus on the subset of mothers who returned to the labor market within 30 days after the end of maternity leave while their last child still requires childcare (34,110). Additionally we restrict the sample to include only mothers with earnings above 2980 Euro per year which corresponds to log earnings larger than 8. Finally, to ensure the identification of common year effects and panel effects separately by treatment status we only consider earnings from the years 2000 to 2008.

The above sample restrictions result in an unbalanced sample of 31,051 mothers that are observed over 4-6 consecutive panel periods, i.e. have no breaks in their employment histories (working at least 360 days the year following the end of the maternity leave). Since we based our analysis on yearly earnings we define the first earnings observation for the year after their return when they report to work at least 360 days.

In Table 1 we present some summary statistics for the sample. Overall 58% of mothers in the sample went on leave under the new leave policy. The average leave in the sample is 21 months with mothers in the treatment group taking almost double the leave time (27 months compared to 15 months). Most mothers have either one child (49.7%) or two children (40.8%), and of the remaining almost all have 3 children. Based on the distribution of the number of children we define a dummy variable for having two children and a dummy variable for having more than 2 children.

	Overall Sample		$x = 0$	$x = 1$
Variable	mean	standard deviation	mean	mean
duration of leave	21.57	6.48	15.04	27.01
z	0.58		0.13	0.95
age mother	30.47	4.88	30.45	30.49
number of children	1.62	0.71	1.62	1.61
working experience (at birth)	9.39	4.58	9.24	9.51
blue collar	0.31		0.32	0.29
same employer	0.74		0.80	0.69
real earnings* base year	20689.44	9840.47	20776.58	20616.81
real earnings* year 1	15997.88	8719.40	17603.46	14659.74

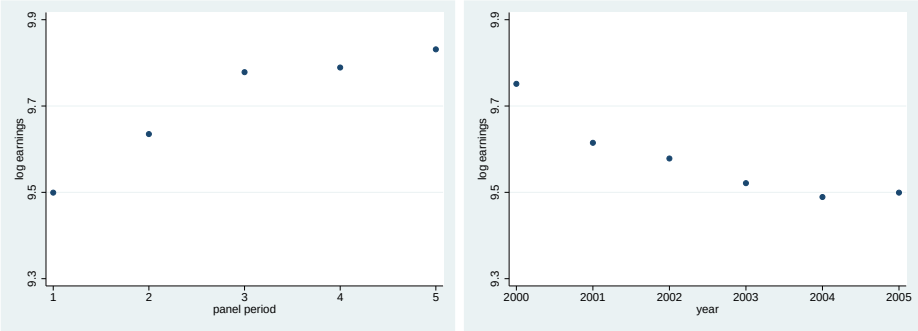
Table 1: Selected sample summary statistics. * in EUR

An interesting point to note is that while overall 58% of the mothers took leave under the new policy, the proportion is 13% for mothers with short leave and 95% for mothers with long leave. Again, this confirms the large positive impact of the increase in the maternity benefit period from 18 months to 30 months on the lengths of leave taken by mothers.

6.2 Model Specifications

We now specify the covariate vectors for the selection and potential outcome models in the analysis of the earnings effects based on the unbalanced sample of mothers described in the previous section under switching regression models and the shared factor model. For the selection model into the long leave treatment we specify the covariate vector $\mathbf{Z}$ to include demographic control variables and controls referring to the labor market experience of mothers before maternity leave: two indicator variables for 2 children or more than two children; an indicator variable for high work experience (above median) before maternity leave; an indicator for blue collar; and a control for earnings before maternity leave for first child (or based on earnings before the birth of last child, if more than 10 years since birth of first child) in terms of indicators for baseline earnings quartiles (base-earn Q2, Q3, Q4). These controls are also included in the potential outcome models. In addition we include an indicator whether a mother returns to the same employer as previous studies of mothers in the US and Britain have indicated that having access to job protected maternity leave has a positive wage effect for mothers by increasing the likelihood that the mother returns to the same employer, decreasing the loss of firm-specific skills and the maintaining of good job matches.

In addition we include flexible controls for the panel periods and a quadratic time trend to account for two specific features of the data, strong panel period and year effects. The two graphs in Figure 2 illustrate the presence of these two data features in the raw data. Due



(a) Earnings panel period (in year 2005) **(b)** Earnings by year (panel period 1)

Figure 2: Average log earnings by panel period and year.

to the structure of the data (see Table 11) we graph the average log earnings for all mothers and a subset of the panel periods and years to ensure a reasonable number of observations in each panel period/year. The first graph indicates that mothers' yearly earnings increase substantially in the years (panel periods) following their return to the labor market. The most likely reason behind this pattern is that mothers increase their working hours as their (youngest) child gets older, partially driven by the better availability of child care options for older children. The panel effects are likely to differ across the two treatment groups, for example due to the fact that mothers with a short leave return when their child is on average younger compared to mothers returning after a long leave. The second graph points to clearly present, and potentially non-linear calender year effects in the data.

To allow for flexible and group-specific panel effects, we include a set of dummies into each potential outcome model to account for the various panel periods. Finally, we include a

quadratic time trend that is common across both potential outcomes as mothers in both groups face the same macro-economic conditions. Hence, the means of potential outcome models are specified as

$$\begin{aligned}\mu(\mathbf{y}_{0i}) &= \mathbf{1}_T\mu + \mathbf{W}_i\boldsymbol{\gamma}_1 + \mathbf{I}_{Pi}\boldsymbol{\gamma}_2 + c\gamma_3 + c^2\gamma_4, \\ \mu(\mathbf{y}_{1i}) &= \mathbf{1}_T(\mu+\kappa) + \mathbf{W}_i(\boldsymbol{\gamma}_1+\boldsymbol{\theta}_1) + \mathbf{I}_{Pi}(\boldsymbol{\gamma}_2+\boldsymbol{\theta}_2) + c\gamma_3 + c^2\gamma_4,\end{aligned} \tag{6.1}$$

where $\mathbf{I}_{Pi}$ refers to the matrix of indicators for panel periods 2 to 6 (panel period 1 is base category) and c is defined as calendar year minus 1999.

Variable selection is implemented on the covariate effects in the selection and earnings models, excluding the common intercept terms. In the earnings models we also exclude the time trend as it proxies for macro-economic conditions, but include κ to test for the presence of a common treatment effect. Selection of κ will be support for the presence of a common treatment effect while selection of elements in $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ indicate the presence of a heterogeneous treatment effect in terms of interaction with covariates.

6.3 Results

For our empirical analysis of the effect of the length of maternity leave on mothers' subsequent earnings, we implemented the prior-posterior analysis of mothers' earnings under the three modeling approaches discussed in the paper, the two switching regression model specifications with a random intercept (SRI) or latent factor (SRF) and the shared factor model (SF). Bayesian inference was done for all models both with and without variable selection for the covariate effects in selection and earnings model. In the discussion below we focus on the model specifications with stochastic variable selection. All results are based on 10,000 runs of the corresponding MCMC algorithm following a burn-in period of 10,000 iterations for which the draws are discarded to allow for convergence of the sampler.

6.3.1 Selection and Earnings Model Estimates with Variable Selection

In this section we focus on the SRF and SF models for a more clear discussion and presentation of results. As the SRF model is a generalization of the SRI model, the two models yield comparable results both in terms of the parameter estimates and inclusion probabilities, see Table 12. We first present the results on the parameter estimates for the selection model into the maternity leave treatment. Table 2 below reports the posterior means and standard deviations as well as the estimated inclusion probabilities for the switching regression model (SRF) and the shared factor model (SF).

In all models we observe a strong positive effect of the policy change and work experience before the maternity leave for selection into long maternity leave, while having base earnings in the highest quartile has a negative effect. Having one child already has a positive effect on selection into long leave under the SF model, while the estimated effect is much lower and not selected under the SRF model, with an estimated inclusion probability of 0.366. Being a blue collar worker is found to have a negative effect on taking up a long leave under the SRF model, while the estimated effect is absolutely smaller with the estimated inclusion probability below 0.5 in the SF model. Having more than two children or baseline earnings in the 2nd and 3rd quartile appear to have no effect on selection into the long maternity leave treatment, the same is found for the interaction of experience and being a blue collar worker.

Table 2: Results selection equation: posterior means (mean), standard deviation (sd, in parentheses) and posterior inclusion probabilities (prob.) of regression effects; *estimated inclusion probability, no selection on intercept; *results for SF with variance adjustment*

	SRF Model		SF Model	
	mean (sd)	prob*	mean (sd)	prob*
intercept	-1.540 (0.029)	–	-1.576 (0.030)	–
z	2.793 (0.026)	1.000	2.825 (0.022)	1.000
child 2	0.021 (0.030)	0.366	0.051 (0.036)	0.724
child ≥ 3	-0.026 (0.045)	0.287	-0.014 (0.035)	0.174
experience	0.091 (0.023)	0.996	0.102 (0.025)	0.998
blue collar	-0.043 (0.042)	0.568	-0.026 (0.037)	0.371
int exp./ blue	-0.018 (0.042)	0.198	-0.017 (0.038)	0.196
base-earn Q2	0.002 (0.011)	0.055	0.002 (0.011)	0.054
base-earn Q3	-0.001 (0.005)	0.024	-0.001 (0.006)	0.026
base-earn Q4	-0.135 (0.026)	1.000	-0.149 (0.027)	1.000

Table 3 reports the estimates of the intercept and the covariate effects in the earnings model. As the latter can vary by treatment state we present two columns of estimates for each model. The first column under "treatment 0" contains the posterior means and standard deviations on the common intercept and covariate effects (μ,$\boldsymbol{\gamma}_1$,$\boldsymbol{\gamma}_2$), while the second column "+ treatment 1" contains the additional effect under the long leave treatment (κ,$\boldsymbol{\theta}_1$,$\boldsymbol{\theta}_2$). For a better reading of the table, estimated inclusion probabilities above 0.5 are indicated by "*". We first turn our attention to the common effects. Overall we observe that the two models agree in terms of which covariates affect the yearly earnings (selected into the model): experience, being a blue collar worker, having base earnings above the 1st the quartile, returning to the same employer and the panel period, as well as the sign of their effect. An interesting feature is the steady increase in the common effects of the panel periods. While the exact magnitude of the effects varies slightly across the models, the effect more than triples from the 2nd to the 6th period. One obvious explanation for this pattern is that mothers increase the hours they work as the (youngest) child gets older and more child care options become available. Since our outcome variable is yearly earnings it subsumes any effects of changes in working hours. As expected we observe strong positive effects of having higher base earnings (especially in the highest two quantiles), and returning to the sample employer, and a negative effect for mothers who are blue collar workers. Interestingly experience seems to have a negative effect on earnings here but that is likely a result of not being able to control for hours and mothers with more experience deciding and being able (to afford) to increase their hours more slowly after returning to work.

The results for additional effects under long treatment (column "+ treatment 1") do not support the simple model with a homogeneous treatment effect. We notice that the steady increase in panel earnings is further magnified under treatment by the increase in the additional panel period effects under long leave. While we notice some variation in the magnitude of the additional panel effects for long leave mothers (as we did in the common panel effects), we again observe a steady increase in the coefficients over the panel periods. The estimated

Table 3: Results Earnings Model: posterior means, standard deviations (in parentheses); inclusion of regression effects based on posterior inclusion probabilities > 0.5 indicated by "*"

	SRF Model				SF Model			
	treatment 0		+ treatment 1		treatment 0		+ treatment 1	
intercept	9.334	(0.012)	-0.125*	(0.012)	9.331	(0.011)	-0.112*	(0.009)
child 2	-0.000	(0.001)	-0.000	(0.001)	-0.000	(0.001)	-0.000	(0.001)
child >3	0.000	(0.001)	0.000	(0.002)	0.000	(0.001)	0.000	(0.002)
exp	-0.087*	(0.008)	0.005	(0.010)	-0.091*	(0.009)	0.010	(0.013)
blue collar	-0.103*	(0.006)	0.000	(0.002)	-0.102*	(0.006)	0.000	(0.001)
int. exp/blue	0.001	(0.004)	0.007	(0.013)	0.000	(0.003)	0.010	(0.015)
base-earn Q2	0.068*	(0.006)	0.000	(0.002)	0.069*	(0.006)	0.000	(0.003)
base-earn Q3	0.292*	(0.010)	-0.049*	(0.012)	0.292*	(0.009)	-0.050*	(0.012)
base-earn Q4	0.615*	(0.010)	-0.117*	(0.012)	0.610*	(0.010)	-0.118*	(0.013)
eq. emp.	0.051*	(0.005)	0.001	(0.003)	0.051*	(0.005)	0.000	(0.002)
panel t=2	0.071*	(0.006)	0.095*	(0.007)	0.072*	(0.004)	0.061*	(0.005)
panel t=3	0.116*	(0.007)	0.118*	(0.006)	0.117*	(0.006)	0.094*	(0.005)
panel t=4	0.162*	(0.009)	0.139*	(0.007)	0.162*	(0.007)	0.107*	(0.005)
panel t=5	0.217*	(0.012)	0.142*	(0.007)	0.214*	(0.009)	0.113*	(0.006)
panel t=6	0.267*	(0.014)	0.169*	(0.008)	0.261*	(0.011)	0.132*	(0.007)
$(\text{year} - 1999)$	0.036 (0.004)				0.034 (0.003)			
$(\text{year} - 1999)^2$	-0.004 (0.0002)				-0.004 (0.0002)			

inclusion probabilities indicate a higher panel effect for long leave mothers in all panel periods. Both models also indicate an additional negative effect under the long leave for the intercept suggesting a common covariate independent negative earnings effect of long leave. Both models also estimate an additional negative effect for mothers with base earnings in the 3rd and 4th quartile. These additional effects contribute to the heterogeneous treatment effect. Finally, the models yield almost identical estimates of the quadratic year effects which are included in the model to capture changes in macroeconomic conditions over time.

6.3.2 Variances, Covariances and Correlation Structures

We next present the results for the correlation between the latent utility associated with the maternity leave treatment and the two potential earnings. All modeling frameworks allow for different correlation patterns between the latent utility of leave and earnings between short and long leave mothers and also across time. As they differ, however, in their specification of the correlation structure between the latent utility and the potential earnings and the correlation structure within the potential earnings due to the role of the latent factor we compare the correlations from the shared factor model with those of the switching regression models marginalized over the random intercept and the latent factor (equation (3.14)) in Table 4 below.

The estimates confirm the presence of unobserved factors that drive both leave decision and earnings. All three models imply similar patterns with a negative correlation under short

Table 4: Marginal correlations $\mathsf{Cor}(y_{j,it}, x_i^*)$ between latent utility x_i^* and outcomes $y_{j,it}$, posterior means, standard deviations (in parentheses)

	SRI Model		SRF Model		SF Model	
t	treatment 0	treatment 1	treatment 0	treatment 1	treatment 0	treatment 1
1	-0.119 (0.012)	0.279 (0.023)	-0.117 (0.012)	0.224 (0.023)	-0.176 (0.009)	0.171 (0.009)
2	-0.162 (0.011)	0.084 (0.022)	-0.157 (0.009)	0.011 (0.026)	-0.207 (0.010)	0.219 (0.011)
3	-0.185 (0.008)	0.192 (0.018)	-0.181 (0.008)	0.141 (0.021)	-0.226 (0.011)	0.239 (0.012)
4	-0.196 (0.005)	0.145 (0.021)	-0.194 (0.005)	0.093 (0.023)	-0.237 (0.012)	0.232 (0.011)
5	-0.203 (0.006)	0.181 (0.021)	-0.197 (0.007)	0.127 (0.023)	-0.235 (0.012)	0.219 (0.011)
6	-0.184 (0.008)	0.112 (0.020)	-0.180 (0.008)	0.068 (0.023)	-0.223 (0.011)	0.210 (0.010)

leave for all periods, and a positive correlation under long leave for all periods. In other words, for short leave mothers we have negative confounding between the utility and earnings under long leave or positive confounding between utility and earnings under short leave, i.e. mothers under short leave gain from short leave. These patterns can be seen as a consequence of the findings from the motherhood wage gap literature that show unobserved factors that affect earnings play a role in fertility decisions.

The magnitude of the correlations is similar for the short leave treatment in all three models, with more pronounced correlations in the shared factor model. Under long leave treatment the shared factor model yields positive correlations around 0.2, whereas the other two models yield only low positive correlations for $t = 2$. The more flexible modeling of the correlation structures in the switching regression models might be responsible for the higher variation across the time periods, as the structure of the shared factor model induces a temporal smoothing of the correlations across time. It should also be noted that, in particular in the two switching regression models, the correlations are identified based on mothers that chose a maternity leave length different from the incentives given by the policy regime in place. This is a very small subset of 772 mothers who chose long leave prior to policy change. Under the SF model the correlations are in part based on the estimates of the factor loadings that are more embedded in the modeling structure with more identification coming from the model in addition to the smoothing from the latent factor.

The estimation results for the remaining parameters of the dependence structure are provided in the Appendix C. The estimated variances Σ_j of the idiosyncratic error terms $\boldsymbol{\epsilon}_{ji}$ in the potential earnings models are almost identical across both models (see Table 13). An interesting feature is the decrease in variance over the first 3-4 panel periods and a slight increase afterwards. The decrease is likely to be driven by a convergence in the hours worked by mothers as their children get older and child care options increase. Similarly, the increase coincides roughly with the onset of the school age, around period 5 under short leave and at period 4 for long leave mothers whose children are on average one year older at their return to the labor market. As school ends midday in Austria with no lunch provided many mothers reduce their working hours again.

Further, in the SRF and SF models, where the correlation across the potential outcomes is captured by a flexible factor structure with the time-varying factor loadings, we observe almost identical values of the absolute values of the factor loadings, see Table 13.

In connection with the essentially identical estimates of the idiosyncratic error variances, this leads to roughly the same covariance matrix Ω_j of the potential earnings vectors marginalized over the random effects or factors across these two models (results for SRF model in Table 14). Under both treatments we observe that the correlation between $y_{j,it}$ and the outcomes in the following period increases as t increases. Assuming that mothers increase their work hours, unaccounted for in our analysis due to data limitations, we would expect to see such a pattern. In comparison, the SRI model imposes compound symmetry on the covariance matrices (see Table 15).

Finally, under the SF model the estimated factor loadings of the potential earnings models further imply a correlation structure across the potential outcomes. Interestingly, the implied covariance structure between the potential earnings, $\mathsf{Cov}(\mathbf{y}_{0i}, \mathbf{y}_{1i}) = \boldsymbol{\lambda}_0 \boldsymbol{\lambda}_1'$ reported in Table 16 yields negative estimates for all covariances. According to standard human capital theory, where the correlations are assumed to be driven by unobserved ability, we would expect a positive sign. High ability mothers would be expected to be earning more under short and long leave relative to a low ability mothers. While it is not clear what drives the negative correlation in our case, there are a number of possible explanations. An obvious possibility is the lack of information on working hours that would distort the correlation if the work hour patterns are different across the two maternity leave treatments for high (low) ability mothers in that high ability mothers would work less hours under the long leave treatment. One possible scenario that could lead to such a pattern is that high ability mothers who have invested a lot into their career before children, continue with that approach and take a short leave and return for many hours or decide to invest a lot into their child's development by taking a long maternity leave and working less hours after they return. Another possible explanation is based on assortative mating, i.e. the idea that high earning mothers are married to high earning men, while lower income mothers are more likely to be married to lower income men. Thus for lower income mothers it might not be a financially viable option to return for only a portion of full-time hours due to the lower family income, while high income mothers can rely on their partner's income.

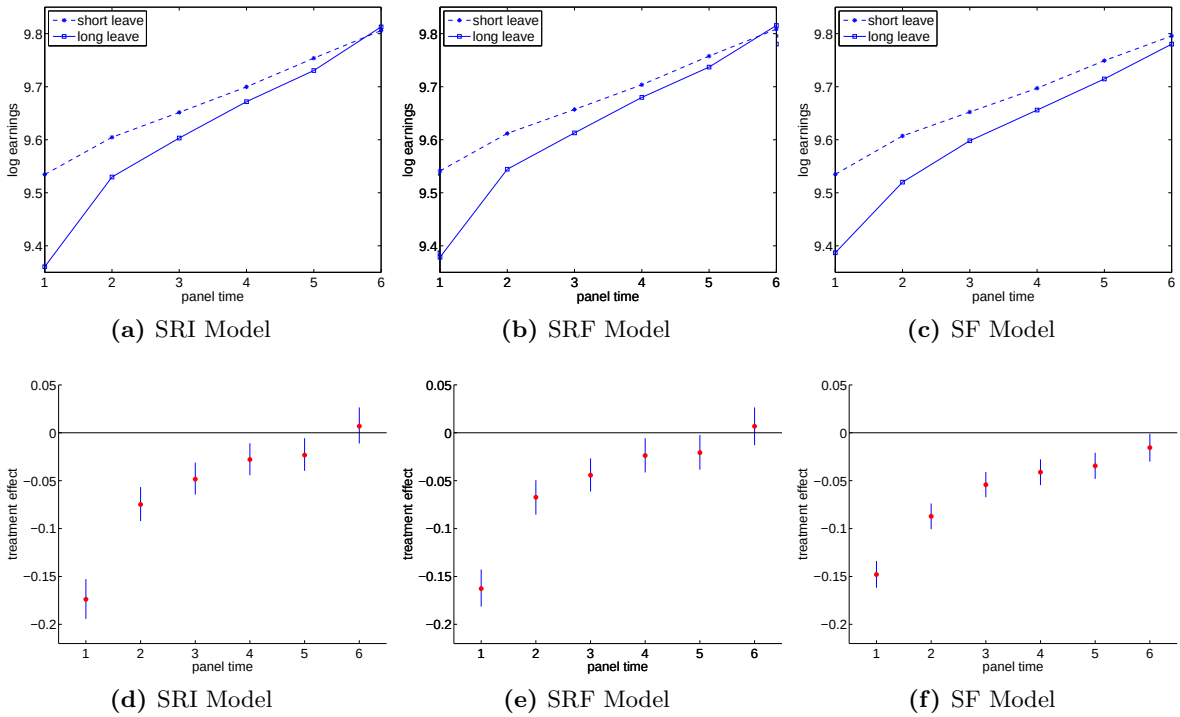
6.3.3 Earnings Dynamics and Treatment Effects

In this section we present results on the potential earnings dynamics and the earnings effects from the three models. All computations are based on the methods described in Subsection 5.2.

We start with a discussion of the dynamics of in-sample ATE and potential log earnings. The upper panel in Figure 3 graphs the estimated posterior expectations $\hat{y}_j(t)$ of potential log earnings dynamics for short leave ($j = 0$) and long leave mothers ($j = 1$) as implied by the three different models. The lower panel in Figure 3 visualizes the posterior distribution of the average treatment effect $ATE(t)$ at panel period t through plots of the 95% credibility intervals.

All three models suggest very similar patterns for the potential earnings and average treatment effect dynamics with the two SR models yielding almost identical results. The key feature is that we observe that mothers with a long leave start out with considerably lower potential earnings than mothers with a short leave in the first period of their return to the labor market, with the gap decreasing over the next 4-5 panel periods. This is reflected in the decreasing negative $ATE(t)$.

We observe some differences in the dynamics of the potential log earnings and ATE dy-



(a) SRI Model **(b)** SRF Model **(c)** SF Model

(d) SRI Model **(e)** SRF Model **(f)** SF Model

Figure 3: Estimated posterior expectations $\hat{y}_j(t)$ of potential log earnings dynamics for short ($j = 0$) and long leave ($j = 1$) mothers (upper panel) and estimated average treatment effect $\widehat{ATE}(t)$ in panel period t (lower panel, red dots) and 95% credibility intervals (vertical bars) under the two switching regression models and the shared factor model.

namics between the SR and SF models. While under the SF model the estimate for the last panel period is negative and the credibility interval does not include zero, under the SR models we observe a positive ATE with the credibility intervals not including negative values. Hence, the SR models indicate a closing of the earnings gap six years after the return to the labour market, while the SF model only suggests a further reduction but a remaining earnings gap.

However, as Table 11 on the distribution of the sample by panel period, year and treatment group in the appendix shows, some caution is required when interpreting treatment effects such as ATE (and TT, TU and MTE presented below) for the 6th panel period. While we observe essentially all mothers under the short leave treatment up to the 6th panel period, we observe the vast majority of the long leave mothers only up to the 5th panel period, and about one third in the last panel period. Given the implementation of the policy change in July 2000, our sample based on the two year window around the policy and that these mothers spend between 1.5 and 2.5 years on leave, we do not observe the last panel period for mothers that gave birth later than 2000 and chose a leave close to 2.5 years. Having less long leave mothers with close to the maximum leave may imply that our treatment effects estimates for period 6 understate the earnings penalty. This may also explain why we observe a substantially steeper reduction in the earnings penalty from the 5th to the 6th period than in previous periods in all three models (most pronounced in SR models).

Finally, the SF model implies a somewhat smoother path of the average potential earnings

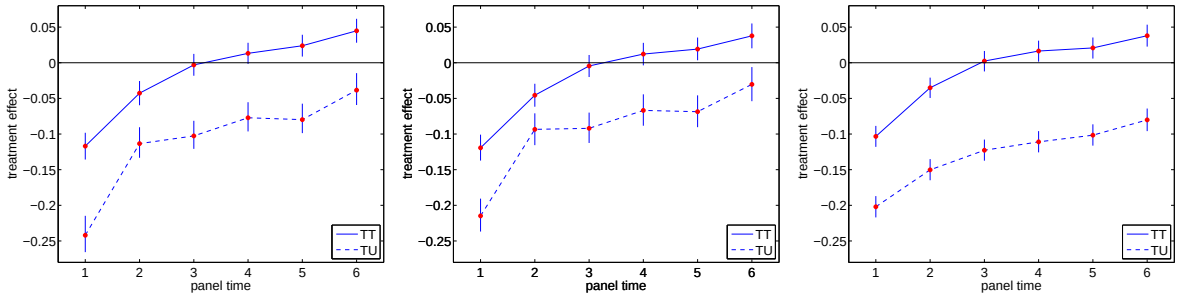
Table 5: Estimated posterior expectations and estimated posterior standard deviations (sd, in parentheses) of the average treatment effect $ATE(t)$ and percentage change in ATE in panel period t under the two switching regression models and the shared factor model.

	ATE log earnings: $\widehat{ATE}(t)$ (sd)			ATE percentage change: mean (sd)		
t	SRI Model	SRF Model	SF Model	SRI Model	SRF Model	SF Model
1	-0.174 (0.010)	-0.163 (0.010)	-0.148 (0.007)	-15.9 (0.9)	-14.9 (0.8)	-13.6 (0.6)
2	-0.075 (0.009)	-0.067 (0.009)	-0.087 (0.007)	-7.1 (0.9)	-6.4 (0.9)	-8.2 (0.6)
3	-0.048 (0.009)	-0.044 (0.009)	-0.054 (0.007)	-4.6 (0.8)	-4.2 (0.8)	-5.2 (0.6)
4	-0.028 (0.008)	-0.024 (0.009)	-0.041 (0.007)	-2.7 (0.8)	-2.2 (0.8)	-3.9 (0.7)
5	-0.023 (0.009)	-0.021 (0.009)	-0.035 (0.007)	-2.2 (0.9)	-1.9 (0.9)	-3.3 (0.7)
6	0.007 (0.009)	0.007 (0.010)	-0.015 (0.007)	0.8 (1.0)	0.8 (1.0)	-1.4 (0.7)

over time which is reflected in more steady but slower reduction in the earnings gap between long and short leave mother shown in the posterior distributions of $ATE(t)$. These differences are to a large extent driven by the different panel effects estimated for both treatment groups under the different models that were reported in Table 3.

These patterns are also evident in the estimates for the average earnings effects in terms of log earnings and percentage changes reported in Table 5. Mothers with long leave return with an expected gap of roughly 0.15-0.17 in log earnings under all three models, which implies roughly an expected earnings gap of 15% in the first year. After 6 years mothers with a long leave have on average 0.8% higher earnings under the two SR models, and a remaining average gap of 1.4% under the SF model.

Next, we investigate the dynamics of in-sample treatment effects for long leave mothers, $TT(t)$, and short leave mothers, $TU(t)$, for the SRF and SF models. Figure 4 shows the posterior mean and 95% credibility intervals (indicated by vertical bars) of $TT(t)$ (solid blue line) and $TU(t)$ (dashed blue line) for all panel periods. As discussed in detail in Section 5,



(a) SRI Model **(b)** SRF Model **(c)** SF Model

Figure 4: Posterior distributions of in-sample treatment effects for long leave mothers ($TT(t)$) and short leave mothers ($TU(t)$) under the SR and SF models; estimated posterior means $\widehat{TT}(t)$ and $\widehat{TU}(t)$ (red dots) and 95% credibility intervals (vertical bars).

the treatment effects TT and TU take into account selection on observables and unobservables and are computed for each group as the average treatment effect, adjusted by a correction

term to account for selection on unobservables. The size and sign of the latter depend on the estimated correlation patterns between the errors in the latent utility of selection and the potential earnings sequences. Given the reported estimated negative correlations under short leave and positive correlations under long leave (Table 4), the formulae for the correction term in (5.3) imply a positive correction for long leave mothers (TT) and a negative correction for short leave mothers (TU). As a consequence we observe TT effects above the TU effects in all panel periods for all models. The estimated posterior means, together with the HDP regions, imply that mothers taking a short leave would have suffered a considerably higher earnings penalty under long leave than those mothers who chose long leave, as a result of selection based on unobservable characteristics. For example, we know from the estimation results that mothers with earnings in the highest quartile before child birth are more likely to decide against a long leave. These mothers are likely to have different types of jobs and work environments and as a result suffer a higher earnings penalty under long leave as indicated.

Further, the results imply that long leave mothers only suffer an earnings penalty in the first panel periods while short leave mothers would have a persistent earnings penalty. $TT(t)$ is positive from the 4th panel period, while $TU(t)$ remains negative until the last period. Under all models both effects start out negative in the first panel period and are decreasing over time. However, we observe some differences in the estimates between the SR and SF models as a result of the different parameterizations of the correlations and estimates under the SR and SF models. The latter implies more pronounced negative effects $TU(t)$ for periods $t = 2$ to 6, i.e. considerably larger and more persistent earnings penalties for short leave mothers under the long leave treatment.

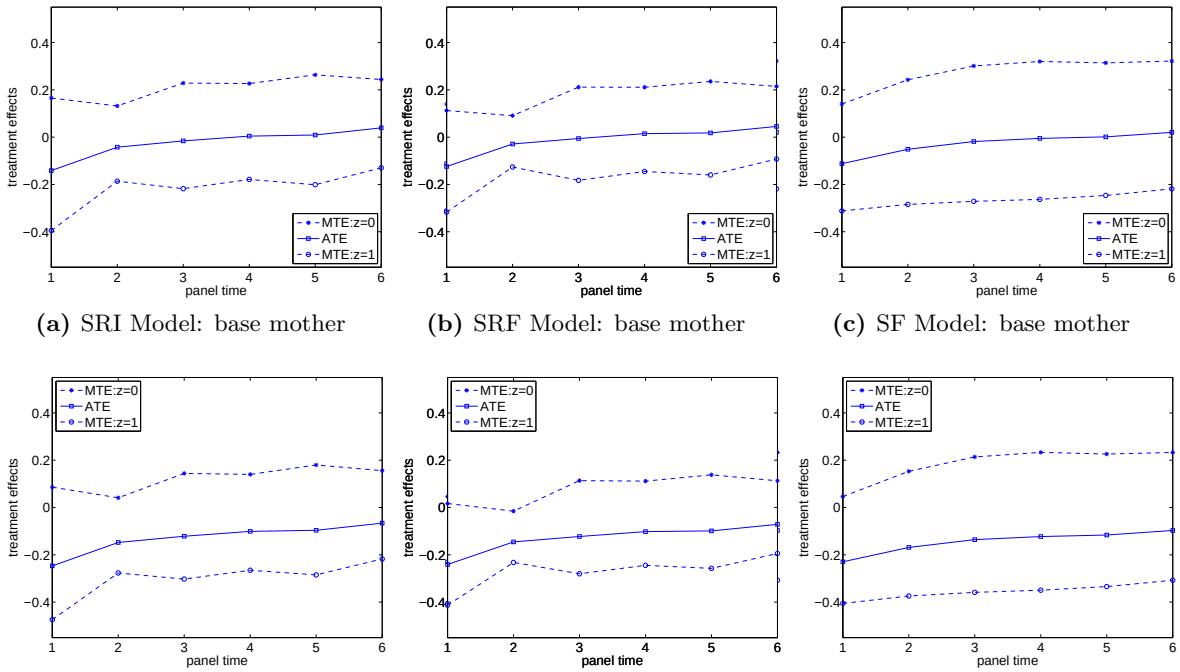
Finally, we investigate the marginal treatment effects for a mother on the margin between short and long leave under the old and new policy. In Figure 5, we plot the estimated posterior expectation $\widehat{MTE}(t|\mathbf{w}, \mathbf{Z})$ of MTE for various mothers which is estimated from the posterior draws as explained in Subsection 5.2, see equation (5.6).

The upper panel plots $\widehat{MTE}(t|\mathbf{w}, \mathbf{Z})$ for a base mother with the following set of characteristics: white collar, one child, low prior experience, baseline earnings in lowest quartile and return to the same employer at the end of the maternity leave. We graph $\widehat{MTE}(t|\mathbf{w}, \mathbf{Z})$ for a base mother under two different values of the unobservables in the latent utility that would make her indifferent between short and long leave under the old policy ($z = 0$) and under the new policy ($z = 1$), respectively.

Given the large effect of the policy change, in the SRF model a base mother would require a large positive unobservables term of $\eta = 1.54$ for her to have a latent utility of zero under the old policy. In comparison, a mother with the same observable characteristics would be indifferent between long and short leave after the policy change ($z = 1$) even with a large negative unobservables term of $\eta = -1.25$. For comparison we also graph $\widehat{MTE}(t|\mathbf{w}, \mathbf{Z})$ for a mother where the unobservables are zero, which is simply equal to the ATE for the same set of covariates $\mathbf{w}$ and $\mathbf{Z}$.

Under all three models, a mother at the margin under the old policy would benefit in terms of earnings from taking long leave as suggested by the positive value of $\widehat{MTE}(t|\mathbf{w}, \mathbf{Z})$ resulting for all 6 panel periods under all three models. Under a large negative unobservable term that would see her indifferent under the new policy, we estimate a large negative value of $\widehat{MTE}(t|\mathbf{w}, \mathbf{Z})$ for all panel periods.

The figure reveals a difference between the SR and SF model in the estimated MTE that is larger than for the estimated ATE. The differences are a consequence of the different modeling



(a) SRI Model: base mother **(b)** SRF Model: base mother **(c)** SF Model: base mother

(d) SRI Mode: Q4 mother **(e)** SRF Model: Q4 mother **(f)** SF Model: Q4 mother

Figure 5: Estimated posterior expectation $\widehat{MTE}(t|\mathbf{w},\mathbf{Z})$ for a base mother (top) and a high earnings mother (bottom) at the margin under the old ($z=0$) and under new the policy ($z=1$), in comparison to estimated posterior expectation $\widehat{ATE}(t|\mathbf{w},\mathbf{Z})$.

of the variance and dependence structures. Under the SR models we observe a smaller gap between the two estimated MTEs and less smooth time dynamics. Under the SF model, the estimated $\widehat{MTE}(t|\mathbf{w},\mathbf{Z})$ suggest large and very persistent negative (positive) earnings effects for a base mother at the margin under the new policy (old policy) that remains at around -20% (over 30%), while the SR models suggest a decrease (increase) over time to about -10% (20%). Overall these results indicate that the policy at the margin pushes mothers with a certain set of observable characteristics into long leave that given their unobservables in the latent utility model suffer a considerable earnings penalty from long leave after their return to the labour market. While we see some increase in $\widehat{MTE}(t|\mathbf{w},\mathbf{Z})$ for those mothers, a considerable earnings gap remains even after 6 panel periods. This is in contrast to the ATE results which suggest a closing of the gap on average.

The above results also highlight the role of unobserved heterogeneity in the treatment effects. The 2nd panel in Figure 5 provides $\widehat{MTE}(t|\mathbf{w},\mathbf{Z})$ for a different set of covariates, with the mother's baseline earnings before leave being in the highest quartile (Q4). Since being in the highest quartile affects both the selection into treatment and the earnings, we observe a downward shifts in the ATE as well as $\widehat{MTE}(t|\mathbf{w},\mathbf{Z})$. Given the negative effect of choosing long leave in the SRF model, mothers are indifferent under the old policy now for $\eta = 1.68$ and under the new policy for $\eta = -1.12$. A high earnings base mother on the margin under the new policy would suffer an even higher earnings penalty over all panel periods under long leave. Being on the margin under the old policy, such a mother has smaller $\widehat{MTE}(t|\mathbf{w},\mathbf{Z})$ with

some effects close to zero or slightly negative in the first two panel periods under some models.

7 Conclusion

In this paper, we investigated several Bayesian treatment effect models for panel outcomes within the potential earnings framework based on the switching regression approach and a factor approach. The proposed models differ in their assumptions regarding key modeling features, namely, first, the dependence between the idiosyncratic shocks in the treatment and in the outcome equations, and, second, the dependence structure across the panel outcomes. We relax the restrictive compound symmetry assumption of the common switching regression model for panel data, by introducing a latent factor rather than a random intercept to capture the dependence of the panel outcomes over time. As a second alternative to the standard modelling framework, we consider a factor approach to model the dependence between the idiosyncratic shocks in the treatment and in the outcome equations jointly.

For practical inference, we employ a fully Bayesian inference based on Markov chain Monte Carlo methods. A key contribution of the paper is the introduction of stochastic variable selection via spike and slab priors in the context of these different treatment effects models, which allows to test which covariates should be included and if heterogenous treatment effects are present, where the effect of a covariate depends on the treatment level.

We derive and estimate a range of dynamic treatment effects for all models under investigation, including not only average treatment effects, but also treatment effects on the treated and untreated as well as marginal treatment effects.

To avoid identifiability issues, we confined ourselves to a single factor. A potentially promising extension would be a latent multi-factor factor model as in Carneiro et al. (2003), where one factor captures dependence between latent utility and panel outcomes and one or more factors capture within panel dependence, however we leave this interesting idea for future research.

In our empirical analysis we applied the different modelling frameworks to investigate the effect of long versus short maternity leave on a mother's earnings in a six-year period following her return to the labor market. Focusing on mothers with a strong attachment to the labour market, we find substantial but decreasing negative earnings effects from long maternity leave on a mother's earnings over the first 5 years after her return. We estimate yearly earnings penalties starting from around 14-16% in the first year, 6-8% in the second year and decreasing to 2-3% in the 5th year, pointing also to a considerable cumulative earnings loss over the first 5 years in the labour market after return from leave. The estimated negative earnings effects results are consistent with the findings of negative earnings effects due to time spent out of the labour market reported in the the motherhood gap literature. They are in contrast to the finding of no effects on (cumulative) annual earnings 5 years after childbirth (approximately 2-3 years after return to the labour market in our analysis) from the 2000 policy change itself reported in in Lalive et al. (2014) for a wider sample of mothers with different degrees of attachment to the labour market. However, such an analysis does not take into account the role of selection on unobservables in the leave decision. Our analysis points to the important role played by selection on unobservables as indicated by the correlation patterns and is also evident in the substantial differences between the treatment effects between short leave mothers (untreated) and long leave mothers (treated) and the marginal treatment effects estimates.

References

Aakvik, A., J. J. Heckman, and E. J. Vytlacil (2005). Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. *Journal of Econometrics 125*, 15–51.

Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Anderson, D., M. Binder, and K. Krause (2002). The motherhood wage penalty: Which mothers pay it and why? *AEA Papers and Proceedings 92:2*, 354–358.

Anderson, D., M. Binder, and K. Krause (2003). The motherhood wage penalty revisited: Experience, heterogeneity, work effort, and work schedule flexibility. *Industrial and Labor Relations Review 56(2)*, 273–294.

Björklund, A. and R. Moffitt (1987). The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, 42–49.

Budig, M. J. and P. England (2001). The wage penalty for motherhood. *American Sociological Review 66*, 204–225.

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty of college choice. *International Economic Review 44*, 361–422.

Chib, S. (2007). Analysis of treatment response data without the joint distribution of potential outcomes. *Journal of Econometrics 140*, 401–412.

Chib, S. and B. H. Hamilton (2000). Bayesian analysis of cross-section and clustered data treatment models. *Journal of Econometrics 97*, 25–50.

Chib, S. and L. Jacobi (2007). Modeling and calculating the effect of treatment at baseline from panel outcome. *Journal of Econometrics 140*, 781–801.

Chib, S. and L. Jacobi (2008). Analysis of treatment response data from eligibility designs. *Journal of Econometrics 144*, 465–478.

Frühwirth-Schnatter, S. and R. Tüchler (2008). Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statistics and Computing 18*, 1–13.

Frühwirth-Schnatter, S. and H. Wagner (2010). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics 154*, 85–100.

George, E. I. and R. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*, 881–889.

George, E. I. and R. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica 7*, 339–373.

Geweke, J. (1996). Variable selection and model comparison in regression. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. Smith (Eds.), *Bayesian Statistics 5*, pp. 609–620. Oxford University Press.

Heckman, J. and S. Navarro-Lozano (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics 86*, 30–57.

Heckman, J., J. L. Tobias, and E. Vytlacil (2001). Four parameters of interest in the evaluation of social programs. *Southern Economic Journal*, 211–223.

Heckman, J. and E. Vytlacil (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. In *Proceedings of the national Academy of Sciences*, Volume 96, pp. 4730–4734.

Heckman, J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies 65*, 261–294.

Heckman, J. J., H. Lopes, and R. Piatek (2014). Treatment effects: a Bayesian perspective. *Econometric Reviews 33*, 36–67.

Ishwaran, H. and S. J. Rao (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics 33*, 730–773.

Koop, G. and D. J. Poirier (1997). Learning about the across-regime correlation in switching regression models. *Journal of Econometrics 78*, 217–227.

Lalive, R., A. Schlosser, A. Steinhauer, and J. Zweimüller (2014). Parental leave and mothers careers: The relative importance of job protection and cash benefits. *Review of Economic Studies 81*, 219–265.

Lalive, R. and J. Zweimüller (2009). How does parental leave affect fertility and return-to-work? Evidence from tow natural experiments. *Quarterly Journal of Economics 124*, 1363–1402.

Lee, L. (1978). Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International Economic Review 19*, 415–433.

Lee, M. (2005). *Micro-Econometrics for Policy, Program, and Treatment Effects*. Oxford: Oxford University Press.

Ley, E. and M. F. J. Steel (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics 24*, 651–674.

Li, M. and J. Tobias (2011). Bayesian inference in a correlated random coefficients model: Modeling causal effect heterogeneity with an application to heterogeneous returns to schooling. *Journal of Econometrics 162*, 345–361.

Lundberg, S. and E. Rose (2000). Parenthood and earnings of married men and women. *Labour Economics 7*, 689–710.

Mitchell, T. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*, 1023 – 1032.

Munkin, M. K. and P. K. Trivedi (2003). Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: an application to the demand for healthcare. *Journal of Econometrics 114*, 197–220.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*, 135–146.

Waldfogel, J. (1998a). The family gap for young women in the United States and Britain: Can maternity leave make a difference? *Journal of Labor Economics 16*, 505–545.

Waldfogel, J. (1998b). Understanding the "family gap" in pay for women with children. *Journal of Economic Perspectives 12*, 137–156.

Zweimüller, J., R. Winter-Ebmer, R. Lalive, A. Kuhn, J.-P. Wuellrich, O. Ruf, and S. Büchi (2009). The Austrian Social Security Database (ASSD). Working paper 0903, NRN: The Austrian center for labor economics and the analysis of the welfare state, Linz, Austria.

WEB APPENDIX

A Details on posterior sampling

A.1 Sampling for the Switching Regression Model

The MCMC scheme for posterior inference in the switching regression model with random intercept involves the following steps:

(1) For $i = 1, \ldots, n$ sample x_i^* from its conditional posterior distribution, which is the normal distribution given in equation (4.5), truncated to the interval I_{x_i}.

(2) Sample indicator variables, regression effects and random intercepts.

(2a) Sample $(\boldsymbol{\nu}, \boldsymbol{\delta})$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ for the joint regression model with multivariate normal error distribution given in equation (3.15). This step is described in detail in Appendix A.3.

(2b) For $i = 1, \ldots, n$ and $x_i = j$ sample the random intercept b_{ji} from the full conditional normal posterior $\mathcal{N}(h_i, H_i)$ with the posterior moments depending on $x_i = j$:

$$H_i = (1/D_j + \mathbf{1}'(\Sigma_j - \boldsymbol{\omega}_j\boldsymbol{\omega}_j')^{-1}\mathbf{1})^{-1},$$
$$h_i = H_i\mathbf{1}'(\Sigma_j - \boldsymbol{\omega}_j\boldsymbol{\omega}_j')^{-1}\tilde{\mathbf{y}}_i,$$

where $\tilde{\mathbf{y}}_i$ denotes the working observations $\tilde{\mathbf{y}}_i = \mathbf{y}_{ji} - \mathbf{W}_{ji}\boldsymbol{\beta} - \boldsymbol{\omega}_j(x_i^* - \mathbf{Z}_i\boldsymbol{\alpha})$.

(3) For $j = 0, 1$ sample D_j from the conditional posterior $\mathcal{G}^{-1}\left(d_{j0} + n_j/2, D_{j0} + \sum_{i:x_i=j} b_{ji}^2/2\right)$ where n_j is the number of subjects with $x_i = j$.

(4) For $j = 0, 1$ and $t = 1, \ldots, T$ sample $\log \sigma_{j,t}^2$ from $p(\log \sigma_{j,t}^2 | \Theta^{SRI} \backslash \sigma_{j,t}^2, \mathbf{b}, \mathbf{x}^*, \mathbf{y}, \mathbf{x})$. Updates are performed in a random order of $\{1, \ldots T\}$.

(5) For $j = 0, 1$ and $t = 1, \ldots, T$ sample $\rho_{j,t}$ from $p(\rho_{j,t} | \Theta^{SRI} \backslash \rho_{j,t}, \mathbf{b}, \mathbf{x}^*, \mathbf{y}, \mathbf{x})$. Updates are performed in a random order of $\{1, \ldots T\}$.

(6) Sample π_α from $\mathcal{B}(1 + k_\alpha, 1 + d_\alpha - k_\alpha)$ and π_β from $\mathcal{B}(1 + k_\beta, 1 + d_\beta - k_\beta)$ where $k_\alpha = \sum \nu_l$ is the number of selected regressors for the latent utility and $k_\beta = \sum \delta_l$ accordingly the number of selected regressors for the potential outcomes.

Note that the full conditionals in sampling steps (4) and (5) only involve subjects i with $x_i = j$, see equations (4.6) and (4.7). In both steps we use the Metropolis-Hastings algorithm, where our proposal distribution is a t-distribution with 10 degrees of freedom with parameters obtained from few maximization steps (currently we use 10 iterations of the SQP algorithm implemented in Matlab). This proposal is truncated to the stationarity region when sampling the correlation parameters, more precisely we propose a value $\rho_{j,t}^*$ from the t-distribution truncated to $\pm\sqrt{0.999 - \sum_{t \neq t^*} \rho_{j,t}^2}$. The average acceptance rate ranges from 0.9287 to 0.9637.

For the SRF model with factor structure in the joint variance-covariance matrix $\Omega_j, j = 0, 1$, step (2b) is replaced by sampling the latent factors and step (3) by sampling the factor loadings from their respective full conditionals. These are standard steps in the linear normal model

$$\tilde{\mathbf{y}}_i = \tilde{b}_{ji}\boldsymbol{\lambda}_j + \boldsymbol{\varepsilon}_{ji}, \qquad \boldsymbol{\varepsilon}_{ji} \sim \mathcal{N}_T\left(\mathbf{0}, \Sigma_j - \boldsymbol{\omega}_j\boldsymbol{\omega}_j'\right), \qquad \text{with } x_i = j.$$

To take into account non-identifiability of the signs of factor loadings and factors this step is concluded by a random sign-switch. Finally, to sample the parameters $\boldsymbol{\sigma}_j$ and $\boldsymbol{\rho}_j$ we condition on the latent factors $\mathbf{b}$ as well as the factor loadings $\boldsymbol{\lambda}_j$, which is the only modification required in sampling steps (4) and (5).

In detail, the sampling steps which require modification for a factor structure in the outcome covariance matrix are as follows:

(2b*) For $i = 1, \ldots, n$ and $x_i = j$ sample the latent factor $\tilde{b}_{ji}$ from the full conditional $\mathcal{N}\left(\tilde{h}_i, \tilde{H}_i\right)$ with moments depending on $x_i = j$:

$$\begin{aligned}\tilde{H}_i &= (1 + \boldsymbol{\lambda}_j'(\Sigma_j - \boldsymbol{\omega}_j\boldsymbol{\omega}_j')^{-1}\boldsymbol{\lambda}_j)^{-1},\\ \tilde{h}_i &= \tilde{H}_i\boldsymbol{\lambda}_j'(\Sigma_j - \boldsymbol{\omega}_j\boldsymbol{\omega}_j')^{-1}\tilde{\mathbf{y}}_i.\end{aligned}$$

(3*) For $j = 0, 1$ sample $\boldsymbol{\lambda}_j$ from $\mathcal{N}(\mathbf{l}_j, \mathbf{L}_j)$, where

$$\begin{aligned}\mathbf{L}_j &= \Big(\sum_{i:x_i=j} \tilde{b}_{ji}^2(\Sigma_j - \boldsymbol{\omega}_j\boldsymbol{\omega}_j')^{-1} + \mathbf{L}_{j0}^{-1}\Big)^{-1},\\ \mathbf{l}_j &= \mathbf{L}_j\Big(\sum_{i:x_i=j} \tilde{b}_{ji}(\Sigma_j - \boldsymbol{\omega}_j\boldsymbol{\omega}_j')^{-1}\tilde{\mathbf{y}}_i + \mathbf{L}_{j0}^{-1}\mathbf{l}_{j0}\Big).\end{aligned}$$

To perform the random sign-switch of the latent factors $\mathbf{b}$ and the factor loadings $(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1)$ sample random variables ξ_j for $j = 0, 1$ with $P(\xi_j = -1) = P(\xi_j = 1) = 0.5$. Set $\boldsymbol{\lambda}_j^{(new)} = \boldsymbol{\lambda}_j\xi_j$ and set $\tilde{b}_{ji}^{(new)} = \tilde{b}_{ji}\xi_j$ for all i, where $x_i = j$ and use $\mathbf{b}^{(new)}, \boldsymbol{\lambda}_0^{(new)}$ and $\boldsymbol{\lambda}_1^{(new)}$ as updated values of the chain. Note that this sign-switch does not change the product $\boldsymbol{\lambda}_j\tilde{b}_{ji}$ for $x_i = j$.

(4*) For $j = 0, 1$ and $t = 1, \ldots, T$ sample $\log\sigma_{j,t}^2$ from $p(\log\sigma_{j,t}^2|\Theta^{SRF}\backslash\sigma_{j,t}^2, \mathbf{b}, \mathbf{x}^*, \mathbf{y}, \mathbf{x})$. Updates are performed in a random order of $\{1, \ldots T\}$.

(5*) For $j = 0, 1$ and $t = 1, \ldots, T$ sample $\rho_{j,t}$ from $p(\rho_{j,t}|\Theta^{SRF}\backslash\rho_{j,t}, \mathbf{b}, \mathbf{x}^*, \mathbf{y}, \mathbf{x})$. Updates are performed in a random order of $\{1, \ldots T\}$.

A.2 Posterior Sampling for the Shared Factor Model

In the shared factor model specified in equations (3.10) to (3.12), the error terms $\upsilon_i, \boldsymbol{\epsilon}_{0i}, \boldsymbol{\epsilon}_{1i}$ are independent. Hence the augmented likelihood including the unobserved latent utilities is given as

$$p(\mathbf{x}, \mathbf{x}^*, \mathbf{y}|\Theta^{SF}, \mathbf{f}) = \prod_{i=1}^{n} p(x_i, x_i^*|\boldsymbol{\alpha}, \lambda_x, f_i) \prod_{i:x_i=0} p(\mathbf{y}_{0i}|\boldsymbol{\beta}, \Sigma_0, \boldsymbol{\lambda}_0, f_i) \prod_{i:x_i=1} p(\mathbf{y}_{1i}|\boldsymbol{\beta}, \Sigma_1, \boldsymbol{\lambda}_1, f_i)$$

Conditional on the latent factors, the models for the latent utilities and the potential outcomes are regression models with the additional regressor f_i. This suggests to sample $(\boldsymbol{\alpha}, \lambda_x)$ as well as $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ in one block.

The complete sampling scheme involves the following steps:

(1) For $i = 1, \ldots, n$ sample the latent factor f_i from the full conditional posterior

$$p(f_i|\Theta^{SF}, x_i^*, \mathbf{y}_{x_i,i}) \propto p(x_i^*, \mathbf{y}_{x_i,i}|\Theta^{SF}, f_i)p(f_i)$$

which is a normal distribution, $\mathcal{N}(f_{n,i}, F_{n,i})$, with the posterior moments depending on $x_i = j$, see equation (4.9).

(2) For $i = 1, \ldots, n$ sample x_i^* from $\mathcal{N}(\mathbf{Z}_i\boldsymbol{\alpha} + \lambda_x f_i, 1)$ truncated to the interval I_{x_i}.

(3) Perform variable selection (i.e. sampling of $\boldsymbol{\nu}$) and sample the regression coefficients $(\boldsymbol{\alpha}, \lambda_x)$ in the latent utility model

$$x_i^* = \mathbf{Z}_i\boldsymbol{\alpha} + f_i\lambda_x + \nu_i, \quad \nu_i \sim \mathcal{N}(0, 1).$$

Note that only elements of $\boldsymbol{\alpha}$ are subject to selection whereas λ_x is not.

(4) Perform variable selection (i.e. sampling of $\boldsymbol{\delta}$) and sampling of regression coefficients $(\boldsymbol{\beta}, \boldsymbol{\lambda_0}, \boldsymbol{\lambda_1})$ in the model for the observed outcomes $\mathbf{y}_{x_i,i}$, $i = 1, \ldots, n$ which is given as

$$\mathbf{y}_{x_i,i} = \mathbf{W}_{x_i,i}\boldsymbol{\beta} + f_i\boldsymbol{\lambda}_{x_i} + \boldsymbol{\epsilon}_{x_i,i}, \quad \boldsymbol{\epsilon}_{x_i,i} \sim \mathcal{N}(\mathbf{0}, \Sigma_{x_i}).$$

(5) To take into account non-identifiability of the signs of factors and factor loadings we perform a random sign-switch of $\mathbf{f}$ and $(\lambda_x, \boldsymbol{\lambda_0}, \boldsymbol{\lambda_1})$, i.e. we sample a random variable ξ with $P(\xi = 1) = P(\xi = -1) = 0.5$, set $\mathbf{f}^{(new)} = \xi\mathbf{f}$, $\lambda_x^{(new)} = \xi\lambda_x$ and $\boldsymbol{\lambda}_j^{(new)} = \boldsymbol{\lambda}_j\xi$ for $j = 0, 1$, and use $\mathbf{f}^{(new)}$, $\lambda_x^{(new)}$ and $\boldsymbol{\lambda}_j^{(new)}$ as updated values of the chain.

(6) For $j = 0, 1$ and $t = 1, \ldots, T$ sample $\sigma_{j,t}^2$ from $\mathcal{G}^{-1}(s_{n,jt}, S_{n,jt})$ where

$$s_{n,jt} = s_{0,jt} + n_j/2 \qquad S_{n,jt} = S_{0,jt} + Se_{jt}/2$$

and

$$Se_{jt} = \sum_{i:x_i=j} (y_{j,it} - \mathbf{W}_{j,it}\boldsymbol{\beta} - f_i\lambda_{j,t})^2.$$

Here n_j is the number of subjects with $x_i = j$ and $\mathbf{W}_{j,it}$ denotes the values of the covariates at panel time t, i.e. row t of the covariate matrix $\mathbf{W}_{ji}$.

(7) Sample π_α from $\mathcal{B}(1 + k_\alpha, 1 + d_\alpha - k_\alpha)$ and π_β from $\mathcal{B}(1 + k_\beta, 1 + d_\beta - k_\beta)$ where $k_\alpha = \sum \nu_l$ is the number of selected regressors for the latent utility and $k_\beta = \sum \delta_l$ accordingly the number of selected regressors for the potential outcomes.

Sampling steps (3) and (4) are the standard sampling steps used for linear regression models with variable selection and are detailed in Appendix A.3.

A.3 Variable selection with spike and slab priors in regression models

Consider a linear regression model

$$\mathbf{y}_i = \mathbf{W}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_i),$$

with independent observations $\mathbf{y}_i, i = 1, \ldots, n$ and regressor matrix $\mathbf{W}_i$ of dimension $T \times d$. By introducing a $d \times 1$ indicator vector $\boldsymbol{\delta}$ we specify a spike and slab prior distribution for the elements of $\boldsymbol{\beta}$ as

$$p(\boldsymbol{\beta}|\boldsymbol{\delta}) = \prod_{j:\delta_j=1} p_{\text{slab}}(\beta_j) \prod_{j:\delta_j=0} p_{\text{spike}}(\beta_j). \tag{A.1}$$

For elements of $\boldsymbol{\beta}$ which are not subject to selection the corresponding indicator δ_j is set to 1, otherwise $p(\delta_j = 1) = \pi$ with hyper-prior $\pi \sim \mathcal{B}(a_0, b_0)$. Variable selection is based on the posterior probabilities for $p(\delta_j = 1|\mathbf{y})$, where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, which can be sampled using MCMC methods. Here we use a Dirac spike $p_{\text{spike}}(\beta_j) = \Delta_0(\beta_j)$ and specify the slab component by $p_{\text{slab}}(\beta_j) = p(\beta_j|\mathcal{N}(0, B_0))$. Sampling $\boldsymbol{\delta}$ conditional on $\boldsymbol{\beta}$ would result in a reducible Markov chain, and therefore it is essential to sample the indicator vector $\boldsymbol{\delta}$ marginalizing over $\boldsymbol{\beta}$, when a Dirac spike is specified.

For sampling from the posterior distribution the sampling scheme therefore consists of the following steps.

1. Update $\boldsymbol{\delta}$ componentwise in a random permutation ϱ of $(1, \ldots, d)$: For $j = 1, \ldots, d$ set $l = \varrho_j$ and sample δ_l from the posterior

   $$p(\delta_l|\boldsymbol{\delta}_{\setminus l}, \mathbf{y}, \pi) \propto p(\mathbf{y}|\boldsymbol{\delta})p(\boldsymbol{\delta}|\pi)p(\pi).$$

   For a linear regression model the marginalized likelihood $p(\mathbf{y}|\boldsymbol{\delta})$ is available analytically as

   $$p(\mathbf{y}|\boldsymbol{\delta}) \propto \frac{|\mathbf{B}_n^{\boldsymbol{\delta}}|^{1/2}}{|\mathbf{B}_0^{\boldsymbol{\delta}}|^{1/2}} \prod_{i=1}^{n} |\mathbf{V}_i|^{-1/2} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i'\mathbf{V}_i^{-1}\mathbf{y}_i) - (\mathbf{b}_n^{\boldsymbol{\delta}})'(\mathbf{B}_n^{\boldsymbol{\delta}})^{-1}\mathbf{b}_n^{\boldsymbol{\delta}}\right),$$

   where

   $$(\mathbf{B}_n^{\boldsymbol{\delta}})^{-1} = (\mathbf{B}_0^{\boldsymbol{\delta}})^{-1} + \sum_{i=1}^{n}(\mathbf{W}_i^{\boldsymbol{\delta}})'\mathbf{V}_i^{-1}\mathbf{W}_i^{\boldsymbol{\delta}}, \tag{A.2}$$

   $$\mathbf{b}_n^{\boldsymbol{\delta}} = \mathbf{B}_n^{\boldsymbol{\delta}} \sum_{i=1}^{n}(\mathbf{W}_i^{\boldsymbol{\delta}})'\mathbf{V}_i^{-1}\mathbf{y}_i, \tag{A.3}$$

   and $\mathbf{W}_i^{\boldsymbol{\delta}}$ consists of those columns j of $\mathbf{W}_i$ where the corresponding indicator $\delta_j = 1$ and $\mathbf{B}_0^{\boldsymbol{\delta}} = B_0\mathbf{I}_k$, where k denotes the number of selected effects $k = \sum_{j=1}^{d}\delta_j$.

2. Sample $\boldsymbol{\beta}^{\boldsymbol{\delta}}$, i.e. the elements of $\boldsymbol{\beta}$ with non-zero indicators from its full conditional posterior $\mathcal{N}_k(\mathbf{b}_n^{\boldsymbol{\delta}}, \mathbf{B}_n^{\boldsymbol{\delta}})$ and set the remaining elements of $\boldsymbol{\beta}$ to zero.

3. Sample π from its full conditional, which is the Beta-distribution $\mathcal{B}(a_0 + k, b_0 + d - k)$.

B Simulation study

We have performed a small simulation study to test the performance of the MCMC samplers and to explore consequences of a mis-specification of the covariance structure of the panel outcomes or the dependence between outcome and latent utility. The details of the simulation design described below were chosen to help illustrate the effect of the mis-specification of the dependence structure on the estimation results.

B.1 Simulation Setup

We have generated three data sets of $n = 50000$ subjects with $T = 4$ panel periods from each of the three models models specified in Section 3 (data 1: SRI, data 2: SRF and data 3: SF). In each case the structural mean of the latent utility (equation (3.1)) is specified as

$$\mu(x_i^*) = \mathbf{Z}_i\boldsymbol{\alpha} = \alpha_{10} + \alpha_{11}v_{1i} + \alpha_{12}v_{2i} + \alpha_2 z_i$$

with $\mathbf{Z}_i = (1, v_{1i}, v_{2i}, z_i)$, where v_{1i} is standard normal and v_{2i} and z_i are binary variables with $p(v_{2i} = 1) = p(z_i = 1) = 0.5$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \alpha_2) = (-0.9, 0.8, 0, 1.5)$. To generate the outcome sequences (as defined in equations (3.2)and (3.3)) we used a linear predictor with v_{1i} and v_{2i} and dummies for the panel time points $t = 2, 3, 4$ as regressors. The common intercept and covariate effects were set at $(\mu, \boldsymbol{\gamma}) = (3, 1, 0, 0.1, 0.15, 0.2)$, and the constant and heterogeneous treatment effects of the covariates at $(\kappa, \boldsymbol{\theta}) = (-0.5, 0, 0.2, -0.1, 0, 0.1)$. The implied average treatment effects for the four panel periods are $(-0.4, -0.5, -0.4, -0.3)$. Error variances were set to $\boldsymbol{\sigma}_0^2 = 0.25\,\mathbf{1}_4$ and $\boldsymbol{\sigma}_1^2 = \mathbf{1}_4$.

For the two data sets generated from the switching regression models we set the correlations at $\boldsymbol{\rho}_0 = (0.6, 0.5, 0.4, 0.3)$ and $\boldsymbol{\rho}_1 = -\boldsymbol{\rho}_0$ to capture a dependence between the latent utility and the potential outcomes that varies over time. The covariance structures for the potential outcomes were defined by setting the random intercept variances to $D_0 = 0.4$ and $D_1 = 0.8$ for data set 1 under the SRI, and by setting the factor loadings to $\boldsymbol{\lambda}_0 = (0.4, 0.35, 0.3, 0.25)$ and $\boldsymbol{\lambda}_1 = (0.7, 0.6, 0.5, 0.4)$ for data set 2 under the SRF. For data set 3, generated under the SF model, we set $\lambda_x = 0.7$ and $\boldsymbol{\lambda}_0 = (0.6, 0.6, 0.5, 0.5)$ and $\boldsymbol{\lambda}_1 = -\boldsymbol{\lambda}_0$. The settings for $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ imply a full covariance matrix for the potential outcomes with the covariances across the potential outcomes varying over time under SRF and SF models, compared to the more restrictive compound symmetry structure under the SRI. The implied comparable marginal correlations between latent utility and potential outcomes in data sets 1 to 3, marginalized over the latent factor or the random intercepts respectively, are

$$\begin{aligned}
&\mathsf{Cor}(x_i^*, \mathbf{y}_{0i}) = (0.37, 0.31, 0.25, 0.29) \qquad &&\mathsf{Cor}(x_i^*, \mathbf{y}_{1i}) = (-0.44, -0.37, -0.30, -0.22)\\
&\mathsf{Cor}(x_i^*, \mathbf{y}_{0i}) = (0.47, 0.41, 0.34, 0.27) \qquad &&\mathsf{Cor}(x_i^*, \mathbf{y}_{1i}) = (-0.49, -0.43, -0.36, -0.28)\\
&\mathsf{Cor}(x_i^*, \mathbf{y}_{0i}) = (0.44, 0.44, 0.40, 0.40) \qquad &&\mathsf{Cor}(x_i^*, \mathbf{y}_{1i}) = (-0.30, -0.30, -0.26, -0.26).
\end{aligned}$$

For each data set Bayesian inference was carried out under each of the three model specifications with variable selection on the regression effects. The reported results are based on 10,000 iterations, following 10,000 burn-in iterations of the corresponding MCMC algorithm described in Section 4. For a faster convergence of the sampler, the first 5000 burn-in iterations are drawn from the full model which enables the MCMC chain to reach regions of higher posterior density without the additional computational burden of variable selection. As common in the variable selection literature we do not perform variable selection on the intercept.

B.2 Simulation Results

B.2.1 Results for Regression Effects

In Table 6 we report the estimates of the regression effects in the potential outcome models for all three data sets (row blocks) under the SRI, SRF and SF models (column blocks). The diagonal cell blocks in the table refer to the estimates when the inference is based on the correct model. Bold numbers indicate biased estimates where the true value is not contained in the 99% credibility interval obtained from the posterior distribution. A star indicates that the inclusion probability is estimated above 0.5, suggesting that the covariate should be included in the model.

The results show that the true parameters are recovered well when the correct model is applied for the inference. However, when the inference is based on an incorrect model specification (off-diagonal cell blocks), the estimated regression effects in the potential outcome models are partially effected by model mis-specification. As shown in the next section this is due to the different assumptions about the dependence structures across the three models.

Table 6: Results outcome equation: posterior means, standard deviations (in parentheses) of regression effects

	SRI Model		SRF Model		SF Model	
	$(\mu, \boldsymbol{\gamma})$	$(\kappa, \boldsymbol{\theta})$	$(\mu, \boldsymbol{\gamma})$	$(\kappa, \boldsymbol{\theta})$	$(\mu, \boldsymbol{\gamma})$	$(\kappa, \boldsymbol{\theta})$
data 1						
μ, κ	2.998 (0.007)	-0.505 (0.015)*	2.997 (0.007)	-0.501 (0.016)*	2.991 (0.006)	**-0.617** (0.013)*
v_1	1.007 (0.004)*	-0.000 (0.001)	1.007 (0.004)*	0.000 (0.001)	**1.019** (0.004)*	0.000 (0.003)
v_2	0.000 (0.002)	0.186 (0.013)*	0.000 (0.002)	0.185 (0.013)*	0.000 (0.002)	0.185 (0.013)*
$t = 2$	0.104 (0.006)*	-0.098 (0.013)*	0.105 (0.006)*	-0.098 (0.014)*	**0.124** (0.004)*	**-0.060** (0.011)*
$t = 3$	0.163 (0.006)*	0.000 (0.002)	0.163 (0.006)*	0.000 (0.003)	**0.203** (0.004)*	**0.073** (0.011)*
$t = 4$	0.216 (0.006)*	0.105 (0.013)*	0.218 (0.006)*	0.099 (0.014)*	**0.280** (0.004)*	**0.198** (0.011)*
data 2						
μ, κ	**2.966** (0.006)	-0.515 (0.013)*	2.998 (0.006)	-0.503 (0.012)*	3.012 (0.005)	**-0.557** (0.009)*
v_1	0.998 (0.003)*	0.000 (0.001)	0.998 (0.003)*	-0.000 (0.001)	1.006 (0.002)*	0.000 (0.000)
v_2	0.000 (0.001)	0.202 (0.009)*	0.000 (0.000)	0.201 (0.009)*	0.000 (0.001)	0.201 (0.009)*
$t = 2$	**0.127** (0.007)*	-0.106 (0.015)*	0.110 (0.006)*	-0.096 (0.014)*	0.108 (0.005)*	-0.090 (0.011)*
$t = 3$	**0.196** (0.006)*	0.004 (0.012)	0.161 (0.006)*	0.000 (0.003)	0.158 (0.005)*	0.000 (0.002)
$t = 4$	**0.249** (0.006)*	**0.145** (0.014)*	0.194 (0.006)*	0.118 (0.014)*	0.208 (0.005)*	**0.145** (0.011)*
data 3						
μ, κ	**2.942** (0.006)	-0.476 (0.017)*	**2.959** (0.006)	-0.476 (0.017)*	2.997 (0.007)	-0.513 (0.010)*
v_1	**0.987** (0.005)*	0.010 (0.012)	**0.984** (0.005)*	0.017 (0.013)*	1.000 (0.003)*	-0.000(0.002)
v_2	0.000 (0.001)	0.195 (0.009)*	0.000 (0.001)	0.195 (0.009)*	0.000 (0.001)	0.195 (0.009)*
$t = 2$	0.090 (0.007)*	-0.064 (0.017)*	0.095 (0.005)*	-0.072 (0.015)*	0.108 (0.005)*	-0.093 (0.010)*
$t = 3$	**0.198** (0.006)*	-0.001 (0.006)	0.157 (0.006)*	-0.000 (0.003)	0.153 (0.004)*	-0.000 (0.002)
$t = 4$	**0.240** (0.006)*	0.123 (0.015)*	0.201 (0.007)*	0.117 (0.016)*	0.205 (0.005)*	0.105 (0.011)*

Here we observe biased estimates of the intercept μ, the constant treatment effect κ and the effects of v_1 and panel time. An exception is the SRF model for data set 1. Being more general than the SRI model, for the data set 1 it yields almost identical results to the SRI model, whereas for data set 2 some panel effects and their modifications are biased under the

SRI model. Further, for both data sets 1 and 2 estimates of intercept modification and panel time effects are biased when employing the SF model. Finally, for data set 3 both the SRI and the SRF model yield biased estimates of the intercept μ, as well as the effect of variable v_1. Mis-specification may affect correct selection of variables as a result of biased estimates. For example, in data set 1 the SF model incorrectly selects the effect modification for $t = 3$, while in data set 3 the effect modification of v_1 is incorrectly selected under the SRF model with the inclusion probability estimated at 0.692. Under the SRI model the smaller bias results in an inclusion probability just below 0.5.

In the case of the selection model, the estimates for the regression effects are almost identical across the three models for each data set, suggesting that the estimation of the parameters as well as variable selection in the selection equation is not affected by the mis-specification. In Table 7 below we report the posterior means and the sampling standard deviations for the regression coefficients in the selection equation. Effects for which the inclusion probability is above 0.5, implying that the corresponding effect is selected into the model are marked with '*'. As the variance of the latent utility σ_x is restricted to 1 in the SR models while the SF model specification implies that $\sigma_x = \sqrt{1 + \lambda_x^2}$, we report the estimation results for $\boldsymbol{\alpha}/\sigma_x$. In data set 3, generated from the SF model, the (true) rescaled regression coefficients are $\boldsymbol{\alpha}/\sqrt{1 + \lambda_x^2} = (-0.74, 0.66, 0, 1.23)$. For data set 3 the posterior mean of factor loading λ_x was

Table 7: Results selection equation: posterior means, standard deviations (in parentheses) of regression effects

data		SRI Model	SRF Model	SF Model
1	intercept	-0.91 (0.010)	-0.90 (0.010)	-0.91 (0.010)
	v_1	0.80 (0.008)*	0.80 (0.008)*	0.80 (0.008)*
	v_2	0.00 (0.003)	0.00 (0.002)	0.00 (0.003)
	z	1.51 (0.014)*	1.50 (0.013)*	1.50 (0.014)*
2	intercept	-0.91 (0.010)	-0.90 (0.009)	-0.90(0.010)
	v_1	0.81 (0.008)*	0.80 (0.007)*	0.80 (0.008)*
	v_2	-0.00 (0.002)	0.00 (0.002)	0.00 (0.001)
	z	1.52 (0.014)*	1.50 (0.012)*	1.50 (0.013)*
3	intercept	-0.72 (0.010)	-0.73 (0.008)	-0.72 (0.010)
	v_1	0.66 (0.007)*	0.66 (0.007)*	0.66 (0.007)*
	v_2	0.00 (0.003)	0.00 (0.002)	0.00 (0.004)
	z	1.22 (0.013)*	1.24 (0.012)*	1.22 (0.013)*

$0.681\ (0.021)$ in the SF model.

B.2.2 Results for Dependence Structures

The main differences between the three models are their assumptions with respect to the dependence structure within the latent outcome vectors and between outcomes and latent utility. We recall that the assumptions of the switching regression and shared factor models with respect to the dependence between the latent utility and potential outcomes are rather contrary: in the shared factor model all correlation between latent utility and potential outcomes is attributed to the latent factor which also determines the dependence structure within the potential outcome vectors. In the switching regression models, the latent variables only capture the dependence within the potential outcomes, whereas dependence between the error terms of the potential outcome vector and the latent utility is captured by the correlations $\boldsymbol{\rho}_j$.

It is therefore more suitable to focus on the marginal correlations between latent utility and the panel outcomes (marginalizing over latent factor or random intercept) when comparing the estimated correlations across the three model specifications. These are given in equation (3.14).

In Table 8 we report the differences between estimated and the true correlations. Estimates, where the true, data generating value is not included in the 99% credibility interval are given in bold.

Table 8: Marginal correlation between latent utility and outcome: difference between posterior means and true values, standard deviations (in parentheses)

		SRI Model		SRF Model		SF Model	
data	t	treatment 0	treatment 1	treatment 0	treatment 1	treatment 0	treatment 1
1	1	-0.001 (0.011)	0.016 (0.011)	-0.002 (0.009)	0.013 (0.011)	**-0.029** (0.008)	**0.159** (0.007)
	2	0.011 (0.011)	0.012 (0.011)	0.011 (0.012)	0.009 (0.013)	**0.031** (0.008)	**0.084** (0.007)
	3	0.024 (0.011)	0.011 (0.011)	0.024 (0.011)	0.003 (0.013)	**0.096** (0.008)	0.008 (0.007)
	4	0.028 (0.012)	-0.008 (0.013)	0.029 (0.012)	-0.006 (0.015)	**0.158** (0.008)	**-0.065** (0.007)
2	1	**-0.103** (0.013)	**0.064** (0.011)	-0.021 (0.011)	0.007 (0.010)	0.013 (0.008)	**0.057** (0.008)
	2	**-0.034** (0.013)	**0.040** (0.012)	0.007 (0.012)	-0.007 (0.012)	**0.034** (0.007)	**0.039** (0.008)
	3	0.032 (0.013)	-0.020 (0.012)	0.022 (0.012)	-0.018 (0.011)	**0.050** (0.007)	**0.035** (0.008)
	4	**0.057** (0.014)	**-0.056** (0.013)	-0.019 (0.013)	0.001 (0.013)	**0.068** (0.007)	-0.008 (0.008)
3	1	**-0.107** (0.008)	**0.037** (0.017)	**-0.075** (0.007)	0.017 (0.168)	-0.009 (0.009)	0.012 (0.008)
	2	**-0.147** (0.008)	0.014 (0.017)	**-0.103** (0.005)	-0.004 (0.182)	-0.007 (0.009)	0.007 (0.008)
	3	**-0.026** (0.006)	-0.034 (0.018)	**-0.074** (0.006)	-0.007 (0.170)	-0.008 (0.009)	0.001 (0.007)
	4	**-0.054** (0.006)	-0.033 (0.019)	**-0.095** (0.010)	0.001 (0.189)	-0.009 (0.009)	0.012 (0.007)

As expected, estimates correspond well to the true values when the appropriate model is used for the analysis. Further, as the SRF model is more flexible than the SRI model with respect to the dependence structure of panel outcomes, it can capture the correlation structure implied by the SRI, yielding essentially the same estimates for data set 1. In the SF model the assumption $\mathsf{Cov}(x_i^*, \mathbf{y}_{ji}) = \lambda_x \boldsymbol{\lambda}_j$ with the covariance being proportional to $\boldsymbol{\lambda}_j$ by factor λ_x can be too restrictive and the estimation of $\boldsymbol{\lambda}_j$ be driven by the outcome covariances. For data sets 1 and 2, where the correlations to the latent utility are not proportional to $\boldsymbol{\lambda}_j$ for t=1,..., 4, the SF model yields biased estimates. Although the SR models allow for more flexible correlation structures, the estimated marginal correlations in data set 3 under these two models also deviate considerably from the true values. In this case the positive definiteness condition for $\mathsf{Cov}(\boldsymbol{\varepsilon}_{ji}, \eta_i)$ restricts the range of possible correlations: Given the dependence structure of $\mathbf{y}_{ji}$, where the latent factor accounts for at least 50% of the composite error, the true marginal correlations between $\mathbf{y}_{ji}$ and x_i^* could be obtained in an SR model only with very high correlations between the idiosyncratic error of the outcomes and the error in the latent utility model. For instance, the value for correlation parameter $\boldsymbol{\rho}_0$ yielding the true marginal correlations would be $\boldsymbol{\rho}_0 \approx (0.69, 0.69, 0.57, 0.57)$ which is far beyond the region $\sum_{t=1}^{T} \rho_{j,t}^2 < 1$ required for positive definiteness.

Table 9 reports the bias (true values – estimated values) for the elements of covariance matrices of the outcome vectors with values not included in the 99% credibility interval given in bold. Differences are small if the correct (data generating) model is applied to analyze the data. Results for data sets 2 and 3 indicate that the compound symmetry structure implied by a random intercept is too restrictive to capture the dependence structure in models with a

Table 9: Difference between true and estimated matrices Ω_0 and Ω_1, standard deviations (in parentheses)

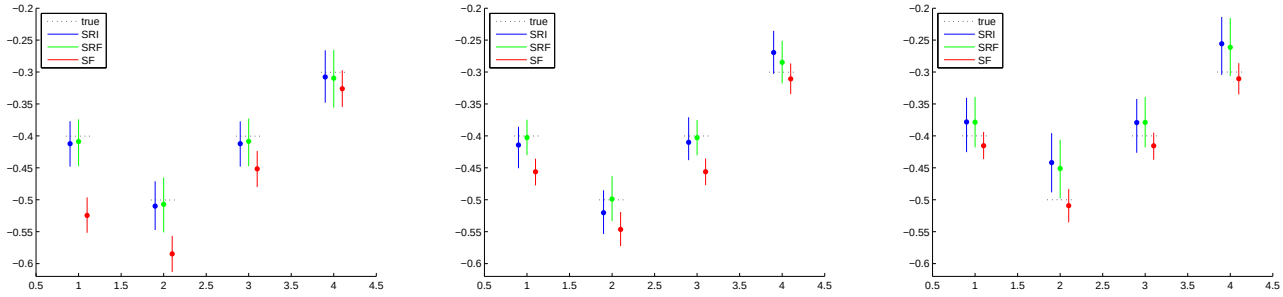
data		SRI Model				SRF Model				SF Model			
1	Ω_0	-0.002	-0.004	-0.004	-0.004	-0.001	-0.001	-0.004	-0.004	-0.001	-0.006	-0.011	-0.013
		(0.005)	(0.004)	(0.004)	(0.004)	(0.006)	(0.005)	(0.005)	(0.004)	(0.006)	(0.005)	(0.005)	(0.005)
			-0.004	-0.004	-0.004		-0.001	-0.003	-0.002		-0.006	-0.011	-0.013
			(0.005)	(0.004)	(0.004)		(0.006)	(0.005)	(0.005)		(0.006)	(0.005)	(0.005)
				-0.002	-0.004			-0.004	-0.005			-0.015	-0.018
				(0.005)	(0.004)			(0.006)	(0.004)			(0.006)	(0.005)
					-0.003				-0.004				-0.020
					(0.005)				(0.006)				(0.006)
	Ω_1	-0.000	-0.005	-0.005	-0.005	0.001	-0.006	-0.012	0.003	0.064	0.026	0.014	0.019
		(0.016)	(0.010)	(0.010)	(0.010)	(0.019)	(0.014)	(0.014)	(0.014)	(0.019)	(0.012)	(0.013)	(0.013)
			-0.001	-0.005	-0.005		-0.005	-0.016	-0.002		0.027	0.004	0.009
			(0.016)	(0.010)	(0.010)		(0.019)	(0.015)	(0.014)		(0.017)	(0.013)	(0.013)
				-0.011	-0.005			-0.022	-0.007			-0.015	-0.034
				(0.015)	(0.010)			(0.018)	(0.014)			(0.018)	(0.013)
					-0.002				0.005				-0.058
					(0.015)				(0.018)				(0.017)
2	Ω_0	**0.027**	**0.037**	**0.017**	**-0.002**	-0.003	-0.004	-0.002	0.001	-0.008	-0.008	-0.007	-0.003
		(0.003)	(0.002)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)	(0.002)	(0.006)	(0.004)	(0.003)	(0.003)
			0.014	0.002	**-0.015**		0.002	-0.002	0.001		-0.002	-0.006	-0.003
			(0.003)	(0.002)	(0.002)		(0.003)	(0.002)	(0.002)		(0.004)	(0.005)	(0.005)
				-0.011	**-0.028**			-0.005	0.001			-0.008	-0.002
				(0.003)	(0.002)			(0.003)	(0.002)			(0.003)	(0.002)
					-0.024				-0.001				-0.005
					(0.003)				(0.002)				(0.003)
	Ω_1	**0.082**	**0.125**	0.055	-0.015	0.004	-0.011	0.007	-0.013	0.029	0.002	0.019	-0.004
		(0.013)	(0.006)	(0.006)	(0.006)	(0.015)	(0.011)	(0.009)	(0.009)	(0.014)	(0.010)	(0.009)	(0.008)
			0.026	0.006	**-0.055**		-0.016	-0.002	-0.018		0.004	0.014	-0.006
			(0.012)	(0.006)	(0.006)		(0.014)	(0.009)	(0.009)		(0.013)	(0.008)	(0.008)
				-0.025	**-0.095**			-0.005	-0.005			0.013	0.006
				(0.012)	(0.006)			(0.012)	(0.007)			(0.012)	(0.006)
					-0.05				0.002				-0.001
					(0.012)				(0.011)				(0.011)
3	Ω_0	**0.058**	**0.077**	**0.017**	**0.017**	**0.018**	**0.017**	**0.016**	**0.016**	-0.000	-0.003	0.000	-0.000
		(0.004)	(0.003)	(0.003)	(0.003)	(0.006)	(0.004)	(0.004)	(0.004)	(0.006)	(0.005)	(0.004)	(0.004)
			0.060	**0.017**	**0.017**		**0.019**	**0.015**	**0.015**		-0.002	-0.002	-0.002
			(0.004)	(0.003)	(0.003)		(0.005)	(0.004)	(0.004)		(0.006)	(0.004)	(0.004)
				-0.023	**-0.033**			**0.014**	**0.015**			0.001	0.000
				(0.004)	(0.003)			(0.004)	(0.003)			(0.005)	(0.004)
					-0.025				**0.012**				-0.003
					(0.003)				(0.005)				(0.005)
	Ω_1	**0.024**	**0.061**	0.001	0.001	0.003	0.005	-0.001	0.013	0.003	0.007	0.001	0.015
		(0.013)	(0.008)	(0.008)	(0.008)	(0.014)	(0.011)	(0.010)	(0.010)	(0.013)	(0.009)	(0.009)	(0.008)
			0.011	0.001	0.001		-0.014	-0.006	0.007		-0.012	-0.006	0.008
			(0.013)	(0.008)	(0.008)		(0.015)	(0.010)	(0.010)		(0.013)	(0.008)	(0.009)
				-0.032	**-0.049**			-0.011	0.003			-0.010	0.003
				(0.013)	(0.008)			(0.013)	(0.009)			(0.012)	(0.008)
					-0.025				0.005				0.008
					(0.013)				(0.013)				(0.012)

latent factor, as deviations of the estimates from the true values can be relatively large.

B.2.3 Results for Average Treatment Effects

Figure 6 shows the true and the estimated average treatment effects for the three data sets that were obtained under the different model specifications. As expected, the estimated effects are almost identical to the true values if the correct model is employed for inference and we again observe that the SRF model yields almost identical estimates to the SRI model for data set 1. In these cases the true values of all the model parameters, including the covariance and correlations parameters, were well recovered.

However, in the remaining cases the biased parameter estimates resulting from a model mis-specification discussed above have implications for the estimation of the average treatment effects. Figure 6 shows that these incorrect model based estimates exhibit larger deviations from the true effects. If the regression effects in the potential outcome models are biased and the dependence structure between outcome and latent utility cannot be captured by the employed model, the estimates of the ATE are negatively affected.



(a) SRI data **(b)** SRF data **(c)** SF data

Figure 6: Posterior distributions of the average treatment effect $ATE(t)$ for $t = 1, \ldots, 4$ under the SR and SF models; estimated posterior means $\widehat{ATE}(t)$ (dots) and 95% credibility intervals (vertical bars) in comparison to the true $ATE(t)$ (vertical dashed line). Panel (a), (b), and (c) refer, respectively, to data 1, data 2, and data 3.

In these cases one or more of the average treatment effects are biased, as can be seen from Table 10. The table reports the difference between the true and estimated average treatment effects. Bold numbers indicate estimates when the true value is not contained in the 99% credibility interval and slanted bold numbers indicate estimates where the true value is not contained in the 95% credibility interval. Our results also indicate that capturing the

Table 10: Average Panel Treatment Effects: difference between posterior mean and true value, standard deviations (in parentheses)

data	t	SRI	SRF	SF
1	1	-0.012 (0.014)	-0.009 (0.014)	**-0.124** (0.011)
	2	-0.001 (0.015)	-0.007 (0.017)	**-0.084** (0.011)
	3	-0.012 (0.014)	-0.008 (0.014)	**-0.051** (0.011)
	4	-0.007 (0.016)	-0.009 (0.018)	***-0.026*** (0.011)
2	1	-0.014 (0.012)	-0.003 (0.011)	**-0.056** (0.008)
	2	-0.020 (0.013)	0.001 (0.013)	**-0.046** (0.010)
	3	-0.010 (0.012)	-0.003 (0.011)	**-0.056** (0.008)
	4	***0.031*** (0.013)	0.015 (0.013)	-0.011 (0.010)
3	1	0.022 (0.017)	0.021 (0.017)	-0.015 (0.009)
	2	**0.058** (0.018)	**-0.049** (0.021)	-0.009 (0.010)
	3	0.028 (0.017)	0.021 (0.017)	-0.016 (0.009)
	4	***0.044*** (0.018)	***0.039*** (0.018)	-0.011 (0.010)

correlation between latent utility and outcome might be more important than the correlation structure within the outcome vectors in the context of the average treatment effects: in data set 2 the SRI model yields only slightly biased estimates compared to those from the SF model.

B.3 Inefficiency factors and estimation issues

The results discussed above show that the sampler performs well, with posterior estimates recovering the true parameter values well if the correct model is used for inference. We now

turn to other aspects of the performances for the three samplers.

For the SF model inefficiency factors are satisfactory for all parameters. As noted in Appendix B.2.1, the marginal variance of the latent utility is $1+\lambda_x^2$ in the SF model, and therefore only parameters in the rescaled probit model can be compared to the SR model. Whereas regression effects in the selection equation as well as the factor loading λ_x can suffer from high autocorrelations, the parameters in the rescaled probit model exhibit small inefficiency factors from 2 to 11.

Inefficiency factors are generally higher in the SR models estimates. While the sampler for the SF model requires only Gibbs steps due to the simpler modeling of the correlation between latent utility and outcomes, the SR models involve MH-steps usually associated with higher autocorrelation in the chain. In particular the inefficiency factors for the correlation parameters are large, leading to inefficient estimation of the marginal correlations (the highest inefficiencies were observed in data set 1 with up to roughly 300 for the SRI and the SRF model). Also, draws of the regression effects $\boldsymbol{\alpha}$ in the selection equation show higher autocorrelations than under the SF model (inefficiency factors up to 40-50 in SRI and the SRF model).

Finally, due to the simpler sampling scheme computation is much faster for the SF model than for the SR models by a factor of roughly 20, with the SRF model being the computationally the most intensive model to fit.

C Application

C.1 Summary Statistics

Table 11 shows the distribution of the data in the sample by (contribution) year and panel period for both treatment groups.

year	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	total*
2000	825/207	0/0	0/0	0/0	0/0	0/0	1,032
2001	5,676/381	825/207	0/0	0/0	0/0	0/0	7,089
2002	5,985/697	5,676/381	825/207	0/0	0/0	0/0	13,771
2003	1,088/4,129	5,985/697	5,676/381	825/207	0/0	0/0	18,988
2004	457/7,179	1,088/4,129	5,985/697	5,676/381	825/206	0/0	26,622
2005	84/4,343	457/7,179	1,088/4,129	5,985/697	5,676/381	820/206	31,026
2006	0/0	84/4,343	457/7,179	1,088/4,129	5,985/694	5,643/379	29,960
2007	0/0	0/0	84/4,343	457/7,179	1,086/4,114	5,943/693	23,899
2008	0/0	0/0	0/0	84/4,343	454/7,109	1,075/4,087	17,152
total	14,115/16,936	14,115/16,936	14,115/16,936	14,115/16,936	13,985/12,504	13,481/5,365	169,539

Table 11: Distribution of mothers in the sample by contribution year and panel period for each treatment group ($x=0$ / $x=1$). total* refers to all observations.

C.2 Priors

Section 4.1 specifies the structure of the prior distribution for each model. Since we implement variable selection on a subset of the regression parameters in the selection and outcome equations, we assume spike and slab priors as described in equations (4.1) and (4.2) for these coefficients. For each regression coefficient we have a binary indicator that takes value 1 with some probability that has a uniform hyper prior $\mathcal{U}[0,1]$. If the indicator takes value 1 the effect is assigned the slab component of the prior, that is assumed a $\mathcal{N}(0,5)$ distribution. If

the indicator takes value zero the effect is assigned a point mass at zero using the Dirac spike Δ_0.

As we do not perform variable selection on the intercept and the year effects we specify standard normal priors. The prior for the intercept is $\mathcal{N}(0, 1000)$ and for the linear calendar and quadratic calendar year effects we use tighter normal priors, $\mathcal{N}(0, 0.1)$.

For the correlation parameters under the SRI and SRF we assume prior independence with $\rho_{j,t} \sim \mathcal{N}(0, 1)$ truncated to the region yielding a positive definite Cholesky factor. For the SRI model we use independent priors for the idiosyncratic variances $\sigma^2_{j,t}$, $t = 1, ...T$ such that $\log \sigma_{j,t} \sim \mathcal{N}(0, 1)$ and specify an inverse Gamma prior for the random intercept variance parameter, $D_j \sim \mathcal{G}^{-1}(3, 1)$. In the SRF model the random factor loadings $\lambda_{j,t}$ are independent with standard Normal priors $\mathcal{N}(0, 1)$.

Finally, in the SF model we assign flat (improper) priors of the form $\mathcal{G}^{-1}(0, 0)$ to the variance parameters $\sigma^2_{j,t}$. For the random factor loadings $\lambda_{j,t}$ and λ_x we use independent standard Normal priors $\mathcal{N}(0, 1)$.

C.3 Results

	SRI results for selection and earnings model			
	selection model		earnings model	
	mean (sd.)	prob*	treatment 0	+ treatment 1
intercept	-1.529 (0.033)	–	9.351 (0.013)	-0.141 (0.013)*
z	2.792 (0.023)	1.000	–	–
child 2	0.025 (0.034)	0.396	-0.000 (0.001)	-0.000 (0.002)
child >3	-0.018 (0.038)	0.217	0.000 (0.001)	0.000 (0.002)
exp	0.072 (0.032)	0.899	-0.096 (0.008)*	0.004 (0.010)
blue collar	-0.066 (0.042)	0.778	-0.120 (0.007)*	0.000 (0.002)
int. exp/blue	-0.010 (0.033)	0.111	0.002 (0.008)	0.016 (0.018)
base-earn Q2	0.001 (0.007)	0.025	0.064 (0.007)*	0.000 (0.002)
base-earn Q3	-0.000 (0.004)	0.018	0.281 (0.011)*	-0.039 (0.016)*
base-earn Q4	-0.131 (0.028)	0.999	0.609 (0.010)*	-0.106 (0.013)*
eq. emp.	–	–	0.040 (0.005)*	0.000 (0.001)
panel t=2	–	–	0.070 (0.006)*	0.099 (0.007)*
panel t=3	–	–	0.117 (0.007)*	0.125 (0.006)*
panel t=4	–	–	0.165 (0.010)*	0.146 (0.007)*
panel t=5	–	–	0.219 (0.012)*	0.151 (0.007)*
panel t=6	–	–	0.271 (0.014)*	0.181 (0.008)*
$(\text{year} - 1999)$	–	–	0.039 (0.004)	
$(\text{year} - 1999)^2$	–	–	-0.004 (0.0002)	

Table 12: SRI: Results for selection and earnings model are reported in terms of posterior means and standard deviations (in parentheses). The inclusion of regression effects based on posterior inclusion probabilities > 0.5 is indicated by "*".

	Covariance parameters in the earnings model							
	SRF Model				SF Model			
t	σ_0^2	${\sigma_1}^2$	$\lvert\lambda_0\rvert$	$\lvert\lambda_1\rvert$	σ_0^2	σ_1^2	$\lvert\lambda_0\rvert$	$\lvert\lambda_1\rvert$
1	0.123 (0.002)	0.102 (0.001)	0.340 (0.004)	0.290 (0.003)	0.123 (0.002)	0.102 (0.001)	0.341 (0.004)	0.291 (0.003)
2	0.072 (0.001)	0.044 (0.001)	0.379 (0.003)	0.353 (0.003)	0.072 (0.001)	0.043 (0.001)	0.381 (0.003)	0.356 (0.003)
3	0.044 (0.001)	0.019 (0.000)	0.411 (0.003)	0.385 (0.002)	0.044 (0.001)	0.019 (0.000)	0.413 (0.003)	0.387 (0.002)
4	0.026 (0.000)	0.028 (0.000)	0.426 (0.003)	0.384 (0.002)	0.026 (0.000)	0.028 (0.000)	0.428 (0.003)	0.387 (0.002)
5	0.028 (0.001)	0.045 (0.001)	0.413 (0.003)	0.366 (0.003)	0.028 (0.001)	0.045 (0.001)	0.415 (0.003)	0.368 (0.003)
6	0.046 (0.001)	0.058 (0.001)	0.395 (0.003)	0.356 (0.004)	0.046 (0.001)	0.058 (0.001)	0.397 (0.003)	0.358 (0.004)

Table 13: Variances Σ_j and absolute factor loadings $|\boldsymbol{\lambda_j}|$ for $j = 0, 1$ in potential earnings models for SRF and SF models; posterior means, standard deviations (in parentheses).

	Covariance matrices under the SRF model											
	treatment 0						treatment 1					
t	1	2	3	4	5	6	1	2	3	4	5	6
1	**0.239**	0.129	0.140	0.145	0.140	0.134	**0.186**	0.102	0.111	0.111	0.106	0.103
2	0.129	**0.216**	0.156	0.162	0.157	0.150	0.102	**0.168**	0.136	0.135	0.129	0.125
3	0.140	0.156	**0.213**	0.175	0.170	0.162	0.111	0.136	**0.167**	0.148	0.141	0.137
4	0.145	0.162	0.175	**0.208**	0.176	0.168	0.111	0.135	0.148	**0.176**	0.140	0.137
5	0.140	0.157	0.170	0.176	**0.199**	0.163	0.106	0.129	0.141	0.140	**0.179**	0.130
6	0.134	0.150	0.162	0.168	0.163	**0.202**	0.103	0.125	0.137	0.137	0.130	**0.185**

Table 14: Posterior means of covariance matrices Ω_0 and Ω_1, marginalized over the random factor. Results are given for the SRF model as the SF model has essentially identical estimates of the covariance matrices elements due to the almost identical absolute values of the factor loadings and idiosyncratic error variances.

	Covariance matrices under the SRI model											
	treatment 0						treatment 1					
t	1	2	3	4	5	6	1	2	3	4	5	6
1	**0.288**	0.164	0.164	0.164	0.164	0.164	**0.241**	0.136	0.136	0.136	0.136	0.136
2	0.164	**0.234**	0.164	0.164	0.164	0.164	0.136	**0.178**	0.136	0.136	0.136	0.136
3	0.164	0.164	**0.207**	0.164	0.164	0.164	0.136	0.136	**0.157**	0.136	0.136	0.136
4	0.164	0.164	0.164	**0.193**	0.164	0.164	0.136	0.136	0.136	**0.167**	0.136	0.136
5	0.164	0.164	0.164	0.164	**0.194**	0.164	0.136	0.136	0.136	0.136	**0.182**	0.136
6	0.164	0.164	0.164	0.164	0.164	**0.211**	0.136	0.136	0.136	0.136	0.136	**0.194**

Table 15: Posterior means of covariance matrices Ω_0 and Ω_1 of the potential earnings sequences for the SRI model, marginalized over the random effects.

t	1	2	3	4	5	6
1	-0.099	-0.121	-0.132	-0.132	-0.126	-0.122
2	-0.111	-0.136	-0.148	-0.147	-0.140	-0.136
3	-0.120	-0.147	-0.160	-0.160	-0.152	-0.148
4	-0.124	-0.152	-0.166	-0.166	-0.158	-0.153
5	-0.121	-0.148	-0.161	-0.160	-0.153	-0.149
6	-0.115	-0.141	-0.154	-0.153	-0.146	-0.142

Table 16: Shared factor model: Posterior mean of the implied covariance $\mathsf{Cov}(\mathbf{y}_{0,i}, \mathbf{y}_{1,i}) = \Omega_{01}$ between the two potential outcome sequences.