Understanding Women's Wage Growth using Indirect Inference with Importance Sampling

Robert M. Sauer*and Christopher Taber†

February 3, 2021

Abstract

The goal of this work is to investigate the effects of time out of the labor market for childcare on women's lifecycle wage growth. We develop a dynamic lifecycle model of human capital, fertility, and labor supply for women. We estimate by indirect inference using importance sampling and formalize the use of this procedure. The results indicate a modest effect of fertility-induced non-employment spells on human capital accumulation. The difference in human capital among prime age women would be approximately 2.4% higher at its peak if the relationship between fertility and working were eliminated, and 4.7% higher if the relationship between marriage and fertility was also eliminated.

Keywords: Gender Premium, Wage Growth, Indirect Inference, Importance Sampling, Simulation Estimation

JEL Codes: C13, C15, C18, C35

*Department of Economics, Royal Holloway, University of London. Email: robertmsauer@protonmail.com
†Department of Economics, University of Wisconsin. Email: taberchris@gmail.com

1 Introduction

The main goal of this work is to understand the relationship between female human capital accumulation and fertility. It is well known that women have a less steep wage profile than men. Presumably some of this difference is due to the fact that women take time out of the labor market during pregnancy and for childcare. We quantify the importance of this effect on human capital accumulation. Specifically, we estimate a Markov model and then simulate the difference in wage growth under the counterfactual that women no longer take time out of the labor market for children and marriage. We find that at its peak, human capital would be 2.4% higher if the fertility effect were eliminated and 4.7% higher if the marriage effect was also eliminated. While these effects are substantial, they are small compared to the raw difference in wages between men and women. This leaves plenty of scope for other channels such as discrimination in the form of glass ceilings.

A second goal of the study is to formalize the use of importance sampling to estimate indirect inference models. We develop a general version of indirect inference and an estimator using importance sampling. We show that this method produces consistent estimates and derive the standard errors for this promising new technique.

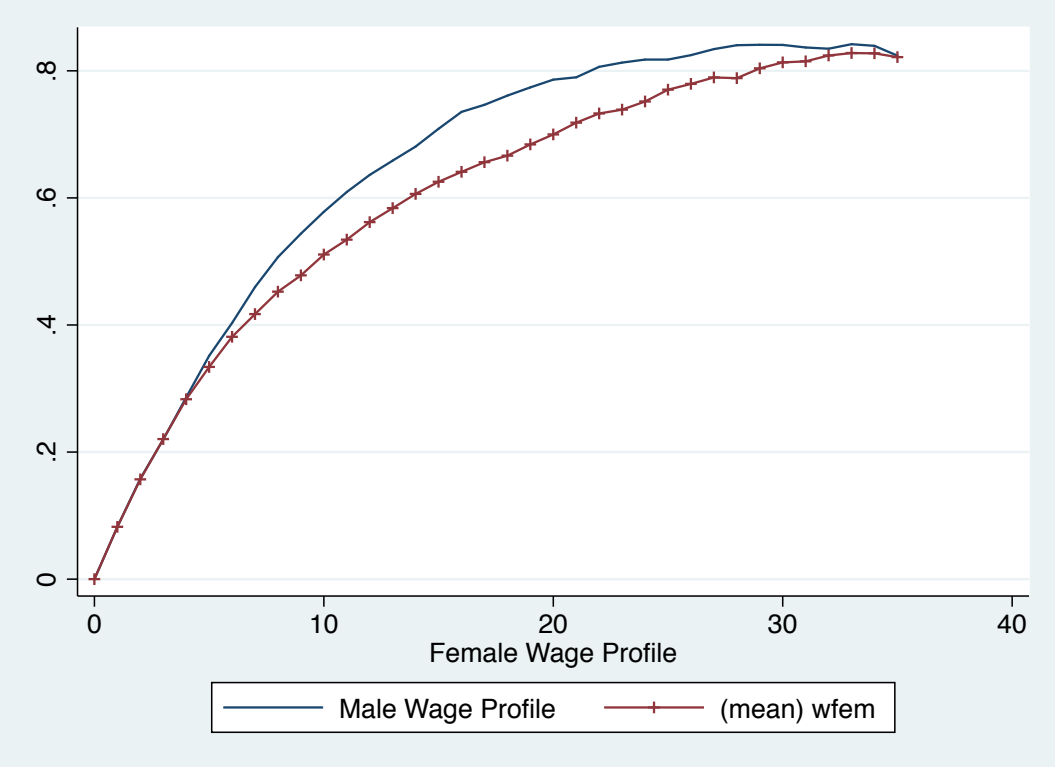
The basic empirical motivation can be seen in Figure 1. We run a regression of log wages on dummy variables for years of potential experience and individual fixed effects for white men and white women using the Survey of Income and Program Participation (SIPP). The predicted profiles are plotted, normalizing log wages at entry to zero. Two things can be seen from the figure. First, as has been previously established,[1] wages increase more quickly for men than for women at the beginning of the lifecycle.[2] Second, while wages diverge in the middle, they eventually converge towards the end of the lifecycle. One possible explanation for this pattern is fertility - when women have children they often leave the labor market and then re-enter as their children age. This could cause wage growth to slow during childrearing and then pick up again after re-entry.

The figure raises a fundamental question in labor economics: what leads to curvature in wage growth? Relatedly, why does wage growth slow more quickly for men than for women? If the curvature is driven by *actual experience* then one might expect this gender pattern. When women re-enter the labor force they have less *actual experience* than men

[1] See e.g. Gladden and Taber (2000) among a large literature.

[2] This difference is smaller than what Gladden and Taber (2000) find for the NLSY though the samples are directly comparable and the SIPP covers a later period.

Figure 1: Male and Female Log Wage Profiles



and thus their wages will grow faster. This is consistent with the wage growth of women with *potential experience* over 20 years being faster than the wage growth of men. In contrast, if it is *potential experience* or *age* that is driving the curvature, then women who re-enter would not see faster wage growth. Our empirical specification below allows for both possibilities and measures their quantitative importance.

The model is estimated using indirect inference which is an increasingly common way to estimate complex econometric models. Similar to simulated method of moments, it is a computationally simple technique since it relies on unconditional simulations of the model to obtain structural estimates. However, one of the main practical problems with indirect inference is the computational difficulty of optimizing the objective function when the structural model contains discrete choices (see, e.g., Magnac, Robin, and Visser, 1995, An and Liu, 2000, or Nagypál, 2007). In this case, a step function arises because a small change in structural parameters causes a jump in the metric of distance between the two sets of auxiliary model parameter estimates. A non-smooth objective function precludes the use of gradient-based numerical optimization methods which can often lead to much faster convergence.

In this paper, we explain how the problem of non-smoothness can be solved using Monte

Carlo importance sampling (see e.g. Kloek and van Dijk, 1978, or Kloek and van Dijk, 1978) in a general class of indirect inference models. We smooth the objective function by making use of importance sampling weights in estimation of the auxiliary model on simulated data. The denominator of the weight is the likelihood contribution of each observation in the simulated sample, at an initial trial vector of structural parameters. The denominator remains fixed during minimum distance iterations. The numerator of the weight is the likelihood contribution at the updated trial vector of parameters. The importance sampling weights can be formed with either the exact likelihood of the structural model or a simulated likelihood in case the former is difficult to construct. We show that this alternative technique which is explained in the context of simulated method of moments by Gourieroux and Monfort (1996) and Ackerberg (2009) can be extended to indirect inference to yield structural parameter estimates that are consistent. While this extension is straight forward, in our view it is a very useful approach which should be more widely applied in estimation by indirect inference.

The rest of this paper is organized as follows. In the next section, we briefly discuss the two relevant literatures. In Section 3, the Markov lifecycle model is presented. In Section 4, we show formally how to incorporate importance sampling into indirect inference. Section 5 presents the data. Section 6 discusses how to implement the procedure in the current context. Section 7 presents the empirical results. Section 8 concludes.

2 Background and Previous Work

2.1 Female Wage Growth

There is a large literature on male-female wage differentials. This study differs from the vast majority of previous work in this area because we focus on wage growth rather than wage levels.

Hill (1979) was one of the first to examine the effect of motherhood on wage levels. She initially finds a 7 percent motherhood-wage penalty for white women, but after controlling for productivity characteristics it nearly disappears. She concludes that "the number of children is a good proxy variable for differential work history and labor force attachment for white women" (p. 591). We use this idea for identification in our model. Becker (1985) suggests that a part of the wage gap observed between single and married mothers arises from the choice by married mothers to work in less intensive and more convenient jobs (p. S54). Korenman and Neumark (1992) find no significant effect on wages of having a first

child, but large effects from the second child (between a 10 and 20 percent penalty).

Waldfogel's (1998a, 1998b) findings suggest a motherhood wage penalty of 4.6 percent for the first child and 12.6 percent for two or more children. She also finds that women who have access to family leave upon childbirth are more likely to return to their pre-childbirth employer and, consequently, receive a wage boost that partially offsets the motherhood wage penalty (75 percent of the wage penalty is eliminated). Anderson, Binder, and Krause (2002) find no evidence (in a panel framework) that reduced work effort is at the root of the wage gap. They estimate the wage gap to be 3 percent for mothers with one child and 6 percent for mothers of two or more children. They posit that the wage gap is largely caused by high costs of flexible work schedules for women holding medium office jobs with standard work hours.

Adda, Dustmann, and Stevens (2017) formulate and estimate a dynamic structural model of female labor supply, marriage and fertility choices and use it decompose the career costs of children into several different components. Using data from Germany, they find that roughly three quarters of the 35% reduction in lifetime income derives from foregone earnings while out of the labor force. The remainder is due to lower wages while working, less work experience and depreciation of skills. In addition, Adda, Dustmann, and Stevens (2017) find that skill depreciation rates are higher in mid-career and differ across occupations.

Loughran and Zissimopoulos (2007) concentrate on the effect of marriage and fertility on the wage growth of men and women. Fixed-effects regressions show that not only does marriage reduce female wage levels, but it also reduces female wage growth by four percentage points. A first birth lowers female wages by between two and three percentage points but does not affect wage growth for males or females.

Daniel, Lacuesta, and Rodríguez-Planas (2013) estimate fixed-effects regressions on Spanish data to explore the effects of childbirth on female wages. The results indicate that, compared to childless women, "mothers to be" experience earnings increases of up to 6 percentage points prior to a first-birth. The earnings advantage is then wiped out. It takes another nine years on average for a mother's earnings to return to pre-birth relative levels (relative to childless women). Roughly half of the earnings loss upon becoming a mother is due to less accumulated work experience, as mothers switch to part-time work or take a leave of absence.

Weiss and Gronau (1981) provides a human capital model showing why wage growth might be lower for women. Polachek (1981) presents a model and evidence that women

choose occupations with lower depreciation of human capital. Like us, Light and Ureta (1995) use a more complicated model for experience. They take advantage of the NLSY79 and the long histories. Baum (2002) looks directly at the effect of work interruptions on wages for women. Wilde, Batchelder, and Ellwood (2010) emphasize the difference between low and high skilled workers in the impact of childbearing.

In addition to Adda, Dustmann, and Stevens (2017) discussed above, our work is related to structural models of fertility, labor supply and wages such as Moffitt (1984), Hotz and Miller (1988), Eckstein and Wolpin (1989), Heckman and Walker (1990), Van Der Klaauw (1996), Altug and Miller (1998), Francesconi (2002) Sheran (2007), Keane and Wolpin (2010), Gayle, Hincapie, and Miller (2018), and Blundell, Costa Dias, Meghir, and Shaw (2016). While we are not explicitly structural, our approach is similar. None of these papers focus on the precise question about fertility and wage growth that we do.

There is also a large literature on the motherhood penalty. Additional papers to the ones discussed above include Waldfogel (1997), Lundberg and Rose (2000), Budig and England (2001), Anderson, Binder, and Krause (2003), Gangl and Ziefle (2009), and Pal and Waldfogel (2014).

2.2 Indirect Inference and Importance Sampling

Indirect inference has become a very important tool for estimation of complex econometric models. Key papers are Smith (1990, 1993) and Gourieroux, Monfort, and Renault (1993). The econometrics is discussed in detail in Gourieroux and Monfort (1996). The basic idea is to estimate auxiliary parameters from the data and match them to the model as we discuss in detail in the next section. The main problem with indirect inference that we address in this paper is one in which the mapping between the underlying parameters and simulated auxiliary parameters is not smooth, which complicates estimation and inference. We show how to use importance weight sampling to smooth the objective function. An alternative approach to importance sampling and smoothing is the method of generalized indirect inference (GII) proposed by Bruins et al. 2018. As it will be easier to discuss that paper after introducing our notation, we defer a detailed discussion of GII until Section 4.3.

Importance sampling has a long history. The use of importance sampling with Monte Carlo to simulate expectations goes back at least to Kloek and van Dijk (1978). It has been used for smoothing objective functions in simulated maximum likelihood (see Keane and Sauer, 2010, for discussion). It was discussed by McFadden (1989) as a way to smooth his

simulated method of moments approach. A particularly relevant paper is Ackerberg (2009) which we discuss in in Section 4.3 after introducing our notation. Our main methodological contribution is to extend the importance sampling methodology to a general class of indirect inference models as well as providing some particular examples. Lee (2012), Han (2016), and Fu and Gregory (2019) have applied this approach based on earlier versions of this paper.

3 The Markov Model for Women's Work, Marriage, and Fertility Patterns over the Lifecycle

The model is a continuous time Markov model in which women transition between several states. Individuals can move into and out of work and into and out of marriage. They also potentially give birth to children which influences other variables. Human capital increases while individuals work and falls when they don't. The state variables are

$$\mathcal{S}_{it} \equiv \{t, L_{it}, M_{it}, H_{it}, K_{it}, \{A_{1it}, .., A_{K_{it}it}\}; E_i, \nu_i\} \tag{1}$$

where t is time since labor market entry (i.e. potential experience), L_{it} is a dummy variable for having a job, M_{it} is a dummy variable for being currently married, H_{it} is human capital, K_{it} is the number of children, and A_{jit} is the age of each child. The last two variables in (1) do not change over time. The first is education E_i, which is observed in the data, and the second is unobserved heterogeneity ν_i. The latter is a two dimensional normal random variable, with one dimension loosely anchored to wages and the other to labor supply.

Our model starts at the point a woman finishes school and we assume they are unmarried with no children. They may have a job, which is determined by a logistic function of (E_i, v_i). The transitions are governed by five different hazard rates; the hazard rate for job arrival amongst the non-employed, $\lambda^J(\mathcal{S}_{it})$, the hazard rate for job destruction (leading to non-employment), $\lambda^N(\mathcal{S}_{it})$, the rate of marriage formation, $\lambda^M(\mathcal{S}_{it})$, for divorce, $\lambda_1^D(\mathcal{S}_{it})$, and finally for birth of children, $\lambda^K(\mathcal{S}_{it})$. With some probability women drop out of the labor market precisely at the time of having children which is specified as a logistic function of (E_i, v_i). The ages of both the woman and her children increase with time. Human capital evolves deterministically as a function of the state variables as described below.

All five hazard rates take the basic form,

$$\log\left(\lambda^R(\mathcal{S}_{it})\right) = X_{it}^R(\mathcal{S}_{it})'\beta_0^R \tag{2}$$

for $R \in \{J, N, M, D, K\}$ where X_{it}^R is a vector of covariates that are functions of the underlying state variables (observable and unobservable). Not all hazards depend on all of

the state variables, rather there is a different subset specified for each outcome. The exact specifications are shown in Table 3a. We discretize the continuous state variables in (2).

The human capital accumulation function allows for curvature of the profile either through age t, or human capital H_{it} using a log-linear functional form. Specifically, for workers, human capital accumulates according to

$$\dot{H} = a\left(\mathcal{S}_{it}\right)\left(\bar{H}_i - H_{it}\right)e^{-\mu_i t} \tag{3}$$

where $\bar{H}_i$ is the maximum level of human capital (and $\dot{H} = \partial H/\partial t$). $\bar{H}_i$ is allowed to vary across people depending on their education using a log-linear functional form. In this specification, as H_{it} approaches $\bar{H}_i$, human capital accumulation slows. The other force that allows for human capital accumulation to slow down is the potential experience term $e^{-\mu_i t}$ (where μ_i varies with education). As discussed above, the distinction between the two is very important for mothers who take time out of the labor force. With a long spell out of the labor force to care for children, at the time of re-entry these mothers will have relatively high t but relatively low H_{it}. So if the first effect is important, mothers should see large wage growth upon re-entering, but for the second, they will not. This intuition is key for distinguishing between these explanations in the data. The key auxiliary parameter is the coefficient on women with children over the age of 18 in the wage growth regression. We put high weight on this parameter to make sure the model fits it very well. We also parameterize $\log\left(a\left(\mathcal{S}_{it}\right)\right)$ to be linear in the state variables. Note that unlike the classic Mincer model, our specification does not allow human capital to fall for older women (when they work). This is consistent with the data (see Figure 1).

When women do not work their human capital depreciates according to the formula,

$$\dot{H} = -\delta H \tag{4}$$

where δ is a single parameter.[3]

Finally, wages depend on human capital as well as some of the other state variables,

$$\log\left(W_{it}\right) = X_{it}\left(\mathcal{S}_{it}\right)'\gamma + H_{it} + \varepsilon_{it}. \tag{5}$$

Since H_{it} is an element of $\mathcal{S}_{it}$, the notation is general enough that we could have incorporated it into $X_{it}\left(\mathcal{S}_{it}\right)'\gamma$. It is shown explicitly in (5) only to clarify that its scale is determined by the wage equation since it is restricted to have a coefficient equal to 1. We also assume that ε_{it} is i.i.d. standard normally distributed with mean zero and variance σ_ε^2.

[3]In a previous version, δ was allowed to be a log linear function of the state variables, but we did not find strong predictors in the data.

4 Indirect Inference with Importance Sampling Weights

4.1 Basic Setup for Indirect Inference

Our framework is very similar to Chapter 4 in Gourieroux and Monfort (1996) but we focus on a narrower (though still large) set of problems for which the importance weight sampling approach is natural. We also focus on the cross section/panel data versions rather than the time series version. We explicitly derive the asymptotic properties using importance weights. The basic properties are quite similar.

We assume that the econometrician observes (Y_i, X_i) for sample $i = 1, ...N$. The observations are i.i.d. and both X_i and Y_i are potentially large dimensional (K_x and K_y). X_i is exogenous in the sense that it is determined outside the model and is i.i.d. coming from underlying distribution Ξ_0.

The data generating process is

$$Y_i \equiv y(X_i, u_i; \theta) \tag{6}$$

where u_i is an i.i.d. vector error term with distribution

$$\Psi(u_i; \theta). \tag{7}$$

Both Ψ and y are known up to parameter $\theta \in \Theta \subset \Re^{K_\theta}$, where the true value is θ_0. This notation is general enough to represent a complicated system with lagged dependent variables and/or equilibrium conditions, but we assume it can be written in reduced form y.[4]

To apply indirect inference, assume the auxiliary model is

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} F\left(\frac{1}{N}\sum_{i=1}^{N} g(X_i, Y_i, \beta), \beta\right) \tag{8}$$

where $\beta \in \mathcal{B} \subset \mathbb{R}^{K_\beta}$. The functions in (8) are $g : \mathbb{R}^{K_x} \times \mathbb{R}^{K_y} \times \mathcal{B} \to \mathbb{R}^{K_g}$ and $F : \mathbb{R}^{K_g + K_\beta} \to \mathbb{R}$. We define the population value of $\widehat{\beta}$ to be $\beta_0 \equiv argmin_\beta F(E\left[g(X_i, Y_i, \beta)\right], \beta)$. These functions are general enough to incorporate many estimators. Simulated Method of Moments is a special case in which the auxiliary parameters are moments: $\widehat{\beta} = \frac{1}{N}\sum_{i=1}^{N} g(X_i, Y_i)$. It can also capture much richer auxiliary parameters. For example, it would be maximum likelihood when g is the negative of the log-likelihood function and F is be the identity function, i.e. $F(x) = x$. It can also incorporate a Generalized Method of Moments type estimator in which

[4] One can think of y as the function used by the computer code that produces simulated data given X_i, u_i, and the parameter value θ. If there are multiple equilibria the code must have some condition for choosing between them. The same mechanism would be incorporated into y.

g is the moments such that $E\left[g(X_i, Y_i, \beta)\right] = 0$ and F is the function $F(g, \beta) = g'Wg$ where W is some weighting matrix. This can incorporate OLS, IV, difference-in-differences, and regression discontinuity. We could also use it to represent a quantile or quantile regression.

To see the idea of indirect inference in this context let Ξ be a potential distribution of X_i, then the data generation process is known up to (Ξ, θ). Define the population functions

$$G(\theta, \beta) \equiv \int \int g(x, y(x, u; \theta), \beta) d\Psi\left(u; \theta\right) d\Xi_0(x) \tag{9}$$

and what Gourieroux, Montfort and Renault (1993) refer to as the binding function in our case is

$$B(\theta) = \underset{\beta}{\text{argmin}} F\left(G(\theta, \beta), \beta\right). \tag{10}$$

Note that $B(\theta_0) = \beta_0$. Identification and estimation of Ξ_0 is straight forward since X_i is observable, so we mostly abstract from it. Essentially what one needs for point identification of θ is that $B(\theta_0)$ is invertible so that knowledge of β_0 is sufficient for knowledge of θ. In that case, since the model is known up to parameter θ, the function $B(\theta)$ is identified. Thus, we could just invert $B\left(\theta_0\right)$ to identify θ_0.

In practice we typically do not have a closed form for B. Instead, we use simulation estimators in order to approximate $B(\theta)$. A typical approach is to generate H different simulated samples each with size S. For each observation we draw $u_{hs}(\theta)$ randomly from the distribution Ψ and calculate X_{hs} from the empirical distribution of X_i.[5] We then define

$$\widetilde{B}_B(\theta) \equiv \frac{1}{H} \sum_{h=1}^{H} \underset{\beta}{\text{argmin}} F\left(\frac{1}{S} \sum_{s=1}^{S} g(X_{hs}, y(X_{hs}, u_{hs}(\theta); \theta); \beta), \beta\right) \tag{11}$$

and choose

$$\widehat{\theta} = \underset{\theta}{\text{argmin}} \left(\widetilde{B}_B(\theta) - \widehat{\beta}\right)' \Omega \left(\widetilde{B}_B(\theta) - \widehat{\beta}\right) \tag{12}$$

where Ω is a weighting matrix. We refer to this as the base approach which is why we use the subscript B.

4.2 Using Importance Sampling Weights

A major problem with the base approach is that often some components of the dependent variable vector, $Y_{hs}(\theta)$ is discrete so that a small change in the parameters leads to jumps

[5] There are different ways to obtain X_{hs}. One possibility is to take $S = N$ and all of the X_i that we see in the data, that is choose $X_{hs} = X_s$. Alternatively we could draw randomly from the empirical distribution of X_i. What is crucial is that the distribution we use converges to Ξ_0.

in $y(X_{hs}, u_{hs}; \theta)$. This causes the objective function to be discontinuous. In principle, with enough simulations we could make this as smooth as we would like, but in practice this can make minimizing the objective function very time consuming. For example, we attempted to estimate our model with 1,580,000 simulations. This was still not sufficiently smooth to use gradient methods or obtain reliable standard errors. Using importance sampling weights smooths the objective function and worked very well for us in practice.[6] This problem has been found in other indirect inference cases as well.[7]

Before describing the method, we introduce notation for a key component of the analysis: Υ_{hs}. This is essentially a superset of the data including additional components such as unobserved heterogeneity and state variables. The empirical economist typically does not get to observe its sample analogue, but since it can be simulated, Υ_{hs} is something on which the empiricist can condition. In a typical problem there are multiple ways to choose Υ_{hs} and finding the best one will be computationally very important. This is essentially what Ackerberg (2009) discusses as a "change in variables."[8]

The data generating process for Y_i ($y(\cdot)$ in equation 6) is augmented to include the intermediate variable Υ_i and is expressed as

$$\begin{aligned} \Upsilon_i =& \upsilon(X_i, u_i; \theta) \\ Y_i =& y_\Upsilon\left(\Upsilon_i, X_i; \theta\right) \end{aligned} \tag{13}$$

where both functions are known up to parameter θ. While the distinction between Υ_i and Y_i may seem arbitrary at this point, its usefulness will become clear in the examples below. Let $\ell(\cdot; X_i, \theta)$ be the likelihood function for Υ_i. The key for this to work well is that ℓ and y_Υ should be differentiable in θ and that ℓ should be easy to compute.

Our importance weighting estimator is the following: Obtain the values of X_{hs} from the empirical distribution of X_i. Generate Υ_{hs} ex-ante without regards to θ using some data generating function leading to likelihood $\ell_0(\Upsilon_{hs}; X_{hs})$. This typically involve simulating using a pre-chosen parameter θ^* which results in $\ell_0(\Upsilon_{hs}; X_{hs}) = \ell(\Upsilon_{hs}; X_{hs}, \theta^*)$, but this is not necessary.

[6] The Bruins et al. (2018) is another approach to smoothing. We describe the differences with that method in Section 4.3

[7] For example, for the model in Taber and Vejlin (2020), the authors could not find a number of simulations that was large enough to allow the use of gradient methods and small enough to be computationally feasible. The authors of Lee (2012), Han (2016), and Fu and Gregory (2019) experienced similar problems.

[8] He denotes it by $u(X_i, \epsilon_i, \theta)$ rather than Υ_i.

Fixing the values of (Υ_{hs}, X_{hs}) we use

$$\widetilde{B}_I(\theta) \equiv \frac{1}{H}\sum_{h=1}^{H} \underset{\beta}{\text{argmin}} F\left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})} g\left(X_{hs}, y_\Upsilon\left(\Upsilon_{hs},X_{hs};\theta\right);\beta\right),\beta\right) \tag{14}$$

and choose

$$\widehat{\theta} = \underset{\theta}{\text{argmin}} \left(\widetilde{B}_I(\theta) - \widehat{\beta}\right)' \Omega \left(\widetilde{B}_I(\theta) - \widehat{\beta}\right). \tag{15}$$

To understand the basic intuition of the approach, suppose that Υ_{hs} has a continuous distribution and ignore the $X's$. Let E_s denote the expected value from the simulation. Since in the simulation Υ_{hs} was drawn from the density ℓ_0,

$$\begin{aligned} E_s\left[\frac{\ell(\Upsilon_{hs};\theta)}{\ell_0(\Upsilon_{hs})} g\left(X_{hs}, y_\Upsilon\left(\Upsilon_{hs};\theta\right);\beta\right)\right] &= \int \frac{\ell(\Upsilon_{hs};\theta)}{\ell_0(\Upsilon_{hs})} g\left(X_{hs}, y_\Upsilon\left(\Upsilon_{hs};\theta\right);\beta\right)\ell_0(\Upsilon_{hs})d\Upsilon_{hs} \\ &= \int g\left(X_{hs}, y_\Upsilon\left(\Upsilon_{hs};\theta\right);\beta\right)\ell(\Upsilon_{hs};\theta)d\Upsilon_{hs} \\ &= G(\theta,\beta). \end{aligned} \tag{16}$$

We approximate this integral using a Monte Carlo procedure where Υ_{hs} is drawn from the distribution $\ell_0(\Upsilon_{hs})$, i.e.,

$$\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};\theta)}{\ell_0(\Upsilon_{hs})} g\left(y_\Upsilon\left(\Upsilon_{hs};\theta\right);\beta\right) \approx G(\theta,\beta). \tag{17}$$

This gives us a consistent estimate of $G(\theta,\beta)$ as S gets large. Critically, as long as $\ell(\Upsilon_{hs};\theta)$ and $y_\Upsilon\left(\Upsilon_{hs};\theta\right)$ are smooth functions of θ, this approximation is a smooth function of θ.

First, note that standard indirect inference is a special case. To avoid jumps in the objective function, researchers typically draw the random variables that determine outcomes first and then fix these values throughout the estimation process. For example, if the distribution of an underlying random variable u_{hs} does not depend on θ, one would draw the u_{hs} once, at the beginning of the estimation procedure, and we would choose $\Upsilon_{hs} = u_{hs}$. In this case, $\ell(\Upsilon_{hs};X_{hs},\theta) = \ell_0(\Upsilon_{hs};X_{hs})$, so the ratio of the likelihoods would just be one and this would be the standard estimator. When u_{hs} does depend on parameters, typically one would draw underlying random variables that do not depend on θ and write u_{hs} as a parametric function of those underlying variables. We discuss this point below.

The key improvement of this approach relative to the base model is that if we choose Υ_{hs} in the appropriate way, $y_\Upsilon\left(\Upsilon_{hs},X_{hs};\theta\right)$ and thus $\widetilde{B}_I(\theta)$ will be continuous and differentiable functions of θ. This makes both estimation and formation of standard errors much easier. To keep our results general enough to cover the base case, in our formal results we do not impose that $y_\Upsilon\left(\Upsilon_{hs},X_{hs};\theta\right)$ is continuous.

In our supplementary Appendix A in Sauer and Taber (2021) we show consistency and asymptotic normality. The asymptotic variance is

$$\left[\frac{\partial B(\theta_0)'}{\partial \theta}\Omega\frac{\partial B(\theta_0)}{\partial \theta'}\right]^{-1}\frac{\partial B(\theta_0)'}{\partial \theta}\Omega F_{\beta\beta}^{-1}VF_{\beta\beta}^{-1}\Omega\frac{\partial B(\theta_0)}{\partial \theta}\left[\frac{\partial B(\theta_0)'}{\partial \theta}\Omega\frac{\partial B(\theta_0)}{\partial \theta'}\right]^{-1} \tag{18}$$

where $F_{\beta\beta} \equiv \frac{d^2F(G(\theta_0,\beta_0),\beta_0)}{d\beta d\beta'}$. V is the variance of $\left(\left[\frac{1}{H}\sum_{h=1}^{H}\widetilde{\vartheta}_{hi}\right]-\vartheta_i\right)$. In addition,

$$\begin{aligned}\vartheta_i \equiv &\left(\frac{\partial G(\beta_0)}{\partial \beta}\frac{\partial^2 F(G(\beta_0),\beta_0)}{\partial G\partial G'}+\frac{\partial^2 F(G(\beta_0),\beta_0)}{\partial \beta\partial G'}\right)(g(X_i,Y_i,\beta_0)-G(\beta_0)) \\ &+\left(\frac{\partial g(X_i,Y_i,\beta_0)}{\partial \beta}-\frac{\partial G(\beta_0)}{\partial \beta}\right)\frac{\partial F(G(\beta_0),\beta_0)}{\partial G}\end{aligned} \tag{19}$$

$$\begin{aligned}\widetilde{\vartheta}_{hi} \equiv &\left(\frac{\partial G(\beta_0)}{\partial \beta'}\frac{\partial F(G(\beta_0),\beta_0)}{\partial G\partial G'}+\frac{\partial F(G(\beta_0),\beta_0)}{\partial \beta\partial G'}\right)(\widetilde{g}_{hi}(\beta_0)-G(\beta_0)) \\ &+\left(\frac{\partial \widetilde{g}_{hi}(\beta_0)}{\partial \beta}-\frac{\partial G(\beta_0)}{\partial \beta}\right)\frac{\partial F(G(\beta_0),\beta_0)}{\partial G}.\end{aligned} \tag{20}$$

4.3 Relationship with Other Work

Gourieroux and Monfort (1996) present a very general indirect inference framework in Chapter 4. They consider a broader definition of the auxiliary estimator than equation (8), which can be written in notation similar to ours as

$$\widehat{\beta} = \underset{\beta}{\text{argmax}} F(Y,X,\beta) \tag{21}$$

where X and Y are matrices of the exogenous and endogenous data. They derive some general properties of the estimator. Our estimator is a special case (though it still includes a large set of auxiliary models) that makes clear where importance sampling enters. We derive the asymptotic properties for our case. Gourieroux and Monfort (1996) also discuss importance sampling for an array of estimators, but not explicitly for the indirect inference estimator.

The main difference between our estimator and Ackerberg (2009) is that we consider a more general estimator. The method of simulated moments is a special case of our model when the auxiliary model is a moment of the data, i.e.,

$$\widehat{\beta} = \frac{1}{N}\sum_{i=1}^{N} g(Y_i,X_i). \tag{22}$$

The main point of Ackerberg (2009) is to emphasize an additional advantage of indirect inference (other than smoothness). Using our notation, evaluating $y(X_i,u_i;\theta)$ can be time

consuming as it might involve solving a dynamic programming problem or for equilibrium. If one can find an appropriate change of variables, Υ_i, the problem can be simplified. If we can write $y_\Upsilon(\Upsilon_i, X_i; \theta)$ in such a way that it does not depend on θ then one does not need to re-solve the model when one does a function evaluation. In our Markov model, calculating y_Υ is not difficult so we have not taken advantage of this feature.

Generalized indirect inference (GII) proposed by Bruins et al. 2018 is an alternative method to smooth the objective function. We briefly, and loosely, introduce GII with our notation. They consider a case in which outcomes are discrete so we can write the outcome for simulation h and s as $y(X_{hs}, u_{hs}; \theta) = (y_0(X_{hs}, u_{hs}; \theta), ..., y_T(X_{hs}, u_{hs}; \theta))$ with

$$y_t(X_{hs}, u_{hs}; \theta) = \underset{j \in 0,...,J-1}{\operatorname{argmax}} \{v_{jt}(X_{hs}, u_{hs}; \theta)\} \tag{23}$$

where j is a discrete option chosen from the set $\{0, ..., J-1\}$ and v_j is a utility function. The discontinuity in the objective function comes from the jump in the value that maximizes it.

They define a smooth function of latent utilities $g(\cdot; \theta)$ such that $\widetilde{y}_t(X_{hs}, u_{hs}; \theta, \lambda) \equiv g(v_{0t}(X_{hs}, u_{hs}; \theta), ..., v_{J-1t}(X_{hs}, u_{hs}; \theta); \lambda)$ converges to $y_t(X_{hs}, u_{hs}; \theta)$ as the smoothing parameter λ goes to zero. GII substitutes $\widetilde{y}_t(X_{hs}, u_{hs}; \theta, \lambda)$ for $y_t(X_{hs}, u_{hs}; \theta)$ in computation of $\widetilde{B}_G(\theta)$ (which can then be used analogously to $\widetilde{B}_B(\theta)$ and $\widetilde{B}_I(\theta)$). Thus, the objective function is smooth and $\widetilde{B}_G(\theta)$ converges to β_0 as λ goes to zero and N goes to infinity, with H and T fixed. The $g(\cdot)$ Bruins et al. (2018) used in their Monte Carlo experiments is the logistic-kernel. The main differences between the two approaches is that importance weighting involves calculating and weighting by the likelihood function while GII depends on the choice of the smoothing parameter.

GII is not suited for our Markov model for two reasons. The first is that the Markov model is in continuous time rather than discrete time. The second is that our model of human capital, which is crucial to the analysis, depends on the full labor market history up to that point. Jumps in previous labor supply lead to jumps in human capital and we do not see a computationally feasible way to use GII to smooth human capital accumulation. Bruins et al. (2018) does discuss how to handle a lagged dependent variable within the context of GII. They employ the same smoothing function that applies to the contemporaneous choice in the Monte Carlo experiments and obtain good results. However, they do not discuss how one would go about smoothing a function of the history of lagged dependent variables as with accumulated capital accumulation. Altonji et al. (2013) use a version of GII that does use the history, but their approach will not work for a continuous state variable like human capital.

4.4 Logit Model

In order to demonstrate the method of indirect inference with importance sampling, we explain how to use it to estimate a logit model using a linear probability model as the auxiliary model. Of course, one would never need to estimate a simple logit model using this technique but it provides a good illustration of the method in a simple case. The true model is

$$Pr(Y_i = 1 \mid X_i) = \Lambda(X_i'\theta_0) \tag{24}$$

where Λ denotes the logit c.d.f..

The auxiliary model is a linear probability model. This can be put into our notation by taking F to be the identity function and choosing

$$g(X_i, Y_i; \beta) = (Y_i - X_i'\beta)^2 . \tag{25}$$

The simulated data is generated in the following way:

1. Choose X_{hs} by drawing randomly from the empirical distribution of X_i[9]
2. Choose some initial logit value θ^*
3. Simulate Y_{hs} so that

$$Y_{hs} = \begin{cases} 1 & \text{with probability } \Lambda\left(X_{hs}'\theta^*\right) \\ 0 & \text{with probability } 1 - \Lambda\left(X_{hs}'\theta^*\right) \end{cases}.$$

In this simple case choose $\Upsilon_{hs} = Y_{hs}$.

For this model, note that

$$\frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})} = \frac{\Lambda\left(X_{hs}'\theta\right)^{Y_{hs}} \left(1 - \Lambda\left(X_{hs}'\theta\right)\right)^{(1-Y_{hs})}}{\Lambda\left(X_{hs}'\theta^*\right)^{Y_{hs}} \left(1 - \Lambda\left(X_{hs}'\theta^*\right)\right)^{(1-Y_{hs})}} \tag{26}$$

so

$$\begin{aligned} \widetilde{B}_I(\theta) =& \frac{1}{H}\sum_{h=1}^{H} argmin_\beta F\left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})} g(X_{hs}, Y_{hs}; \widehat{\beta}), \widehat{\beta}\right) \\ =& \frac{1}{H}\sum_{h=1}^{H} argmin_\beta \frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})}\left(Y_{hs} - X_{hs}'B(\widehat{\beta})\right)^2 \\ =& \frac{1}{H}\sum_{h=1}^{H}\left(\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})} X_{hs}X_{hs}'\right)^{-1}\left(\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})} X_{hs}Y_{hs}\right). \end{aligned} \tag{27}$$

[9] As discussed above, sampling could be with or without replacement. Of course, if it is done without replacement and the simulation sample is larger than the original one, one would have to replenish it after running through the full sample.

Clearly this is just H weighted regressions with weights $\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}$. Also, since the weight is differentiable in θ, so is $\widetilde{B}_I(\theta)$.

To see why this gives a consistent estimate note that

$$\begin{aligned}\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}X_{hs}X_{hs}' &\underset{S\to\infty}{\overset{p}{\to}} E\left(\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}X_{hs}X_{hs}'\right) \\ &=E\left(X_{hs}X_{hs}'E\left[\frac{Y_{hs}\Lambda(X_{hs}'\theta)+(1-Y_{hs})(1-\Lambda(X_{hs}'\theta))}{Y_{hs}\Lambda(X_{hs}'\theta^*)+(1-Y_{hs})(1-\Lambda(X_{hs}'\theta^*))}\mid X_{hs}\right]\right) \\ &=E\left(X_{hs}X_{hs}'\left[\frac{\Lambda(X_{hs}'\theta)}{\Lambda(X_{hs}'\theta^*)}\Lambda(X_{hs}'\theta^*)+\frac{(1-\Lambda(X_{hs}'\theta))}{(1-\Lambda(X_{hs}'\theta^*))}(1-\Lambda(X_{hs}'\theta^*))\right]\right) \\ &=E(X_iX_i')\end{aligned} \tag{28}$$

and at the true value $\theta=\theta_0$

$$\begin{aligned}\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta_0)}{\ell_0(\Upsilon_{hs};X_{hs})}X_{hs}Y_{hs} &\underset{S\to\infty}{\overset{p}{\to}} E\left(X_{hs}E\left[Y_{hs}\frac{Y_{hs}\Lambda(X_{hs}'\theta_0)+(1-Y_{hs})(1-\Lambda(X_{hs}'\theta_0))}{Y_{hs}\Lambda(X_{hs}'\theta^*)+(1-Y_{hs})(1-\Lambda(X_{hs}'\theta^*))}\mid X_{hs}\right]\right) \\ &= E\left(X_{hs}\left[\frac{\Lambda(X_{hs}'\theta_0)}{\Lambda(X_{hs}'\theta^*)}\Lambda(X_{hs}'\theta^*)\right]\right) \\ &= E(X_iY_i).\end{aligned} \tag{29}$$

Thus, the simulator yields a consistent estimate as S grows (i.e. $\text{plim}(B(\theta_0)=\beta_0)$.

4.5 Discrete Time Markov Model and Monte Carlo Results

In this second illustration of the technique, the model is a discrete time Markov model of d_{it} which is binary (0 or 1). Everyone begins with $d_{i0}=0$. Then the transition model is

$$\begin{aligned}Pr(d_{it+1}=1\mid X_i,d_{it}=0)=&\Lambda(X_i'\theta_0+u_{i0}) \\ Pr(d_{it+1}=1\mid X_i,d_{it}=1)=&\Lambda(X_i'\theta_1+u_{i1})\end{aligned} \tag{30}$$

where the distribution of $u_i=(u_{i0},u_{i1})$ is $G(\cdot;\theta_3)$.

We don't observe people throughout the lifecycle and define F_i as the first period in which we observe data for individual i and L_i as the last period. Thus, for each individual we observe $(X_i,d_{iF_i},...,d_{iL_i})$.

Note that because of the initial conditions problem and the unobserved heterogeneity, the likelihood of the data is computationally intensive when F_{1i} is large. To economize on notation let

$$\varrho(d_t,d_{t+1};\theta,X_i,u)\equiv\Lambda(X_i'\theta_{d_t}+u_{d_t})^{d_{t+1}}(1-\Lambda(X_i'\theta_{d_t}+u_{id_t}))^{1-d_{t+1}}. \tag{31}$$

The the likelihood function of the data would be

$$\int \sum_{d_1=0}^{1} ... \sum_{d_{F_i-1}=0}^{1} \varrho\left(0, d_1; \theta, X_i, u_i\right) \varrho\left(d_1, d_2; \theta, X_i, u_i\right) \times .... \times \varrho\left(d_{F_i-1}, d_{iF_i}; \theta, X_i, u_i\right) \times \quad (32)$$
$$\varrho\left(d_{iF_i}, d_{iF_i+1}; \theta, X_i, u_i\right) \times ... \times \varrho\left(d_{iL_i-1}, d_{iL_i}; \theta, X_i, u_i\right) dG\left(u_i; \theta_3\right).$$

Here we can simplify the likelihood used in the importance weights substantially by choosing $\Upsilon_{hs} = (X_{hs}, u_{hs}, d_{hs1}, ..., d_{hsL_{hs}})$ for simulation hs. Then we no longer need to integrate when computing the likelihood function,

$$\varrho(\Upsilon_{hs}; X_{hs}, \theta) = \varrho\left(0, d_{hs1}; \theta, X_{hs}, u_{hs}\right) \varrho\left(d_{hs1}, d_{hs2}; \theta, X_{hs}, u_{hs}\right) \times .... \times \varrho\left(d_{hsL_{hs}-1}, d_{hsL_{hs}}; \theta, X_{hs}, u_{hs}\right) \quad (33)$$

To get a sense of the performance difference between different methods, we used a simple version of this model as the basis of a Monte Carlo study. We specify a Markov model with no unobserved heterogeneity, $F_i = 2$ and $L_i = 3$. As an auxiliary model, we consider three regressions (linear probability models): d_{i2_i} on X_i (initial condition) and a linear probability model of d_{i3} on X_i conditional on the two different values of d_{i2}. For each run, we consider the same auxiliary model but simulate it in three different ways: the base model, the GII procedure, and by importance weighting (i.e. $\widetilde{B}_B\left(\theta\right), \widetilde{B}_G\left(\theta\right)$, and $\widetilde{B}_I\left(\theta\right)$). We also consider two different estimators: a gradient based estimator (LFGBS) and Nelder-Mead from the Optim package in Julia. We tried two different dimensions of X_i: 5 and 15 (which results in 12 and 32 parameters with two set of parameters and intercepts) where the X_i are drawn from a joint normal distribution. The parameters θ are uniformly distributed. We also used 3 different simulation sizes: 20,000, 100,000, and 500,000 for all cases with $H = 1$.

The results are presented in Table 1. For each case, we present 3 summary statistics: the mean squared error, the computation time (relative to using Nelder-Mead in the base model), and the mean of the minimized objective function. Despite the simplicity of the model, for a Monte Carlo study it takes a fair amount of time to estimate the base model with 1 iteration of the six different estimators taking roughly 24 hours on an Amazon AMD processor when we use 500,000 simulations.

Comparing the base estimator with importance weighting, one can see the advantages. First, notice that even with 500,0000 simulations the gradient based estimator does not work as it does not find the minimum of the function consistently. However, the importance weighting estimator works very well. Comparing the indirect inference estimator to the simplex estimated standard estimator, both perform well in terms of mean squared error though the gradient indirect inference estimator performs slightly better in some cases.

Table 1
Monte Carlo Results of Different Approaches
Markov Model
(500 Monte Carlo Iterations in all Cases)

	Standard Ind. Infer		Ind. Infer. Importance Weights		Generalized Ind. Infer.	
	Simplex	Gradient	Simplex	Gradient	Simplex	Gradient
5 Covariates, 20,000 Simulated Individuals						
Mean Squard Error	0.0076	0.1544	0.0073	0.0073	0.0076	0.0076
Average Time	1.0000	0.0384	0.8122	0.2149	0.5649	0.9332
Objective×100	0.0144	1.332	0.0175	0.0175	0.0143	0.0143
5 Covariates, 100,000 Simulated Individuals						
Mean Squard Error	0.0016	0.152	0.0016	0.0016	0.0016	0.0016
Average Time	1.0000	0.0966	0.7684	0.1927	0.5694	0.8590
Objective	0.0031	1.2145	0.0036	0.0036	0.0032	0.0032
5 Covariates, 500,000 Simulated Individuals						
Mean Squard Error	0.0004	0.1095	0.0004	0.0004	0.0004	0.0004
Average Time	1.0000	1.4961	1.0773	0.2495	0.7556	1.0494
Objective	0.0009	0.8153	0.0009	0.0009	0.0009	0.0009
25 Covariates, 20,000 Simulated Individuals						
Mean Squard Error	0.6376	2.1482	0.1564	0.1567	0.1741	0.1749
Average Time	1.0000	0.0153	2.1253	0.0242	1.3869	0.0499
Objective	2.7592	3.3255	0.0972	0.0972	0.0509	0.0509
25 Covariates, 100,000 Simulated Individuals						
Mean Squard Error	0.1178	0.8163	0.0283	0.0283	0.0281	0.0282
Average Time	1.0000	0.0337	1.1810	0.0088	0.7428	0.0175
Objective	0.2591	2.7680	0.0215	0.0215	0.0154	0.0154
25 Covariates, 500,000 Simulated Individuals						
Mean Squard Error	0.0077	0.7641	0.0064	0.0064	0.0058	0.0058
Average Time	1.0000	0.1751	1.4627	0.0105	0.8522	0.0152
Objective	0.0070	2.4191	0.0053	0.0053	0.0046	0.0046

The most important result is that indirect inference/gradient is much faster than the base/ simplex method - roughly five times faster with 12 parameters and 100 times faster with 52 parameters (in the 100,000 and 500,000 simulation cases).

For the Generalized Indirect Inference model, since calculating the optimal smoothing parameter is time consuming we only do it once for each model (i.e. 6 times for the covariate/simulations specification). We do not follow precisely Bruins et al. 2018 but something very close that fits our approach and seemed to work well. While GII sees less time savings in the smaller model it is also much faster and converges better than the simplex method in the base model.

Given that this is only one specification, and also a very simple model, we cannot come to general conclusions in the comparison between the importance weighting procedure and GII. However, it is clear that for large problems there are substantial computational advantages to estimating with a smoothed model and using a gradient based method.

5 Data and Auxiliary Model

We use data from the Survey of Income and Program Participation (SIPP). Alternative data sets we could have used are the National Longitudinal Survey of Youth 1979 (NSLY79) as well as the older National Longitudinal Surveys of Young Women and Mature Women (NLSW). We would not argue that SIPP clearly dominates the NLSY79, but rather there are tradeoffs between the two datasets and the vast majority of previous work discussed above has focused on the NLSY79 or NLSW. The advantage of the NLSY79 is it is a much longer panel, but the disadvantage is that it contains a small number of individuals (at most around 6,000 women which gets smaller over time due to attrition from the survey).

The SIPP is a very large data set with short panels - we use observations from almost 100,000 different women. The challenge with the SIPP is that one does not observe the full lifecycle profile for each woman. One must piece together the panel data for different people at different ages. This requires an econometric model, and we use our Markov model of work, fertility and marriage. Estimating such a model by maximum likelihood is extremely difficult given the severe initial conditions problem. Indirect inference is a feasible alternative.

We estimate the model using the last four panels of the Survey of Income and Program Participation 1996, 2001, 2004, and 2008.[10] This survey interviews individuals every four months and we only use data from the survey month. The sample includes white women who are 18 years or older and have at most 35 years of potential experience. Table 2 presents summary statistics of the main variables used in the analysis. Details of the data are discussed in supplementary Appendix B (Sauer and Taber, 2021).

The auxiliary model is constructed using the following auxiliary parameters (the full description along with their empirical values is in Supplementarly Appendix C, Sauer and Taber, 2021):

- Regression of log wages on potential experience dummies and state variables with individual fixed effects

[10] We do not use earlier years because the nature of the survey changed around 1996. These later panels are substantially longer than the previous ones.

Table 2
Summary Statistics
White Women 18-65
Survey of Income and Program Participation

Variable	Mean	Standard Deviation
Potential Experience	18.028	10.021
Employed	0.728	0.445
log(Wage)	2.642	0.589
Education	13.529	2.412
Currently Married	0.604	0.489
Currently Not Married and Divorced	0.164	0.37
Number Children < 18	0.958	1.168
Number Children < 7	0.344	0.677
Number of Children	1.546	1.353
Any Children	0.717	0.451
Age Youngest	8.147	8.679
Age Difference Oldest/Youngest	5.699	4.038
Had Baby	0.009	0.094
Number of Cells	726484	
Number of Women	97354	

- Within and between variance of the error term from the previous regression
- Regression of estimated wage fixed effect on education
- Linear probability regression of working on potential experience dummies and state variables with individual fixed effects
- Between variance of the error term from the previous regression
- Regression of wage fixed effect on work fixed effect
- Linear probability regressions of whether a woman is currently married/currently unmarried and divorced from second wave of survey on potential experience dummies
- Linear probability regressions of whether an unmarried woman gets married/becomes unmarried between waves, on potential experience dummies and state variables
- Fraction of mothers who are married at childbirth
- Regression of having a child on one year lagged work status wages of mothers who work (with other covariates)

- Age difference between youngest and oldest child
- Linear probability regression of any children/two children/number of kids, on potential experience dummies and state variables
- Linear probability regression of working in one wave conditional on working in the previous wave, on potential experience dummies, state variables, and work fixed effect
- Fraction of mothers who work in interview before giving birth
- Regression of wage gains between periods for women who are employed between periods
- Change in log wages for women with non-employment spells divided by difference in potential experience dummies.

The key parameters are the effects of the number of children on various outcomes and can be seen in the "data" part of the tables and figures. Most surprisingly, in the fixed effects wage regression (Supplementary Appendix Table C1, Sauer and Taber, 2021), there is no evidence of a children penalty relative to many of the papers cited previously. This is in large part because this is a very short panel. We also find that having young children is an important determinant of working in the fixed effect regression shown in Supplementary Appendix Table C2 (Sauer and Taber, 2021), but it goes away as the children age. We also include age 7 in this regression but it is statistically and economically insignificant - the coefficient is -.0044. For this reason, in the model we impose that only children less than seven influence labor supply decisions. There are also substantial effects of marriage on labor supply in this fixed effect regression.

Another key parameter is children over 18 in the wage growth regression. This variable is the key to identification of experience versus age effects. This is a proxy for actual experience since we know women who have more children have less experience (making use of the Markov model in the sense that it tells us how much less experience). The coefficient is interacted with education. The results are in Supplementary Appendix Table C7 (Sauer and Taber, 2021). The effect at 12 years of school is positive (and for college even larger) indicating that women at the same age who have had more children have faster wage growth. This is evidence that actual experience matters for the curvature. The magnitude of this effect will dictate the magnitude of the actual experience effect in the model.

We included other children variables in the log wage growth equation but did not find significant results so we do not include them here, and do not incorporate motherhood

directly into the human capital production function.

6 Implementation of Approach in Practice

To formally describe Υ_{hs} for our model, some new notation is introduced. First, since we take $H = 1$ in practice, we abstract from including h in the subscripts. Let N_s^w be the number of work transitions and the dates (in terms of actual experience) of these transitions be $\tau_{s1}^w, ..., \tau_{sN_s^w}^w$. Let L_{s0} be labor force status upon labor market entry. Given L_{s0}, we can keep track of the state so we know the direction of the transition. Similarly for marriage, let N_s^m be the number of marriage transitions and $\tau_{s1}^m, ..., \tau_{sN_s^m}^m$ be their dates. Analogously, let N_s^k be the number of children and $\tau_{s1}^k, ..., \tau_{sN_s^k}^k$ the dates when the children were born. Note that knowledge of the dates when children were born and labor force transitions also tells us if women stopped working right after childbirth. Finally, we write the measurement term on wages as $\varepsilon_{st} = \sigma_\varepsilon \epsilon_{st}$ where ϵ_{st} is standard normal. Let ϵ_s be the vector of these objects for the periods in which the wage is observed by the econometrician. Then we take our Υ_s to be

$$\Upsilon_s = \left\{ L_{s0}, \tau_{s1}^w, ..., \tau_{sN_s^w}^w, \tau_{s1}^m, ..., \tau_{sN_s^m}^m, \tau_{s1}^k, ..., \tau_{sN_s^k}^k, \nu_s, \epsilon_s \right\} \tag{34}$$

What is crucial for our approach is that the likelihood function $\ell(\Upsilon_s \mid X_s; \theta)$ is smooth as a function of θ and the rest of the variables used to produce the auxiliary model are smooth in θ after conditioning on Υ_s.

Note that the labor market, marriage, and birth transitions are perfectly pinned down by Υ_s, but human capital and thus wages is not. The reason is that conditional on the work transitions, human capital is a smooth function of the parameters so there is no reason to include human capital as part of Υ_s. Note as well that we could have included the ε_{st} in Υ_s rather than ϵ_{st}. In the former case, the importance weights would change with σ_ε, parameterizing it with ϵ_{st} instead, they do not. Our experience is that the procedure works better in the latter case for reasons which are discussed below.

Writing the likelihood in terms of Υ_{hs} simplifies its computation over the likelihood of the underlying data for two reasons: integrating over unobserved heterogeneity and solving the initial conditionsl problem. In this case, the difficult integration problem arises due to the initial conditions problem. Unobserved heterogeneity by itself is not that computationally difficult because the measurement error ϵ_s is *i.i.d.* over periods, and ν_s is only a two dimensional and permanent object. It is important to point out that there are many other

models for which the initial conditions problem might not exist, but integrating over the distribution of unobservables would be much more complicated making maximum likelihood computationally infeasible. For example, we could relax the $i.i.d$ assumption on ϵ_i, allow ν_i to evolve, or include an additional transitory error term that affects human capital accumulation each period, as in Fan, Seshadri, and Taber (2019). In our empirical problem the main difficulty with maximum likelihood is the initial conditions problem, but the general approach is useful for a much broader set of applications.

In practice, since the model is complicated, if the estimation procedure runs long enough so that the parameters change substantially, the likelihood can get very small for many observations. As a result, the weight $\ell(\Upsilon_{hs}; X_{hs}, \theta)/\ell_0(\Upsilon_{hs}; X_{hs})$ becomes approximately zero for a large number of the observations. In theory, there is no problem with this, as if S is large enough, the law of large numbers still works. However, as a practical matter, one is using effectively a much smaller sample to approximate the auxiliary moments. Note as well that if one simulates the model using parameter value θ_0 then $\ell_0(\Upsilon_{hs}; X_{hs}) = \ell(\Upsilon_{hs}; X_{hs}, \theta_0)$, so if we simulate at this parameter value, the weights are all equal to one. Putting these together, it is useful for this method to occasionally re-simulate the model with intervening estimates.

After experimenting with different approaches, we settled on the following iterative procedure. Start with some initial value and then:

- Let $\widehat{\theta}_{j-1}$ be the last estimated value of the parameter (or initial value when $j = 1$)
- Simulate the model with estimate $\widehat{\theta}_{j-1}$ and let $\ell(\Upsilon_s; X_s, \widehat{\theta}_{j-1})$ be the likelihood of this simulated data with this parameter
- Use a Newton method and to minimize the distance between the auxiliary and simulated parameters
$$\left(\widetilde{B}_I(\theta) - \widehat{\beta}\right)' \Omega \left(\widetilde{B}_I(\theta) - \widehat{\beta}\right) \tag{35}$$
with at most 100 Newton steps where
$$\widetilde{B}_I(\theta) = argmin_\beta F \left(\frac{1}{S} \sum_{s=1}^{S} \frac{\ell(\Upsilon_s; X_s, \theta)}{\ell(\Upsilon_s; X_s, \widehat{\theta}_{j-1})} g\left(X_s, y_\Upsilon\left(\Upsilon_s, X_s; \theta\right); \beta\right), \beta \right) \tag{36}$$
and Ω is a diagonal weighting matrix
- Let the parameter that minimizes this be $\widehat{\theta}_j$.
- Repeat this procedure until the fit stops improving

Once fit stops improving, we switch to a simplex method without importance weighting. We use our current best guess of θ as a starting value and then choose our estimate to minimize the unweighted objective function

$$\left(\widetilde{B}_B(\theta) - \widehat{\beta}\right)' \Omega \left(\widetilde{B}_B(\theta) - \widehat{\beta}\right) \tag{37}$$

where $\widetilde{B}_B(\theta)$ is the base simulator defined in equation (11).

The reason for the final step with the simplex method is that if we stopped at iteration j the parameter estimate is $\widehat{\theta}_j$ but the model that simulated the data for that estimator was $\widehat{\theta}_{j-1}$. This means that for counterfactuals, if we just simulate from $\widehat{\theta}_j$ we will not get exactly the same simulation as our final estimates. Using the standard method in the final step guarantees that the fit of the data from our final estimates comes from $\widehat{\theta}_j$ alone. As a practical matter this final stage did not lead to a substantial change in the parameter values but was time consuming.

The weights for the parameters of the auxiliary model were chosen in a somewhat ad hoc manner. We chose a diagonal weighting matrix Ω where for most auxiliary parameters we divided by the variance of the estimated parameter. The problem with this default approach, or more generally efficient weighting, is that it does not put the proper weight on the moments we are most interested in fitting. For example, most of our regression models contain a full set of potential experience dummy variables which gives 35 parameters, but there are only a few variables related to fertility (the fixed effect wage regression has a single one). This means that the statistical criterion will put much more weight on fitting the experience profile because this is 35 parameters rather than fertility which is only 1. We adjust for this by overweighting the fertility parameters. While ad hoc, we think it provides a better objective function than a pure statistical one. The precise scales are presented in Supplementary Appendix C (Sauer and Taber, 2021).

7 Empirical Results

In the model, not all state variables affect all outcomes. The specification was chosen in order to fit the data in as parsimonious a way as possible and is motivated by patterns found in the data. In Tables 3a-3d we present the estimated parameters of the model. The parameters themselves are more easily understood through simulations rather than on an individual basis. For this reason, we do not offer a detailed discussion of them. Most of the parameters have the signs one would expect.

Table 3a
Model Estimates: Hazard Estimates

Covariate	Get Married	Get Divorced	Find Job	Leave Job	Have Kid
Education	-0.002	-0.076	0.252	-0.204	-0.114
	(0.024)	(0.134)	(0.013)	(0.016)	(0.024)
ν_1	-0.021	-0.124			-0.023
	(0.057)	(0.115)			0.059)
ν_2	-0.035	0.827	2.024	-1.406	0.099
	(0.067)	(0.323)	(0.075)	(0.072)	(0.134)
Married			-0.633	0.100	2.416
			(0.084)	(0.078)	(0.159)
Number of Kids < 18		0.018			
		(0.332)			
Number of Kids < 7			-0.334	0.298	
			(0.043)	(0.047)	
Working					-1.393
					(0.253)
Number of Kids=1					-0.221
					(0.197)
Number of Kids=2					-1.885
					(0.219)
Number of Kids>2					-5.396
					(1.719)
Number of Kids× Education					-0.144
					(0.609)
Age Youngest					0.006
					(0.015)
Potential Experience ≤ 5	-2.107	-3.450	1.761	-1.347	-1.372
	(0.047)	(0.591)	(0.124)	(0.137)	(0.184)
$5 \leq$ Potential Experience ≤ 10	-1.828	-3.621	2.406	-0.913	-1.984
	(0.100)	(0.699)	(0.122)	(0.126)	(0.222)
$10 \leq$ Potential Experience ≤ 15	-2.810	-3.552	1.673	-0.934	-2.835
	(0.388)	(0.763)	(0.143)	(0.129)	(0.260)
$15 \leq$ Potential Experience ≤ 20	-2.921	-3.497	1.180	-0.910	-3.073
	(0.403)	(0.800)	(0.132)	(0.140)	(0.334)
$20 \leq$ Potential Experience ≤ 25	-3.634	-4.515	0.653	-1.441	-4.434
	(0.845)	(0.984)	(0.189)	(0.196)	(0.884)
$25 \leq$ Potential Experience ≤ 30	-3.623	-4.534	0.606	-1.169	-4.446
	(0.957)	(0.987)	(0.168)	(0.172)	(0.901)
Potential Experience > 30	-3.550	-4.523	0.279	-1.101	-4.452
	(1.034)	(1.158)	(0.218)	(0.229)	(1.342)

Table 3b
Model Estimates: Work Probability

Covariate	Initial	After Birth of Child
Intercept	-1.954	1.650
	(0.651)	(0.680)
Education	0.113	0.088
	(0.260)	(0.263)
ν_2	0.073	0.010
	(0.461)	(0.738)

Table 3c
Model Estimates: Human Capital and Wages

Covariate	a	λ	$\bar{H}$	Wages
Intercept	-3.161	-19.182	0.214	
	(0.303)	(0.005)	(0.141)	
Education	0.428	-2.604	-0.054	0.034
	(0.088)	(0.005)	(0.036)	(0.015)
ν_1				0.422
				(0.008)
Married	-0.387			0.016
	(0.247)			(0.005)
Number of Kids < 18				0.009
				(0.004)
Number of Kids < 7				0.003
				(0.004)

Table 3d
Model Estimates: Additional Parameters

δ	0.060
	(0.020)
σ_ε	0.290
	(0.023)
ρ_{12}	0.938
	(0.004)

The fit of the model is presented in Supplemental Appendix Tables C1-C7 and Figure C1-C5 (Sauer and Taber, 2021). One can see in the tables that with only a few exceptions, the fit is excellent. We also attempt to fit the lifecycle profile of working, wages, marriage

and children. Given the coarseness of the model, the relationship between hazard rates and potential experience does not fit perfectly, but in general the overall lifecycle patterns are fit very well.

The first issue of main interest is the determinants of the curvature of human capital, which is important for understanding the shape of female wage growth. Recall that our baseline model is

$$\dot{H} = a\left(\mathcal{S}_{it}\right)\left(\bar{H}_i - H_{it}\right)e^{-\mu_i t}, \tag{38}$$

where curvature can come from two different sources. The first source is from the term $\left(\bar{H} - H_{it}\right)$ which leads to human capital slowing down as it approaches $\bar{H}$. The second source is from the μ_i term which leads to human capital slowing down as workers age. The former is analogous to curvature due to "actual experience" while the latter is analogous to curvature due to "potential experience." As mentioned previously, we believe this difference is identified by the coefficient on "kids greater than 18" and its interaction with education in the wage growth regression. One can see from Supplementary Appendix Table C7 (Sauer and Taber, 2021) that these are matched well.[11]

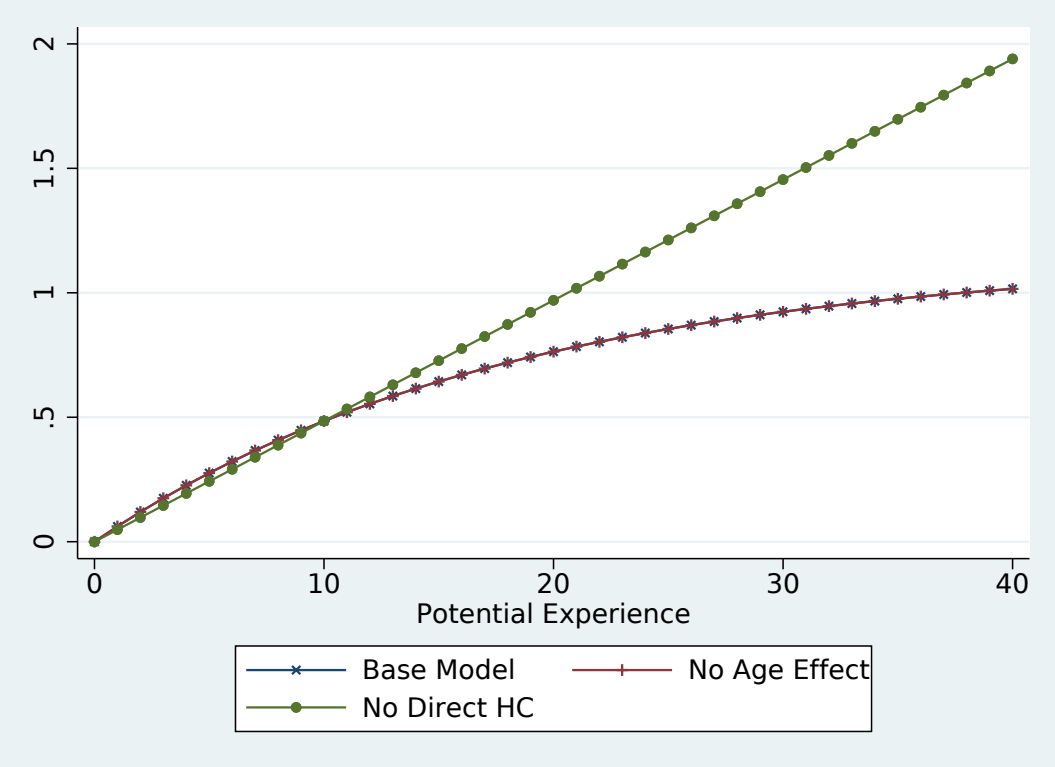
To better understand this distinction, we graph two alternative versions of the human capital production function:

$$\begin{aligned}
\text{Baseline}: \quad \dot{H} &= \bar{a}\left(\bar{H}_i - H_{it}\right)e^{-\bar{\mu}t} \\
\text{Model A}: \quad \dot{H} &= \bar{a}^A\left(\bar{H} - H_{it}\right) \\
\text{Model B}: \quad \dot{H} &= \bar{a}^B e^{-\bar{\mu}t}
\end{aligned}$$

for married women with the average amount of education (13.5 years of education). Since the point of this exercise is to explore curvature rather than rate of wage growth, $\bar{a}^A$ and $\bar{a}^B$ are calibrated to values that keep human capital growth in the first ten years the same as in the base case. The first model eliminates the age effect through μ while the second eliminates the curvature in the human capital production function. Figure 2 presents the results in which Model A is labeled "No Age Effect," and Model B is labeled "No Direct Human Capital." Note that the distinction between the Baseline and Model A is barely visible while Model B is distinctly different. Clearly μ_i is largely irrelevant and the curvature derives from human capital accumulation.

[11] Note that the standard errors on these parameters are large but partly because the level and interaction with education are collinear. The p-value on the joint test that both are zero is 0.002.

Figure 2: Curvature in Human Capital



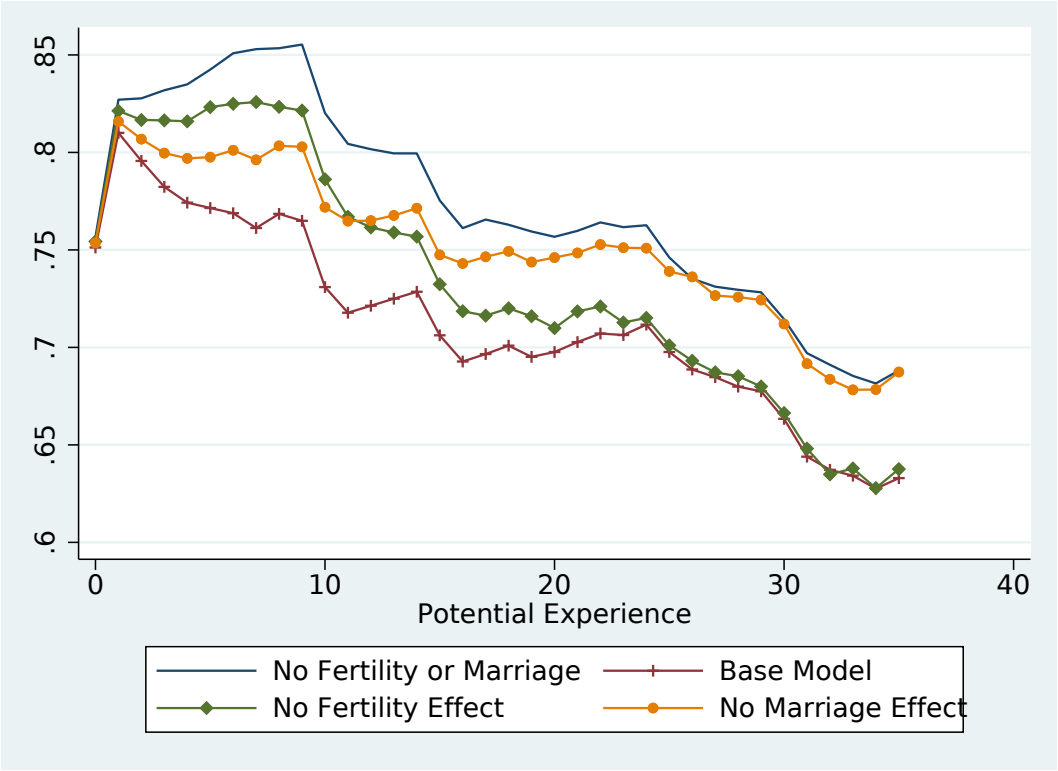
Next we turn to the main aim of this paper and simulate a model in which the relationship between fertility and work is relaxed to see how that affects human capital accumulation. That is, we compare our baseline model to a counterfactual in which children at home have no effect on working. Specifically, the effect of "Number of Kids < 7" on finding a job and leaving a job are set to zero (see Table 3a) and the probability of leaving the work force immediately upon having a child is also set to zero (Table 3b). We also find large effects of marriage on labor supply (Table 3a) so we also consider a counterfactual in which we eliminate this effect. Our third counterfactual eliminates both the marriage and fertility effects.

The direct effects on labor supply are presented in in Figure 3. First, examining the fertility effects, one can see that eliminating these would lead to a a considerable increase in labor force participation during the prime child-bearing years that dissipates as women age. The difference peaks at around 10 years of potential experience at a level of roughly 10%. In our view, this is a substantial effect, but it is not enormous. This is not that surprising based on the raw data.[12] Many women stop working while they have young children, but

[12]One can see in the fixed effect work regression in Table C2 (Sauer and Taber, 2021) that the coefficients on young children are of a similar magnitude.

most do not. The marriage pattern is (not surprisingly) quite different over the lifecycle. It is substantially smaller early in the lifecycle, but persists much longer.

Figure 3: Labor Supply Counterfactual

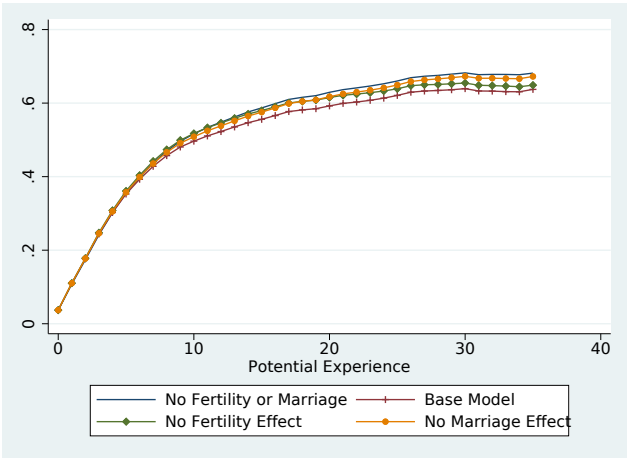


We next examine the effect of these labor supply effects on human capital accumulation. These simulations are not analogous to those in Figure 1, as to be in the actual wage regression, a woman needs to be working. This means that the shape of the profile in Figure 1 depends not just on human capital accumulation but also on selection into who is working. Since our counterfactual involves a change in working, there will be a selection effect that will influence the profile. We can avoid this problem when we simulate the model because we can simulate a counterfactual wage and a level of human capital for everyone - those working and those not working. This is what we present.

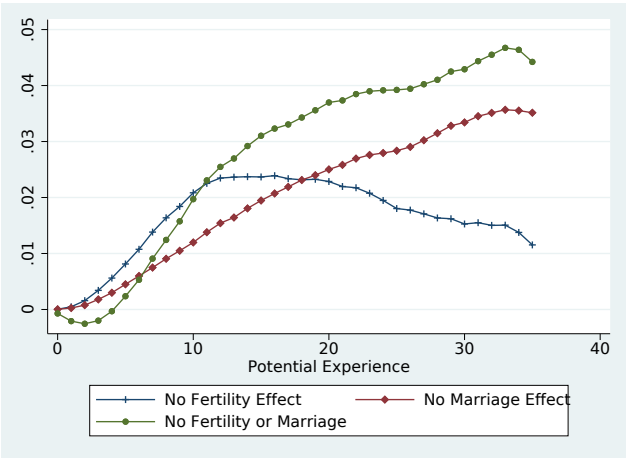
Figure 4a presents a simulation of the level of human capital (H_{it} in our model) at different ages. The labels are analogous to the counterfactuals presented in Figure 3. One can see that the loss in labor supply from fertility and marriage does suppress human capital but it is relatively modest. To put this simulation in a more familiar context, we calculate the difference in human capital between each of the three counterfactuals and the baseline and plot it in Figure 4b. For the fertility counterfactual, the difference in wages peaks around experience levels 10-20 at a difference of somewhat over .024. This suggests that, on

Figure 4: Human Capital Counterfactual

(a) Human Capital Levels



(b) Human Capital Differences



average, wages of women at these ages would be about two percent larger if there was no effect of fertility on labor supply. Again, this is a non-trivial effect, but when compared to the difference in log wages between men and women it is quite modest.[13] When adding the marriage effect as well, one gets more substantial effects of up to 4.7%. While these effects go in the right direction, they are small relative to the wage growth differences between men and women leaving room for other channels such as discrimination (perhaps in the form of glass ceilings) to be important.

8 Conclusion

In this paper, a continuous time Markov model of female work, marriage, and fertility is estimated using data from the Survey of Income and Program Participation. The model provides a good fit of the data. Two different types of counterfactuals are simulated with the estimated parameters. The first seeks to understand whether the curvature in female wage growth is determined primarily by curvature in the human capital accumulation function as a function of previous human capital, or if it is primarily driven by age. The results strongly suggest that curvature in the human capital production function is the driving force. The second counterfactual attempts to uncover the extent to which dropping out of the labor

[13]The two percent motherhood wage penalty we find is greater than the near zero penalty found in Hill (1979) and Korenman and Neumark (1992), but less than the penalty amongst mothers with two or more children found in Waldfogel (1998a), Waldfogel (1998b), and Anderson, Binder, and Krause (2002). Our estimate is similar in magnitude to the wage penalties for the first child found in Loughran and Zissimopoulos (2007) and Miller (2011). Our other estimates are less directly comparable to previous findings but are consistent with results in Adda, Dustmann, and Stevens (2017) and Braga (2013).

force amongst females, for fertility or marriage related reasons, suppresses human capital accumulation. Our finding is that it does so to a modest extent. Wages among prime age women would be approximately 2.4% higher if the relationship between fertility and working were eliminated and up to an additional 4.7% higher if the marriage effect were also eliminated.

The study also illustrates how to use importance sampling weights to smooth the objective function for indirect inference with discrete endogenous variables. Our procedure requires calculating the likelihood contribution for each observation in the sample at an initial trial vector of structural parameters. This constitutes the denominator of the weight, which remains fixed during minimum distance iterations. The numerator of the weight is the likelihood contribution at the updated vector of trial parameters. At each iteration, the likelihood ratio is the importance sampling weight used in estimation of the auxiliary model. The importance sampling weights can be formed with either the exact likelihood of the structural model or a simulated likelihood in case the former is difficult to construct. Our Monte Carlo study suggests potentially large gains in speed can be gained from smoothing the objective function in such a way.

References

Ackerberg, D. A. (2009). A new use of importance sampling to reduce computational burden in simulation estimation. *Quantitative Marketing and Economics 7*(4), 343–376.

Adda, J., C. Dustmann, and K. Stevens (2017). The career costs of children. *Journal of Political Economy 125*(2), 293–337.

Altonji, J. G., A. A. Smith, and I. Vidangos (2013). Modeling earnings dynamics. *Econometrica 81*(4), 1395–1454.

Altug, S. and R. Miller (1998). The effect of work experience on female wages and the effect of work experience on female wages and labour supply. *Review of Economic Studies*.

An, M. Y. and M. Liu (2000). Using indirect inference to solve the initial-conditions problem. *The Review of Economics and Statistics 82*(4), 656–667.

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

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

Baum, C. L. (2002). The efffect of work interruptions on women's wages. *LABOUR 16*(1), 1–37.

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

Blundell, R., M. Costa Dias, C. Meghir, and J. Shaw (2016, September). Female labour supply, human capital and welfare reform. *Econometrica*, 1705–1753.

Braga, B. (2013). Schooling, experience, career interruptions, and earnings.

Bruins, M., J. Duffy, M. Keane, and A. Smith (2018, July). Generalized indirect inference for discrete choice models. *Journal of Econometrics 205*(1), 177–203.

Budig, M. J. and P. England (2001). The wage penalty for motherhood. *American Sociological Review 66*(2), 204–225.

Daniel, F.-K., A. Lacuesta, and N. Rodríguez-Planas (2013). The motherhood earnings

dip: Evidence from administrative records. *Journal of Human Resources 48*(1), 169–197.

Eckstein, Z. and K. I. Wolpin (1989). Dynamic labour force participation of married women and endogenous work experience. *The Review of Economic Studies 56*(3), 375–390.

Fan, X., A. Seshadri, and C. Taber (2019, December). Estimation of a life-cycle model with human capital, labor supply and retirement.

Francesconi, M. (2002). A joint dynamic model of fertility and work of married women. *Journal of Labor Economics 20*(2), 336–380.

Fu, C. and J. Gregory (2019, March). Estimation of an equilibrium model with externalities: Combining the strengths of structural models and quasi-experiments. *Econometrica 87*(2), 387–421.

Gangl, M. and A. Ziefle (2009). Motherhood, labor force behavior, and women's careers: An empirical assessment of the wage penalty for motherhood in britain, germany, and the united states. *Demography 46*(2), 341–369.

Gayle, G.-L., A. Hincapie, and R. Miller (2018). Life-cycle fertility and human capital accumulation.

Gladden, T. and C. Taber (2000). Wage progression among less skilled workers. In Card and Blank (Eds.), *Finding Jobs, Work and Welfare Reform*, pp. 160–192. Russel Sage.

Gourieroux, C. and A. Monfort (1996). *Simulation-Based Econometric Methods*. Oxford University Press.

Gourieroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics 8*(S1), S85–S118.

Han, J. (2016). *Three Essays on Life-Cycle Labor Supply and Human Capital Formation*. Ph. D. thesis, University of Wisconsin-Madison.

Heckman, J. J. and J. R. Walker (1990). The relationship between wages and income and the timing and spacing of births: Evidence from swedish longitudinal data. *Econometrica 58*(6), 1411–1441.

Hill, M. S. (1979). The wage effects of marital status and children. *The Journal of Human Resources 14*(4), 579–594.

Hotz, V. J. and R. A. Miller (1988). An empirical analysis of life cycle fertility and female labor supply. *Econometrica 56*(1), 91–118.

Keane, M. P. and R. M. Sauer (2010). A computationally practical simulation estimation algorithm for dynamic panel data models with unobserved endogenous state variables*. *International Economic Review 51*(4), 925–958.

Keane, M. P. and K. Wolpin (2010). The role of labor and marriage markets, preference heterogeneity, and the welfare system in the life cycle decisions of black, hispanic, and white women. *International Economic Review 51*(3), 851–892.

Kloek, T. and H. K. van Dijk (1978). Bayesian estimates of equation system parameters: An application of integration by monte carlo. *Econometrica 46*(1), 1–19.

Korenman, S. and D. Neumark (1992). Marriage, motherhood, and wages. *The Journal of Human Resources 27*(2), 233–255.

Lee, Y. (2012). *Labor Supply Effects of the Earned Income Tax Credit with Labor Supply Restrictions*. Ph. D. thesis, University of Wisconsin-Madison.

Light, A. and M. Ureta (1995). Early-career work experience and gender wage differentials. *Journal of Labor Economics 13*(1), 121–154.

Loughran, D. and J. Zissimopoulos (2007). Why wait? the effect of marriage and childbearing on the wages of men and women. *Journal of Human Resources*.

Lundberg, S. and E. Rose (2000). Parenthood and the earnings of married men and women. *Labour Economics 7*(6), 689–710.

Magnac, T., J.-M. Robin, and M. Visser (1995). Analysing incomplete individual employment histories using indirect inference. *Journal of Applied Econometrics 10*(S1), S153–S169.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica 57*(5), 995–1026.

Miller, A. R. (2011). The effects of motherhood timing on career path. *Journal of Population Economics 24*(3), 1071–1100.

Moffitt, R. (1984). Profiles of fertility, labour supply and wages of married women: A complete life-cycle model. *The Review of Economic Studies 51*(2), 263–278.

Nagypál, É. (2007). Learning by doing vs. learning about match quality: Can we tell them apart? *The Review of Economic Studies 74*(2), 537.

Pal, I. and J. Waldfogel (2014, August). Re-visiting the family gap in pay in the united states.

Polachek, S. W. (1981). Occupational self-selection: A human capital approach to sex differences in occupational structure. *The Review of Economics and Statistics 63*(1), 60–69.

Sauer, R. M. and C. Taber (2021). *Supplemental Appendix to Understanding Women's Wage Growth using Indirect Inference with Importance Sampling.*

Sheran, M. (2007). The career and family choices of women: A dynamic analysis of labor force participation, schooling, marriage, and fertility decisions. *Review of Economic Dynamics 10*(3), 367–399.

Smith, A. (1990). *Three Essays on the Solution and Estimation of Dynamic Macroeconomic Models*. Ph. D. thesis, Duke University.

Smith, A. A. (1993). Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics 8*(S1), S63–S84.

Taber, C. and R. Vejlin (2020, May). Estimation of a roy/search/compensating differential model of the labor maket. *Econometrica 88*(3), 1031–1069.

Van Der Klaauw, W. (1996). Female labour supply and marital status decisions: A life-cycle model. *The Review of Economic Studies 63*(2), 199–235.

Waldfogel, J. (1997). The effect of children on women's wages. *American Sociological Review 62*(2), 209–217.

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*(3), 505–545.

Waldfogel, J. (1998b). Understanding the "family gap" in pay for women with children. *The Journal of Economic Perspectives 12*(1), 137–156.

Weiss, Y. and R. Gronau (1981). Expected interruptions in labour force participation and sex-related differences in earnings growth. *The Review of Economic Studies 48*(4), 607–619.

Wilde, E., L. Batchelder, and D. Ellwood (2010). The mommy track divides: The impact of childbearing on wages of women of differeing skill levels. *NBER Working Paper 16582*.