NBER WORKING PAPER SERIES

INDIRECT INFERENCE WITH IMPORTANCE SAMPLING:
AN APPLICATION TO WOMEN'S WAGE GROWTH

Robert M. Sauer
Christopher R. Taber

Working Paper 23669
http://www.nber.org/papers/w23669

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2017

The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Indirect Inference with Importance Sampling: An Application to Women's Wage Growth
Robert M. Sauer and Christopher R. Taber
NBER Working Paper No. 23669
August 2017
JEL No. C51,J16

ABSTRACT

This paper has two main parts. In the first, we describe a method that smooths the objective function in a general class of indirect inference models. Our smoothing procedure makes use of importance sampling weights in estimation of the auxiliary model on simulated data. The importance sampling weights are constructed from likelihood contributions implied by the structural model. Since this approach does not require transformations of endogenous variables in the structural model, we avoid the potential approximation errors that may arise in other smoothing approaches for indirect inference. We show that our alternative smoothing method yields consistent estimates. The second part of the paper applies the method to estimating the effect of women's fertility on their human capital accumulation. We find that the curvature in the wage profile is determined primarily by curvature in the human capital accumulation function as a function of previous human capital, as opposed to being driven primarily by age. We also find a modest effect of fertility induced nonemployment spells on human capital accumulation. We estimate that the difference in wages among prime age women would be approximately 3% higher if the relationship between fertility and working were eliminated.

Robert M. Sauer
Royal Holloway College
University of London
Egham, Surrey TW20 0EX
United Kingdom
robert.sauer@rhul.ac.uk

Christopher R. Taber
Department of Economics
University of Wisconsin -Madison
1180 Observatory Dr
Social Sciences Building #6448
Madison, WI 53706-1320
and NBER
ctaber@ssc.wisc.edu

1 Introduction

Indirect inference is an increasingly common way to estimate complex econometric models. Similar to the simulated method of moments, it is a computationally practical technique since it relies on unconditional simulations of the model to obtain structural estimates. In contrast to the simulated method of moments, indirect inference involves estimating a reduced-form auxiliary model. The auxiliary model is estimated on the actual data as well as simulated data from the structural model. Structural estimates are found by minimizing the distance between the two sets of reduced-form auxiliary parameter estimates.

A nice property of indirect inference is that consistent structural estimates are obtained even if the the auxiliary model is not an exact reduced form of the structural model. The only requirement for identification and consistent estimation is that each structural parameter has an independent effect on at least one parameter of the auxiliary model. The specification of the auxiliary model has implications for efficiency only. When the auxiliary model is an exact reduced form of the structural model, indirect inference is analogous to maximum likelihood (Smith, 1993, 1990, Gourieroux, Monfort, and Renault, 1993, Gallant and Tauchen, 1996, and Gourieroux and Monfort, 1996).

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 often 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 leading to slow convergence and difficulties in obtaining standard errors. Moreover, non-differentiability of the objective function may lead to estimates that are consistent but not asymptotically normal (see Gallant and Tauchen, 1996, and Hansen, Heaton, and Luttmer, 1995, in the context of GMM). A main purpose of this paper is to propose a flexible strategy for this problem that involves the use of importance sampling.

In this paper, we explain how the problem of non-differentiability 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 is underused in the empirical literature using indirect inference models.

In order to deal with the problem of non-differentiability, Bruins et al. (2017) propose an alternative method, called generalized indirect inference (GII), that replaces the the discrete endogenous variables in the structural model with a logistic-kernel of simulated latent utilities (see also McFadden, 1989). Consistent structural parameter estimates are obtained as the smoothing parameter in the logistic kernel goes to zero. However, considerable approximation error and small sample bias may arise for any non-zero smoothing parameter, especially when discrete outcomes accumulate in the structural model. This arises in models that contain variables such as endogenous work experience and tenure (see Altonji et al., 2013). We view our approach as an additional tool rather than as an alternative to GII. For some problems GII may work better, but we suspect that there are others for which our approach is preferable.

The second main contribution of this paper is empirical. We are motivated by the finding that the wage profile of men is much steeper than that of women (see figure 1). We apply indirect inference with importance sampling in this context by estimating a continuous time Markov model of female work, marriage, and fertility using data from the Survey of Income and Program Participation. Our specific goal is to understand the importance of child care related non-employment spells in explaining the gender gap in wage growth. In the model, women move into and out of work as well as into and out of marriage. They also have children. Importantly, the number of children influence labor supply which affects human capital. Note that if the curvature of the female wage profile is purely due to previous human capital, when mothers re-enter the labor market they can expect rapid wage growth because they have relatively low levels of human capital. By contrast, if the flattening is due to age, and they re-enter mid-life, they would expect slower wage growth.

We have two main empirical findings. We show that the curvature in the wage profile is determined primarily by curvature in the human capital accumulation function as a function of previous human capital, as opposed to being driven primarily by age. We then measure the extent to which women's dropping out of the labor force for fertility 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 3% higher if fertility did not affect women's labor supply.

The rest of this paper is organized as follows. In the next section, we provide a more detailed background on the method of Bruins et al. (2017) and on importance sampling in order to put the alternatives in context. Section 3 defines the basic setup for indirect inference. In section 4, we describe our alternative method in more detail. Section 5 provides several examples. Section 6 includes our empirical work and section 7 concludes.

2 Background and Previous Work on Smoothing in Indirect Inference

Indirect inference has become a very important tool for estimation of econometric models. We avoid a long survey here as it has been discussed widely. Key papers are Smith (1993,1990) 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. That is, the auxiliary model is estimated on the real data and also estimated on data simulated from the model. The underlying parameters are chosen to minimize the distance between the model estimated auxiliary parameters and simulated auxiliary parameters. We discuss this in more detail below. The problem addressed in this paper is that the mapping between the underlying parameters and simulated auxiliary parameters is often not smooth which complicates estimation and inference. We show how to use importance weight sampling to smooth this relationship.

Another approach to smoothing is provided by the method of generalized indirect inference (GII) proposed by Bruins et al. 2017. It replaces the discrete outcome simulated by the structural model with a logistic-kernel transformation of simulated latent utilities. The resulting objective function is smooth because the latent utilities are smooth functions of the structural parameters, and the logistic-kernel is a smooth function of the latent utilities.

In order to illustrate GII formally, let $\widetilde{y}^h(\beta) \equiv \left\{\widetilde{y}^h_{it}(\beta)\right\}$, $h = 1, ..., H$, denote H statistically independent simulated choice sequences for a given trial vector of structural parameters, β, using the same set of observed exogenous variables $x \equiv \{x_{it}\}$. The auxiliary model is estimated separately for each $\left\{\widetilde{y}^h_{it}(\beta)\right\}$ yielding $\widetilde{\theta}_h(\beta)$. That is, $\widetilde{\theta}_h(\beta)$ solves:

$$\widetilde{\theta}_h(\beta) = \arg\max_{\theta} \mathcal{L}\left(\widetilde{y}^h(\beta); x, \theta\right)$$

where $\mathcal{L}(\cdot)$ is the likelihood associated with the auxiliary model.[1] Since $\widetilde{y}^h(\beta)$ is a step

[1] Auxiliary parameter estimates may be obtained using other statistical methods. Maximum likelihood is referred to without loss of generality.

function, so is $\widetilde{\theta}_h(\beta)$. The same is true for

$$\widetilde{\theta}(\beta) = \frac{1}{H}\sum_{h=1}^{H}\widetilde{\theta}_h(\beta)$$

whose distance from $\widehat{\theta}$ is minimized to obtain structural parameters estimates via indirect inference.

$\widehat{\theta}$ is obtained as

$$\widehat{\theta} = \arg\max_{\theta}\mathcal{L}(y; x, \theta)$$

using the observed choice sequence in the data $y \equiv \{y_{it}\}$.

Note that structural parameter estimates produced by indirect inference are consistent as long as $\widehat{\theta}$ and $\widetilde{\theta}(\beta_0)$ converge to the same "pseudo" true value $\theta_0 = h(\beta_0)$ as the sample size N grows. h is referred to as the "binding" function by Gourieroux, Monfort, and Renault (1993). Underlying GII, and our alternative method, is the insight that the estimation procedures applied to the observed and simulated data sets can differ, as long as they both yield consistent estimates of the same vector of pseudo true parameter values (see also Genton and Ronchetti, 2003). Our methods differ in how we smooth the function $\widetilde{\theta}(\beta)$, and avoid the need to optimize a step function.

In order to illustrate how GII smooths $\widetilde{\theta}(\beta)$, denote $\widetilde{u}^h_{it}(\beta)$ as individual i's set of latent utilities in time t for the first $J-1$ choice options in simulated data set m. Define a smooth function of latent utilities $g\left(\widetilde{u}^h_{it}(\beta), j; \lambda\right)$ such that $g\left(\widetilde{u}^h_{it}(\beta), j; \lambda\right)$ converges to $\widetilde{y}^h_{itj}(\beta)$ (the simulated choice out of the $J-1$ alternatives in time t) as the smoothing parameter λ goes to zero. GII substitutes $g\left(\widetilde{u}^h_{it}(\beta), j; \lambda\right)$ for $\widetilde{y}^h_{itj}(\beta)$ in computation of $\widetilde{\theta}(\beta)$. Thus, the objective function is smooth and $\widetilde{\theta}(\beta_0)$ converges to θ_0 as λ goes to zero and N goes to infinity, with H and T fixed.

The $g\left(\widetilde{u}^h_{it}(\beta), j; \lambda\right)$ Bruins et al. (2017) use in their Monte Carlo experiments is the logistic-kernel,

$$g\left(\widetilde{u}^h_{it}(\beta), j; \lambda\right) = \frac{\exp\left(\widetilde{u}^h_{itj}(\beta)/\lambda\right)}{1+\sum_{k=1}^{J-1}\exp\left(\widetilde{u}^h_{itk}(\beta)/\lambda\right)}.$$

Monte Carlo experiments on four specific types of discrete choice models show that their method, combined with very flexible auxiliary models, yields structural parameter estimates with little bias and an efficiency level close to that produced by simulated maximum likelihood (using the GHK algorithm).[2]

[2] The four discrete choice models are, i) a two-alternative dynamic probit model with serially correlated errors, ii) a two-alternative dynamic probit model with serially correlated errors and a lagged dependent variable, iii) the same model as in (ii) with an initial conditions problem, and iv) a static three-alternative probit model with contemporaneously correlated errors. The model with the initial conditions problem is not estimated by SML due to complications arising from the initial conditions problem.

In models where past simulated choices are transformed into a state variable which determines current period choices, the $g(\cdot)$ functions corresponding to previous period choices can be accumulated over time. This is a straightforward extension of the Bruins et al. (2017) approach. However, this strategy can further exacerbate the inherent measurement error problem (see Altonji, Smith, and Vidangos, 2013 and Guvenen and Smith, 2014).

Importance sampling has a long history, and 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) who discusses cases in which it can greatly simplify simulated method of moments and simulated maximum likelihood models. Our main methodological contribution is to extend this methodology to a general class of indirect inference models as well as providing some additional examples. Lee (2012), Han (2016), and Fu and Gregory (2016) have applied our approach.

3 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 our importance weight sampling is natural. We also focus on the cross section and panel data version 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) which are i.i.d. and both X_i and Y_i are potentially large dimensional (K_x and K_y). Here 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.

We write the data generating process as

$$Y_i \equiv y(X_i, u_i; \theta)$$

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

$$\Psi(u_i; \theta).$$

Both Ψ and y are known up to parameter $\theta \in \Theta \subset \Re^{K_\theta}$ and we write the true value as θ_0. We mean this notation to be general enough to represent a complicated system with lagged dependent variables and/or equlibrium, but we assume it can be written in reduced form using y.[3]

To apply indirect inference assume that we can write our auxiliary model as

$$\widehat{\beta} = argmin_{\beta} F\left(\frac{1}{N}\sum_{i=1}^{N} g(X_i, Y_i, \beta), \beta\right)$$

where $\beta \in \mathcal{B} \subset \mathbb{R}^{K_\beta}$. Here we take g to be a function $g : \mathbb{R}^{K_x} \times \mathbb{R}^{K_y} \times \mathcal{B} \to \mathbb{R}^{K_g}$ while $F : \mathbb{R}^{K_g + K_\beta} \to \mathbb{R}$. We have written this in a general enough way that it can incorporate a typical M-estimator such as maximum likelihood. In that case g would be the negative of the log-likelihood function and F would be degenerate. It can also incorporate a Generalized Method of Moments type estimator in which g would be the moments so that $E\left[g(X_i, Y_i, \beta)\right] = 0$ and F would be the function:

$$F(g, \beta) = g'Wg$$

where W is some weighting matrix. We could also use it to represent a quantile or quantile regression.

Define the population value of $\widehat{\beta}$ to be

$$\beta_0 \equiv argmin_{\beta} F(E\left[g(X_i, Y_i, \beta)\right], \beta).$$

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

$$G(\theta, \beta) \equiv \int\int g(x, y(x, u; \theta), \beta) dF\left(u \mid \theta\right) d\Xi_0(x)$$

and what Gourieroux, Montfort and Renault (1993) refer to as the binding function

$$B(\theta) = argmin_{\beta} F\left(G(\theta, \beta), \beta\right).$$

Note that $B(\theta_0) = \beta_0$. Identification of Ξ_0 is straight forward since X_i is observable, which is why we mostly abstract from it. Essentially what one needs for identification is that this equation 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 known. Thus, we could just invert β to obtain an estimate of θ_0.

In practice we typically do not have a closed form for B. Instead, we need to use simulation

[3] 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 mulitiple equlibria the code must have some form of choosing between them. The same mechanism would be incorporated into y.

estimators in order to approximate $B(\theta)$. A typical approach is to generate H difference simulated samples each with size S. For each observation we draw u_{hs} randomly from the distribution Ψ calculate X_{hs} from the empirical distribution of X_i.[4] We then define

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

and choose

$$\widehat{\theta} = argmin_{\theta}\left(\widetilde{B}(\theta) - \widehat{\beta}\right)' \Omega \left(\widetilde{B}(\theta) - \widehat{\beta}\right)$$

where Ω is a weighting matrix.

4 Indirect Inference with Importance Sampling Weights

A major problem with this procedure is that often some components the dependent variable vector, $Y_{hs}(\theta)$ is discrete so that a small change in the parameters can lead to discontinuous jumps in $y(X_i, u_i;\theta)$. This leads the objective function to be discontinuous as well. In principle, with enough simulations we could make this as smooth as we would like, but in practice this can be a major problem and minimizing the objective function can be difficult. The Bruins et al. (2017) approach is to smooth $Y_{hs}(\theta)$. There are two drawbacks to this approach. First, one must choose a smoothing parameter. Second, it will not work for all cases.[5] Hence, we describe an alternative method.

We assume that there is missing data, so the econometrician does not generally get to observe the full set of potential data, but only a subset of it. The key component of our analysis is Υ_i which one can think of as a superset of the data. The actual empirical economist does not get to observe Υ_i, but the data simulator can observe this. In a typical problem there are multiple ways to choose Υ_i, and finding the best one will be computationally very important. This is essentially what Ackerberg (2009) discusses as a "change in variables."[6] He makes some stronger assumptions on these variables than we do which give particularly nice computational results. We are somewhat more general, but the the basic idea is similar.

[4] 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.

[5] At least there are models in which we have not figured out how to use their approach. Our example 5.4 and empirical example are cases.

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

We express the data generation process as:

$$\begin{aligned}\Upsilon_i \sim & \ell(\cdot \mid X_i;\theta)\\ Y_i = & y_\Upsilon\left(\Upsilon_i, X_i;\theta\right)\end{aligned}$$

where the likelihood function ℓ and the data generating function y_Υ are known up to parameter θ. While the distinction between Υ_i and Y_i may seem arbitrary at this point, its usefulness should become clear in the examples below. What will be important to get the model to be well behaved is that ℓ and y_Υ should be differentiable in θ and that ℓ should be relatively easy to compute.

Our approach 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 the distribution $\ell_0(\Upsilon_{hs}; X_{hs})$.

1. For any given θ Calculate $\ell(\Upsilon_{hs}; X_{hs}, \theta)$ as the likelihood function of Υ_{hs} given θ

   (a) Calculate

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

Now choose

$$\widehat{\theta} = argmin_\theta \left(\widetilde{B}(\theta) - \widehat{\beta}\right)' \Omega \left(\widetilde{B}(\theta) - \widehat{\beta}\right).$$

First, note that standard indirect inference as it typically is practiced is a special case of this. To avoid jumps in the objective function researchers typically draw the random variables that determine outcomes first and then fix these values through the estimation. For example if the distribution of an underlying random variable u_{hs} does not depend on θ, one would draw the u_{hs} one time at the beginning of the program 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 $\widetilde{B}(\theta)$ will be continuous and differentiable functions of θ as long as ℓ and F are continuous and differentiable functions. 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.

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

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

Thus, using importance sampling gives a consistent estimate of the function $G(\theta, \beta)$. Importantly, we will approximate this integral using a Monte Carlo procedure where we draw Υ_{hs} from the distribution $\ell_0(\Upsilon_{hs})$, then

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

but most important, as long as $\ell(\Upsilon_{hs};\theta)$ and $y_\Upsilon\left(\Upsilon_{hs};\theta\right)$ are smooth functions of θ, then this approximation is a smooth function of θ.

4.1 Consistency

We first show consistency of the estimator.

Define

$$\begin{aligned} \widehat{G}(\beta) \equiv& \frac{1}{N}\sum_{i=1}^{N} g(X_i, Y_i, \beta) \\ \widetilde{G}_h(\theta,\beta) \equiv& \frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})} g(X_{hs}, Y_{hs}; \beta). \end{aligned}$$

We need the following assumptions:

Assumption 1 *$\widehat{G}(\beta)$ converges uniformly in probability to $G(\theta_0, \beta)$.*

The key aspect of this is that g is well behaved so that this convergence is uniform. Note that we are being general enough not to require the expressions to be differential in the underlying function θ but are assuming that the auxiliary model that we estimate on the actual data is simple.

The next are standard regularity assumptions as well as a condition for identification of $B(\theta)$.

Assumption 2 *Θ and $\mathcal{B}$ are compact.*

Assumption 3 *For each $\theta \in \Theta$, $B(\theta)$ is a singleton. F,g, and B are continuous.*

Presumably one could relax the assumption of point identification of $B(\theta)$ allowing this to be a set and modify the objective function so that the set $B(\widehat{\theta})$ is close to the set $\widehat{\beta}$. This seems straight forward, but we do not know of an empirical researcher that has done this, so we focus on the point identified case.

Next we have the identification assumption for θ :

Assumption 4 *If $\theta_1 \neq \theta_2$ then*

$$B(\theta_1) \neq B(\theta_2).$$

If this assumption were relaxed we would no longer obtain point identification, but would instead obtain set identification.

Assumption 5 *We can write $\Upsilon_{hs} = \left\{\Upsilon^d_{hs}, \Upsilon^c_{hs}\right\}$ where Υ^d_{hs} is discrete taking on values $\Upsilon^d_{(1)}, ..., \Upsilon^d_{(K_\Upsilon)}$ and Υ^c_{hs} is continuous. For every $\theta \in \Theta$, the support of Υ_{hs} generated by $\ell(\Upsilon_{hs}; X_{hs}, \theta)$ is a subset of or equal to the support of Υ_{hs} generated by $\ell_0(\Upsilon_{hs}; X_{hs})$.*

This assumption makes the likelihood function easy to write down. We could easily extend the results to accommodate other specific cases.

Assumption 6 *For each simulation $h = 1, .., H$,*

$$\frac{1}{S}\sum_{s=1}^{S} \frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})} g(X_{hs}, Y_{hs}; \beta))$$

converges uniformly in probability over β and θ to

$$E_s\left(\frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})} g(X_{hs}, Y_{hs}; \beta); \Xi_0, \ell_0\right)$$

where E_s represents the expected value when the data are generated from a simulation in which X_{hs} is drawn from Ξ_0 and Υ_{hs} is drawn from $\ell_0(\Upsilon; X_{hs})$.

There are really three separate aspects of this assumption. First that we are drawing Υ_{hs} from $\ell_0(\Upsilon; X_{hs})$ which is a fundamental part of the importance weight sampling approach. The second is that convergence is uniform which is standard. The third aspect reflects how X_{hs} is chosen. For calculating the asymptotic distribution we will need to put more structure on this, but here we just require that asymptotically it is drawn from the true distribution. There are many ways to do this, and we will discuss this in the next section.

Theorem 1 *Under Assumptions 1-6, $\widehat{\theta}$ converges in probability to θ.*

4.2 Asymptotic Distribution

We now explicitly define $G_j, \widehat{G}_j$, and $\widetilde{G}_j$ to be the j^{th} element of $G, \widehat{G}$, and $\widetilde{G}_h$ respectively. We first assume the following regularity conditions. These are weak assumptions that are standard and will hold in typical applications. The first is standard.

Assumption 7 *$B(\theta)$ is differentiable with*

$$B_\theta \equiv \frac{dB(\theta_0)}{d\theta'}$$

and $B_\theta' \Omega B_\theta$ is of full rank and θ_0 is an interior point.

The second is an assumption about stochastic equicontinuity. In small samples our estimator is potentially discontinuous in θ but it converges to a smooth function.

Assumption 8 *For and δ_N,*

$$\sup_{\|\theta-\theta_0\| \le \delta_N} \frac{\sqrt{N}\left\|\widetilde{B}(\theta) - \widetilde{B}(\theta_0) - B(\theta) + B(\theta_0)\right\|}{1+\sqrt{N}\|\theta-\theta_0\|} \xrightarrow{p} 0.$$

Finally we use,

Assumption 9 *$\widehat{G}$ and $\widetilde{G}$ are differenable in β. Letting the notation $\overset{U_p}{\to}$ denote uniform* *convergence in probability,*

$$\begin{aligned}
\frac{\partial \widehat{G}_j\left(\beta\right)}{\partial \beta} &\xrightarrow{U_p} \frac{\partial G_j\left(\theta_0,\beta\right)}{\partial \beta} \\
\frac{\partial^2 \widehat{G}_j\left(\beta\right)}{\partial \beta \partial \beta'} &\xrightarrow{U_p} \frac{\partial^2 G_j\left(\theta_0,\beta\right)}{\partial \beta \partial \beta'} \\
\frac{\partial \widetilde{G}_j\left(\theta_0,\beta\right)}{\partial \beta} &\xrightarrow{U_p} \frac{\partial G_j\left(\beta\right)}{\partial \beta} \\
\frac{\partial^2 \widetilde{G}_j\left(\theta_0,\beta\right)}{\partial \beta \partial \beta'} &\xrightarrow{U_p} \frac{\partial^2 G_j(\beta)}{\partial \beta \partial \beta'}
\end{aligned}$$

and all of these objects are continuous in their arguments.

The differentiability rules out some interesting cases like quantile regression. Extending this to allow for more complicated cases should be straight forward but our main goal is to provide the formula for the asymptotic variance in the typical case rather than the most general case.

Let $\beta_0 = B\left(\theta_0\right)$, we also need

Assumption 10 *F is two time (totally) continuously differential and define*

$$F_{\beta\beta} \equiv \frac{d^2 F\left(G\left(\theta_0,\beta_0\right),\beta_0\right)}{d\beta d\beta'}$$

and assume $F_{\beta\beta}$ is of full rank.

We assume that the X_{hs} are composed of actual values that we see in the data. Let M_{hi} be the total number of times X_i is used for each simulated data set h. This can pick up two important cases. In one case we let each simulated data set be the same size as the actual data ($S = N$) and each value of X_i is used once so $X_{hs} = X_s$. In this case $M_{hi} = 1$ for every i. The other case is one in which X_{hs} is drawn from the empirical distribution. In this case M_{hi} is a random variable taking integer values with expected value S/N. Of course it can cover other cases as well, for example if $S = 2N$ and each observable is used twice.

To simplify notation let Υ_{him} denote the m^{th} simulation using observation i for sample h. Define

$$\widetilde{g}_{hi}\left(\beta\right) \equiv \frac{N}{S}\sum_{m=1}^{M_{hi}} \frac{\ell\left(\Upsilon_{him};X_i,\theta_0\right)}{\ell_0(\Upsilon_{him};X_i)} g\left(X_i, y_{\Upsilon}\left(\Upsilon_{him},X_i;\theta_0\right),\beta\right)$$

Notice that this means that

$$\begin{aligned}\widetilde{G}_h(\theta_0,\beta) =& \frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}g(X_{hs},Y_{hs},\beta)\\ =& \frac{1}{N}\sum_{i=1}^{N}\widetilde{g}_{hi}\left(\beta\right).\end{aligned}$$

Define

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

Let V be the variance of $\left(\left[\frac{1}{H}\sum_{h=1}^{H}\widetilde{\vartheta}_{hi}\right]-\vartheta_i\right)$.

Theorem 2 *Under Assumptions ..., $\sqrt{N}\left(\widehat{\theta}-\theta\right)$ converges in distribution to a normal random variable with expected value 0 and variance*

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

5 Examples

We present a number of different examples starting with very simple ones. One wouldn't need to use our methodology in the first two cases, but they are nice for demonstrating the basic ideas. The third and fourth examples are more complicated problems for which this approach would be well suited.

5.1 Example 1: Bernoulli Random Variables

The easiest way to see our approach is with a Bernoulli random variable. We repeat the statement above: this is a simple problem that one would never use indirect inference to estimate, but works nicely for showing the issue in standard indirect inference (or Simulated Method of Moments) and how the approach works. Since many problems involve simulating discrete random variables, this is the essence of the problem.

Suppose that

$$Y_i = \begin{cases} 1 & \text{with probability } \rho_0 \\ 0 & \text{with probability } 1-\rho_0 \end{cases}.$$

To estimate the model in a standard way one would just use the sample mean which is the maximum likelihood estimator

$$\widehat{\rho}_{mle} = \bar{Y}.$$

Given this, if one were to think about estimating p using indirect inference the most natural auxiliary model would be the sample mean $\bar{Y}$. Focusing on the case in which we only simulate once ($H = 1$), the standard way to estimate using indirect inference would be to draw a sample of S random variables from a uniform distribution, $u_1, ..., u_S$. We could then simulate $\bar{Y}$ for any given level of p as

$$\widetilde{Y}_1(\rho) = \frac{1}{S}\sum_{s=1}^{S} 1(u_s < \rho).$$

We then choose our indirect inference estimator

$$\widehat{\rho}_1 = argmin_p \left(\bar{Y} - \widetilde{Y}_1(\rho)\right).$$

The problem is that $\widetilde{Y}(p)$ is a step function so solving for ρ and finding standard errors is problematic.

To simulate using importance sampling weights, rather than simulate u_s we simulate Y_s directly from some initial probability, say μ, so Y_s is drawn from a Bernoulli distribution with probability μ. The simplest simulator for the auxiliary model in this case is

$$\widetilde{Y}_2(\rho) = \frac{1}{S}\sum_{s=1}^{S} Y_s \frac{\ell(Y_s;\rho)}{\ell_0(Y_s)}$$

where the likelihood functions are

$$\begin{aligned} \ell(Y_s;\rho) &= Y_s\rho + (1-Y_s)(1-\rho) \\ \ell_0(Y_s) &= Y_s\mu + (1-Y_s)(1-\mu). \end{aligned}$$

To see why $\widetilde{Y}_1(\rho)$ is consistent, let $\overline{Y}_s$ be the fraction of the simulated sample for which $Y_s = 1$ and divide the numerator and denominator by S, then

$$\begin{aligned} \widetilde{Y}_2(\rho) =& \frac{1}{S}\sum_{s=1}^{S} Y_s \frac{\ell(Y_s;\rho)}{\ell_0(Y_s)}. \\ =& \overline{Y}_s \frac{\rho}{\mu} \\ \xrightarrow{p}& \rho \end{aligned}$$

because $\overline{Y}_s \xrightarrow{p} \mu$ as S gets large.

The estimator is

$$\widehat{\rho}_2 = argmin_p \left(\bar{Y} - \widetilde{Y}_2(\rho)\right)$$

so the solution is

$$\widehat{\rho}_2 = \bar{Y}\frac{\mu}{\overline{Y}_s}.$$

To see how this fits into our notation above with H potentially greater than one, we would choose:

$$\begin{aligned} g(X_i, Y_i; \beta) &= Y_i \\ F(d,\beta) &= (d-\beta)^2 \\ \Upsilon_i &= Y_i. \end{aligned}$$

Then

$$\begin{aligned} \widetilde{B}(p) \equiv& \frac{1}{H}\sum_{h=1}^{H} argmin_\beta \left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};\rho)}{\ell_0(\Upsilon_{hs})}Y_{hs} - \beta\right)^2 \\ =& \frac{1}{H}\sum_{h=1}^{H}\frac{1}{S}\sum_{s=1}^{S}\frac{\rho}{\mu}Y_{hs}. \\ =& \frac{\rho}{\mu}\frac{\sum_{h=1}^{H}\sum_{s=1}^{S} Y_{hs}}{H+S} \end{aligned}$$

5.2 Example 2: Logit Model

Our next example is a logit model. The true model is

$$Pr(Y_i = 1 \mid X_i) = \Lambda(X_i'\theta_0)$$

where Λ denotes the logit cdf. We use the linear probability model as our auxiliary model. We can put this into our notation by ignoring F[7] and choosing

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

We generate the simulated data in the following way.

1. Choose X_{hs} by drawing randomly from the empirical distribution of X_i[8]

2. Choose some initial logit value θ^*

3. Simulate Y_{hs} so that

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

Once again we will choose $\Upsilon_{hs} = Y_{hs.}$

For this model note that

$$\begin{aligned} W_{hs}\left(\theta\right) &\equiv \frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})} \\ &= \frac{Y_{hs}\Lambda\left(X_{hs}'\theta\right) + \left(1 - Y_{hs.}\right)\left(1 - \Lambda\left(X_{hs}'\theta\right)\right)}{Y_{hs}\Lambda\left(X_{hs}'\theta^*\right) + \left(1 - Y_{hs.}\right)\left(1 - \Lambda\left(X_{hs}'\theta^*\right)\right)} \end{aligned}$$

so

[7] That is F is just the identity function.

[8] As discussed above this could be with or without replacement. Of course, if you did it without replacement and your simulation sample is larger than your original one, you would have to replenish it once you have run through the full sample.

$$
\begin{aligned}
B(\theta) \equiv & \frac{1}{H}\sum_{h=1}^{H} argmin_{\beta} F\left(\frac{1}{S}\sum_{s=1}^{S} W_{hs}\left(\theta\right) g(X_{hs}, Y_{hs}; \widehat{\beta}), \widehat{\beta}\right)\\
= & \frac{1}{H}\sum_{h=1}^{H} argmin_{\beta} \frac{1}{S}\sum_{s=1}^{S} W_{hs}\left(\theta\right)\left(Y_{hs} - X'_{hs}B(\widehat{\beta})\right)^{2}\\
= & \frac{1}{H}\sum_{h=1}^{H}\left(\sum_{s=1}^{S} W_{hs}\left(\theta\right) X_{hs}X'_{hs}\right)^{-1}\left(\sum_{s=1}^{S} W_{hs}\left(\theta\right) X_{hs}Y_{hs}\right).
\end{aligned}
$$

Clearly this is just H weighted regressions with weights $W_{hs}(\theta)$. Also, since $W_{hs}(\theta)$ is differentiable in θ, so is $B(\theta)$. To see why this works note that

$$
\begin{aligned}
\frac{1}{S}\sum_{s=1}^{S} W_{hs}\left(\theta\right) X_{hs}X'_{hs} \underset{S\to\infty}{\overset{p}{\rightarrow}} & E\left(W_{hs}\left(\theta\right) X_{hs}X'_{hs}\right)\\
= & E\left(X_{hs}X'_{hs}E\left[\frac{Y_{hs}\Lambda\left(X'_{hs}\theta\right) + \left(1 - Y_{hs}\right)\left(1 - \Lambda\left(X'_{hs}\theta\right)\right)}{Y_{hs}\Lambda\left(X'_{hs}\theta^{*}\right) + \left(1 - Y_{hs}\right)\left(1 - \Lambda\left(X'_{hs}\theta^{*}\right)\right)} \mid X_{hs}\right]\right)\\
= & E\left(X_{hs}X'_{hs}\left[\frac{\Lambda\left(X'_{hs}\theta\right)}{\Lambda\left(X'_{hs}\theta^{*}\right)}\Lambda\left(X'_{hs}\theta^{*}\right) + \frac{\left(1 - \Lambda\left(X'_{hs}\theta\right)\right)}{\left(1 - \Lambda\left(X'_{hs}\theta^{*}\right)\right)}\left(1 - \Lambda\left(X'_{hs}\theta^{*}\right)\right)\right]\right)\\
= & E\left(X_{i}X'_{i}\right)
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{1}{S}\sum_{s=1}^{S} W_{hs}(\theta_0)X_{hs}Y_{hs} \underset{S\to\infty}{\overset{p}{\rightarrow}} & E\left(X_{hs}E\left[Y_{hs}\frac{Y_{hs}\Lambda\left(X'_{hs}\theta_0\right) + \left(1 - Y_{hs}\right)\left(1 - \Lambda\left(X'_{hs}\theta_0\right)\right)}{Y_{hs}\Lambda\left(X'_{hs}\theta^{*}\right) + \left(1 - Y_{hs}\right)\left(1 - \Lambda\left(X'_{hs}\theta^{*}\right)\right)} \mid X_{hs}\right]\right)\\
= & E\left(X_{hs}\left[\frac{\Lambda\left(X'_{hs}\theta_0\right)}{\Lambda\left(X'_{hs}\theta^{*}\right)}\Lambda\left(X'_{hs}\theta^{*}\right)\right]\right)\\
= & E(X_iY_i).
\end{aligned}
$$

Thus, this procedure will give a consistent estimate (i.e. plim$(B(\theta_0) = \beta_0)$.

5.3 Example 3: Dynamic Labor Force Model

Now we consider a more complicated case in which one might actually want to use this approach. The main issues and modeling is analogous to the previous two examples. In this dynamic labor force model, we assume that X_{it} is observable and v_i is unobservable. We will think of d_{it} as employment and that we have panel data on employment, and when the worker is employed we also observe the wage. The underlying data process is

$$
\begin{aligned}
d_{it} =& \begin{cases} 1\left(X_{it}'\delta_0 + \alpha_{v0}v_i + \alpha_{\varepsilon 0}\varepsilon_{it} + \eta_{it} \geq 0\right) & d_{it-1} = 0 \\ 1\left(X_{it}'\delta_1 + \alpha_{v1}v_i + \alpha_{\varepsilon 1}\varepsilon_{it} + \eta_{it} \geq 0\right) & d_{it-1} = 1 \end{cases} \\
log(w_{it}) =& X_{it}'\gamma + \alpha_3 v_i + \varepsilon_{it} \\
\eta_{it} \sim& \text{ Logistically Distributed} \\
v_i \sim& N(0,1) \\
\varepsilon_{it} \sim& N(0, \sigma_\varepsilon^2)
\end{aligned}
$$

where the idiosyncratic part of the error term ε_{it} is i.i.d.. Note that the terms α_{v0} and $\alpha_{\varepsilon 0}$ allow for sample selection bias on wages in a reduced form manner. The model starts in period 1 and we assume that $d_{i0} = 0$ (i.e. workers enter the labor market without a job). However, the data is not collected until some point τ and for each person we observe the data from τ to T.

This is a problem for which Indirect Inference is well suited. Maximum likelihood is very computationally heavy here because of the "initial conditions" problem. The data start at time τ so we may observe individuals working at time τ (i.e. $d_{i\tau} = 1$), but we don't know how they got there. In pure maximum likelihood one would have to integrate through all of the paths that could lead one from $d_{i0} = 0$ because we don't observe the sequence $d_{i1}, ..., d_{i\tau-1}$. However, this is not a problem when we simulate the model. Since we simulate those objects as well we can incorporate them into Υ_{hs} which simplifies the problem substantially.

An interesting aspect of this model relative to our previous cases is that it is smooth in its determination of $\log(w_{it})$ but not in the determination of d_{it}. For this reason we only need to smooth part of the model. First we define

$$\varepsilon_{it} = \sigma_\varepsilon \epsilon_{it}$$

and draw ϵ_{hs} ex-ante. Since the model with be continuous in σ_ε,y and thus the objective will also be continuous in σ.

We choose

$$\Upsilon_i = (v_i, d_{i1}, ..., d_{iT}, \epsilon_{i1}, ..., \epsilon_{iT}).$$

Note that the sequence of d begins at 1 rather than τ so to calculate the likelihood functions for the weights we avoid the initial conditions problem.

In practice there are many ways one could choose to estimate the auxiliary model. We do not take a stand on precisely what that is, but leave it in a general form. Instead, we describe how to simulate the model. We propose the following algorithm: We first draw the original data. For every h and s, Draw υ_{hs} from a standard normal distribution

1. Then for every $t = 1, ..., T$

   (a) Draw ϵ_{hst} from a standard normal distribution

      i. Simulate d_{hst} from the model

      $$Pr(d_{hst} = 1 \mid d_{hst-1} = 0) = \Lambda\left(X'_{it}\delta^*_{d_{hst-1}} + \alpha^*_{vd_{hst-1}}\upsilon_i + \alpha^*_{\varepsilon d_{hst-1}}\sigma^*_{\varepsilon}\epsilon_{it}\right)$$

      where $(\delta^*_0, \delta^*_1, \alpha^*_{v0}, \alpha^*_{v1}, \alpha^*_{\varepsilon 0}, \alpha^*_{\varepsilon 1}, \sigma^*_{\varepsilon})$ are chosen from somewhere

   (b) Calculate the likelihood function which is

   $$\begin{aligned}\ell_0(\Upsilon_{hs}; X_{hs}) =& \phi(\upsilon_{hs})\prod_{t=1}^{T}\left[\phi(\epsilon_{hst})\Lambda\left(X'_{it}\delta^*_{d_{hst-1}} + \alpha^*_{vd_{hst-1}}\upsilon_i + \alpha^*_{\varepsilon d_{hst-1}}\sigma^*_{\varepsilon}\epsilon_{hst}\right)^{d_{st}}\right.\\ &\left.\times\left(1 - \Lambda\left(X'_{it}\delta^*_{d_{hst-1}} + \alpha^*_{vd_{hst-1}}\upsilon_i + \alpha^*_{\varepsilon d_{hst-1}}\sigma^*_{\varepsilon}\epsilon_{hst}\right)\right)^{1-d_{st}}\right].\end{aligned}$$

2. Calculate objective function for a given $\theta = (\delta_0, \delta_1, \alpha_{v0}, \alpha_{v1}, \alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}, \sigma_{\varepsilon})$, for each hs

   (a) Calculate the likelihood which is very similar to above

   $$\begin{aligned}\ell(\Upsilon_{hs}; X_{hs}, \theta) =& \phi(\upsilon_{hs})\prod_{t=1}^{T}\left[\phi(\epsilon_{hst})\Lambda\left(X'_{it}\delta_{d_{hst-1}} + \alpha_{vd_{hst-1}}\upsilon_i + \alpha_{\varepsilon d_{hst-1}}\sigma_{\varepsilon}\epsilon_{hst}\right)^{d_{st}}\right.\\ &\left.\times\left(1 - \Lambda\left(X'_{it}\delta_{d_{hst-1}} + \alpha_{vd_{hst-1}}\upsilon_i + \alpha_{\varepsilon d_{hst-1}}\sigma_{\varepsilon}\epsilon_{hst}\right)\right)^{1-d_{st}}\right]\end{aligned}$$

   (b) Calculate the weight as

   $$W_{hs}(\theta) = \frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})}$$

   (c) Generate log wages for each t as

   $$log(w_{hst}) = X'_{hst}\gamma + \alpha_3\upsilon_{hs} + \sigma_{\varepsilon}\epsilon_{hst}$$

(d) We then define:

$$Y_{hs} \equiv (d_{hs\tau}, ..., d_{hsT}, d_{hs\tau} log(w_{i\tau}), ..., d_{hsT} log(w_{hsT})).$$

3. Given knowledge of g and F, solve for the auxiliary parameter

$$\widehat{B}(\theta) \equiv \frac{1}{H} \sum_{h=1}^{H} argmin_{\beta} F \left(\frac{1}{S} \sum_{s=1}^{S} \frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})} g(X_{hs}, Y_{hs}; \beta), \beta \right).$$

Notice that in this case the ratio of the likelihoods will simplify. We needed to smooth part of the model but not all. The aspects of the likelihood related to the parts we didn't need to smooth (ν_i, ϵ_{it}) will drop out. That is, the term $\phi(v_{hs}) \prod_{t=1}^{T} \phi(\epsilon_{hst})$ will cancel.[9]

5.4 Example 4: Continuous Time Transition Model with Wages

Now we consider a framework very similar to the previous one, except that time is continuous rather than discrete. We assume that the data we observe is analogous to most sample designs - at certain points of time we observe current employment status and the current wage but do not know what happened in between. This is a reasonable way to think about most data sets as one can sometimes try to extract more information on employment histories, but this data relies on respondents memory which may not be accurate. This methodology can be extended to other types of data gathering as well, but we stick with this one as it is most closely related to the above model. A major difference between this model and the previous one is that we can not figure out how to use the Bruins et al. (2017) approach for this model.

We take the continuous time analogy to the model above with a constant (over time) but heterogeneous (over workers) hazard rate of finding a job

$$e^{X_i' \beta_0 + \alpha_0 v_i}$$

and an analogous rate for losing one's job

$$e^{X_i' \beta_1 + \alpha_1 v_i}$$

[9] Monte Carlo experiments on a set of similar but slightly simpler dynamic choice models show that our smoothing technique has excellent small sample properties as well. These results are available upon request.

with

$$v_i \sim N(0,1).$$

For simplicity we abstract from time varying aspects of the wage affecting employment, but this should be straightforward to do after modeling the stochastic process. Wages are the same as the previous case

$$\begin{aligned} log(w_{it}) =& X_{it}'\gamma + \alpha_3 v_i + \varepsilon_{it} \\ \varepsilon_{it} \sim& \ N(0, \sigma_\varepsilon^2) \end{aligned}$$

where we will assume that we observe wages and employment at the interval values $\tau, \tau+1, ...,$ and T.

Consider how one would simulate this model. First, we draw v_i from a standard normal distribution. Now for a given guess of $(\beta_0 \ ,\alpha_0, \beta_1, \alpha_1)$ we know the two hazard rates. Starting at zero we draw the length of the first unemployment spell. Letting U_{hs1}^u be a uniform (0,1) random variable this can be written as

$$\ell_{hs1}^u = - \ log(1 - U_{hs1}^u) e^{-X_i'\beta_0 - \alpha_0 v_i}.$$

If this unemployment spell is less than T, we then draw the length of the next employment spell

$$\ell_{hs1}^e = - \ log(1 - U_{hs1}^e) e^{-X_i'\beta_1 - \alpha_1 v_i}$$

where U_{hs1}^e is uniform. We then check whether the cumulated time $(\ell_{hs1}^u + \ell_{hs1}^u) > T$. If so we stop, if not we draw another unemployment spell. We keep iterating on this procedure until we draw a spell that extends beyond T. For notational purposes let J_{hs} be the number of unemployment spells we observe and let T_{hs}^u be a dummy variable indicating that the last unemployment spell was right truncated at T (with $T_{hs}^u = 0$ implying that the last employment spell was right truncated at T). Also, we take the length of the last employment or unemployment spell to be its right truncated value (as oppose to its final realized value). We will then define

$$\Upsilon_{hs} = (v_{hs}, \epsilon_{hs1}, ..., \epsilon_{hsT}, J_{hs}, T_{hs}^u, \ell_{hs1}^u, ..., \ell_{hsJ_{hs}}^u, \ell_{hs1}^e, ..., \ell_{hsJ_{hs}-T_{hs}^u}^e).$$

Given this information for each $t = \tau, ..., T$ we can check whether the individual is employed at t. We don't explicitly write down the function $y_\Upsilon(\Upsilon_i, X_i; \theta)$ as it is relatively simple to understand yet relatively cumbersome to represent as a mathematical expression.

Wages can be simulated in exactly the same way as in the previous expression.

The only additional thing one would need to apply the methodology is the likelihood function which needs to be evaluated at the initial draw and on any subsequent draw. To simplify the expression let

$$\begin{aligned}
\lambda^u_{hs}\left(\theta\right) \equiv & e^{X'_{hs}\beta_0+\alpha_0 v_{hs}} \\
\lambda^e_{hs}\left(\theta\right) \equiv & e^{X'_{hs}\beta_1+\alpha_1 v_{hs}}.
\end{aligned}$$

Then one can write the likelihood function as

$$\begin{aligned}
\ell(\Upsilon_{hs}; X_{hs}, \theta) =& \phi(v_{hs}) \prod_{t=1}^{T} \phi(\epsilon_{hs}) \\
& \times \left[T^u_{hs} \prod_{j=1}^{J_{hs}-1} \left[e^{-\lambda^u_{hs}(\theta) t \ell^u_{hsj}} \lambda^u_{hs}\left(\theta\right) e^{-\lambda^e_{hs}(\theta)\ell^e_{hsj}} \lambda^e_{hs}\left(\theta\right) \right] e^{-\lambda^u_{hs}(\theta)\ell^u_{hsJ_{js}}} \right. \\
& \left. + \left(1 - T^u_{hs}\right) \prod_{j=1}^{J_{hs}-1} \left[e^{-\lambda^u_{hs}(\theta) t \ell^u_{hsj}} \lambda^u_{hs}\left(\theta\right) e^{-\lambda^e_{hs}(\theta)\ell^e_{hsj}} \lambda^e_{hs}\left(\theta\right) \right] e^{-\lambda^u_{hs}(\theta)\ell^u_{hsJ_{js}}} \lambda^u_{hs}\left(\theta\right) e^{-\lambda^e_{hs}(\theta)\ell^e_{hsJ_{js}}} \right].
\end{aligned}$$

Notice one important feature of this approach. When we simulate this model, we have data from all periods, not just after τ. This means we don't have to worry about the initial conditions problem when calculating this likelihood function.

6 Empirical Application: Fertility and Female Human Capital Accumulation

We next develop and estimate an empirical model using the importance sampling methodology. The main goal of this empirical exercise is to understand the relationship between human capital accumulation and fertility for women. 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 to have and care for children. Our main goal is to quantify the importance of this effect on human capital accumulation.

A second closely related goal is to understand the source of the curvature of the lifecycle wage profile. Wages increase rapidly at the beginning of the lifecycle and then flatten out in the middle. In a learning-by-doing model it could be due either to declining returns from learning or from the aging process. The idea of the first is that once I have mastered my job, my wages stop rising. This would correspond to curvature in actual experience while the other would be curvature in potential experience. These two types of curvature can also captured by the Ben-Porath (1967) model. As workers get closer to retirement, the incentive to invest declines which leads to an age affect but there will also be a direct human capital effect as conditional on age, the incentive to invest is lower when human capital is lower. We do not explicitly allow for endogenous Ben-Porath investment, but we do allow the curvature to result from both previous human capital and from age. For women who take time out of the labor market to have children, the distinction is very important. If it is purely previous human capital, when mothers re-enter the labor market they can expect rapid wage growth because they have relatively low levels of human capital. By contrast, if the flattening is due to age, and they re-enter mid-life, they would expect slower wage growth.

To address these questions we use data from the Survey of Income and Program Participation (SIPP). Alternative data sets that researchers have used to study female wage growth is the National Longitudinal Survey of Youth 1979 (NSLY79) as well as the older National Longitudinal Surveys of Young Women and Mature Women (NLSW). We don't want to argue that SIPP is clearly better than the NLSY79, but rather that there are tradeoffs between the two and most previous work (see discussion in next section) 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 is much smaller number of individuals (at most around 6000 women which gets smaller over time due to attrition from the survey). The SIPP is a very large data set with short panels - we will use observations from almost 100,000 different women. The challenge with the SIPP is that since we do not observe the full lifecycle profile for any women, we must piece the panel data of people at different ages. This requires an econometric model and we propose a Markov model of work, fertility and marriage. Estimating such a model by maximum likelihood is extremely difficult given the severe initial conditions problem with this data. For this reason indirect inference is a more feasible way to address the problem. However, given that the main state variables of interest (work, marriage, number of children) are discrete, smoothness in the objective function will be an issue. This makes indirect inference with importance weight sampling ideal for this problem.

The basic empirical motivation can be seen in Figure 1. We run a regression of log wages on dummy variables for potential experience and individual fixed effects for white men and for white women. We plot the predicted profile normalizing log wages at entry to zero. Two things can be seen from the figure. First, as has been established,[10] wages increase more quickly for men than for women during the beginning of the lifecycle.[11] Secondly, while wages diverge in the middle, they eventually converge towards the end of the lifecycle. One possible explanation for this pattern is labor supply and fertility - when women have

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

[11] 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.

children they tend to leave the labor market and then re-enter as their children age. This could cause wage growth to slow during this time, but then pick up after re-entry. This raises the fundamental question in labor economics we mentioned before: what leads to the curvature in wage growth? That is, wage growth slows more quickly for men than for women. If the curvature is driven by "actual experience" then one would expect this. When women re-enter the labor force they have less actual experience than men and thus their wages will grow faster. This could explain why wage growth of women with potential experience over 20 experience faster wage growth then men. By contrast if it is potential experience or age that is driving the curvature, then women who re-enter will not see faster wage growth. Our empirical specification below allows for both possibilities to measure their quantitative importance.

6.1 Literature

There are a large number of papers looking at male-female wage differentials. We differ from the vast majority of papers as we focus on female wage growth rather than levels. We mention some relevant work.

Hill (1979) was one of the first to examine the effect of motherhood on wages. She uses one wave of the PSID and finds a 7 percent motherhood wage penalty for white women when productivity characteristics are excluded. After adding productivity characteristics, the motherhood wage penalty nearly disappears. The driving factor in the wage differences is intermittent work amongst mothers. 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). Married men do not typically make such trade-offs. Korenman and Neumark (1992) use NLSY data and 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). Using panel data methods this effect disappears. However, cross-sectional IV estimates imply that working continuously following childbirth will not eliminate the motherhood wage gap.

Using the NLSY and estimating wage equations with fixed-effects, 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) use the NLSY and 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 programming 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 the German Socio-Economic Panel (GSOEP) and other sources, 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. Since selection into occupations is based partially on expected fertility outcomes, a portion of the career costs of children are incurred prior to children being born. They also show that fertility leads to changes in the ability composition of working women over the life-cycle and find that fertility explains a substantial part of the gender wage-gap, especially for women in their mid-thirties.

Loughran and Zissimopoulos (2007) concentrate on the effect of marriage and fertility on the wage growth of men and women. Fixed-effects regressions using NLSY data show that not only does marriage reduce female wage levels, but it also reduces female wage growth by four percentage points. The wage growth of men is reduced by two 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. The findings are consistent with male careers being accommodated more than female careers within the couple. This can lead to lower wage growth for married women, even before children are born. The arrival of children further reduces female wage levels as they reduce labor supply or drop out of the labour force. Since marriage and childbearing at young ages can cause substantial decreases in lifetime earnings, this may be an important factor in explaining why marriage and fertility has been delayed since the mid-1960s.

Daniel, Lacuesta, and Rodríguez-Planas (2013) estimate fixed-effects regressions on Spanish data, also controlling for firm-level heterogeneity, 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.

Using NLSY data and a variety of regression techniques, Braga (2013) finds that more educated workers benefit from faster wage growth due to accumulation of work experience but suffer greater wage losses from spells of unemployment. He uncovers this by estimating regression models where earnings depend on work experience, past unemployment and non-participation periods, interacted with schooling. The data is restricted to non-black males but qualitatively similar results are obtained for blacks and women.

Other important papers include Weiss and Gronau (1981), which 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 and Miller (2012), and Blundell, Costa Dias, Meghir, and Shaw (2015). 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).

6.2 The Markov Model

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 generally increases while individuals work and falls when they don't. The state variables are

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

where t is time since labor market entry (e.g. potential experience), L_{it} is a dummy variable for having a job, M_{it} is a dummy variable for marriage, H_{it} is human capital, K_{it} is the number of children the woman has given birth to, and the A_{jit} are the ages of each of the children. The last two variables 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 assumed to have a standard normal distribution.

The transitions are governed by five different hazard rates; the hazard rate for job arrival for the non-employed, $\lambda^J\left(\mathcal{S}_{it}\right)$, the hazard rate for job destruction (leading to non-employment), $\lambda^N\left(\mathcal{S}_{it}\right)$, the rate of marriage formation, $\lambda^M\left(\mathcal{S}_{it}\right)$, for divorce. $\lambda_1^D\left(\mathcal{S}_{it}\right)$, and finally for births of children, $\lambda^K\left(\mathcal{S}_{it}\right)$. The ages of both the woman and her children increase with time and human capital evolves deterministically as a function of the state variables. Wages are a function of the state variables and an i.i.d. error term.

All five hazard rates take the basic form.

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

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). We discretize all of the continuous state variables in this expression. In particular, we allow the child age range to differ for children less than seven and from eight to seventeen. We use three age groups, potential experience is ten or less, between 11 and 20, and older than 20. The specific variables are listed in the tables below. Of course, the fertility variables are key and we will estimate how they influence labor supply and thus human capital accumulation.

We choose a human capital accumulation function that allows for curvature of the profile either through age t, or human capital H_{it}. Specifically, for workers we allow human capital to accumulate according to

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

where $\bar{H}$ is the maximum level of human capital (and $\dot{H} = \partial H/\partial t$). One can see that as H_{it} approaches $\bar{H}$, human capital accumulation slows down. The other force that may allow for human capital to accumulation to slow down is the potential experience term $e^{-\mu t}$. 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, they will have relatively high t but relatively low H_{it}. So if the first effect is important, they should see large wage growth upon re-entering, but with higher μ they will not. The key moment for identifying this parameter is the wage growth for women with children over the age of 18. We put high weight on this parameter to make sure the model fits it very well. We parameterize $\log\left(a\left(\mathcal{S}_{it}\right)\right)$ to be linear in state variables. Note that our specification also does not allow human capital to fall for older women (when they work). This is consistent with our data-see Figure 1.

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

$$\dot{H} = -\,\delta H$$

where δ is a parameter.[12]

Finally we allow wages to depend on human capital as well as the other state variables

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

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

6.3 Data and Auxiliary model

We estimate the model using the last four panels of the Survey of Income and Program Participation 1996, 2001, 2004, and 2008.[13] This survey interviews individuals every four months and we only use data from the survey month. This data has the advantage of very large samples and a panel structure. The large sample size is important as identification for many of our parameters is quite subtle. Panel data is essential as well. We use white women who are 18 years or older and have at most 35 years of potential experience. Table 1 presents summary statistics of the main variables we use in our analysis. Details of the data are discussed in Appendix B.

We construct our auxiliary model using the following auxiliary parameters. The full list can be seen in the tables with more detail contained in Appendix C.Regression of log wages on potential experience dummies and state variables with individual fixed effects

- Within and between variance of the error term from the regression
- Regression of estimated fixed effect on education
- Linear probability regression of whether a woman is married in the initial period we see her, on potential experience dummies and state variables

[12] In a previous version, we allowed it to be a log linear function of state variables, but we did not find sufficiently strong predictors of this in the data.

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

- Linear probability regression of whether an unmarried woman gets married between waves, on potential experience dummies and state variables
- Linear probability regression of whether a married woman stays married between waves, on potential experience dummies and state variable
- Fraction of mothers who are married at childbirth
- Regression of having a child on 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 whether a woman works in the initial period we see her, 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 and state variables
- Linear probability regression of working in one wave conditional on not working in the previous wave, on potential experience dummies and state variables
- Fraction of mothers who work in interview before giving birth
- Per person regression of mean number of observations individual works on wage fixed effect
- 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.

These auxiliary parameters can be seen in Tabes 3a-3f under the column labeled data. The key parameters are the effects of the number of children on various outcomes. In the fixed effects wage regression, we see little evidence of a children penalty relative to many of the papers mentioned above. This is in large part because this is a very short panel. Another key parameter is the children over 18 in the wage growth regression. 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 children directly into the human capital production function. We see large effects of fertility on labor supply. We discuss many of the other auxiliary parameters when we examine the results of the model.

6.4 Estimation in Practice

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 the law of large numbers still works. However, as a practical matter, one is using essentially 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 evaluate at this parameter value, the weights are all equal to one. We used the following iterative approach to deal with this problem. At iteration j we have estimated $\widehat{\theta}_j$ At iteration $j+1$ re-simulate the model using θ_j so that $\ell_0(\Upsilon_{hs}; X_{hs}) = \ell(\Upsilon_{hs}; X_{hs}, \theta_j)$

- Use a Newton method to minimize the distance between the auxiliary and simulated parameters with at most 100 steps
- Let the parameter that minimizes this be $\widehat{\theta}_{j+1}$. The value of the standard simulated objective function may not continue improving as j increases. In practice we find that it does improve for a while, and then stops. At that point we use a simplex method starting with the value of $\widehat{\theta}_j$ that minimizes the unweighted objective function. We then iterate between the simplex and Newton methods until convergence. In practice we find this works well. To estimate the standard errors we use the Importance Sampling approach.

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 the default approach to doing this 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 only a few variables picking up fertility (the wage regression has two). 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 two. 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 design is presented in Appendix C.

In the model we simulate, Movement in and out of work

- Movement in and out of marriage
- Birth of children
- Human capital
- Wages. We start the model when individuals enter the market after school and we assume they are unmarried, without a job, and without children. Note that the discreteness in the movement in and out of work, in and out of marriage, and having children all lead to discrete jumps in the state space when they occur. By contrast, human capital and the subsequent wage are smooth functions of the parameters conditional on the state variables. To formally define Υ_i we define some new notation. Let N_i^w be the number of work transitions and define the date (in terms of actual experience) of these transitions to be $\tau_{i1}^w, ..., \tau_{iN_i^w}^w$. Note that since individuals start nonemployed we can keep track of the state so we know the direction of the transition. Similar for marriage we define N_i^m to be the number of marriage transitions and $\tau_{i1}^m, ..., \tau_{iN_i^m}^m$ their dates. Similarly let N_i^k be the number of children and $\tau_{i1}^k, ..., \tau_{iN_i^k}^k$ the dates when they were born. Finally, as in the examples above, we take $\varepsilon_{it} = \sigma_\varepsilon \epsilon_{it}$ where ϵ_{it} is standard normal. Let ϵ_i be the vector of these objects for the periods in which the wage is observed by the econometrician. Then we take our Υ_i to be

$$\Upsilon_i = \left\{ \tau_{i1}^w, ..., \tau_{iN_i^w}^w, \tau_{i1}^m, ..., \tau_{iN_i^m}^m, \tau_{i1}^k, ..., \tau_{iN_i^k}^k, \epsilon_i \right\}.$$

What is crucial for our approach is that the likelihood function $\ell(\Upsilon_i \mid X_i; \theta)$ is smooth as a function of θ and the rest of the variables used to produce the auxiliary model are smooth in θ once we condition on Υ_i.

6.5 Empirical Results

In Tables 2a-2c we present the estimated parameters of the model. The parameters themselves are difficult to interpret on their own but can be more easily understood through the simulations that follow. Most of the parameters have the signs one would expect.

We present the fit of the model in Tables 3a-3d, Figures 2a-2d, and in Appendix Figures A1-A8. One can see in Tables 3a-3d that with only a few exceptions, the fit of the model for these parameters is excellent. We also try to fit the profile of wages, marriage, children, and working across the lifecycle. Figures 2a-2d show that the fit for wages is excellent and the fit for the other three is good. We put less weight on fitting the profile in the log wage regression than the profile result. This fit is shown in Appendix Figure A1. The main goal is to fit the levels though we also fit the transitions which are shown in appendix Figures A2-A8. Given the coarseness of our model, the relationship between hazard rates and potential experience we can not fit perfectly and in some cases one can see that the model is too course to fit some of the details in the data. Since our main goal is the overall wage profile we do not think this is problematic.

Our first issue of interest is the curvature in human capital which is important for understanding the shape of women's wage growth for the reasons discussed above. Recall that our baseline model is

$$\dot{H} = a\left(\mathcal{S}_{it}\right)\left(\bar{H} - H_{it}\right)e^{-\mu t}.$$

In this case curvature can come from two different sources. The first is the term $\left(\bar{H} - H_{it}\right)$ that leads to human capital slowing down as it approaches $\bar{H}$. The second is from the μ term in which human capital will slow 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 think this difference is identified by the coefficient on kids greater than 18 in the wage growth regression and we put a lot of weight on this particular auxiliary covariate. One can see from Table 3e that it is matched quite well.[14] The reason for using this moment is analogous to why this distinction is important. If women take a lot of time out of the labor market when having children and then re-enter they will have relatively low H_{it} and relatively high t. If the curvature primarily comes from μ they will not accumulate a lot of human capital when they re-enter the labor market. However, if the curvature comes from $\left(\bar{H} - H_{it}\right)$ they will see their wages rise relatively rapidly because $\left(\bar{H} - H_{it}\right)$ will be relatively large. To understand this, we graph three alternative versions

[14] Our estimate is lower than the estimate in the data, though well within the 95% confidence region. A larger number would mean even higher growth for women re-entering the market which would suggest an even smaller roll for μ, so this would just reinforce our main result.

of the human capital production function:

$$\begin{aligned}
\text{Model A}: \ \dot{H} &= a^{A}\left(\mathcal{S}_{it}\right)\left(\bar{H} - H_{it}\right) \\
\text{Model B}: \dot{H} &= a^{B}\left(\mathcal{S}_{it}\right)e^{-\mu t} \\
\text{Model C}: \dot{H} &= a^{C}\left(\mathcal{S}_{it}\right).
\end{aligned}$$

In all cases we adjust the value of $a\left(\mathcal{S}_{it}\right)$ to keep human capital growth in the first ten years the same as in the base case. Figure 3 presents the results in which Model A is labeled "No Age Effect," Model B is labeled "No Direct Human Capital," and Model C is labeled "Neither Age nor Hum. Cap." We show the progression of the function for married women with 4 years of college. It is clear here that our estimated value of μ of 0.002 is sufficiently small that the direct human capital effect is much more important.[15]

Next we simulate a model in which we relax the relationship between fertility and work and see how that affects human capital accumulation. That is, we compare our base model to a counterfactual in which children at home have no effect on working. Specifically, the effect of "Number of Kids $<$ 18", "Number of Kids $<$ 7" and "Any Kids $<$ 7" on finding a job and leaving a job are set to zero (the five numbers are shown in Table 2a). The direct effect can be seen in Figure 4a in which there is a considerable increase in labor force participation during the prime child bearing years. The difference peaks at around 10 years of potential experience at a level of roughly 10% (85% compared to 75%). It is worth pointing out that this is a substantial effect, but it is not enormous. This is not that surprising. From the raw data, one can see in the initial work regression in Table 3d, that the coefficient on "Number of Kids $<$ 7" is of a similar magnitude. Many women stop working while they have young children, but most do not.

We next examine the effect of this labor supply decline on human capital accumulation and wages. These simulations will not be completely 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 affect the profile. We 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. Figure 4b presents a simulation of the level of log wages (H_{it} in our model) at different ages. The line labeled "Base Model" is a simulation using the estimated parameters while the line labeled "Fertility Doesn't Affect Work" presents a case in which human capital acquisition speeds up because women's labor

[15] One can also see that the standard error on this parameter is quite large. This is really due to the fact that the value of the parameter is essentially zero so the asymptotic approximation is poor.

supply is no longer affected by working (as in Figure 4a). One can see that the loss in labor supply does suppress human capital. The difference peaks at experience level 18 where it is 0.79 in the baseline and 0.82 in the counterfactual. While the sign is as expected, the magnitude of the difference is modest in comparison to the difference in wage growth between men and women. We next exclude the motherhood penalty by disallowing the direct effect of children on wages (coefficients on wages in Table 2b). The effect here is quite modest.

To put this simulation in a more familiar context we calculate the difference in log wages for both the counterfactual and the baseline and plot it in Figure 3c. The difference in wages peaks around experience levels 15-20 at a difference of somewhat over 0.032. This suggests that on average, wages of women at these ages would be about three 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.[16]

7 Conclusion

In this paper, we show how to use importance sampling weights for indirect inference that maintains the discreteness of endogenous variables in the model. Thus, we eliminate the approximation error and potential small sample bias that may arise with the use of GII from Bruins et al. (2017). 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.

We apply our new approach to estimating a continuous time Markov model of female work, marriage, and fertility using data from the Survey of Income and Program Participation. The model provides a reasonably good fit of the data. We then simulate two different types of counterfactuals. The first attempts to see whether the curvature in the female wage profile 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. Our results strongly

[16] The three 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).

suggest that curvature in the human capital production function is the driving force. Our second counterfactual attempts to uncover the extent to which women's dropping out of the labor force for fertility related reasons suppresses human capital accumulation. Our finding is that it does to a modest extent. Wages among prime age women would be approximately 3% higher if the relationship between fertility and working were eliminated.

References

Ackerberg, D. A. (2009). A new use of importance sampling to reduce computational burden in simulation estimation. *QME 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.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *The Journal of Political Economy*, 352–365.

Blundell, R., M. Costa Dias, C. Meghir, and J. Shaw (2015, March). Female labour supply, human capital and welfare reform.

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

Bruins, M., J. Duffy, M. Keane, and A. Smith (2017, March). Generalized indirect inference for discrete choice models.

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.

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 (2016, March). Estimation of an equilibrium model with externalities: Combining the strengths of structural models and quasi-experiments.

Gallant, A. R. and G. Tauchen (1996). Which moments to match? *Econometric Theory 12*(4), 657–681.

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. and R. Miller (2012). Life-cycle fertility and human capital accumulation.

Genton, M. G. and E. Ronchetti (2003). Robust indirect inference. *Journal of the American Statistical Association 98*(461), 67–76.

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.

Guvenen, F. and A. A. Smith (2014). Inferring labor income risk and partial insurance from economic choices. *Econometrica 82*(6), 2085–2129.

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

Hansen, L. P., J. Heaton, and E. G. J. Luttmer (1995). Econometric evaluation of asset pricing models. *The Review of Financial Studies 8*(2), 237–274.

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.

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.

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*.

Table 1
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
Married/Spouse Present	0.591	0.492
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	

Table 2a

Model Estimates: Hazard Estimates

	Get	Get	Find	Leave	Have
Covariate	Married	Divorced	Job	Job	Kid
Education	-0.005	-0.111	0.050	-0.189	-0.162
	(0.015)	(0.048)	(0.005)	(0.013)	(0.011)
ν	0.016	-0.288	-0.395	-0.539	0.130
	(0.021)	(0.107)	(0.053)	(0.044)	(0.034)
Married			-0.245	-0.067	0.130
			(0.040)	(0.042)	(0.034)
Number of Kids < 18		0.037	0.071	0.051	
		(0.100)	(0.020)	(0.020)	
Number of Kids < 7			-0.181	0.205	
			(0.031)	(0.026)	
Any Kids < 7		-0.241			
			(0.047)		
Working					-0.826
					(0.100)
Number of Kids=1					-0.314
					(0.153)
Number of Kids=2					-1.752
					(0.150)
Number of Kids>2					-5.102
					(3.048)
Number of Kids$\times$ Education					0.307
					(0.817)
Age Youngest					-0.055
					(0.021)
Potential Experience ≤ 10	-2.250	-3.682	-0.158	-1.544	-2.271
	(0.037)	(0.312)	(0.032)	(0.087)	(0.086)
$10 \leq$ Potential Experience ≤ 20	-2.840	-3.449	-0.612	-1.971	-2.930
	(0.169)	(0.309)	(0.094)	(0.122)	(0.107)
Potential Experience > 20	-3.561	-4.218	-1.168	-2.194	-4.154
	(0.522)	(0.656)	(0.104)	(0.098)	(0.208)

Table 2b

Model Estimates: Human Capital and Wages

Covariate	Human Capital (a)	Wages
Intercept	-3.069	
	(0.331)	
Education	0.307	0.026
	(0.111)	(0.015)
ν_i		0.446
		(0.010)
Married	-0.119	0.014
	(0.147)	(0.006)
Number of Kids < 18		-0.004
		(0.003)
Number of Kids < 7		-0.002
		(0.003)

Table 2c

Model Estimates: Additional Parameters

δ	0.055
	(0.016)
μ	0.002
	(0.013)
$\bar{H}$	0.987
	(0.061)
σ_ε	0.290
	(0.010)

Table 3a

Fit of Model: Wages

	log(wage)		Fixed Effects	
Covariate	with fixed effects		Themselves	
	Model	Data	Model	Data
Education			0.107	0.114
				(0.001)
Married	0.020	0.009		
		(0.006)		
Number of Kids < 18	-0.007	-0.000		
		(0.003)		
Number of Kids < 7	-0.003	-0.002		
		(0.003)		

Table 3b

Fit of Model: Marriage

	Initial		Get		Get	
Covariate	Married		Married		Divorced	
Education	0.015	0.015	0.000	-0.000	0.001	0.001
		(0.001)		(0.000)		(0.000)
$\widehat{\nu}_i$	0.039	0.045	0.002	0.000	0.004	0.003
		(0.004)		(0.001)		(0.000)
$\widehat{\nu}_i$ Missing	0.041	0.041	-0.002	-0.002	0.003	0.003
		(0.004)		(0.001)		(0.000)
Number of Kids < 18					-0.000	-0.000
						(0.000)
Pot. Exp. Dummies	Yes	Yes	Yes	Yes	Yes	Yes

Table 3c

Fit of Model : Fertility

	Any		Two		Number	
Covariate	Kids		Kids		Kids	
	Model	Data	Model	Data	Model	Data
Education	-0.026	-0.026	0.002	0.002	-0.112	-0.110
		(0.001)		(0.001)		(0.002)
Pot. Exp. Dummies	Yes	Yes	Yes	Yes	Yes	Yes

Table 3d

Fit of Model : Work

Covariate	Initial Work		Start Work		Keep Working	
	Model	Data	Model	Data	Model	Data
Education	0.038	0.037	0.007	0.007	0.008	0.008
		(0.001)		(0.000)		(0.000)
Married	-0.001	-0.006	-0.035	-0.032	0.002	0.002
		(0.003)		(0.003)		(0.001)
Any Kids ≤ 6			-0.026	-0.022		
				(0.004)		
Number of Kids < 18	-0.022	-0.020	0.002	0.001	-0.001	-0.001
		(0.002)		(0.001)		(0.000)
Number of Kids < 7	-0.087	-0.094	-0.025	-0.021	-0.013	-0.012
		(0.003)		(0.002)		(0.001)
$\widehat{\nu}_i$					0.039	0.039
						(0.001)
Exp. Dummies	Yes	Yes	Yes	Yes	Yes	Yes

Table 3e

Fit of Model : Wage Growth

Covariate	Wage Growth×100 Continuously Employed		Wage Growth ×100 Nonemployment Spell	
	Model	Data	Model	Data
Education	0.023	0.024		
		(0.015)		
Married	-0.092	-0.173		
		(0.070)		
Total Kids	0.097	0.121		
		(0.038)		
Change in Potential Experience			-0.864	-0.921
				(0.429)
Pot. Exp. Dummies	Yes	Yes	Yes	Yes

Table 3f

Fit of Model : Additional Auxiliary Parameters

	Model	Data
Within Variance Log Wages	0.068	0.067
		(0.001)
Between Variance Log Wages	0.288	0.263
		(0.002)
Praction Married When Giving Birth	0.768	0.769
		(0.005)
Regression had Kid on Wage Residual	0.008	0.007
		(0.002)
Age Difference Youngest/ Oldest	5.771	5.770
		(0.021)
Working Before Giving Birth	0.658	0.664
		(0.007)
Working on $\widehat{\nu}_i$	0.078	0.078
		(0.002)

Figure 1: Male and Female Log Wage Profiles

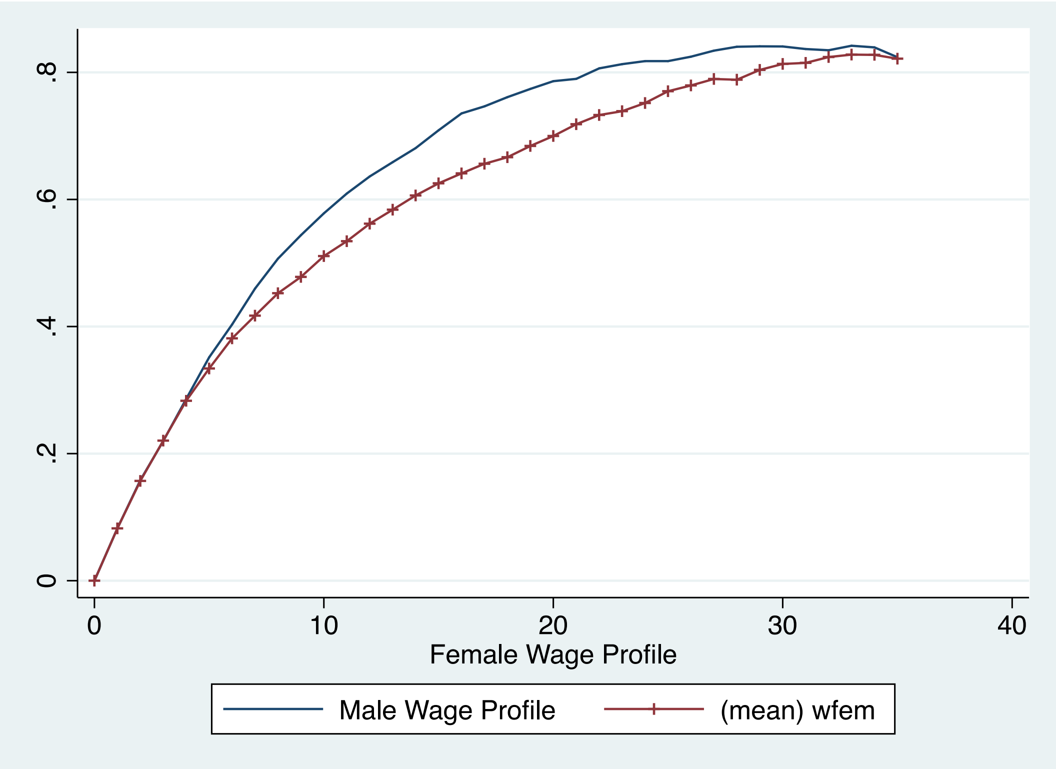


Figure 2a: Fit of model: Wage Growth Employed

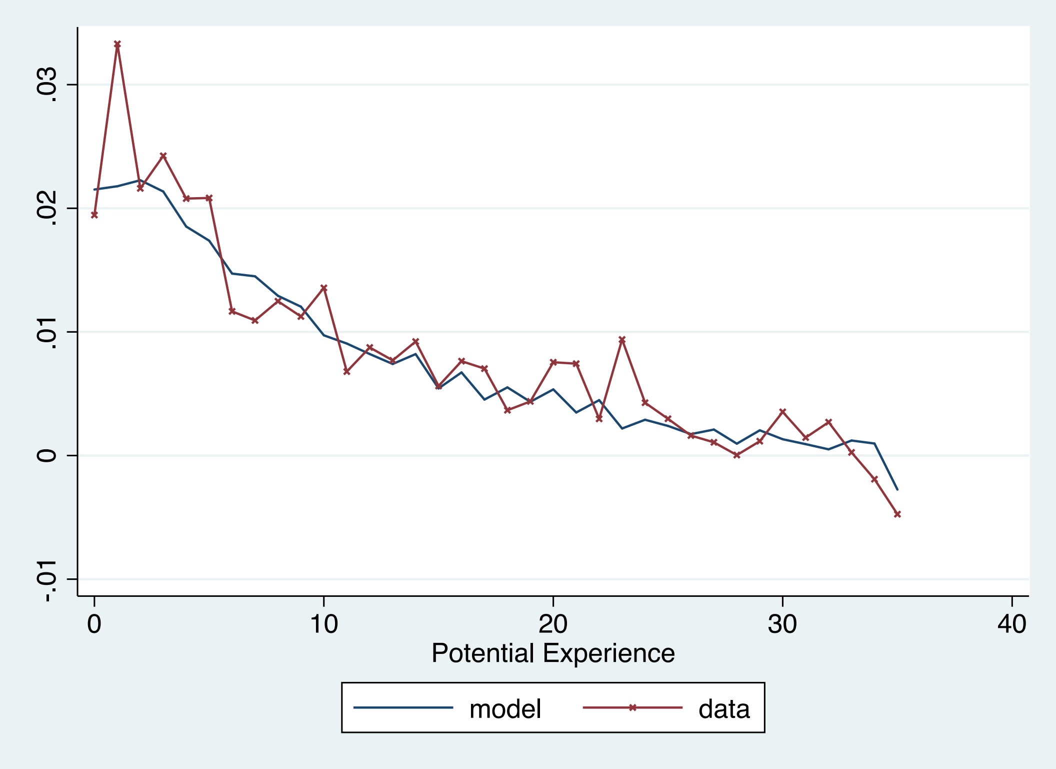


Figure 2b: Fit of model: Initial Marriage

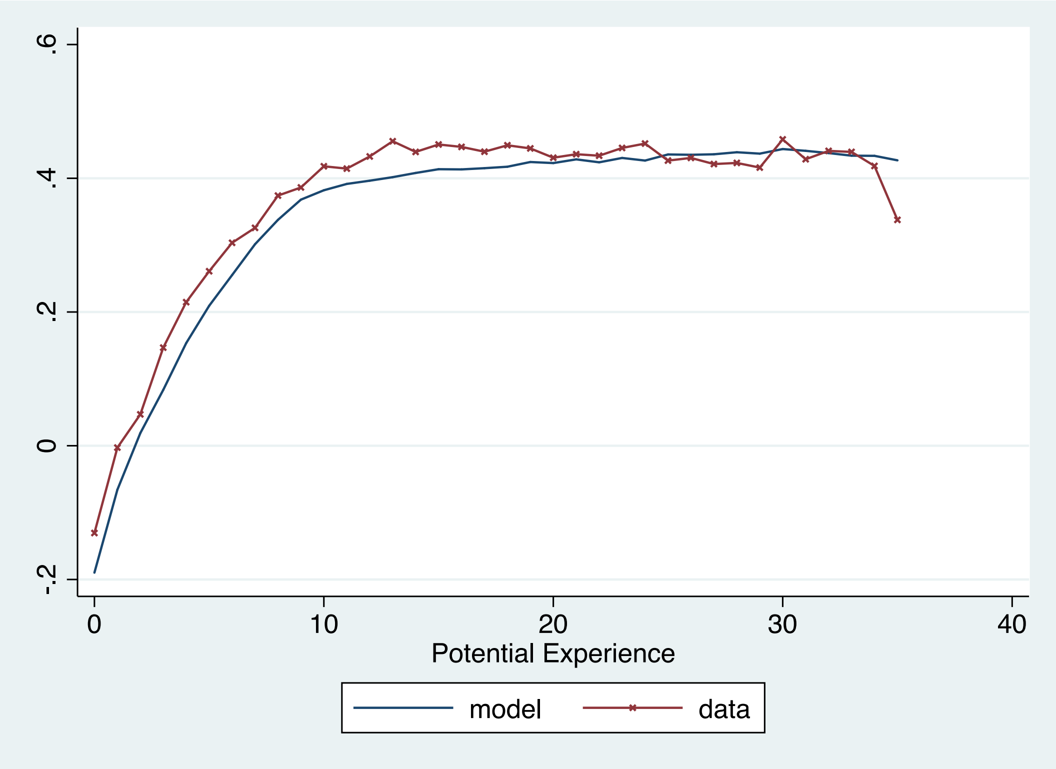


Figure 2c: Fit of model: Number Children

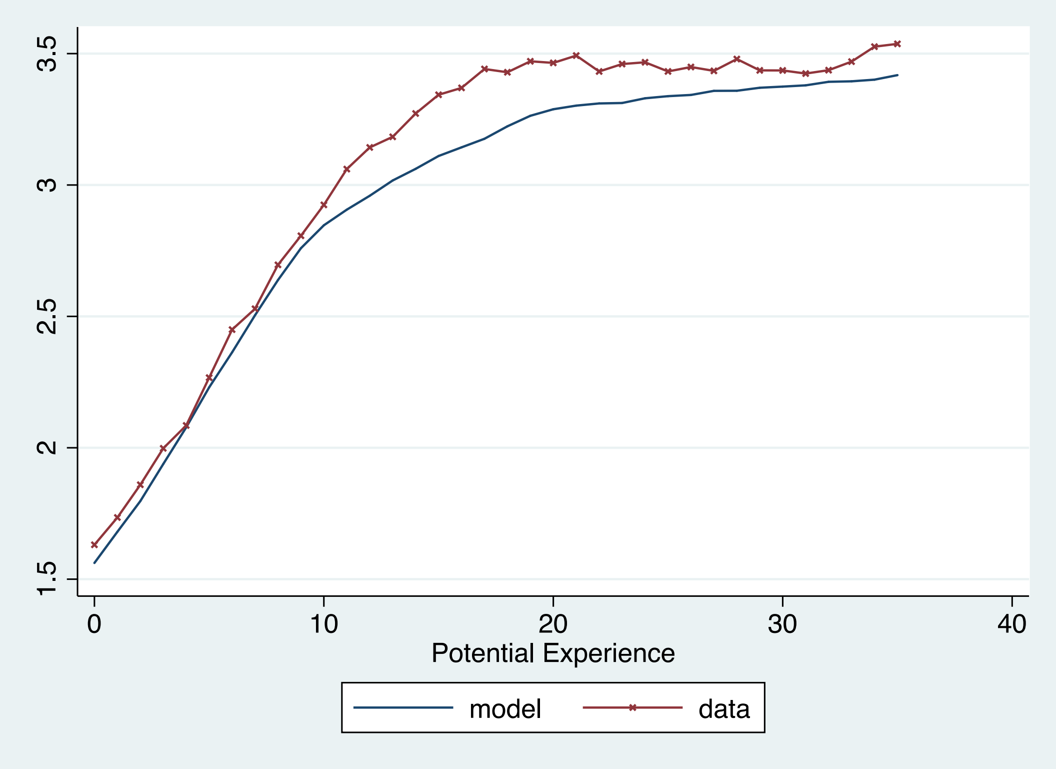


Figure 2d: Fit of model: Initial Work

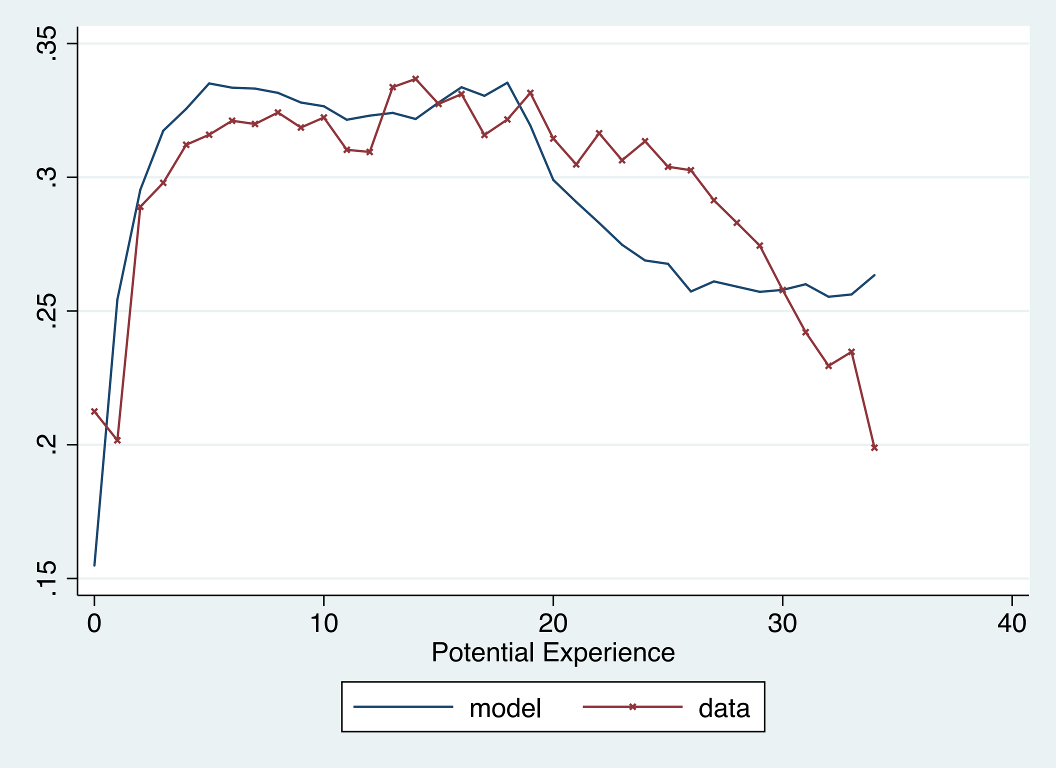


Figure 3: Human Capital Accumulation under Alternative Curvature Parameters

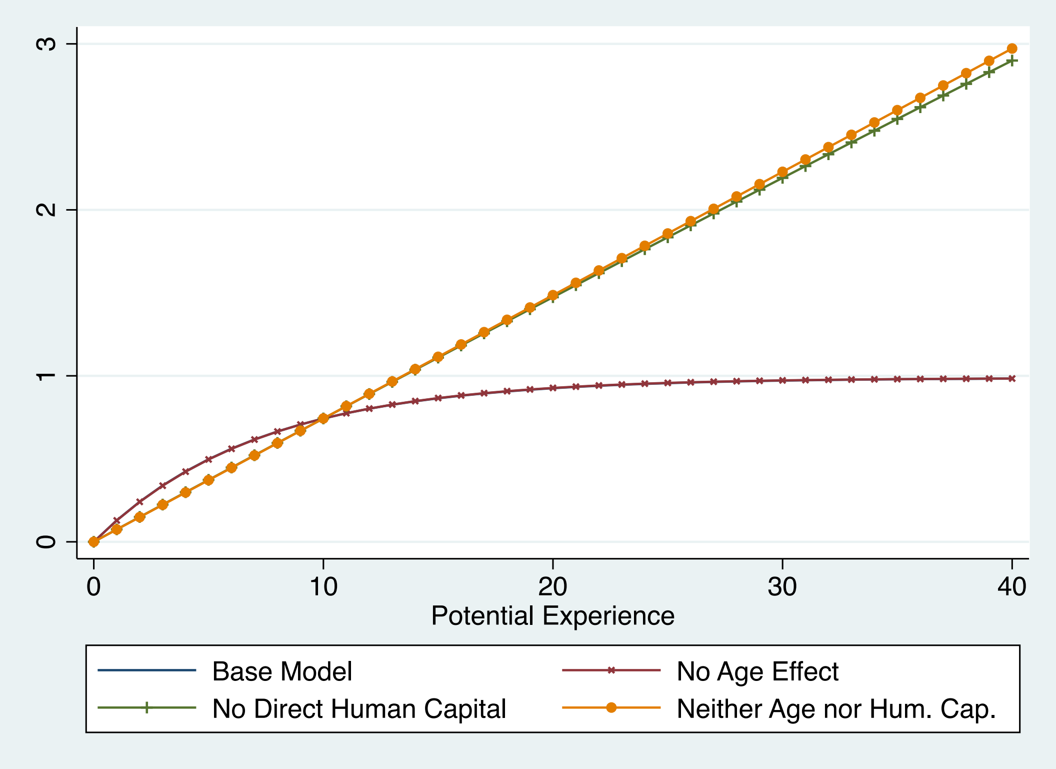


Figure 4a: Labor Supply when Fertility Doesn't Affect Labor Supply

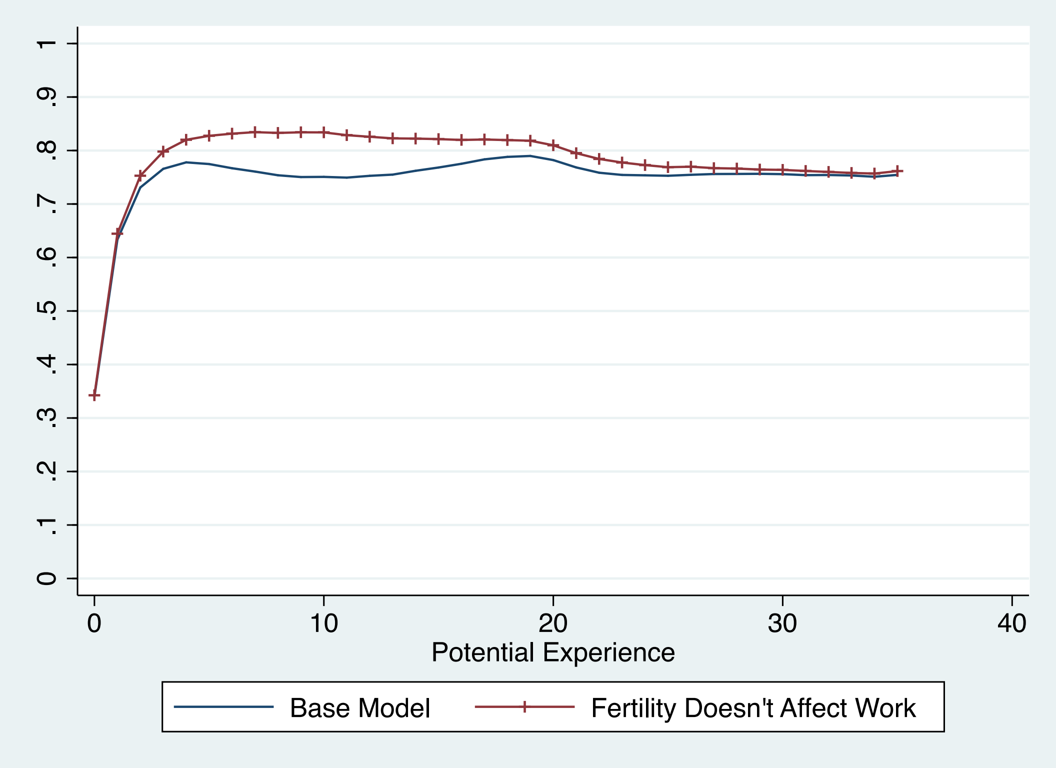


Figure 4b: Log Wages when Fertility Doesn't Affect Labor Supply

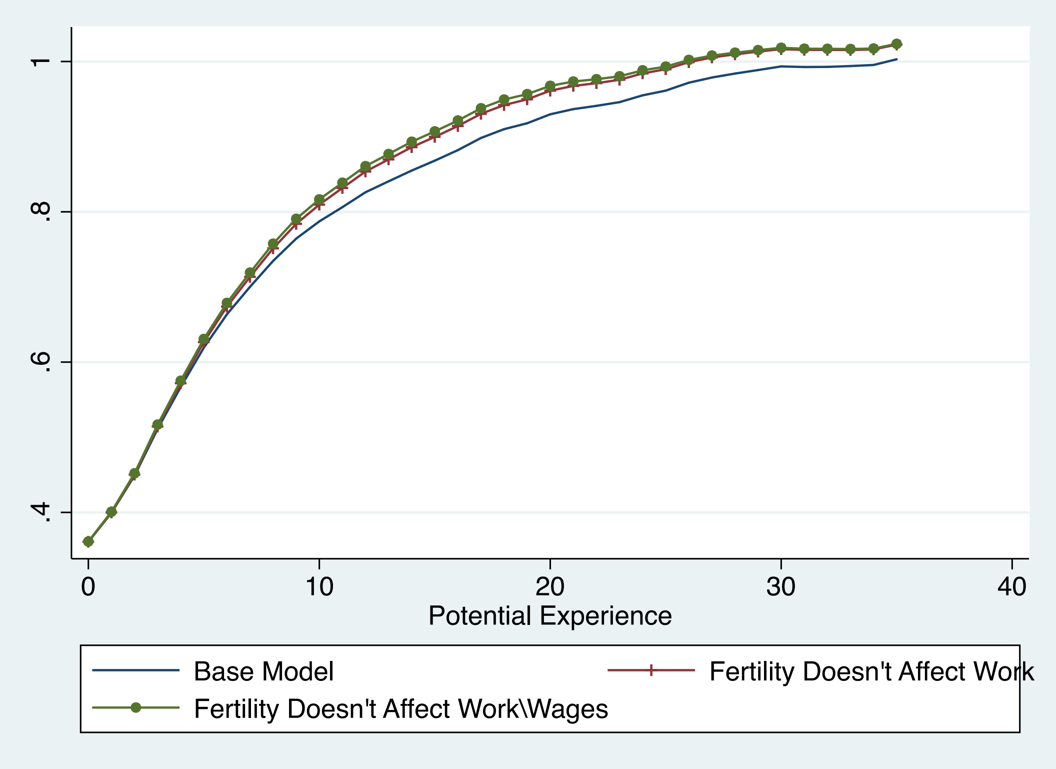
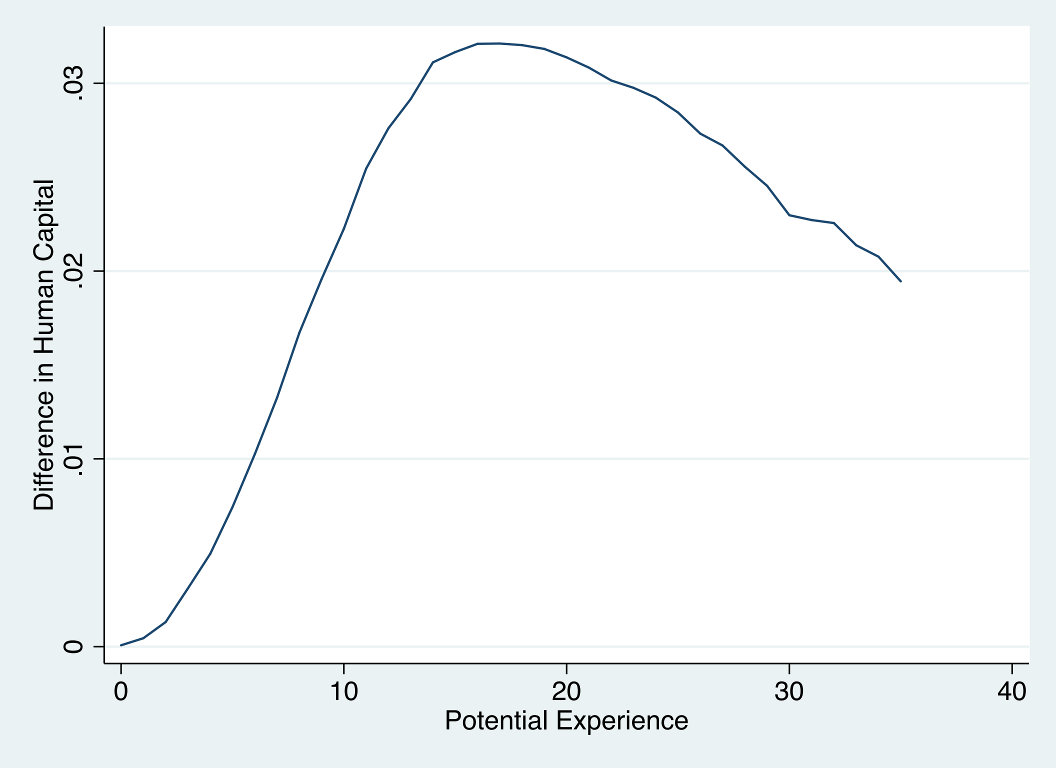


Figure 4c: Difference in Human Capital when Fertility Doesn't Affect Labor Supply



Appendix A: Proofs

A.1 Proof of Consistency Theorem

We verify the four conditions for consistency from Newey and McFadden Theorem 2.1.

Following their notation we define

$$Q_0(\theta) \equiv -\left(B(\theta) - B(\theta_0)\right)' \Omega \left(B(\theta) - B(\theta_0)\right).$$

Their first assumption is that $Q_0(\theta)$ is maximized at θ_0. This follows since it is negative at any other value and zero when evaluated at θ_0 by Assumption 4.

Their second assumption is that Θ is compact which we assume directly in Assumption 2.

Their third assumption is that Q_0 is continuous which follows directly from Assumption 3.

Finally we need that

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

converges uniformly to Q_0.

We know that given assumptions 1,2, and 6 the standard argument for consistency of M-estimators gives

$$\widehat{\beta} \xrightarrow{p} B(\theta_0).$$

Thus, what remains is that we need to show that $\widetilde{B}(\theta)$ converges uniformly to $B(\theta)$.

First note that when Υ_{hs} is simulated from $\ell_0(\Upsilon_{hs}; X_{hs})$

$$\begin{aligned}
E_s\left(\frac{\ell(\Upsilon_{hs}; X_{hs}, \theta)}{\ell_0(\Upsilon_{hs}; X_{hs})} g(X_{hs}, Y_{hs}; \beta)\right) =& \int\int \sum_{j=1}^{K_\Upsilon} \frac{\ell(\Upsilon^d_{(j)}, \Upsilon^c; X_{hs}, \theta)}{\ell_0(\Upsilon^d_{(j)}, \Upsilon^c; X_{hs})} g(X_{hs}, Y_{hs}; \beta) \ell_0(\Upsilon^d_{(j)}, \Upsilon^c; X_{hs}) d\Upsilon^c d\Xi_0(x) \\
=& \int\int g(X_{hs}, Y_{hs}; \beta) \ell(\Upsilon^d_{(j)}, \Upsilon^c; X_{hs}, \theta) d\Upsilon^c d\Xi_0(x) \\
=& G(\theta, \beta)
\end{aligned}$$

where E_s is the expected value from the simulator.

For any $\varepsilon > 0$, the following three inequalities hold with probability approaching 1.

$$\sup_{\theta\in\Theta}\left[\frac{1}{H}\sum_{h=1}^{H}F\left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}g(X_{hs},Y_{hs};\widetilde{B}(\theta)),\widetilde{B}(\theta)\right)\right.$$
$$\left.-\frac{1}{H}\sum_{h=1}^{H}F\left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}g(X_{hs},Y_{hs};B(\theta)),B\left(\theta\right)\right)\right] < \frac{\varepsilon}{3}$$

$$\sup_{\theta\in\Theta}\left[\frac{1}{H}\sum_{h=1}^{H}argmin_\beta F\left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}g(X_{hs},Y_{hs};\widetilde{B}(\theta)),\widetilde{B}(\theta)\right) - F\left(G(\Xi_0,\theta,\widetilde{B}(\theta)),\widetilde{B}(\theta)\right)\right] <\frac{\varepsilon}{3}$$

and

$$\sup_{\theta\in\Theta}\left[F\left(G(\theta,B(\theta))\right) - \frac{1}{H}\sum_{h=1}^{H}argmin_\beta F\left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}g(X_{hs},Y_{hs};B(\theta)),B(\theta)\right)\right] <\frac{\varepsilon}{3}.$$

The first one comes from the fact that $\widetilde{B}(\theta)$ maximizes the objective function and the second two come from assumption 6.

So with probability approaching one

$$\sup_{\theta\in\Theta}\left[F\left(G(\theta,B(\theta))\right),B(\theta)) - F\left(G(\theta,\widetilde{B}(\theta))),\widetilde{B}(\theta)\right)\right]$$
$$\leq \sup_{\theta\in\Theta}\left[F\left(G(\theta,B(\theta)),B(\theta)\right) - \frac{1}{H}\sum_{h=1}^{H}argmin_\beta F\left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}g(X_{hs},Y_{hs};B(\theta)),B(\theta)\right)\right]$$
$$+\sup_{\theta\in\Theta}\left[\frac{1}{H}\sum_{h=1}^{H}argmin_\beta F\left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}g(X_{hs},Y_{hs};B(\theta)),B(\theta)\right)\right.$$
$$\left.-\frac{1}{H}\sum_{h=1}^{H}argmin_\beta F\left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}g(X_{hs},Y_{hs};\widetilde{B}(\theta)),\widetilde{B}(\theta)\right)\right]$$
$$+\sup_{\theta\in\Theta}\left[\frac{1}{H}\sum_{h=1}^{H}argmin_\beta F\left(\frac{1}{S}\sum_{s=1}^{S}\frac{\ell(\Upsilon_{hs};X_{hs},\theta)}{\ell_0(\Upsilon_{hs};X_{hs})}g(X_{hs},Y_{hs};\widetilde{B}(\theta)),\widetilde{B}(\theta)\right) - F\left(G(\theta,\widetilde{B}(\theta)),\widetilde{B}(\theta)\right)\right]$$
$$<\varepsilon.$$

Since F and G are continuous and Θ and $\mathcal{B}$ are both compact, for any δ define

$$\varepsilon^*(\delta) \equiv \inf_{\theta\in\Theta,\beta\in\mathcal{B}|\|\beta-B(\theta)\|\geq\delta} F\left(G(\theta,B(\theta)),B(\theta)\right) - F\left(G(\theta,\beta),\beta\right).$$

Then choose $\varepsilon = \varepsilon^*(\delta)$. That means with probability approaching 1, $\sup_{\theta\in\Theta}\inf_{\beta\in\mathcal{B}|\|\beta-B(\theta_0)\|\geq\delta}\sup_{\theta\in\Theta}\Big[F\,(G($ $\varepsilon^*(\delta)$ so with probability approaching 1, $\sup_{\theta\in\Theta}\left\|\widetilde{B}\left(\theta\right)-B(\theta)\right\| < \delta$.

The fact that $\widetilde{B}(\theta)$ converges uniformly to $B(\theta)$ and that $\widehat{\beta} \xrightarrow{p} B(\theta_0)$ means that $\left(\widetilde{B}(\theta) - \widehat{\beta}\right)' \Omega \left(\widetilde{B}(\theta) - \widehat{\beta}\right)$ converges uniformly in probability to $-\left(B(\theta) - B(\theta_0)\right)' \Omega \left(B(\theta) - B(\theta_0)\right).$

Thus we have verified all of the conditions of Newey and McFadden Theorem 2.1.

A.2 Proof of Asymptotic Distribution

We follow Newey and McFadden Theorem 7.2.

Our estimator satisfies their basic conditions to apply the theorem. Assumption (i) holds since $B(\theta_0) - B(\theta_0) = 0$. Assumption 7 guarantees that (ii) and (iii) hold, and 8 guarantee that (v) holds.

To prove the result we need to derive the asymptotic distribution of $\sqrt{N}\left(\widetilde{B}(\theta_0) - \widehat{\beta}\right).$

Consider $\widehat{\beta}$. Then the first order condition comes from totally differentiating the objective function

$$0 = \frac{dF\left(\widehat{G}\left(\widehat{\beta}\right), \widehat{\beta}\right)}{d\beta}.$$

Let $\beta_0 = B\left(\theta_0\right).$ With the mean value theorem we get

$$0 = \left[\frac{d^2F\left(\widehat{G}\left(\overline{\beta}\right), \overline{\beta}\right)}{d\beta d\beta'}\right]\left(\widehat{\beta} - \beta_0\right) + \frac{dF\left(\widehat{G}\left(\beta_0\right), \beta_0\right)}{d\beta}.$$

Let G_j and $\widehat{G}_j$ be the j^{th}elements of G and $\widehat{G}$ respectively, then

$$\begin{aligned}
\frac{d^2F\left(\widehat{G}\left(\overline{\beta}\right),\overline{\beta}\right)}{d\beta d\beta'} =& \sum_{j=1}^{K_g}\frac{\partial F\left(\widehat{G}\left(\overline{\beta}\right)\right)}{\partial G_j}\frac{\partial^2\widehat{G}_j\left(\overline{\beta}\right)}{\partial\beta\partial\beta'} \\
&+\frac{\partial\widehat{G}\left(\overline{\beta}\right)'}{\partial\beta}\left(\frac{\partial^2F\left(\widehat{G}\left(\overline{\beta}\right),\overline{\beta}\right)}{\partial G\partial G'}\frac{\partial\widehat{G}\left(\overline{\beta}\right)}{\partial\beta'}+\frac{\partial^2F\left(\widehat{G}\left(\overline{\beta}\right),\overline{\beta}\right)}{\partial G\partial\beta'}\right) \\
&+\frac{\partial F\left(\widehat{G}\left(\overline{\beta}\right),\overline{\beta}\right)}{\partial\beta\partial G'}\frac{\partial\widehat{G}\left(\overline{\beta}\right)}{\partial\beta'}+\frac{\partial^2F\left(\widehat{G}\left(\overline{\beta}\right),\overline{\beta}\right)}{\partial\beta\partial\beta'} \\
\overset{U_p}{\to}& \sum_{j=1}^{K_g}\frac{\partial F\left(G\left(\beta_0\right)\right)}{\partial G_j}\frac{\partial^2G_j\left(\beta_0\right)}{\partial\beta\partial\beta'} \\
&+\frac{\partial G\left(\beta_0\right)}{\partial\beta}\left(\frac{\partial^2F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G\partial G'}\frac{\partial G\left(\beta_0\right)}{\partial\beta'}+\frac{\partial^2F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G\partial\beta'}\right) \\
&+\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta\partial G'}\frac{\partial G\left(\beta_0\right)}{\partial\beta'}+\frac{\partial^2F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta\partial\beta'} \\
=& F_{\beta\beta}
\end{aligned}$$

and using the fact that β_0 solves

$$\begin{aligned}
0 =& \frac{dF\left(G\left(\beta_0\right),\beta_0\right)}{d\beta} \\
=& \frac{\partial G\left(\beta_0\right)}{\partial\beta}\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}+\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta}
\end{aligned}$$

then adding and subtracting terms including the term in the above expression and using the mean value theorem

$$\begin{aligned}
&\sqrt{N}\frac{dF\left(\widehat{G}\left(\beta_0\right),\beta_0\right)}{d\beta}\\
=&\sqrt{N}\left(\frac{\partial\widehat{G}\left(\beta_0\right)'}{\partial\beta}\frac{\partial F\left(\widehat{G}\left(\beta_0\right),\beta_0\right)}{\partial G}+\frac{\partial F\left(\widehat{G}\left(\beta_0\right),\beta_0\right)}{\partial\beta}\right)\\
=&\sqrt{N}\left(\frac{\partial\widehat{G}\left(\beta_0\right)}{\partial\beta}'\frac{\partial F\left(\widehat{G}\left(\beta_0\right),\beta_0\right)}{\partial G}-\frac{\partial\widehat{G}\left(\beta_0\right)}{\partial\beta}'\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}\right)\\
&+\sqrt{N}\left(\frac{\partial\widehat{G}\left(\beta_0\right)'}{\partial\beta}\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}-\frac{\partial G\left(\beta_0\right)'}{\partial\beta}\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}\right)\\
&+\sqrt{N}\left(\frac{\partial F\left(\widehat{G}\left(\beta_0\right),\beta_0\right)}{\partial\beta}-\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta}\right)\\
=&\sqrt{N}\left(\frac{\partial G\left(\beta_0\right)'}{\partial\beta}\frac{\partial^2 F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G\partial G'}\frac{1}{N}\sum_{i=1}^{N}\left(g\left(X_i,Y_i,\beta_0\right)-G\left(\beta_0\right)\right)\right)\\
&+\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial g\left(X_i,Y_i,\beta_0\right)'}{\partial\beta}-\frac{\partial G\left(\beta_0\right)'}{\partial\beta}\right)\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}\right)\\
&+\sqrt{N}\left(\frac{\partial^2 F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta\partial G'}\frac{1}{N}\sum_{i=1}^{N}\left(g\left(X_i,Y_i,\beta_0\right)-G\left(\beta_0\right)\right)\right)+o_p(1)\\
=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\left(\frac{\partial G\left(\beta_0\right)}{\partial\beta}'\frac{\partial^2 F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G\partial G'}+\frac{\partial^2 F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta\partial G'}\right)\left(g\left(X_i,Y_i,\beta_0\right)-G\left(\beta_0\right)\right)\right.\\
&\left.+\left(\frac{\partial g\left(X_i,Y_i,\beta_0\right)'}{\partial\beta}-\frac{\partial G\left(\beta_0\right)'}{\partial\beta}\right)\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}\right]+o_p(1)\\
=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\vartheta_i+o_p(1).
\end{aligned}$$

Next we derive the asymptotic distribution for $\widetilde{B}\left(\theta_0\right)$. This follows an analogous but slightly more complicated derivation. First define

$$\widetilde{B}_h\left(\theta\right)\equiv argmin_\beta F\left(G(\theta,\beta)\right)$$

then the first order condition and mean value theorem gives for each $h=1,...,H$

$$0=\frac{d^2F\left(\widetilde{G}_h\left(\theta_0,\overline{\beta}\right),\overline{\beta}\right)}{d\beta d\beta'}\left(\widetilde{B}_h\left(\theta_0\right)-\beta_0\right)+\frac{dF\left(\widetilde{G}_h\left(\theta_0,\beta_0\right),\beta_0\right)}{d\beta}$$

where

$$\begin{aligned}
\frac{d^2 F\left(\widetilde{G}_h\left(\theta_0,\overline{\beta}\right),\overline{\beta}\right)}{d\beta d\beta'} =& \sum_{j=1}^{K_g} \frac{\partial F\left(\widetilde{G}_h\left(\theta_0,\overline{\beta}\right),\overline{\beta}\right)}{\partial G_j'} \frac{\partial^2 \widetilde{G}_{hj}\left(\theta_0,\overline{\beta}\right)}{\partial\beta\partial\beta'} \\
&+ \frac{\partial \widetilde{G}_h\left(\theta_0,\overline{\beta}\right)'}{\partial\beta}\left(\frac{\partial^2 F\left(\widetilde{G}_h\left(\theta_0,\overline{\beta}\right),\overline{\beta}\right)}{\partial G\partial G'}\frac{\partial\widetilde{G}_h\left(\theta_0,\overline{\beta}\right)}{\partial\beta'} + \frac{\partial^2 F\left(\widetilde{G}_h\left(\theta_0,\overline{\beta}\right),\overline{\beta}\right)}{\partial G\partial\beta'}\right) \\
&+ \frac{\partial F\left(\widetilde{G}_h\left(\theta_0,\overline{\beta}\right),\overline{\beta}\right)}{\partial\beta\partial G'}\frac{\partial\widetilde{G}_h\left(\theta_0,\overline{\beta}\right)}{\partial\beta'} + \frac{\partial^2 F\left(\widetilde{G}_h\left(\theta_0,\overline{\beta}\right),\overline{\beta}\right)}{\partial\beta\partial\beta'} \\
\xrightarrow{U_p}& \sum_{j=1}^{K_g} \frac{\partial F\left(G\left(\beta_0\right)\right)}{\partial G_j'}\frac{\partial^2 G_j\left(\beta_0\right)}{\partial\beta\partial\beta'} \\
&+ \frac{\partial G\left(\beta_0\right)'}{\partial\beta}\left(\frac{\partial^2 F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G\partial G'}\frac{\partial G\left(\beta_0\right)}{\partial\beta'} + \frac{\partial^2 F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G\partial\beta'}\right) \\
&+ \frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta\partial G'}\frac{\partial G'\left(\beta_0\right)}{\partial\beta} + \frac{\partial^2 F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta\partial\beta'} \\
=& F_{\beta\beta}
\end{aligned}$$

where $\widetilde{G}_{hj}$ is the j^{th}element of $\widetilde{G}_h$.

Analogously to above

$$\begin{aligned}
&\sqrt{N}\frac{dF\left(\widetilde{G}_h\left(\theta_0,\beta_0\right),\beta_0\right)}{d\beta}\\
=&\sqrt{N}\left(\frac{\partial\widetilde{G}_h\left(\theta_0,\beta_0\right)}{\partial\beta}\frac{\partial F\left(\widetilde{G}_h\left(\theta_0,\beta_0\right),\beta_0\right)}{\partial G}+\frac{\partial F\left(\widetilde{G}_h\left(\theta_0,\beta_0\right),\beta_0\right)}{\partial\beta}\right)\\
=&\sqrt{N}\left(\frac{\partial\widetilde{G}_h\left(\theta_0,\beta_0\right)}{\partial\beta}\frac{\partial F\left(\widetilde{G}_h\left(\theta_0,\beta_0\right),\beta_0\right)}{\partial G}-\frac{\partial\widetilde{G}_h\left(\theta_0,\beta_0\right)}{\partial\beta}\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}\right)\\
&+\sqrt{N}\left(\frac{\partial\widetilde{G}_h\left(\theta_0,\beta_0\right)}{\partial\beta}\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}-\frac{\partial G\left(\beta_0\right)}{\partial\beta}\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}\right)\\
&+\sqrt{N}\left(\frac{\partial F\left(\widetilde{G}_h\left(\theta_0,\beta_0\right),\beta_0\right)}{\partial\beta}-\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta}\right)\\
=&\sqrt{N}\left(\frac{\partial G\left(\beta_0\right)}{\partial\beta'}\frac{\partial F\left(G\left(\beta_0\right),B\left(\theta_0\right)\right)}{\partial G\partial G'}\frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{g}_{hi}\left(\beta_0\right)-G\left(\beta_0\right)\right)\right)\\
&+\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial\widetilde{g}_{hi}\left(\beta_0\right)}{\partial\beta}-\frac{\partial G\left(\beta_0\right)}{\partial\beta}\right)\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}\right)\\
&+\sqrt{N}\left(\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta\partial G'}\frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{g}_{hi}\left(\beta_0\right)-G\left(\beta_0\right)\right)\right)+o_p(1)\\
=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\left(\frac{\partial G\left(\beta_0\right)}{\partial\beta'}\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G\partial G'}+\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial\beta\partial G'}\right)\left(\widetilde{g}_{hi}\left(\beta_0\right)-G\left(\beta_0\right)\right)\right.\\
&\left.+\left(\frac{\partial\widetilde{g}_{hi}\left(\beta_0\right)}{\partial\beta}-\frac{\partial G\left(\beta_0\right)}{\partial\beta}\right)\frac{\partial F\left(G\left(\beta_0\right),\beta_0\right)}{\partial G}\right]+o_p(1)\\
=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\widetilde{\vartheta}_{hi}+o_p(1).
\end{aligned}$$

And so

$$\begin{aligned}
\sqrt{N}\left[\widehat{B}(\theta_0)-\widehat{\beta}\right]=&F_{\beta\beta}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\left[\frac{1}{H}\sum_{h=1}^{H}\widetilde{\vartheta}_{hi}\right]-\vartheta_i\right)+o_p(1)\\
\sim&N\left(0,F_{\beta\beta}^{-1}VF_{\beta\beta}^{-1}\right).
\end{aligned}$$

Then

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right)\xrightarrow{d}N\left(0,\left[\frac{\partial B(\theta_0)'}{\partial\theta}\Omega\frac{\partial B(\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial B(\theta_0)'}{\partial\theta}\Omega F_{\beta\beta}^{-1}VF_{\beta\beta}^{-1}\Omega\frac{\partial B(\theta_0)}{\partial\theta}\left[\frac{\partial B(\theta_0)'}{\partial\theta}\Omega\frac{\partial B(\theta_0)}{\partial\theta'}\right]^{-1}\right).$$

Appendix B: Data

As mentioned in the text, we use white women from the last four panels of the Survey of Income and Program Participation. We first measure potential experience in months and use anyone from 1 month to 35 years of potential experience. The variable used in the data is annualized. The SIPP is asked every four months. We only use data from the month of the interview.
We detail construction of the variables

- Potential Experience: For older workers we don't know exactly when they graduated school. We assume that they graduate in June of the year they turn a)16 if education is less than 12, b) 18 if their education is exactly 12, c) 20 if their education is more than 12 but less than 16, and d) 22 if their education is larger than 22.

- Employment: We define employment to be 1 for individuals who work some during the month and are never unemployed during the month.

- Education: We take the maximum of the education variable in each wave which is completed education. We convert to numeric variables as, 0 if less than first grade, 2.5 if education is first through fifth grade, 5.5 if it is fifth or sixth, 7.5 if it is seventh or eighth, the numeric grade completed through high school, 12 if high school or equivalent, 13 if a vocational certificate, 13.5 if a vocational associate degree, 14 if an academic vocational degree, 16 if a four year graduate, 17 if a masters degree, and 18 if professional degree or higher.

- Log wage: Wage is constructed as the hourly rate of pay for people who are paid by the hour and monthly earnings divided by (weeks worked×usual hours per week). If one worked every week of the month we use 4.3 as the number of weeks. It is deflated to 2008 dollars using the personal consumption expenditures price index. We drop observations with a real wage below 1$ or above 300$.

- Married: What we really use for married is whether the spouse is present in the household. Using the household interview status code, we use epnspous to match the respondent to spouse, and our married variable is a dummy variable for whether that spouse is in the household in that particular wave.

- Kids ages: To measure kids that are present we use the household data to match mothers to their children. From the individual records we collect their birthdates and use that to keep track of the number of children less than seven and less than eighteen.

We also use this to construct the difference in age between the oldest and youngest child. We construct a dummy variable for having a baby when a baby enters the household between waves (thus, if they never lived in the household with the mother any any wave they would not be recorded as children).

- Total Number of Children ever had: From topical module wave 2, variable tmomchl. We also sometimes use a dummy variable for whether this is bigger than zero. We also occasionally use a variable that is kids greater than 18. To get this we combine data from the topical module for total children and then subtract the number less than 18 at wave 2. We then use the ages in the household roster to determine that in the other waves.

Appendix C: Auxiliary Model

We have a total of 432 auxiliary parameters. We detail them in this appendix and document how much weight is given to each. All of them start with the inverse of the variance of estimated parameters and are multiplied by the weights listed below.

- Regression of log wages on experience and other variables (38 parameters)

  - 35 Experience dummies (all weight=1)
  - Number of Children $<$ 18 (weight 101)
  - Number of Children $<$ 7 (weight 101)
  - Dummy variable for Married (weight 101)

- Within and between variance of residual from fixed effect regression (2 parameters)

  - Let T_i be the number of wage observations we have from individual i. Construct residuals from the fixed effect regression above (where the fixed effect is included in the regression). Define these as ω_{it} and order them as $\omega_{i1}, ..., \omega_{iT_i}$. Define

$$\begin{aligned} \overline{\omega}_i \equiv & \frac{1}{T_i} \sum_{t=1}^{T_i} \log\left(\omega_{it}\right) \\ \overline{\omega} = & \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \log\left(\omega_{it}\right)}{\sum_{i=1}^{N} T_i}. \end{aligned}$$

We can then decompose the total variance into

$$\frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \left[\log\left(\omega_{it}\right) - \overline{\omega}\right]^2}{\sum_{i=1}^{N} T_i} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \left[\log\left(\omega_{it}\right) - \overline{\omega}_i\right]^2}{\sum_{i=1}^{N} T_i} + \frac{\sum_{i=1}^{N} T_i \left[\overline{\omega}_i - \overline{\omega}\right]^2}{\sum_{i=1}^{N} T_i}$$

where the first part is the within variance and the second is the between variance.

  - Both have weight=1

- Regression of fixed effect on education

  - coefficient on education (weight=1)
  - Intercept from regression not matched (weight=0)

- Linear probability model of being married in the first period observed (39 parameters)

  - 36 experience dummies (all weight=1)
  - education (weight=101)
  - estimate of fixed effect (weight=101)
  - dummy variable for fixed effect missing not matched (weight=0)

- Linear probability of marriage conditional on not married in previous period (38 parameters)

  - 35 experience dummies (all weight=1)
  - education (weight=101)
  - estimate of fixed effect (weight=101)

  - dummy variable for fixed effect missing not matched (weight=0)

- Linear probability of marriage conditional on married in the previous period (39 parameters)

  - 35 experience dummies (all weight=1)
  - education (weight=101)
  - estimate of fixed effect (weight=101)
  - dummy variable for fixed effect missing not matched (weight=0)
  - number of kids≤ 18(weight=101)

- fraction of people having a baby who are married (1 parameter)

  - weight=101

- regression of having a kid on wage (1 parameter)

  - current wage for women working (weight=101)
  - also in regression: intercept, education, married, number of kids <7, number of kids ≤ 18, potexp, potexp^2 (not matched, weight=0)

- Age difference between oldest and youngest child (1 parameter)

  - weight=101

- regression of dummy variable for any kids (37 parameters)

  - 36 age dummies (weight=1)

  - education (weight=101)

- regression of dummy variable for exactly 2 kids (37 parameters)

  - 36 age dummies (weight=1)
  - education (weight=101)

- regression of number of kids (37 parameters)

  - 36 age dummies (weight=1)
  - education (weight=101)

- Linear probability model of working in the first period observed (40 parameters)

  - experience dummy variables (36 parameters)
  - education (weight=1)
  - married (weight=101)
  - number of kids <7 (weight=101)
  - number of kids < 18 (weight=101)

- Linear probability model of finding a job conditional on not working in the previous period (40 parameters)

  - experience dummy variables (35 parameters)
  - education (weight=1)
  - married (weight=101)
  - any kids <7 (weight=101)
  - number of kids <7 (weight=101)

  - number of kids < 18 (weight=101)

- Linear probability model of keeping one's job conditional on working in previous period (40 parameters)

  - 35 experience dummy variables (weight=1)
  - education (weight=1)
  - fixed effect (weight=101)
  - married (weight=101)
  - number of kids < 7 (weight=101)
  - number of kids ≤ 18 (weight=101)

- fraction of women who were working when having a kid (1 parameter)

  - weight=101

- regression of work on fixed effect (1 parameter)

  - fixed effect (weight=101)
  - intercept not matched (weight=0)

- Regression of wage growth for continuously employed (39 parameters)

  - 36 experience dummy variables (weight=11)
  - education (weight=801)
  - married (weight=801)
  - number of kids ≥ 18 (weight=801)

- wage growth overlapping non-employment spell (1 parameter)
  - change in potential experience (weight=801)
  - That is, we look at people who had a wave where they were working followed by one or more waves of non-employment, followed by a wave in which they were working. We regress the difference in wages post and pre non-employment spell on the length of the time in between.

Figure A1: Fit of model: Fixed Effect Profile

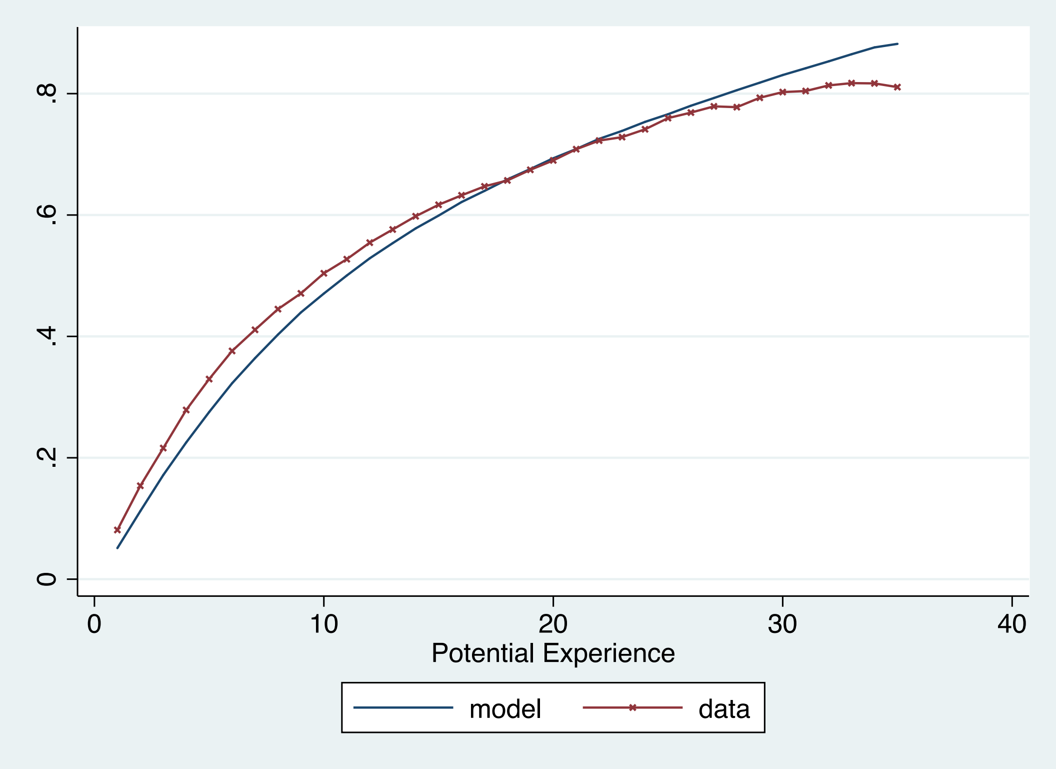


Figure A2: Fit of model: Get Married

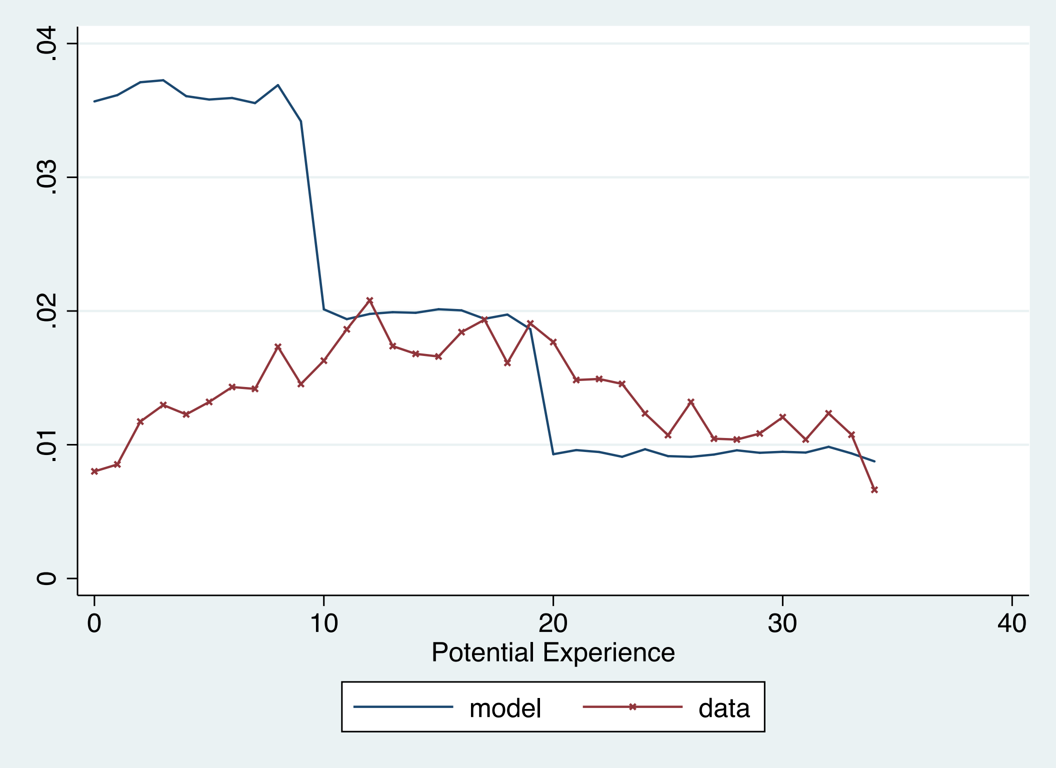


Figure A3: Fit of model: Stay Married

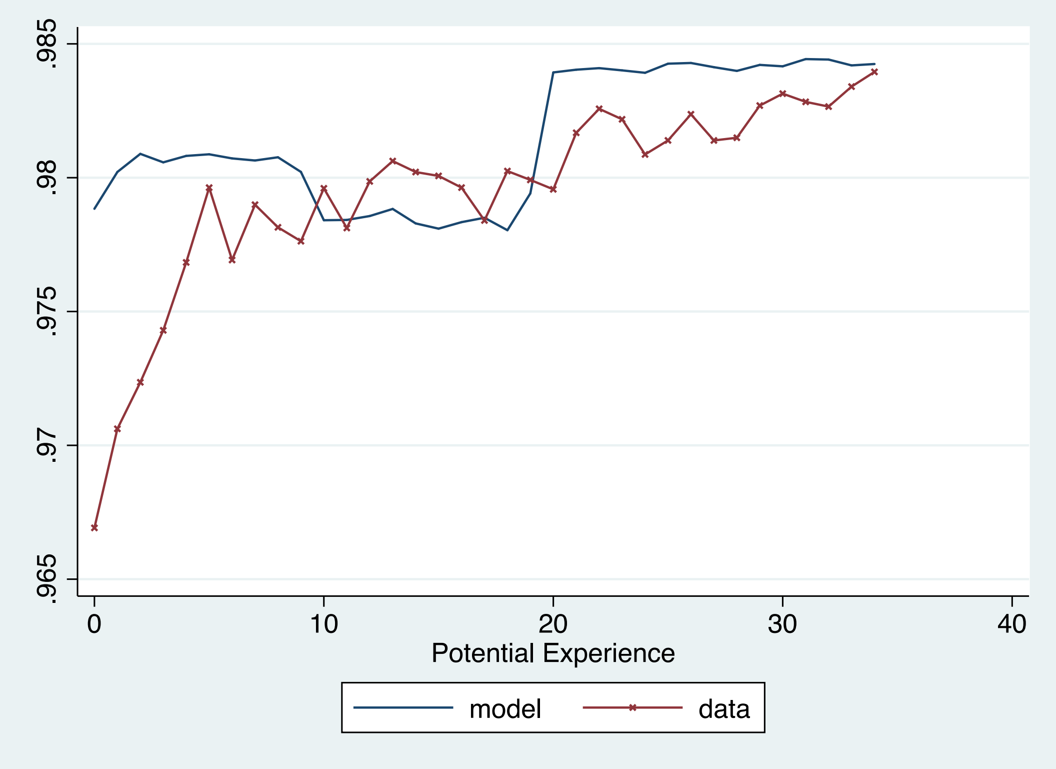


Figure A4: Fit of model: Any Children

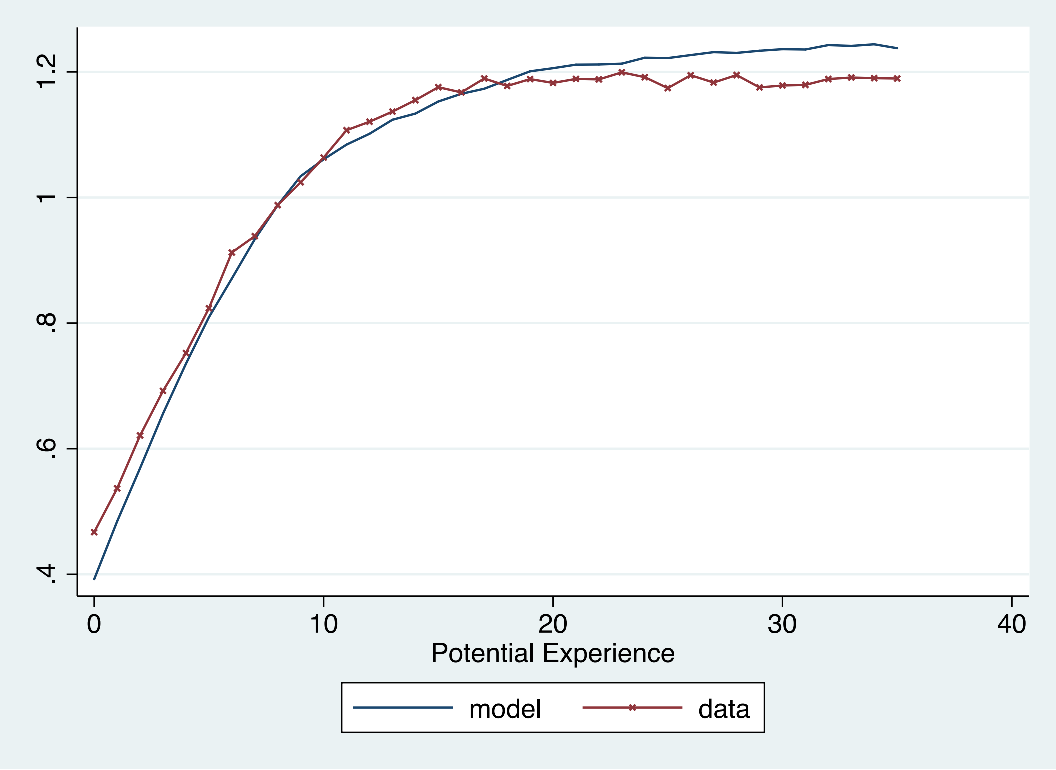


Figure A5: Fit of model: Stay Married

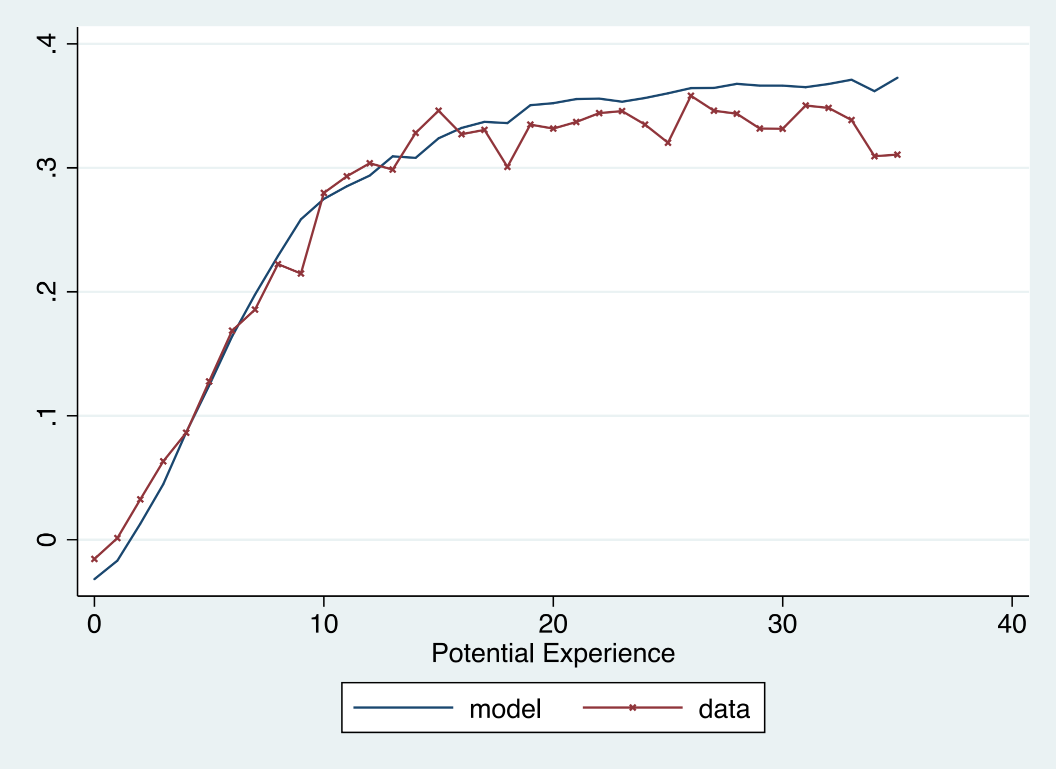


Figure A6: Fit of model: Stay Employed

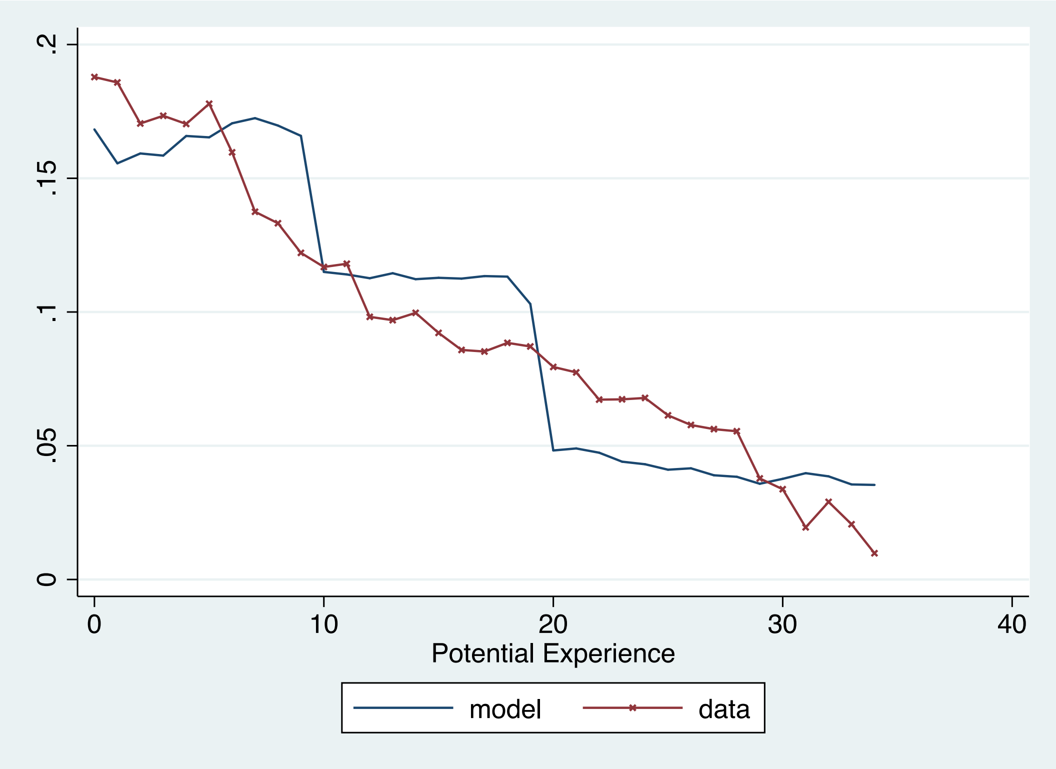
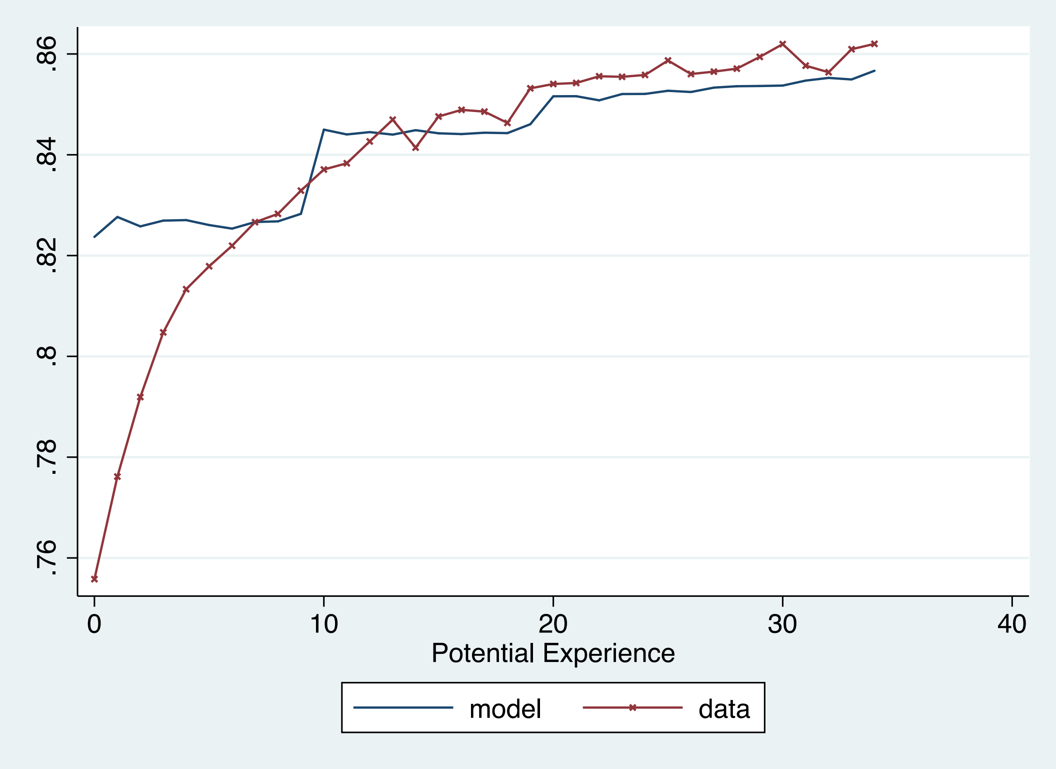


Figure A7: Fit of model: Non-employed to Employed


Potential Experience
model
data