A Factor Model Approach for a Conditional Choice Probability Estimator

Yujung Hwang

Introduction

- Traditional Conditional Choice Probability (CCP) estimator has used only longitudinal choice information for identification and estimation (e.g., Arcidiacono and Miller, 2011)
- Such an approach is not ideal for DDCM in the presence of time-varying discrete latent types
  (i) DDCM w/ hidden action, (ii) DDCM w/ a law of motion for latent types
- This paper proposes a DDCM-CCP + Factor Model Approach (e.g. Cunha, Heckman, Schennach, 2010)

Proxies For Unobs. Heterogeneity

Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

Search

TOPICS ▾ DATA & DOCUMENTATION ▾ RESEARCH ▾ PARTICIPANTS ABOUT US ▾ HELP & SUPPORT ▾

MAIN SURVEY

Home > Data and documentation > Main survey > Variable search

Variable search | Datafiles | Index terms | Questionnaire modules

Index terms

Our Index Terms cover all the thematic areas in the Study. Use the Index to identify the variables most relevant to your research interests and to find other variables with related data throughout the dataset.

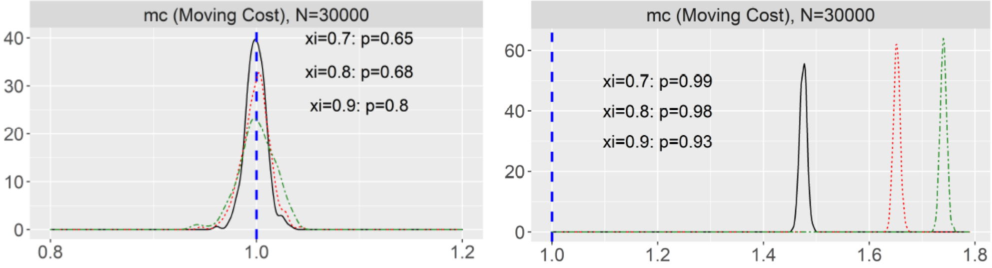
- Caring
- Child Care
- Children
- Coded variables
- Cognitive ability
- Computers and Computing
- Consents
- Crime
- Derived Variables
- Education: Background and attainments
- Education: Expectations
- Education: Recent Education and Training
- Employment History: Labour Force Status Spells
- Employment History: Reasons for Leaving and Taking Jobs
- Employment History: Size, Sector and Duties
- Employment: Attitude to Work and Incentives
- Financial Management: Loan Repayments
- Financial Management: Material Wellbeing
- Financial Management: Pensions
- Financial Management: Personal Spending
- Financial Management: Problems
- Financial Management: Savings and Bank Accounts
- Friendship
- Gender Roles
- Geographic Location
- Geographic Mobility
- Harassment
- Health behaviour
- Health: Accidents, Illness
- Health: Childrens Health
- Health: Drinking
- Health: Effect on Daily Life, Employment
- Health: Height and Weight
- Incomes: Benefits and Allowances and Pensions
- Incomes: Grants for Education
- Incomes: Household Income
- Incomes: Rents, Savings, Investments
- Incomes: Windfalls
- Interview Characteristics and Conditions
- Key Linking Variable
- Leisure Activity
- Life Events
- Marital and Cohabitation History
- Marital Status
- Neighbourhood and Residence
- Newspaper Readership
- Parents and Children: Parenting Styles
- Person identifiers
- Personality traits
- Physical Characteristics

Why Factor Model Approach?

1. Often $\exists$ many proxies on a single trait $\rightarrow$ curse of dimensionality or ad-hoc choice
2. Constructing a consistent summary index (e.g., principal component or z-score) is difficult in an unbalanced panel of proxies
3. Ignoring measurement errors in proxies + Hotz-Miller (1993) leads to substantial bias in the estimates (Monte Carlo Example)

Monte Carlo Simulation: Schelling's Segregation Model, Moving Cost Parameter

Left: Factor Model Approach, Right: Hotz-Miller assuming no ME



Blue: True Parameter Black: 90% Proxy Precision,

Red: 80% Proxy Precision, Green: 70% Proxy Precision

Paper Summary

1. Identification
   - Proxies (noisy msrments of types) help identification ($\Leftrightarrow$ Hu and Shum, 2012)
   - There is a better survey design for the identification
2. Estimation: Extension of Arcidiacono and Miller (2011) in the presence of proxies
3. (Monte Carlo Simulation: Both single-agent and general equilibrium model. Treating Proxies as a Measurement-Error Free Observable State Variable leads to substantial bias (e.g. first principal component, z-scores))
4. Empirical Application: Model of labor supply and mental health (also see Hwang 2019; Wang 2021)

Literature

- Measurement Error

  Hu (2008), Hu and Schennach (2008), Allman, Matias, Rhodes (2009), Cunha, Heckman, Schennach (2010), Hu (2017)

- Identification of a Dynamic Discrete Choice Model

  Hu and Shum (2012), Hu and Sasaki (2018), Berry and Compiani (2023)

- Conditional Choice Probability Estimator

  Hotz and Miller (1993), Hotz, Miller, Sanders, Smith (1994), Arcidiacono and Jones (2003), Arcidiacono and Miller (2011), Bajari, Benkard, Levine (2007), Aguirregabiria and Mira (2007), Pesendorfer and Schmidt-Dengler (2008)

Identification

DDCM Model

- observable discrete choice j_{it}
- state variables $\Omega_{it} = \{X_{it}, U_{it}\}$, X_{it} observable, U_{it} (discrete) unobservable (evolves by hidden action or law of motion), proxies for U_{it}, $\{Y_{mit}\}_{m=1}^{K_t}$

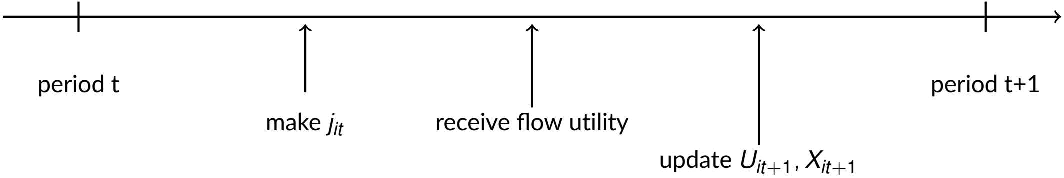


Figure: Timeline of the Model

Summary of Identification Results Under Different Survey Designs

Three proxies (that satisfies Hu (2008), a.k.a Conditional Independence and Rank condition) are measured in the same period,

Case 1 ..and one of them is available in consecutive periods

Case 2 ..and one of them is available, p-periods apart

Case 3 ..and none of them are measured again

As we move from case 1 to 3, we need more restrictions on model dynamics

Summary of Identification Case 1

- Assume the repeatedly measured proxy in consecutive periods has a rank $\geq$ cardinality of U_{it}. Then, 2 periods of data are sufficient to identify the reduced-from choice probabilities, initial type probability, and 1-period transition probability
- $\Leftrightarrow$ Hu and Shum (2012) derived the identification without proxies using $T = 4$ periods of data and invertibility assumption on model dynamics.

Summary of Identification Case 2

- When a proxy exists p-periods apart, with similar logic, we can show identification of reduced-form choice probabilities, initial probability, and p-period transition probability

$$f(U_{it+p}|U_{it}, \{X_{it+l}, j_{it+l}\}_{l=0}^{l=t+p-1})$$

- We need assumptions on p-period transition probability to identify a unique 1-period transition probability

Summary of Identification Case 3

- When no proxy is repeatedly measured, we can show identification (of reduced-form choice probabilities, initial type probability, and 1-period transition probability) with $T = 2$ periods of data under strong assumptions on model dynamics
  - Number of observable choice is greater than cardinality of unobservable state var $J \geq \overline{u}$
  - For $\forall X_{it}$, both CCP matrix $[p(j_{it}|X_{it}, U_{it})]$ of size $J \times \overline{u}$ and the joint density matrix matrix $[f(j_{it}, X_{it}, U_{it})]$ of size $J \times \overline{u}$ have a full rank, equal to $\overline{u}$.

Estimator: Extension of Arcidiacono and Miller (2011)

New Estimator Summary

- Idea of Arcidiacono and Jones (2003), Arcidiacono and Miller (2011): Estimate the CCP $p^j(X_{it}, U_{it})$ using EM-algorithm in the first-step
- Factor Model Approach : Construct the likelihood from both observed discrete choices and proxies of U_{it}
  - New per-period likelihood

$$\widehat{L}_{ist}^{(k)} = \underbrace{\widehat{p}_j^{(k)}(\Omega_{ist})}_{\text{AM(2011)}} \underbrace{l_Y(Y_{it}|U_{it} = s_t, \widehat{\theta}_y^{(k)})}_{\text{Proxy likelihood (New!)}} \tag{1}$$

- Extra Estimation Steps
  - Proxy likelihood estimation $l_Y(Y_{it}|U_{it} = s_t, \widehat{\theta}_y^{(k)})$
  - Unobs. State Transition $\widehat{p}_u^{(k)}(U_{it+1} = s_{t+1}|j_{it}, X_{it}, U_{it} = s_t)$

Estimation Algorithm: First Stage

Step 1 Set $\overline{u}$ and initialize type probability and type transition probability

Step 2 Maximization step (M-step)

Step 3 Expectation step (E-step)

Step 4 Iterate between M- and E- step until convergence

- After the first stage, we get proxy measurement parameters θ_Y, X_{it} transition parameters θ_f, type probability q_{ist}, reduced-form type transition probabilities $q_{is't+1|s}$, reduced-form CCPs p_j, p_u, initial type probability π_{is}

Estimation Algorithm: First Stage M-step

Step 2.1 Estimate reduced-form CCP

$$\widehat{p}_j^{(k)}(\Omega_{it}) = \frac{\sum_{i=1}^{N} \widehat{q}_{ist}^{(k)} I(j_{it} = j) I(X_{it} = x)}{\sum_{i=1}^{N} I(X_{it} = x)} \tag{2}$$

Step 2.2 Estimate $\widehat{\theta}_y^{(k)}$ by maximizing the likelihood of proxies Y_{it} given $\widehat{q}_{ist}^{(k)}$

$$\underset{\widehat{\theta_y}^{(k)}}{argmax} \quad \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{\bar{u}} \widehat{q}_{ist}^{(k)} log l_Y(Y_{it} | U_{it} = s, \widehat{\theta_y}^{(k)}) \tag{3}$$

Estimation Algorithm: First Stage M-step

Step 2.3 Estimate $\widehat{\theta}_f^{(k)}$ by maximizing the pseudo log-likelihood of observable state transition probability given $\widehat{q}_{ist}^{(k)}$

$$\underset{\widehat{\theta}_f^{(k)}}{argmax} \quad \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{\bar{u}} \widehat{q}_{ist}^{(k)} logf(X_{it}|j_{it-1}, X_{it-1}, U_{it}=s, \widehat{\theta}_f^{(k)}) \tag{4}$$

Step 2.4 Estimate $\widehat{\mathbb{P}}_u^{(k)}$ by computing the weighted nonparametric forward probability $\widehat{q}_{is't+1|s}^{(k)}$ given $\widehat{q}_{ist}^{(k)}$

$$\widehat{p}_u^{(k)}(j_{it}, \Omega_{it}) = \frac{\sum_{i=1}^{N} \widehat{q}_{is't+1|s}^{(k)} q_{ist}^{(k)} I(j_{it}=j)I(X_{it}=x)}{\sum_{i=1}^{N} I(j_{it}=j)I(X_{it}=x)} \tag{5}$$

Estimation Algorithm: Second Stage E-step

- Deterministically update $\widehat{q}_{ist}^{(k+1)}, \widehat{q}_{is't+1|s}^{(k+1)}$ using Bayes rule and constructed likelihoods based on estimates from M-step

$$
\begin{aligned}
\widehat{L}_i^{(k)}(U_{it}=s) &= \sum_{s_0=1}^{\overline{u}} \cdots \sum_{s_{t-1}=1}^{\overline{u}} \sum_{s_t=s}^{s} \sum_{s_{t+1}=1}^{\overline{u}} \cdots \sum_{s_T=1}^{\overline{u}} \pi_{is0} \Pi_{t=1}^{T} \\
&\quad \Big\{ \widehat{L}_{ist}^{(k)} \widehat{p}_u^{(k)}(U_{it+1}=s_{t+1} | j_{it}, X_{it}, U_{it}=s_t) \\
&\quad \times f(X_{it+1} | j_{it}, X_{it}, U_{it+1}=s_{t+1}, \widehat{\theta}_f^{(k)})) \Big\} \\
\widehat{L}_{ist}^{(k)} &= \widehat{p}_j^{(k)}(\Omega_{ist}) l_Y(Y_{it} | U_{it}=s_t, \widehat{\theta}_y^{(k)})
\end{aligned}
$$

Estimation Algorithm: First Stage, E-step

$$
\begin{aligned}
\widehat{L}_i^{(k)} &= \sum_{s_t=1}^{\overline{u}} \widehat{L}_i(U_{it} = s_t) \\
\widehat{q}_{ist}^{(k+1)} &\equiv \frac{\widehat{L}_i^{(k)}(U_{it} = s)}{\widehat{L}_i^{(k)}} \\
\widehat{q}_{is'\,t+1|s}^{(k+1)} &= \frac{\widehat{p}_u^{(k)}(U_{it+1} = s' | j_{it}, X_{it}, U_{it} = s) \left[\sum_{U_{it+2}=1}^{\overline{u}} \cdots \sum_{U_{i\mathcal{T}}=1}^{\overline{u}} \widehat{p}_u^{(k)}(U_{it+2} | j_{it+1}, X_{it+1}, U_{it+1} = s') \left(\Pi_{t'=t+2}^{\mathcal{T}} \widehat{L}_{ist'} \right) \right]}{\sum_{z=1}^{\overline{u}} \widehat{p}_u^{(k)}(U_{it+1} = z | j_{it}, X_{it}, U_{it} = s) \left[\sum_{U_{it+2}=1}^{\overline{u}} \cdots \sum_{U_{i\mathcal{T}}=1}^{\overline{u}} \widehat{p}_u^{(k)}(U_{it+2} | j_{it+1}, X_{it+1}, U_{it+1} = s') \left(\Pi_{t'=t+2}^{\mathcal{T}} \widehat{L}_{ist'} \right) \right]}
\end{aligned}
$$

Estimation Algorithm: First Stage, E-step

Step 3.2 Update $\widehat{\pi}^{(k+1)}(X_{i0})$ given $\widehat{q}_{is0}^{(k+1)}$

$$\widehat{\pi}^{(k+1)}(X_{i0} = x) = \frac{\sum_{i=1}^{N} q_{is0}^{(k+1)} I(X_{i0} = x)}{\sum_{i=1}^{N} I(X_{i0} = x)} \tag{8}$$

CCP Estimation Second Step

- Like other CCP estimators, the second stage estimation can be done in various ways, and it exploits the fixed-point constraints (hidden action VS law of motion) imposed by the model

$$
\begin{aligned}
\mathbb{P} &= \Psi(\Omega_{it}, \mathbb{P}, \theta) \\
&= \begin{bmatrix} \Psi_1(\Omega_{it}, \mathbb{P}, \theta) \\ \Psi_2(\Omega_{it}, j_{it}, \mathbb{P}, \theta) \end{bmatrix} \\
&= \begin{bmatrix} l_j(j_{it} = 1 | \Omega_{it}, \mathbb{P}, \theta) \\ \vdots \\ l_j(j_{it} = J | \Omega_{it}, \mathbb{P}, \theta) \\ l_u(U_{it+1} = 1 | \Omega_{it}, j_{it}, \mathbb{P}, \theta) \\ \vdots \\ l_u(U_{it+1} = \overline{u} | \Omega_{it}, j_{it}, \mathbb{P}, \theta) \end{bmatrix}
\end{aligned}
$$

Estimator's Asymptotic Property

$$\max_{\lambda_2} \quad \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{\bar{u}} \widehat{q}_{ist} log(\widetilde{\Psi}(j_{it}, \widehat{\mathbb{P}}_u)) \tag{9}$$

$$\widetilde{\Psi}(j_{it}, \widehat{\mathbb{P}}_u) = \Pi_{j'=1}^{J} \Psi_{1j'}^{1(j_{it}=j')} \times \Pi_{u'=1}^{\bar{u}} \Psi_{2u'}^{(\widehat{\mathbb{P}}_u)_{u'}} \tag{10}$$

- AM(2011) estimator's root-N consistency and asymptotic normality hold for this estimator as well

Estimation: Labor Supply and Mental Health

Labor Supply and Mental Health

- Mental health both directly and indirectly affect labor supply
  - Bad mental health increases disutility when working
  - Bad mental health affects other determinants, e.g. marital status, wage rate etc
- Labor supply (and its determinants) may affect mental health dynamics
- Use proxies for mental health (GHQ measures)
- Use BHPS-UKHLS data

Model

- An agent is characterized by their current period log labor income w_{it}, cumulative work experience n_{it}, mental health u_{it}, marital status m_{it} and age a_{it}. That is, the state variables are $\Omega_{it} = \{w_{it}, n_{it}, u_{it}, m_{it}, a_{it}\}$.
- Mental health $u_{it} \in \{1, \cdots, \overline{u}\}$. High values of u_{it} mean more severe distress.
- Every period, the agent chooses a labor supply $j_{it} \in \{0, 1\}$

Model

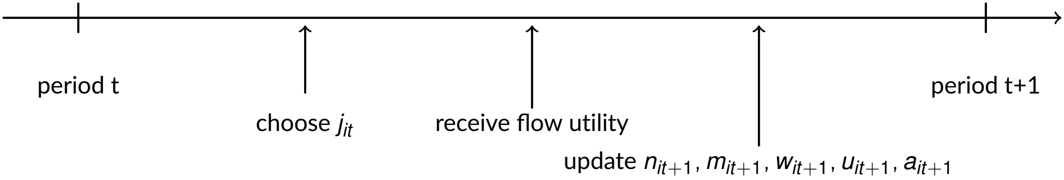


Figure: Timeline of the Model

Model

- Flow utility

$$u(w_{it}, u_{it}, j_{it}) = \Big\{\theta_1 w_{it} + \theta_2 + \sum_{u'=2}^{\overline{u}} \theta_{3u'} \mathbb{1}(u_{it} = u') + \theta_4 m_{it} + \theta_5 a_{it} + \theta_6 n_{it} \\ + \theta_7 \mathbb{1}(n_{it} = 0)\Big\} j_{it}$$

- Value function

$$\begin{aligned} v_j(\Omega_{it}) &= u(w_{it}, u_{it}, j_{it}) + \beta E\left[V(\Omega_{it+1}) | \Omega_{it}, j_{it}\right] \\ V(\Omega_{it}) &= E\left[\max\{v_0(\Omega_{it}) + \epsilon_{0it}, v_1(\Omega_{it}) + \epsilon_{1it}\}\right], \quad \epsilon_{jit} \sim Gumbel(0, 1) \end{aligned}$$

Model

- Terminal period: *cumulative work experience n_{iT} is a valuable asset*

$$V_T(\Omega_{iT}) = \left(\sum_{u'=2}^{\bar{u}} \iota_{1u'} \mathbb{1}(u_{iT} = u') + \iota_2 m_{iT} + \iota_3 \right) n_{iT}$$

- Wage transition: *random walk + deterministic depreciation & hetero. wage growth*

$$w_{it+1} = w_{it} + \beta_1^w + j_{it} \left\{ \beta_2^w + \beta_3^w n_{it} + \beta_4^w a_{it} + \beta_5^w m_{it} + \sum_{u'=2}^{\bar{u}} \beta_{6u'}^w \mathbb{1}(u_{it} = u') \right\} + \nu_{it},$$

$$\nu_{it} \sim N(0, \sigma_\nu^2)$$

Model

- Mental health transition

$$\log\frac{P(u_{it+1}=k)}{P(u_{it+1}=1)} = \beta_0^{uk}+\beta_1^{uk}j_{it}+\beta_2^{uk}w_{it}+\beta_3^{uk}n_{it}+\beta_4^{uk}a_{it}+\beta_5^{uk}m_{it} + \sum_{u'=2}^{\overline{u}}\beta_{6u'}^{uk}\mathbb{1}(u_{it}=u')$$

If $\beta_1^{uk}=\beta_2^{uk}=\beta_3^{uk}=\beta_4^{uk}=\beta_5^{uk}=0$, then it's Markov chain

- Marriage transition

$$\log\frac{P(m_{it+1}=1)}{P(m_{it+1}=0)} = \beta_0^{m}+\beta_1^{m}j_{it}+\beta_2^{m}w_{it}+\beta_3^{m}n_{it}+\beta_4^{m}a_{it}+\beta_5^{m}m_{it} + \sum_{u'=2}^{\overline{u}}\beta_{6u'}^{m}\mathbb{1}(u_{it}=u')$$

Data

- BHPS-UKHLS 1991-2019
- Non-White males without a college degree, aged between 30-50 (Emp. rate 88% VS White Male College Educated 98%)
- Drop part-time workers
- Two-stage estimator and Hotz, Miller, Sanders, Smith (1994) in the second stage [No Finite Dependence], num of simulation = 100

BHPS-UKHLS Sample Characteristics (Non-White Male Without a College Degree, Aged Between 30 and 50), Year 1991-2019

Variable	Obs	Mean	Std. Dev.
Whether employed	4311	0.89	0.31
Age	4311	40.79	5.53
Married	4311	0.77	0.42
Vocational degree	4311	0.20	0.40
A-level equivalents	4311	0.30	0.46
Below A-level	4311	0.27	0.44
Other qualification	4311	0.15	0.35
No qualification	4311	0.08	0.27
Real monthly labor income (unit: 1000 pounds, CPI 2015=100)	4311	2.27	1.11
Individual-Year Observation	4311		
Individual Observation	744		

General Health Questionnaires(GHQ) in BHPS-UKHLS

Acronym	Questionnaire	Scale	Wave
	The next questions are about how you have been feeling over the last few weeks. Have you recently...		
scghqa	• been able to concentrate on whatever you're doing?	4-scale	every wave
scghqb	• lost much sleep over worry?	4-scale	every wave
scghqc	• felt that you were playing a useful part in things?	4-scale	every wave
scghqd	• felt capable of making decisions about things?	4-scale	every wave
scghqe	• felt constantly under strain?	4-scale	every wave
scghqf	• felt you couldn't overcome your difficulties?	4-scale	every wave
scghqg	• been able to enjoy your normal day-to-day activities?	4-scale	every wave
scghqh	• been able to face up to problems?	4-scale	every wave
scghqi	• been feeling unhappy or depressed?	4-scale	every wave
scghqj	• been losing confidence in yourself?	4-scale	every wave
scghqk	• been thinking of yourself as a worthless person?	4-scale	every wave
scghql	• been feeling reasonably happy, all things considered?	4-scale	every wave

Mental Health Type Estimation

- Set the number of latent types $\overline{u} = 4$ (Can reject $\overline{u} = 3$)
- Can estimate the $f(Y_{it}|U_{it})$ before applying the CCP estimator (because there are more than three proxies in one wave)
- Not all proxies are equally informative

Estimates : $(M_k)_{zj} = [P(y_{kit} = j | u_{it} = z)]$

'scghqk' been thinking of yourself as a worthless person?

$$\begin{bmatrix} 0.97 & 0.02 & 0.01 & 0.00 \\ 0.64 & 0.35 & 0.01 & 0.00 \\ 0.35 & 0.36 & 0.27 & 0.02 \\ 0.17 & 0.10 & 0.41 & 0.31 \end{bmatrix}$$

'scghqa' been able to concentrate on whatever you're doing?

$$\begin{bmatrix} 0.19 & 0.75 & 0.04 & 0.02 \\ 0.04 & 0.91 & 0.05 & 0.01 \\ 0.03 & 0.49 & 0.45 & 0.03 \\ 0.02 & 0.20 & 0.47 & 0.31 \end{bmatrix}$$

Estimates : $(M_k)_{zj} = [P(y_{kit} = j | u_{it} = z)]$

'scghqk' been thinking of yourself as a worthless person?

$$\begin{bmatrix} \boxed{0.97} & 0.02 & 0.01 & 0.00 \\ 0.64 & 0.35 & 0.01 & 0.00 \\ 0.35 & 0.36 & 0.27 & 0.02 \\ 0.17 & 0.10 & 0.41 & 0.31 \end{bmatrix}$$

'scghqa' been able to concentrate on whatever you're doing?

$$\begin{bmatrix} \boxed{0.19} & 0.75 & 0.04 & 0.02 \\ 0.04 & 0.91 & 0.05 & 0.01 \\ 0.03 & 0.49 & 0.45 & 0.03 \\ 0.02 & 0.20 & 0.47 & 0.31 \end{bmatrix}$$

Estimates : $(M_k)_{zj} = [P(y_{kit} = j | u_{it} = z)]$

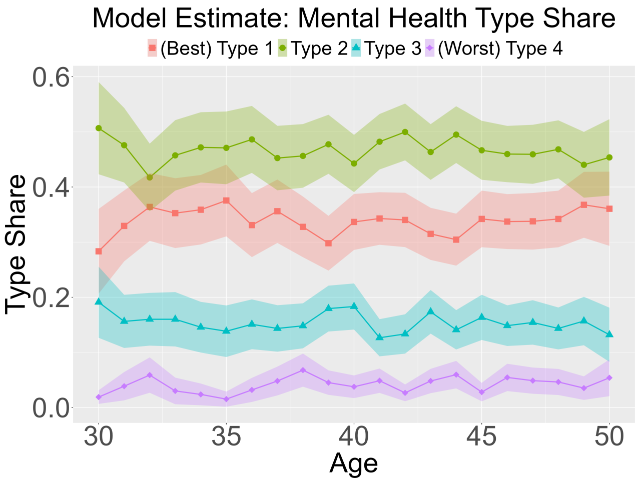
'scghqe' felt constantly under strain?

$$\begin{bmatrix} 0.66 & 0.30 & 0.02 & 0.01 \\ 0.07 & 0.82 & 0.10 & 0.00 \\ 0.04 & 0.31 & 0.57 & 0.08 \\ 0.02 & 0.05 & 0.32 & 0.60 \end{bmatrix}$$

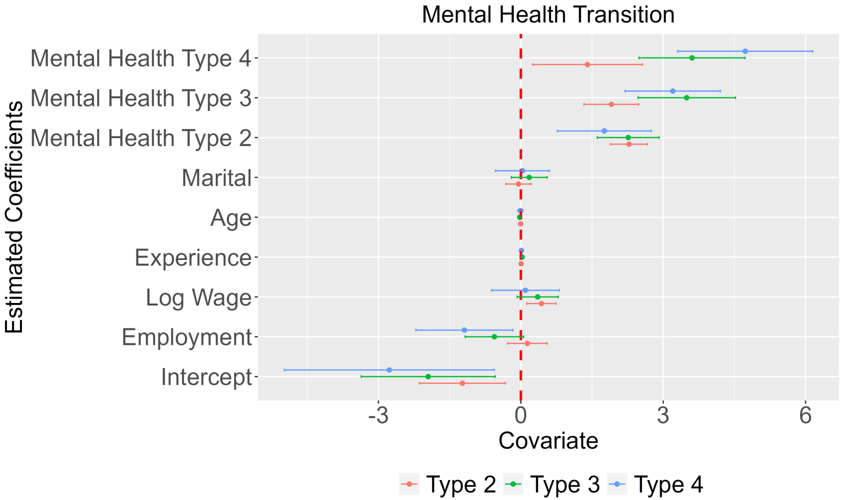
'scghqi' been feeling unhappy or depressed?

$$\begin{bmatrix} 0.94 & 0.04 & 0.01 & 0.00 \\ 0.35 & 0.63 & 0.02 & 0.00 \\ 0.14 & 0.38 & 0.45 & 0.03 \\ 0.11 & 0.07 & 0.35 & 0.46 \end{bmatrix}$$

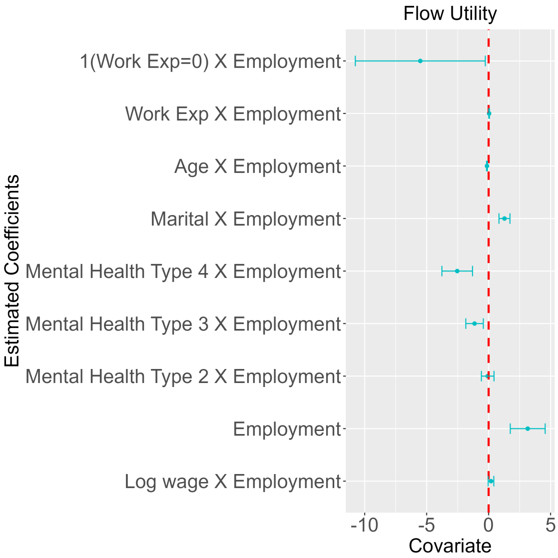
Estimates : Mental Health Types



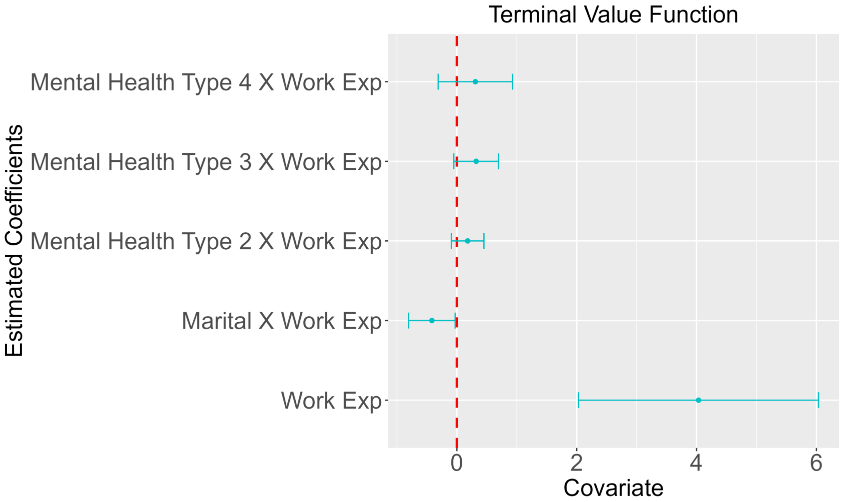
Estimates : Mental Health Transition



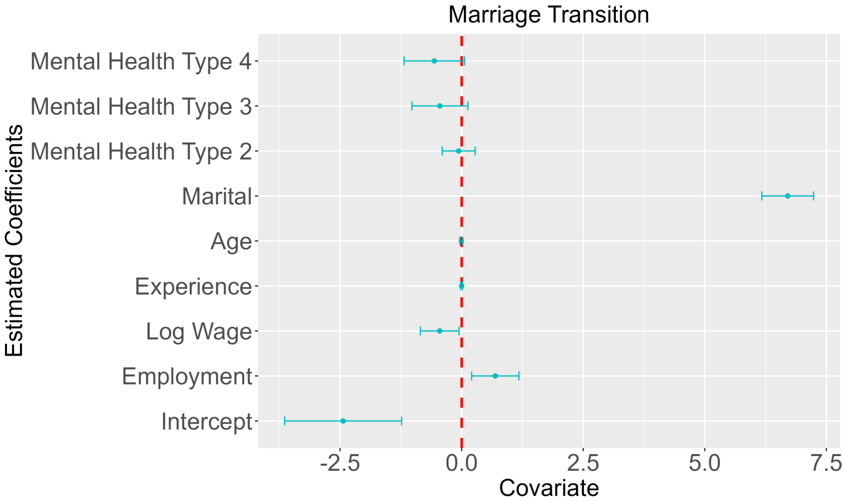
Estimates : Flow Utility



Estimates : Terminal Value Function

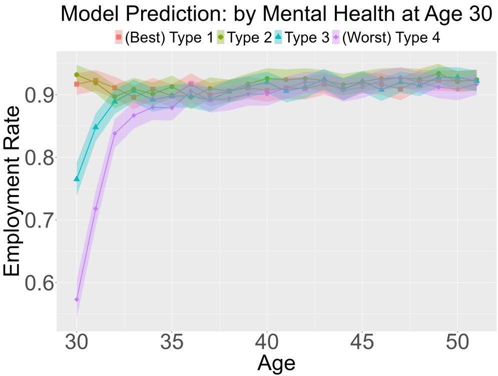


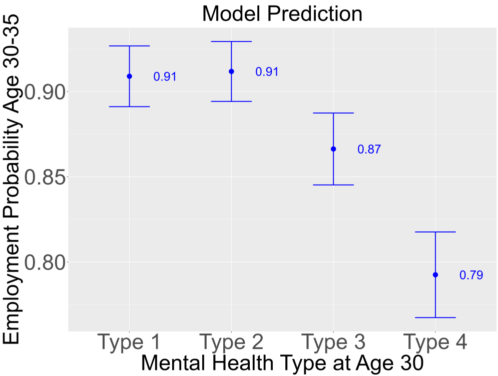
Estimates : Marriage Transition



Estimates : Wage Transition

Model Prediction: Employment Probability by Mental Health at Age 30





Band mental health (Type 4) in age 30 $\rightarrow$ 19,360 GBP $\downarrow$ lifetime income ($\approx$ 71% of avg. annual labor income)

Conclusion

This paper proposed a CCP + Factor model approach for DDCM

- Proxies (ideally, three in one wave and another in consecutive periods) give substantial advantage in the identification

- This paper extended Arcidiacono and Miller (2011)'s CCP estimator to pool information from Proxies and validated its performance through Monte Carlo simulations & demonstrated it using an empirical application of a model of labor supply and mental health