Artificial Self-insurance for Heterogeneous Households

Using GQ-learning

Noah Williams, Ivy X. Yang

Department of Economics
University of Miami

May 16, 2025

UNIVERSITY OF MIAMI
MIAMI HERBERT
BUSINESS SCHOOL

Contribution

Merge causal inference methods with reinforcement learning

This method could

- Relax assumptions on laws of motion while keeping casual interpretation.
- Use causal structural estimates to find decision rules.
- Use decision rules to simulate counterfactuals for policy analysis.
- Use sparse-on-time individual panel data with higher frequency and high dimension macro time series.

References

- HANK model: compute policy functions at a steady state to respond to shocks: Melcangi and Sterk (2024), Acharya et al. (2023), Luetticke (2021), Kaplan et al. (2018).
- Reinforcement learning applications: Cong et al. (2022), Rao and Jelvis (2023), or simply a policy function generator: Fatih and Paolo (2022), Moll (2024).
- Reduce the high-dimensional macro state space: Bayer et al.,(2024),Fernandez-Villaverde et al. (2021), Han et al. (2021), and Payne et al., 2024.

Based on the idea of

1. **Reinforcement learning:** Sutton and Barto (2013)
2. **G-estimation:** Robins et al. (1992a), Lewis and Syrgkanis (2021), Yang (2025).
3. **Adaptive learning:** Williams (2024), Evans and Honkapohja (2001), and Marcet and Sargent (1989)

Basic Algorithm

The basic algorithm includes:

- G-estimation, a method that predicts future micro-level states using past choices and current macro state.
- Q-learning, a reinforcement learning algorithm that derives policy functions through reward optimization.

GQ-learning combines the two, which is an algorithm for reinforcement learning.

G-estimation

Aims at estimating the non-linear law of motion $f(\cdot)$:

$$S_{i,t+1} = f(a_{i,t}, S_{i,t}, L_t | x_i) \tag{1}$$

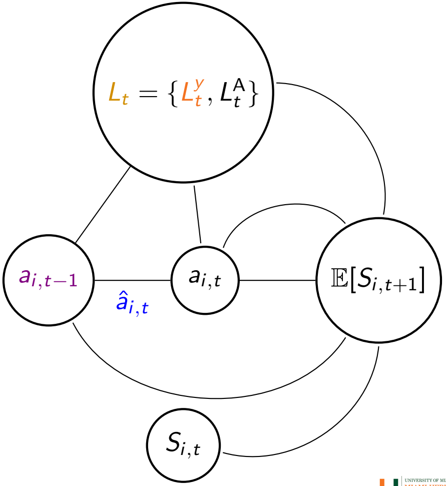
- $S_{i,t+1} = \{w_{i,t+1}, y_{i,t+1}\}$, micro-level state variables include wealth and income.
  - State variables.
- $a_{i,t} = \{\pi_{i,t}, c_{i,t}\}$, choice variables (action space) include portfolio choice $\pi_{i,t}$ and the consumption-to-wealth ratio $c_{i,t}$.
  - Action space.
- $L_t = \{L_t^y, L_t^A\}$, macro-level state includes the source of shock we are interested in L_t^y, and a component L_t^A.
  - Environment.
- x_i time invariant household specific variables.

Reference:Robins et al. 1992a,Robins, 1986.

G-estimation

G-estimation Stage 1:

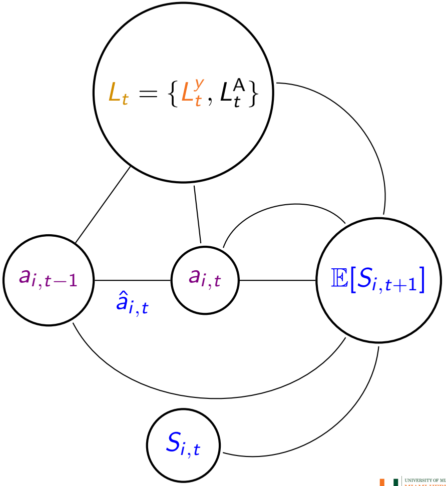
$$\mathbb{E}[\hat{a}_{i,t} | a_{i,t-1}, x_i, L_t^y, L_t^A)] \\ = \eta_0 + \eta_1 a_{i,t-1} + \eta_2 x_i + \eta_3 L_t^y + \eta_4 L_t^A$$



G-estimation

G-estimation Stage 2: Structural estimation for the future state

$$
\begin{aligned}
\mathbb{E}[S_{i,t+1} | a_{i,t}, a_{i,t-1}, x_i, S_{i,t}, L_t^y, L_t^A] \\
= & \beta_0 + \underbrace{\beta_1 a_{i,t-1} + \beta_2 S_{i,t} + \beta_3 x_i + \beta_4 [L_t^y + L_t^A]}_{\text{immediate effect}} \\
& + \beta_5 \underbrace{(1 + L_t^y + L_t^A)\hat{a}_{i,t}}_{\text{adjustment term of expectation}} \\
& + \psi \underbrace{(1 + L_t^y + L_t^A)a_{i,t}}_{\text{treatment effect}}
\end{aligned}
$$



Baseline model to Q-learning

The law of motion:

$$w_{i,t+1}(L_t) = \underbrace{[(1 + r^f) + \pi_{i,t}(r^s_{i,t}(L_t) - r^f)]}_{1+r_{i,t}}(w_{i,t}(L_t) - C_{i,t}) + y_{i,t}(L_t) \tag{2}$$

UNIVERSITY OF MIAMI
MIAMI HERBERT
BUSINESS SCHOOL

Baseline model to Q-learning

The law of motion:

$$w_{i,t+1}(L_t) = [\underbrace{(1+r^f) + \pi_{i,t}(r^s_{i,t}(L_t) - r^f)}_{1+r_{i,t}}](w_{i,t}(L_t) - C_{i,t}) + y_{i,t}(L_t) \tag{2}$$

Baseline model

$$V(S_{i,t}, L_t) = \max_{a_{i,t}}\{u(C_{i,t}) + \beta\mathbb{E}_t[V(S_{i,t+1}, L_{t+1})]\} \tag{3}$$

Baseline model to Q-learning

The law of motion:

$$w_{i,t+1}(L_t) = \underbrace{[(1+r^f) + \pi_{i,t}(r^s_{i,t}(L_t) - r^f)]}_{1+r_{i,t}}(w_{i,t}(L_t) - C_{i,t}) + y_{i,t}(L_t) \tag{2}$$

Baseline model

$$V(S_{i,t}, L_t) = \max_{a_{i,t}}\{u(C_{i,t}) + \beta\mathbb{E}_t[V(S_{i,t+1}, L_{t+1})]\} \tag{3}$$

The estimated law of motion:

$$\begin{aligned} S_{i,t+1} &= f(a_{i,t}, S_{i,t}, L_t|x_i) \\ L_{t+1} &= g(L_t, \epsilon_{t+1}) \end{aligned}$$

Given an arbitrary action $\bar{a}$, the Q-learning process is:

$$Q(\bar{a}, S_{i,t}, L_t) = u(\bar{c}) + \beta\mathbb{E}_t \max_a Q[a, f(\bar{a}, S_{i,t}, L_t|x_i), L_{t+1}]$$

Algorithm

A toy model - linear G-estimation with the temporal difference (TD) method of Q-learning

Set an initial Q-value for each agents $Q_{i,0}$

$\downarrow$

Algorithm

A toy model - linear G-estimation with the temporal difference (TD) method of Q-learning

Set an initial Q-value for each agents $Q_{i,0}$

$\downarrow$

Agents (households) choose a pair of choice variables from the action space $a_{i,t} = \{c_j, \pi_k\}$.

$\downarrow$

Algorithm

A toy model - linear G-estimation with the temporal difference (TD) method of Q-learning

Set an initial Q-value for each agents $Q_{i,0}$

$\downarrow$

Agents (households) choose a pair of choice variables from the action space $a_{i,t} = \{c_j, \pi_k\}$.

$\downarrow$

With the current $a_{i,t}$, we could predict $w_{i,t+1}$ with the G-estimation, and update the Q-function:

$$\max_{c,\pi} Q_t(c, \pi, w_{t+1}(c_j, \pi_k)) = \max_{j,k} Q_t(j, k, n)$$

$\downarrow$

Algorithm

A toy model - linear G-estimation with the temporal difference (TD) method of Q-learning

Set an initial Q-value for each agents $Q_{i,0}$

$\downarrow$

Agents (households) choose a pair of choice variables from the action space $a_{i,t} = \{c_j, \pi_k\}$.

$\downarrow$

With the current $a_{i,t}$, we could predict $w_{i,t+1}$ with the G-estimation, and update the Q-function:

$$\max_{c,\pi} Q_t(c, \pi, w_{t+1}(c_j, \pi_k)) = \max_{j,k} Q_t(j, k, n)$$

$\downarrow$

Solve this to get the counterfactual optimal action $a^*_{i,t} = \{c^*_i, \pi^*_k\}$

Algorithm

The iteration method used here is the Temporal-Difference (TD). The Q-function is updating on a 3-D grid by ϵ-greedy methods

- **if exploitation** - choose actions according to the maximization.

$$Q_{t+1}\left(c_t,\pi_t,w_t\right)=Q_t\left(c_t,\pi_t,w_t\right)+\gamma*\left[u\left(c_t,w_t\right)+\beta Q_t\left(c^*,\pi^*,w_{t+1}\right)-Q_t\left(c_t,\pi_t,w_t\right)\right]$$

- **if exploration** - choose actions randomly from the grid.

$$Q_{t+1}\left(c_t,\pi_t,w_t\right)=Q_t\left(c_t,\pi_t,w_t\right)+\gamma*\left[u\left(c_t,w_t\right)+\beta Q_t\left(c_j,\pi_k,w_{t+1}\right)-Q_t\left(c_t,\pi_t,w_t\right)\right]$$

Stop when $\left|\left(Q_{n,t+1}-Q_{n,t}\right)/Q_{n,t}\right|<$ threshold - converges.

GQ-algorithm application to the real-world data

How do households make portfolio choices facing income and macro shocks?

Data:

Micro-level (PSID) 2009-2021 $S_{i,t}$, $a_{i,t}$, x_i

- Wealth and labor income $S_{i,t} = \{w_{i,t}, y_{i,t}\}$
- Stock value and consumption/wealth ratio $a_{i,t} = \{\pi_{i,t}, c_{i,t}\}$
- Baseline feature x_i

Macro-level (FRED) $L_t = \{L_t^y, L_t^s, r_t^f, L_t^A\}$

- High-dimensional macroeconomic variables L_t^A.
- Deflated S&P500 index L_t^s
- 3-Month Treasury Bill Secondary Market Rate r_t^f
- Aggregate income L_t^y
  "Real Rersonal Income Excluding Current Transfer Receipt".

G-estimation

Table: The second stage regression of G-estimation with PSID (2009-2021)

	$w_{i,t+1}$	$y_{i,t+1}$
$\pi_{i,t-1}$	$-4.252e+06$	$1.416e+04$
Last portfolio choice	$(4.04e+06)$	$(9.3e+04)$
$w_{i,t}$ or $y_{i,t}$	-0.10	0.69^{***}
Current states	(0.19)	(0.16)
x_i	71.78	-1.48
Initial inheritance	(82.37)	(1.85)

w is not well fitted: $R^2(w) = 0.711$ and adjusted $R^2(w) = 0.421$

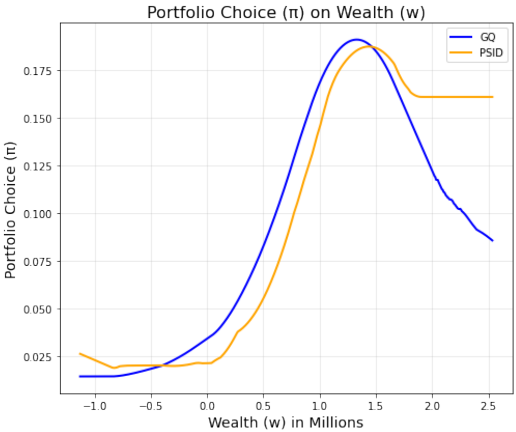
y is well fitted: $R^2(y) = 0.988$ and adjusted $R^2(y) = 0.976$

	$w_{i,t+1}$	$y_{i,t+1}$
L_t^y	-854.48	-10.38
Real personal income	(701.67)	(16.15)
L_t^s	-2428.33	-7.01
$S\&P500_t$	(2658.62)	(60.15)
r_t^f	$-3.212e+8^{*}$	$-5.06e+5$
risk free rate	$(1.36e+8)$	$(3.19e+6)$
L_t^{A1}	$6.269e+8^{*}$	$9.89e+5$
component 1	$(2.65e+05)$	$(6.22e+6)$
L_t^{A2}	$3.786e+06$	$2.101e+04$
Component 2	$(3.43e+05)$	$(7.78e+04)$
$(1+L_t)\times\hat{\pi}_{i,t}$	-3070.33	32.35
Effect of expectation	(3369.29)	(75.61)
$(1+L_t)\times\pi_{i,t}$	29.04	-3.94
Joint effect	(116.669)	(2.35)

Households' wealth levels are more affected by the macroeconomic factors, while labor income is more persistent.

Policy functions by GQ-learning compared with the real data

Reward function: $U(C) = \frac{C^{1-\gamma}}{1-\gamma}$, $\gamma = 0.85$, discount factor $\beta = 0.95$.



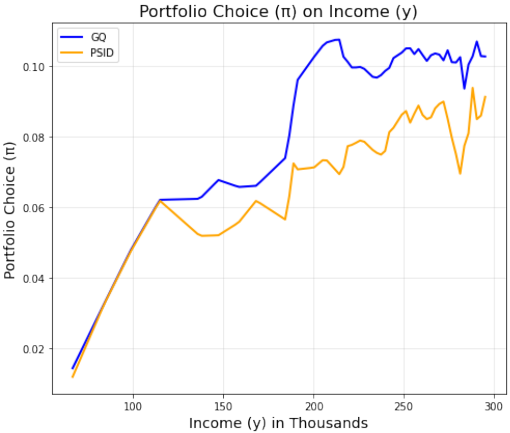


Figure: Recovery Policy Functions for Portfolio Choice.

UNIVERSITY OF MIAMI
MIAMI HERBERT
BUSINESS SCHOOL

GQ-learning results: Consumption ratio $c_{i,t}$

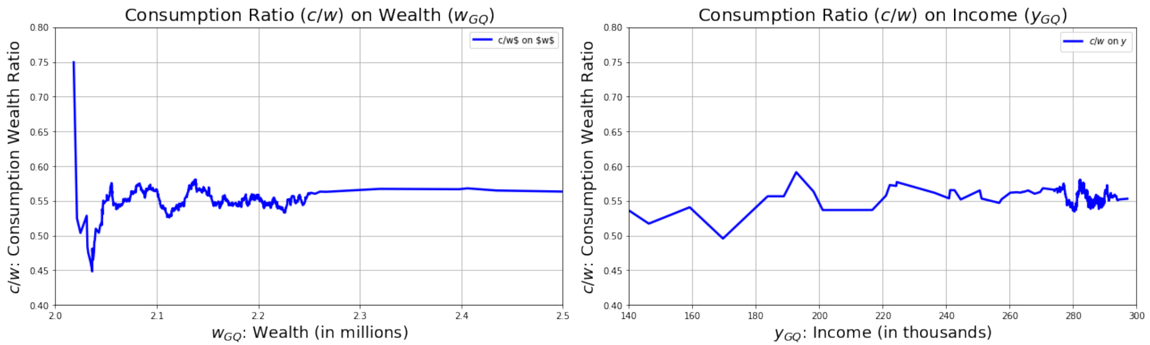
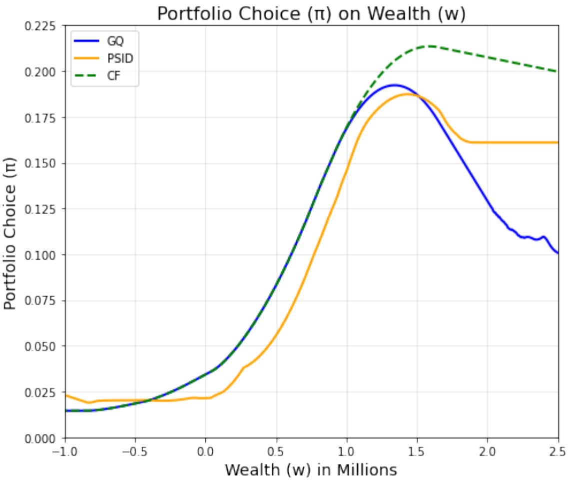


Figure: Recovery Policy Functions for Consumption Ratio.

This is not read from the data

π of a counterfactual low income level



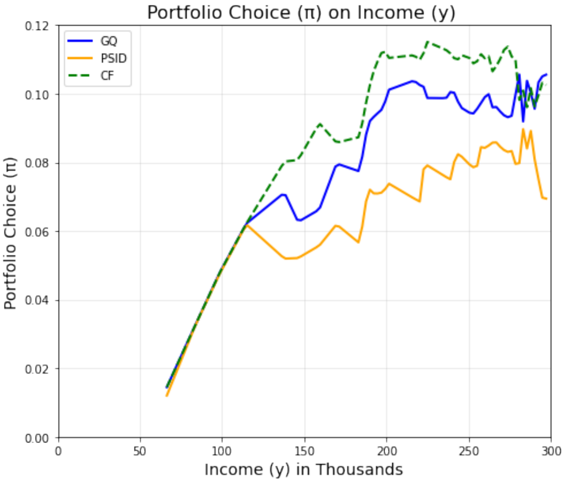
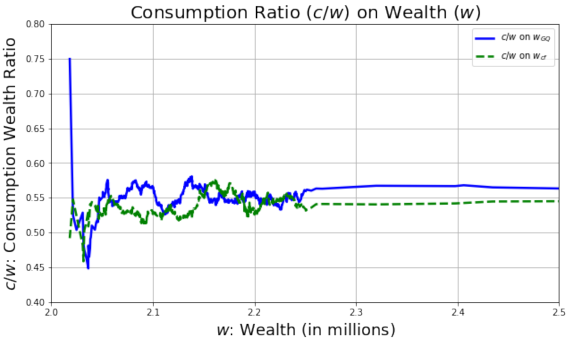


Figure: Aggregate income level is 8 trillion (16.2 trillion in Jan.2024)

c/w of a counterfactual low income level



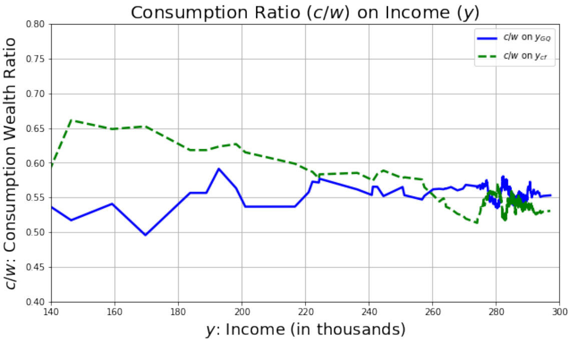


Figure: Aggregate income level is 8 trillion (16.2 trillion in Jan.2024)

π of a counterfactual decrease real risk-free rate

Figure: If cut r_f by 1% (the mean of r_f is 1.18%)

c/w of a counterfactual decrease real risk-free rate

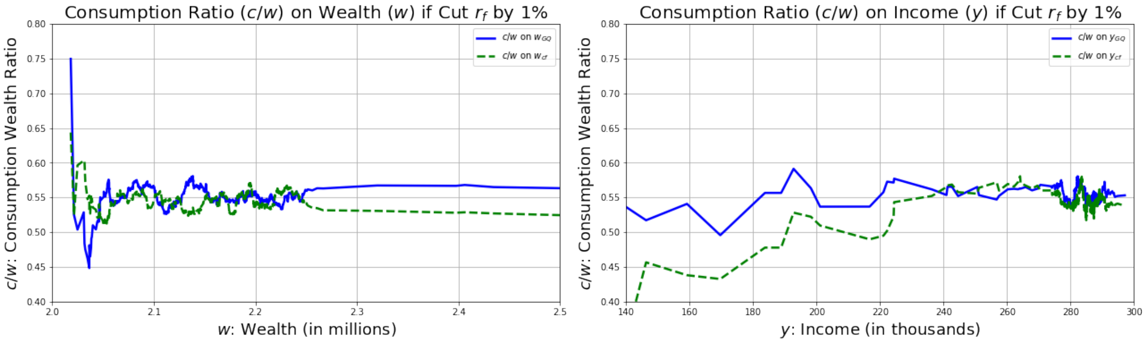


Figure: If cut r_f by 1% (the mean of r_f is 1.18%)

π of a counterfactual increase real risk-free rate

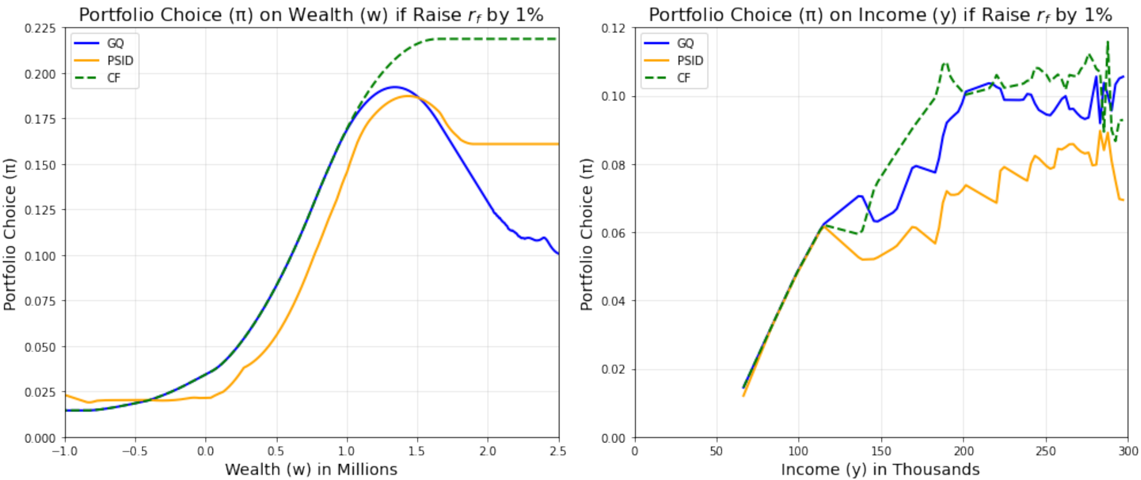
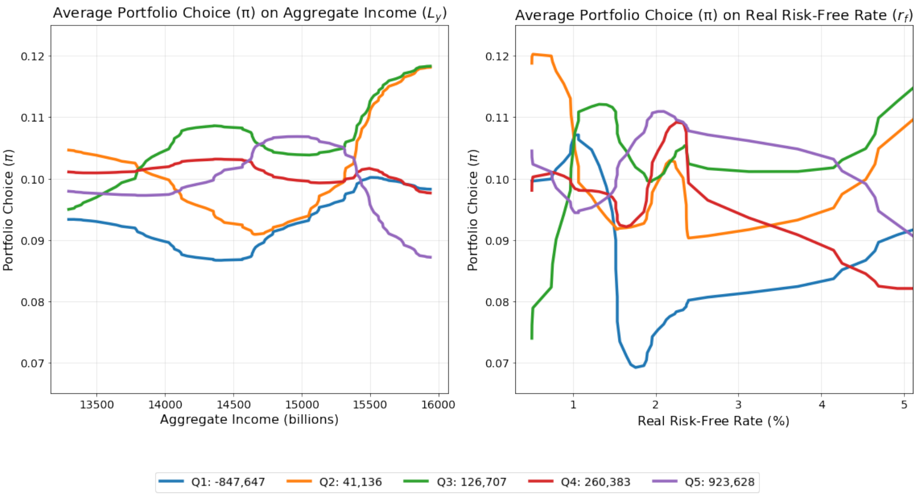


Figure: If raise r_f by 1% (the mean of r_f is 1.18%)

Observe choices at finer granular



Households are in quintile based on their wealth in 2009.

The End