Lecture 4: Simulated Method of Moments and Indirect Inference

Ciprian Domnisoru
Aalto University

Method of Simulated Moments- example

- Consider a dynamic discrete female labor supply model:
  $\max_{\{C_{it}, L_{2it}, A_{it+1}\}_{t=0}^{T}} E_0 \sum_{t=0}^{T} \beta^t U(C_{it}, L_{2it}; Z_{it})$ subject to a lifetime budget constraint, an income process for men and one for women.
- Suppose we are interested in estimating women's labor supply preferences.
- For example, we may want to understand the labor supply implications of providing public free child care.
- GMM may not be possible because we may not be able to derive closed form orthogonality conditions $E[h(w_i, \theta_0)] = 0$
- Instead, use the conditional mean, variance, and quantiles of female labor supply in the data to attempt to recover the parameters of interest.
- We will use the model to generate such moments given draws for men's incomes, women's wages, and family demographics, and match those moments to the empirical ones.

Simulated method of moments

1. select initial parameter vector (female labor supply preferences)
2. solve model on the support of income, wages, demographics
3. draw from distributions of income, wage, demographics
4. simulate model given draws
5. estimate artificial moments of interest
6. match artificial moments to empirical ones
7. repeat until match of moments is acceptable (some difference criterion between empirical and artificial moments stops the execution of the code)

- random draws of income, wage, demographics must be S multiples of N
- the seed of the random number generator must be constant over iterations: ensures that artificial moments change over iterations only due to θ
- Efficiency loss for small S.

Background

- Paper: Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments by Eisenhauer et al. (2015).
- Simulated method of moments (SMM) estimates parameters by fitting a vector of empirical moments to their theoretical counterparts.
- The performance is compared to standard maximum likelihood (ML) estimation
- SMM has been used to estimate
  - Models of job search (Flinn and Mabli, 2008)
  - Educational and occupational choices (Adda et al., 2011, 2013)
  - Household choices (Flinn and Del Boca, 2012)
  - Stochastic volatility models (Andersen et al., 2002; Raknerud and Skare, 2012)
  - Dynamic stochastic general equilibrium models (Ruge-Murcia, 2012)

Alternatives to ML

"Alternative estimation methods have been proposed to overcome the rigidities and complexities of ML estimation. Most require the analyst to characterize the likelihood function but simplify its computation. One of the most popular methods, simulated ML (SML), substitutes the exact likelihood function with a simulated one. An example is the Hajivassiliou-Geweke-Keane (HGK-SML) estimator (Geweke, 1989; Hajivassiliou and McFadden, 1998; Keane, 1994) used for multinomial probit estimation. Approximations of the dynamic programming problem have often been combined with SML in models with a large state space (Keane and Wolpin, 1994). Another popular method is the conditional choice probabilities (CCP) algorithm first proposed by Hotz and Miller (1993) and recently extended to allow for unobserved heterogeneity (Arcidiacono andMiller, 2011). Using CCP, a consistent estimator of the model parameters can be derived without the need of the full solution of the dynamic programming problem. The CCP method, however, restricts the flexibility of the estimable models by imposing assumptions that limit the expectation formation of agents and restrict the stationarity of the environment. SMM is a more general alternative to ML estimation."

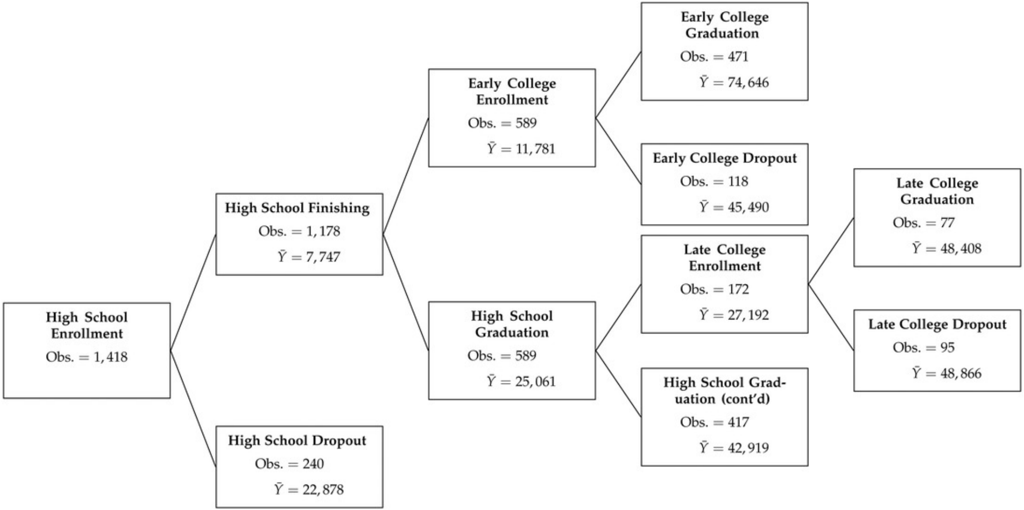
The set up

- Simplified dynamic discrete choice model of schooling based on a sample of white males from the National Longitudinal Survey of Youth of 1979 (NLSY79) using ML
- Model is more restrictive compared to standard dynamic discrete choice models (Keane and Wolpin, 1997, 2001)
  - Limited number of choices and the timing of decisions and outcomes
  - Agents are restricted to binary choices
  - Allows evaluating likelihood analytically without the need for simulation or interpolation
- The results of SMM are compared to standard maximum likelihood (ML) estimation using Monte Carlo simulations

Using SMM in different environments

- SMM can be used for any model, regardless of how complex or difficult to compute the likelihood may be, as long as it is possible to simulate it
- SMM is consistent and asymptotically normal under the conditions presented in the literature (Gourieroux and Monfort, 1997).
- There are numerous choices to be made
  1. Moments used in the estimation
  2. The number of replications used to compute the simulated moments
  3. The moment weighting matrix
  4. The algorithm used for optimization
- The effect of these choices on the performance of the estimator are unclear. This paper proposes diagnostic tests.

Decision tree of the student



Dynamic model of educational choices (1/3)

- The agents (students) are expected to move from one schooling state to the next and have two choices at each decision note
- The value of each state is determined by its immediate rewards and costs and by the expected future value of all feasible states made available by a choice
- Agents are restricted to binary choices. This assumption allows the authors to evaluate the likelihood without any simulation, which provides a clean comparison of ML estimation with simulation-based alternatives.
- Ex post, the agent receives per period rewards defined as the difference between per period earnings $Y(s^{'})$, and the costs $C(s^{'}, s)$ associated with moving from state s to state s' (includes monetary and psychic costs). Can only identify the differences in the costs for two alternative states: normalize the cost of one of the exits to zero.

$$R(s^{'}) = Y(s^{'}) - C(s^{'}, s) \tag{1}$$

Dynamic model of educational choices (2/3)

- We collect the subset of states with a costly exit in S^c. Denote $\hat{s}'$ costly exits, and $\tilde{s}'$ zero cost exits.
- We assume earnings and costs are separable functions of observed covariates $X(s)$ and $Q(\hat{s}', s)$ for costs.
- There is a stochastic component $(U_Y(s), U_C(\hat{s}^{'}, s))$ to each of them. Earnings are expressed as

$$Y(s) = \mu_s(X(s)) + U_Y(s) \tag{2}$$

- The costs of going from state s to state $\hat{s}^{'}$ are defined as

$$C(\hat{s}^{'}, s) = K_{\hat{s}^{'}, s}(Q(\hat{s}^{'}, s)) + U_C(\hat{s}^{'}, s) \tag{3}$$

- **Some variables in $Q(\hat{s}^{'}, s)$ and $X(s^{'})$ might be the same. Their distinct elements constitute the exclusion restrictions.**

Dynamic model of educational choices (3/3)

- We assume a factor structure on the unobservables by postulating that a low-dimensional vector of latent factors θ is the sole source of dependency among the unobservables of the model

$$U_Y(s) = \theta^{'}\alpha_s + \varepsilon(s) \qquad U_C(\hat{s}^{'}, s) = \theta^{'}\phi_{\hat{s}^{'},s} + \eta(\hat{s}^{'}, s) \tag{4}$$

- The individual-specific factors θ are known to agents but unknown to the econometrician, whereas the idiosyncratic shocks $\varepsilon(s)$ and $\eta(\hat{s}^{'}, s)$ are unknown to the econometrician (assumed drawn from a normal distribution at each state) and only known by the agents at different stages of the decision process.
- This formulation allows for unobservable correlations in outcomes and choices across states through θ and the factor loadings α_s and $\phi_{\hat{s}^{'},s}$, which vary by states.

Value function and returns to education

- Given the available information set $\mathscr{I}(s)$ (see p.334), the agent's value function at state s is defined as

$$V(s) = Y(s) + \max_{s' \in \Omega(s)} \left\{ \frac{1}{1+r} (-C(\hat{s}', s) + \mathbb{E}[V(\hat{s}') | \mathscr{I}(s)]) \right\} \tag{5}$$

- We define ex ante and ex post net returns to schooling (including per period earnings and costs associated with each educational choice and option value of future opportunities.
- Ex ante net returns are defined before the unobservable components of future earnings are realized. They depend on agents' expectations and determine their choices.
- **Standard methods for computing rates of return such as Mincer coefficients or internal rates of returns ignore costs and option values of future opportunities.** They are only interpretable for terminal choices and ex post realized earnings streams.

Option values of schooling

- A high school enrollee makes the decision whether to graduate or drop out.
- A benefit of graduating is the option to start college and it is defined as the difference between the value of taking the optimal choice when moving from $s^{'}$ and the fallback value of the zero cost exit $\tilde{s}^{''}$

$$OV(s',s) = \frac{1}{1+r}\mathbb{E}\left[\mathbb{E}\left[\max_{S''\in\Omega(s')}\left\{V(s'') - C(s'',s')\right\} \mid \mathscr{I}(s')\right] - \mathbb{E}\left[V(\tilde{s}'') \mid \mathscr{I}(s')\right] \mid \mathscr{I}(s)\right] \tag{6}$$

$$= \frac{1}{1+r}\mathbb{E}\left[\underbrace{\max_{S''\in\Omega(s')}\left\{V(s'') - C(s'',s')\right\}}_{\text{value of options arising from s'}} - \underbrace{V(\tilde{s}'')}_{\text{fallback value}} \mid \mathscr{I}(s)\right]$$

- We define the option value contribution $OVC(s^{'},s) = \frac{OV(s^{'},s)}{\mathbb{E}[V(s^{'})|I(s)]}$ as the relative share of the option value in the overall value of a state. This component is not reported in standard calculations of the Mincer return or the internal rate of return

Return to education

- Ex ante and ex post returns do not necessarily agree
  - Agents cannot predict their future earnings.
  - Decisions that are optimal for an agent ex ante might be suboptimal ex post.
  - To account for this, the percentage of agents experiencing regret is calculated, that is, those agents for whom the ex post and ex ante returns do not agree in sign.
- The option value makes a sizable contribution to the overall value of the state
  - Early college enrollment has the highest option value as graduation yields a large gain in early earnings
- Investigation of the impact of 50% reduction in tuition costs on agents' college-going decision
  - Educational choices of 50,000 simulated students compared between the baseline regime and the policy alternative.
  - Reducing tuition increased high school attendance rates by one percentage point due to increased option value.
  - About half of the new early enrollees will graduate while only a quarter of the late enrolleess will do so.

Model fit

- "Comparing the fit of the model with cross-section moments is a weak criterion for a dynamic model. A more exacting criterion is to predict sequences of educational choices."
- χ^2 (P-value $\leq \alpha$: The observed data are statistically different from the expected values (Reject H0)). p-value of a joint test of the relative share of agents for each state for all realizations of selected covariates.

TABLE 2
CONDITIONAL MODEL FIT (P VALUES)

State	Number of Children	Baby in Household	Parental Education	Broken Home
High school dropout	0.77	0.26	0.37	0.03
High school finishing	0.88	0.73	0.55	0.35
High school graduation	0.91	0.94	0.65	0.91
High school graduation (cont'd)	0.95	0.33	0.40	0.85
Early college enrollment	0.46	0.54	0.01	0.15
Early college graduation	0.06	0.86	0.00	0.14
Early college dropout	0.33	0.27	0.54	0.75
Late college enrollment	0.80	0.23	0.90	0.60
Late college graduation	0.90	0.39	0.90	0.60
Late college dropout	0.89	0.42	0.91	0.76

Comparison of ML and SMM (1/2)

- Synthetic sample of 5,000 agents is simulated using the baseline estimates of the structural parameters.
- “We disregard our knowledge about the true structural parameters and estimate the model on the synthetic sample by ML and SMM to compare their performance in recovering the true structural objects."
- Within-sample model fit is compared, as well as the accuracy of the estimated returns to education and policy predictions.
- The measurement equation is a function describing the measurement process in terms of all quantities that can contribute a significant uncertainty to the measurement result.

$$M(j) = X(j)^{'}\kappa_j + \theta^{'}\gamma_j + \nu(j) \qquad \forall j \in M \tag{7}$$

$$Y(s) = X(s)^{'}\beta_s + \theta^{'}\alpha_s + \varepsilon(s) \qquad \forall s \in S \tag{8}$$

$$C(\hat{s}^{'}, s) = Q(\hat{s}^{'}, s)^{'}\delta_{\hat{s}^{'},s} + \theta^{'}\phi_{\hat{s}^{'},s} + \eta(\hat{s}^{'}, s) \quad \forall s \in S^{c} \tag{9}$$

Comparison of ML and SMM (2/2)

- All unobservables of the model are normally distributed:

$$\eta(\hat{s}^{'},s) \sim N(0,\sigma_{\eta(\hat{s}^{'},s)})\forall s \in S^{c} \qquad \varepsilon(s) \sim N(0,\sigma_{\varepsilon(s)})\forall s \in S \tag{10}$$

$$\theta \sim N(0,\sigma_{\theta})\forall \theta \in \Theta \qquad \nu(j) \sim N(0,\sigma_{\nu(s)})\forall j \in \mathcal{M} \tag{11}$$

- The unobservables $(\varepsilon(s),\eta(\hat{s}^{'},s),\nu(j))$ are independent across states and measures. The two factors θ are independently distributed
  - This still allows for unobservable correlations in outcomes and choices through the factor components θ

The ML approach

- For each agent, we define an indicator function $G(s)$ that takes value 1 if the agent visits state s.
- Let $\psi \in \Psi$ denote a vector of structural parameters and Γ the subset of the states visited by the agent i. We collect $D = \{\{X(j)\}_{j \in M}, \{X(s), Q(\hat{s}', s)\}_{s \in S}\}$ all observed agent characteristics.
- The likelihood for observation i is given by

$$\int_{\underline{\Theta}} \left[\prod_{j \in M} \underbrace{f(M(j)|D, \theta; \psi)}_{\text{Measurement}} \prod_{s \in S} \left\{ \underbrace{f(Y(s)|D, \theta; \psi)}_{\text{Outcome}} \underbrace{Pr(G(s) = 1|D, \theta; \psi)}_{\text{Transition}} \right\}^{\mathbb{1}\{s \in \Gamma\}} \right] dF(\theta) \tag{12}$$

  Where Θ is the support of θ

- After taking the logarithm of the equation above and summing across all agents, we obtain the sample log likelihood. We use **Gaussian quadrature** to evaluate the integrals of the model. We maximize the sample log likelihood using the Broyden-Fletcher-Goldfarb-Shanno **(BFGS) algorithm.**

The SMM approach (1/2)

- The goal in SMM approach is to choose a set of structural parameters ϕ to minimize the weighted distance between selected moments from the observed sample and a sample from a structural model
- The criterion function takes the following form

$$\Lambda(\psi) = [\check{f} - \hat{f}(\psi)]' W^{-1} [\check{f} - \hat{f}(\psi)] \tag{13}$$

- $\check{f}$ represents a vector of moments computed on the observed data, $\hat{f}(\psi)$ denotes an average vector of moments calculated from R simulated data sets, and W is positive definite weighting matrix
- We define $\hat{f}(\psi)$ as

$$\hat{f}(\psi) = \frac{1}{R} \sum_{r=1}^{R} \hat{f}_r(u_r; \psi) \tag{14}$$

The SMM approach (2/2)

- The simulation of the model involves the repeated sampling of the components $u_r = \{\{\varepsilon(s), \eta(\hat{s}', s)\}_{s \in S}\}$ determining the agents's outcomes and choices.
- The simulation is repeated R times for fixed ψ to obtain an average vector of moments and the model is solved using backward induction to compute each single set of moments.
- The random components u_r are drawn at the beginning of the estimation procedure and remain fixed throughout
  - Avoids chatter in the simulation for alternative ψ, where changes in the criterion function could be due to either ψ or u_r (McFadden, 1989).
- To implement the criterion function, it is necessary to choose a set of moments, the number of replications, a weighting matrix, and an optimization algorithm.
- The model does not aim to minimize the score of an auxiliary model but a quadratic form in the difference between the moments on the simulated and observed data.

Choice of moments: Heckman at al. (2014b)

- The authors develop a sequential schooling model that is a halfway house between a reduced form treatment effect model and a fully formulated dynamic discrete choice model such as ours.
- They approximate the underlying dynamics of the agents' schooling decisions by including observable determinants of future benefits and costs as regressors in the current choice.
- Moments:
  - dynamic versions of Linear Probability (LP) models for each transition
  - mean and standard deviation of within-state earnings
  - the parameters of ordinary least squares (OLS) regressions of earnings on covariates to capture the within-state benefits to educational choices
  - state frequencies
- 440 moments to estimate 138 free structural parameters.

Results

- We set the number of replications R to 30 and thus simulate a total of 150,000 educational careers for each evaluation of the criterion function
- The weighting matrix W is a matrix with the variances of the moments on the diagonal and zero otherwise. We determine the latter by resampling the observed data 200 times.
- Due to the choice of weighting matrix, the equation can be written as

$$\Lambda(\psi) = \sum_{i=1}^{I} \left(\frac{\check{f}_i - f_i(\hat{\psi})}{\hat{\sigma}_i}\right)^2 \tag{15}$$

  Where I is the total number of moments, f_i denotes moment i, and $\hat{\sigma}_i$ its bootstrapped standard deviation. Note the fact that the criterion function has the form of a standard nonlinear least-squares problem.
- "Our criterion is not a smooth function of the model parameters. Small changes in the structural parameters cause some simulated agents to change their educational choices, resulting in discrete jumps in our set of moments (Smith and Keane, 2004). Thus we cannot use gradient based methods for optimization and rely on derivative-free alternatives instead."

Comparison

- ML and SMM estimations are compared to learn whether the paper's version of SMM is a good substitute for ML
  1. First, comparison of basic model fit statistics
  2. Second, studying the estimates for the returns to education and performing a counterfactual policy exercise
  3. Finally, exploring alternative choices for
     - the set of moments
     - weighting matrix
     - number of replications
     - optimization algorithms.

Comparison

- The state frequencies are matched very well in both cases. Goodness of fit test statistics reported.
- Similar approximations of gross returns, but SMM overestimates net returns: "In contrast to the gross returns, these returns include the current costs and the systematic part of all future costs of educational choices. SMM is unable to detect the systematic differences in the cost faced by agents." Despite encouraging values for model fit criteria, SMM fails to accurately estimate the net return to educational choices.
- "Based on the ML results, all policy predictions line up with the underlying truth. This is only partly true for the SMM estimation, where the predicted graduation rate for those induced to enroll in college late is too optimistic. Only a quarter will actually graduate, whereas the SMM results forecast about half. The SMM's failure to distinguish between the systematic and unsystematic cost components driving educational choices translates into (partly) flawed policy conclusions as well."

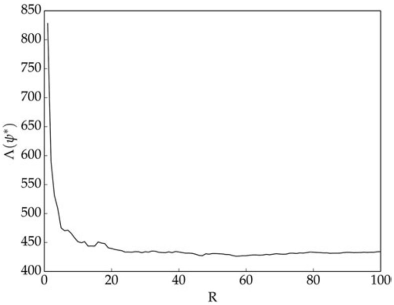
Alternative moments

TABLE 8
SET OF MOMENTS

	Cross-Section Moments	Dynamic (Panel) Moments		
Sets	Base	Base	Alt. A	Alt. B
Outcome models				
Means	√	√	√	√
Standard deviations	√	√	√	√
Ordinary least squares	√	√	√	√
Correlations				√
Choice models				
State frequencies	√	√	√	√
Linear probability				
Cross section	√			
Dynamic		√	√	√
Probit				
Dynamic			√	√
Correlations				√
Overall statistics				
Number of moments	222	440	690	868
Number of replications	50	50	50	50
Weighting matrix	diagonal variance matrix			
Algorithm	POUNDerS			
Quality of fit measures				
$\Lambda(\hat{\psi})$	130.69	383.49	666.57	798.33
$\Lambda(\psi^*)$	222.12	434.07	685.94	847.64

NOTES: Alt. = Alternative.

Number of replications



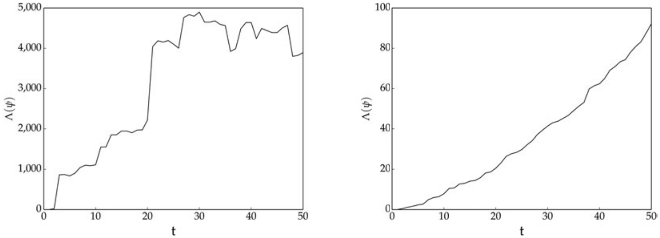
NOTE: Investigation using estimation sample of 5,000 agents with varying number of replications.

FIGURE 8

ROLE OF REPLICATIONS

Alternative weighting matrices

Choosing the identity matrix for W results in multiple local minima, whereas using the variances smoothes the overall surface of the criterion function.



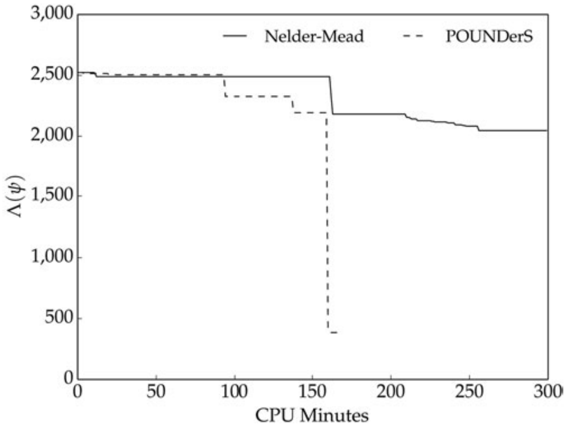
(a) Identity Matrix (b) Inverse Variances on Diagonal

NOTE: Investigation using estimation sample of 5,000 agents with one replication and alternative weighting matrices.

FIGURE 9

ALTERNATIVE WEIGHTING MATRICES

Optimization algorithm



NOTE: Investigation using estimation sample of 5,000 agents with 30 replications and all tuning parameters of the algorithms set to their default values.

FIGURE 10

OPTIMIZATION ALGORITHMS

Conclusions

- We are unable to improve the SMM results by using alternative tuning parameters.
- Our discussion cautions that inspection of model fit statistics alone does not guarantee accurate economic implications. For our model, large unobserved variation in educational choices translates into a noisy criterion function, which leaves SMM unable to recover the true returns to education.
- Our results do not discredit SMM as a useful tool for the estimation of complex economic models. Our results are highly model dependent, but our diagnostics are not. We now outline a Monte Carlo exercise that allows SMM users to build confidence in their particular implementation in any applied setting.
  - fit the observed data with SMM to produce a set of structural parameters.
  - simulate a synthetic sample using these parameters
  - fit the synthetic data with SMM to produce new parameters
  - compare the two sets of parameters
- This exercise showcases the performance of the estimator in a favorable setting as the model is correctly specified. If the structural parameters in the synthetic dataset are successfully recovered, this is encouraging but does not provide a definite proof of the performance in the observed data. A failure, however, offers reason for concern.

Indirect Inference

Simplified indirect inference example

- When models are complex, estimate a set of **auxiliary models** using the empirical data- the auxiliary models could just be linear or non-linear statistical relationships between the observable variables, without any causal or structural interpretation.
- Suppose the dgp is $y_i = exp(\mathbf{x_i}\gamma) + u_i,\ u_i \sim N[0, \sigma^2]$
- Choose an auxiliary model $y_i = x_i\beta + \varepsilon_i, \varepsilon_i \sim N[0, \sigma_\varepsilon^2]$
- The function that connects the parameters of the auxiliary model to those of the dgp is called the binding function,
- Note the **binding function** is $\gamma = E[y|x]^{-1}\beta$
- Given data and $\hat{\beta}$ and a draw $u^{(0)}$, generate $y_i^{(1)}$ using
$$y_i^{(1)} = exp(\mathbf{x_i}\hat{\beta}) + u_i^{(0)}$$
- Obtain a revised estimator $\hat{\beta}^{(1)}$ using OLS and $y_i^{(1)}$, which in turn is used to obtain another set of pseudo-observations. Holding $u^{(0)}$ fixed, repeat until $(\hat{\beta}^{(s)} - \hat{\beta})'\Omega(\hat{\beta}^{(s)} - \hat{\beta})$ approaches some desired lower threshold.
- The resulting estimate of γ is an indirect estimator.

Indirect inference example: Guvenen and Smith (2010)

Rather than selecting a set of *moments* upon which to base our estimation (as is typically done with the method of simulated moments), indirect inference allows us to focus on *a set of dynamic equations*. In this paper, we show through a Monte Carlo study that indirect inference yields excellent results when these dynamic equations are chosen to mimic the structural equations of a *simplified* (or perturbed) version of the full structural model that we wish to estimate. Specifically, if we assume quadratic utility and no borrowing constraints in the model above, we can derive exact structural equations that link consumption choice to innovations in the beliefs (about β^i and z_t^i). In Section 2.2 we use these equations to characterize the kinds of variation in consumption and income data that would be informative about different structural parameters (Propositions 1 and 2). But because this is a perturbed model, we *cannot* estimate these equations directly (e.g., via GMM) to obtain consistent estimates of our structural parameters. These equations *can*, however, form the basis of an "auxiliary model" by means of which indirect inference can deliver consistent estimates. Loosely speaking, the indirect inference estimator is obtained by choosing the values of the structural parameters so that the estimated model and the US data "look as similar as possible" when viewed through the "lens" of the auxiliary model.

Indirect inference example

2.2, beliefs are updated over the life cycle using past income realizations, and prior beliefs are likely to be correlated with future income realizations, which leads us to the following feasible regression:[27]

$$c_t = \mathbf{a}'\mathbf{X}_{c,t} + \epsilon_t^c = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t+1} + a_4 y_{t+2} + a_5 \overline{y}_{1,t-3}$$
$$a_6 \overline{y}_{t+3,R} + a_7 \Delta y_{1,t-3} + a_8 \Delta y_{t+3,R} + a_9 c_{t-1} + a_{10} c_{t-2} + a_{11} c_{t+1} + a_{12} c_{t+2} + \epsilon_t^c, \qquad (22)$$

where c_t and y_t are the log of consumption and income, respectively; $\Delta y_{\tau_1,\tau_2}$ and $\overline{y}_{\tau_1,\tau_2}$ are, respectively, the average of the growth rate and the average of the level of log income from time τ_1 to τ_2; and $\mathbf{a}$ and $\mathbf{X}_{c,t}$ denote the vectors of coefficients and regressors. The use of logged variables in this regression seems natural given that the utility function is CRRA and income is log-normal. This regression captures the predictions made by the HIP and RIP models discussed above, by adding past and future income growth rates as well as past and future income levels. Leads and lags of consumption are also added to capture the dynamics of consumption around the current date.

To complete the auxiliary model, we add a second equation with y_t as the dependent variable, and use all the income regressors above as left-hand-side variables:

$$y_t = \mathbf{b}'\mathbf{X}_{y,t} + \epsilon_t^y = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + b_3 y_{t+1} + b_4 y_{t+2} + b_5 \overline{y}_{1,t-3} + b_6 \overline{y}_{t+3,R}$$