Estimating Tax Complexity Over Time: A Maximum Likelihood Approach

John T.H. Wong

Agenda

1. Structural definition of complexity
2. Parameters that characterize the tax system
3. Derivation of CDF
4. Identification
5. Likelihood function
6. Final output

What do we mean by complexity?

- We want to start with a baseline degree of simplicity. Suppose every person with the same income pays the same amount of tax.
- Let's just consider two households for now, the i-th and j-th households, where $j \neq i$:

$$\frac{T_i}{y_i} = \frac{T_j}{y_j},$$

where

- $y_i =$ pre-tax income.
- $T_i =$ tax liability.

What do we mean by complexity?

Now suppose we deviate from this system, such that household i pays more than j as a share of pre-tax income:

$$\frac{T_i}{y_i} > \frac{T_j}{y_j}.$$

This is equivalent to saying that household i has less post-tax income than j, as a share of pre-tax income:

$$\frac{y_i - T_i}{y_i} < \frac{y_j - T_j}{y_j}.$$

The more complex a tax system is, the more likely that this inequality is true. Thus, we want to we specify:

$$\Pr\{\frac{y_i - T_i}{y_i} < \frac{y_j - T_j}{y_j}\}.$$

What parameters determine post-tax income?

We want to know $\Pr\{\frac{y_i - T_i}{y_i} < \frac{y_j - T_j}{y_j}\}$.

You might wonder: what if the deviation is caused by tax progressivity? Is that "complexity"?

A post-tax income function

Assume that post-tax income is determined by Heathcote, Storesletten, and Violante's (2017) transfer function:[1]

$$y_i - T_i = \lambda y_i^{1-\tau},$$

where

- λ is a parameter for the tax system's flatness.
- τ is a parameter for progressivity.

[1] Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. "Optimal Tax Progressivity: An Analytical Framework." Q. J. Econ. 132, no. 4 (November 2017): 1693–1754.

What parameters determine post-tax income?

$$y_i - T_i = \lambda y_i^{1-\tau}$$

Sidenote: Some Interpretations

The two parameters need to be jointly interpreted.
Suppose $\tau = 1$, then post-tax income is simply:

$$y_i - T_i = \lambda,$$

i.e., that everyone has the same post-tax income λ. Alternative interpretations are that the tax system is fully egalitarian, or that it is purely confiscatory beyond a prescribed income level.

What parameters determine post-tax income?

$$y_i - T_i = \lambda y_i^{1-\tau}$$

Some Interpretations (cont.)

Suppose $\tau = 0$, then post-tax income is:

$$y_i - T_i = \lambda y_i,$$

i.e., that everyone keeps the same proportion of their pre-tax income, independent of their pre-tax income level. Alternative interpretation is that the tax system is flat, with the flat rate of $1 - \lambda$ charged to all taxpayers.

What parameters determine post-tax income?

We want to make modifications

First, the function should be stochastic, not deterministic:

$$
\begin{aligned}
y_i - T_i &= \lambda y_i^{1-\tau} \\
&\to \lambda y_i^{1-\tau} e^{\epsilon_i}
\end{aligned}
$$

where

- $\epsilon_i \sim N(0, \sigma^2)$, which characterizes random error in how post-tax income is determined.

Second, household size should determine post-tax income:

$$
\to \lambda(\frac{A_i^{\theta}}{a_i}) y_i^{1-\tau} e^{\epsilon_i}
$$

where

- a_i denotes the count of working adults in the household.
- A_i denotes the count of all members in the household.
- $\theta \lesseqgtr 1$ implies a household size penalty/benefit.

CDF derivation

Recall we left off with $\Pr\{\frac{y_i - T_i}{y_i} < \frac{y_j - T_j}{y_j}\}$. By the end, we want to derive a cumulative distribution function (CDF).

Plug in the post-tax income function $\lambda(\frac{A_i^\theta}{a_i})y_i^{1-\tau}e^{\epsilon_i}$:

$$\Pr\{\frac{\lambda(\frac{A_i^\theta}{a_i})y_i^{1-\tau}e^{\epsilon_i}}{y_i} < \frac{\lambda(\frac{A_j^\theta}{a_j})y_j^{1-\tau}e^{\epsilon_j}}{y_j}\}.$$

CDF derivation

Take the logarithm on both sides:

$$\Pr\{\ln \lambda + \theta \ln A_i - \ln a_i + (1-\tau) \ln y_i + \epsilon_i - \ln y_i$$
$$< \ln \lambda + \theta \ln A_j - \ln a_j + (1-\tau) \ln y_j + \epsilon_j - \ln y_j\}.$$

λ cancels out. Factorize y_i, y_j:

$$\Pr\{\theta \ln A_i - \ln a_i - \tau \ln y_i + \epsilon_i < \theta \ln A_j - \ln a_j - \tau \ln y_j + \epsilon_j\}.$$

CDF derivation

Rearrange $\epsilon_i - \epsilon_j$ to one side:

$$\Pr\{\epsilon_i - \epsilon_j < \tau(\ln y_i - \ln y_j) + \ln a_i - \ln a_j - \theta \ln A_i + \theta \ln A_j\}$$

$$\implies \Pr\{\epsilon_i - \epsilon_j < \tau \ln(\frac{y_i}{y_j}) + \ln(\frac{a_i}{a_j}) - \theta \ln(\frac{A_i}{A_j})\}$$

Obtain standard-normal CDF:

$$\Pr\{\frac{\epsilon_i - \epsilon_j}{\mathrm{SE}(\epsilon_i - \epsilon_j)} < \frac{\tau \ln(\frac{y_i}{y_j}) + \ln(\frac{a_i}{a_j}) - \theta \ln(\frac{A_i}{A_j})}{\mathrm{SE}(\epsilon_i - \epsilon_j)}\}$$

CDF derivation

What is $\mathrm{SE}(\epsilon_i - \epsilon_j)$?

Recall that $\epsilon_i \sim N(0, \sigma^2)$.
Because ϵ_i and ϵ_j are iid:

$$\mathrm{Var}(\epsilon_i - \epsilon_j) = 2\sigma^2$$

$$\implies \mathrm{SE}(\epsilon_i - \epsilon_j) = \sqrt{2}\sigma.$$

Therefore,

$$\implies \Pr\{\frac{\epsilon_i - \epsilon_j}{\sqrt{2}\sigma} < \frac{\tau \ln(\frac{y_i}{y_j}) + \ln(\frac{a_i}{a_j}) - \theta \ln(\frac{A_i}{A_j})}{\sqrt{2}\sigma}\}$$

$$\equiv \Phi[\frac{\tau \ln(\frac{y_i}{y_j}) + \ln(\frac{a_i}{a_j}) - \theta \ln(\frac{A_i}{A_j})}{\sqrt{2}\sigma}].$$

Identification

- Note that τ, θ, **and** σ are all identified.
- The CDF changes if we multiply all the parameters by δ:

$$\Phi[\frac{\tau \ln(\frac{y_i}{y_j}) + \ln(\frac{a_i}{a_j}) - \theta \ln(\frac{A_i}{A_j})}{\sqrt{2}\sigma}] \neq \Phi[\frac{\delta\tau \ln(\frac{y_i}{y_j}) + \ln(\frac{a_i}{a_j}) - \delta\theta \ln(\frac{A_i}{A_j})}{\delta\sqrt{2}\sigma}].$$

Interpreting the parameters

$$\Phi[\frac{\tau \ln(\frac{y_i}{y_j}) + \ln(\frac{a_i}{a_j}) - \theta \ln(\frac{A_i}{A_j})}{\sqrt{2}\sigma}]$$

- σ captures deviation from flat taxation that are **not** explained by (i) progressivity, which is captured by τ and (ii) household size (θ).
- Essentially, we have allowed progressivity and household size treatment to be policy variables that can dynamically change.

Likelihood function

- We just specified the probability that two households have a different tax rate.
- For the likelihood function, we want to make make one (and only one) pairwise comparison of every household.
- Suppose there are four observations $i = 1, 2, 3, 4$.
- We would compare the 1st household to the 2nd and then the 3rd to the 4th.
  - Or we can compare the 1st to the 3rd, and then the 2nd to the 4th. So forth.
- Therefore, for $n = 4$, there would be $n/2$ independent pairwise comparisons of $\epsilon_i - \epsilon_j$.

Likelihood function

Define the following variables:

$$P_{ij} \equiv \Phi[\frac{\tau \ln(\frac{y_i}{y_j}) + \ln(\frac{a_i}{a_j}) - \theta \ln(\frac{A_i}{A_j})}{\sqrt{2}\sigma}],$$

$$I_{ij} \equiv 1_{\{}\frac{T_i}{y_i} > \frac{T_j}{y_j}\}.$$

Assume households are sorted randomly. The log-likelihood function we want to maximize is then:

$$\ln L = \sum_{i=1}^{\lceil n/2 \rceil - 1} I_{ij} \cdot \ln P_{ij} + (1 - I_{ij}) \cdot \ln(1 - P_{ij}),$$

where $\lceil n/2 \rceil$ rounds the index up to the higher integer if the median index is a decimal (i.e., n is even) and $j = \lceil n/2 \rceil - 1 + i$.

Complexity over time

If we conduct maximum likelihhod estimation separately for each given fiscal year $t = 1, \dots, T$,

we can estimate the following vectors:

$$\vec{\tau} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_T \end{bmatrix} \ \vec{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_T \end{bmatrix} \ \vec{\sigma} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_T \end{bmatrix},$$

the last of which gives us a time series of tax complexity.