Estimating Tax Complexity Over Time

John T.H. Wong

Intuition

- How complex is the tax system?
- We want to start with a baseline degree of simplicity. Let's say we start with a flat-tax system where everyone pays the same average tax rate.
- Let's just consider two households for now, the i-th and j-th households, where $j \neq i$.
- Under a flat tax system:

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

where

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

We can also start from everyone having an equal average tax rate conditional on income, $\frac{T_i/y_i}{y_i} = \frac{T_j/y_j}{y_j}$. Or we can start from an equal lump sum tax, $T_i = T_j$. It does not substantially change our analysis.

Intuition

Now suppose we deviate from the flat-tax 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 determines 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:

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

where

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

What determines 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 determines 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 determines post-tax income?

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

Some Interpretations (cont.)

At $\tau \in (0, 1)$, the tax system is somewhere between the two extremes:

- Progressivity increases with τ.
- λ is a "income-neutral progressivity amplifier."

What determines post-tax income?

We want to make modifications

First, the function should be stochastic, not deterministic:

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

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:

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

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.

Probability of unequal tax rates

Recall we left off with $\Pr\{\frac{y_i - T_i}{y_i} < \frac{y_j - T_j}{y_j}\}$.

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

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

Probability of unequal tax rates

Take the logarithm on both sides:

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

λ cancels out. Factorize y_i, y_j:

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

Probability of unequal tax rates

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:

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

Probability of unequal tax rates

What is the standard error?

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

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

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 (for example, if the legislature changes the tax code).

Likelihood function

- We just specified the probability that two households have a different tax rate.
- For the likelihood function, we want to make pairwise comparisons of every household in a fiscal year.
- Suppose there are four observations $i = 1, 2, 3, 4$.
- We would compare the 1st household to the 2nd, 3rd, 4th
    - Then the 2nd to the 3rd and 4th
    - Finally the 3rd to the 4th.
- Therefore, for $n = 4$, there would be $3 + 2 + 1$ pairwise combinations of $\epsilon_i - \epsilon_j$.
- More generally, for a given household, we only want to compare it to households indexed higher.

Likelihood function

Define the following variables:

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

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

The log-likelihood function is then:

$$\ln L = \sum_{i=1}^{n-1} \sum_{j>i}^{n} \; I_{ij} \cdot \ln P_{ij} + (1 - I_{ij}) \cdot \ln(1 - P_{ij}).$$

- Note that information from actual tax liability (T_i) is piped in through I_{ij}.
- The total number of combinations is a triangular number: $(n-1) + (n-2) + \cdots + 1 = \frac{n(n+1)}{2}$. Or is it?

Likelihood function

Issue 1: Independence

- Recall that $P_{ij} = \Pr(\frac{T_i}{y_i} < \frac{T_j}{y_j})$.
- Suppose we know that:

$$\frac{T_4}{y_4} < \frac{T_1}{y_1} < \frac{T_2}{y_2} < \frac{T_3}{y_3}.$$

- Then after comparing the 1st household to the 2nd, 3rd, and 4th,
  - Comparing the 4th household to the 2nd (or 3rd) reveals no additional information.
    - i.e., P_{24} is no longer independent given P_{12} and P_{14}.
  - Whereas, we still want to compare the 2nd household to the 3rd.

Likelihood function

Issue 1: Independence (cont.)

- One solution is to specify a tournament using notation.
- We sort those with higher average tax rate than household 1 into bracket A, and lower in bracket B.
  - We randomly select a household i in bracket A, and do pairwise comparisons with all other households in bracket A.
    - Those higher than i goes into bracket A1, those lower goes into bracket A2.
  - We repeat for random household j in bracket B, creating brackets B1 and B2.
  - Then we create brackets A1A, A1B, A2A, A2B, B1A... so forth. Tedious.

Likelihood function

Issue 1: Independence (cont.)

- Instead, we can order households by their average tax rates, i.e.:

$$i = 1, 2, \ldots, n \text{ is indexed such that } \frac{T_1}{y_1} < \frac{T_2}{y_2} < \cdots < \frac{T_n}{y_n}.$$

- Then, pairwise comparisons between household 1 and all other households would reveal no information about comparisons between household 2 and all households ranked higher than 2, and so forth.

Likelihood function

Issue 1: Independence (cont.)

When i is defined as such,

$$I_{ij} = 0 \text{ for all } i.$$

Thus, the log-likelihood function simplifies:

$$\ln L = \sum_{i=1}^{n-1} \sum_{j>i}^{n} \; I_{ij} \cdot \ln P_{ij} + (1 - I_{ij}) \cdot \ln(1 - P_{ij})$$

$$\rightarrow \sum_{i=1}^{n-1} \sum_{j>i}^{n} \ln(1 - P_{ij}).$$

Likelihood function

Issue 2: Convergence

Plug in the full term for P_{ij} and we'll notice a second issue:

$$\ln L = \sum_{i=1}^{n-1} \sum_{j>i}^{n} \ln\{1 - \Phi[\frac{\tau \ln(\frac{y_i}{y_j}) + \ln(\frac{a_i}{a_j}) - \theta \ln(\frac{A_i}{A_j})}{\sqrt{2}\sigma}]\}.$$

- Aren't the optimal solutions for $[\tau \quad \theta \quad \sigma]$ just $[-\infty \quad \infty \quad 0]$?
  - Well, no, because $\ln(\frac{y_i}{y_j})$ and $\ln(\frac{A_i}{A_j})$ can be negative or positive.
  - But even if τ and θ do not converge, then σ will still converge to ∞ to minimize $\Phi(\cdot)$ and maximize L.

Likelihood function

Issue 2: Convergence (cont.)

- The solution is: when we do our pairwise comparisons, we start with the last household (i.e., the one with the highest average tax rate).
- Because $I_{ni} = 1$ for all $i < n$, this will guarantee that both $\Phi(\cdot)$ and $1 - \Phi(\cdot)$ enters into our likelihood function.

The log-likelihood function is modified for the final time:

$$\ln L = \sum_{i=1}^{n-1} \sum_{j>i}^{n} \ln(1 - P_{ij})$$

$$\rightarrow \sum_{k=1}^{n-1} \ln P_{nk} + \sum_{i=1}^{n-2} \sum_{j>i}^{n-1} \ln(1 - P_{ij}).$$

Estimates over time

If we loop a MLE for each given fiscal year $t = 1, \ldots, T$,

we can estimate the following vectors:

$$\vec{L} = \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_T \end{bmatrix} \quad \vec{\tau} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_T \end{bmatrix} \quad \vec{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_T \end{bmatrix} \quad \vec{\sigma} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_T \end{bmatrix},$$

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