Code documentation

César Garro-Marín*

December 20, 2021

1 Estimating θ

1.1 Assuming scales

We normalize all the skill questions to range between zero and one. I define as the simple average of the skill questions involved:

$$S_{\theta,m} = \frac{1}{||m||} \sum_{l=1}^{||m||} s_{mli}$$

where s_{mji} is the individual i's answer to the skill question l. Next, I aggregate the dataset to at the occupation-year level and I estimate θ_i using the regression:

$$d \ln f_\theta(J) - \overline{d \ln f_\theta(J)} \quad = \quad \pi_\theta + \sum_i \beta_{\theta,m}(S^\star_{\theta,m}(J) - \overline{S^\star_{\theta,m}(J)}) + \nu_\theta(J)$$

The implied θ_i are given by,

- Unweighted: see log file
- Weighted: see log file

*Boston University, email: cesarlgm@bu.edu

Table 1: Estimates of $\beta_{\theta,m}$

	Unweighted			Weighted		
	Low	**Mid**	**High**	**Low**	**Mid**	**High**
	(1)	(2)	(3)	(4)	(5)	(6)
dm_manual	-0.48	-2.19**	0.40	2.29***	1.91	-0.15
	(0.62)	(0.68)	(0.41)	(0.55)	(1.53)	(0.59)
dm_routine	-0.32	-0.68	-0.13	1.72**	2.48	0.00
	(0.66)	(1.05)	(0.50)	(0.63)	(1.64)	(.)
dm_abstract	-2.28*	-3.75*	0.35	0.00	0.00	0.22
	(0.99)	(1.52)	(0.57)	(.)	(.)	(0.68)
dm_social	0.00	0.00	0.00	3.50***	4.24*	-0.28
	(.)	(.)	(.)	(1.01)	(1.76)	(0.55)
n_occupations	42	10	59	42	10	59
N	100	25	170	100	25	170
r2	0.10	0.34	0.01	0.15	0.33	0.01

1.2 Sanity check

1.2.1 Simulating data

If the model were true in the data, it generate data following the two equations below:

$$\sum_{m=1}^{I} \theta_j S_{\theta,m}(J) = 1 \tag{1}$$

$$d\ln f_\theta(J) - \overline{d\ln f_\theta(J)} = \frac{\varepsilon}{\varepsilon - 1} \sum_i \left[\theta_i S^\star_{\theta,i}(J) - \theta_i \overline{S^\star_{\theta,i}(J)}\right] d\ln A_i \tag{2}$$

How do I simulate the data:

Equation (2) is simple. Choose ε and $d\ln A_i$ and generate data following:

$$d\ln f_\theta(J) - \overline{d\ln f_\theta(J)} = \frac{\varepsilon}{\varepsilon - 1} \sum_i \left[\theta_i S^\star_{\theta,i}(J) - \overline{S^\star_{\theta,i}(J)}\right] d\ln A_i + \nu_\theta(J)$$

Making equation (3) work seems more involved. This equation implies that:

$$S_{\theta,M}(J) = \frac{1}{\theta_M}\left(1 - \sum_{m=1}^{M-1} \theta_i S_{\theta,m}(J)\right) + \eta_\theta(J)$$

I will start simple:

- Assume a matrix of θ_i.
- Why am I complicating myself with this? Simply generate the data following the above equation.
- See if my algorithm works in finding the solution.

1.3 Estimating scales

Estimation uses two key equations from the model:

$$\sum_{m=1}^{I} \theta_j S_{\theta,m}(J) = 1 \tag{3}$$

$$d\ln f_\theta(J) - \overline{d\ln f_\theta(J)} = \frac{\varepsilon}{\varepsilon - 1} \sum_i \left[\theta_i S^\star_{\theta,i}(J) - \theta_i \overline{S^\star_{\theta,i}(J)}\right] d\ln A_i \tag{4}$$

our current procedure to estimate the model parameter is choose scales c_{jml} and scale weights α_{jm} to minimize the MSE from equation (3):

$$\min_{\alpha_{jm}, c_{jml}} \frac{1}{N} \left[\sum_{m=1}^{I} \theta_j S_{\theta,m} - 1\right]^2 \quad \text{s.t. } S_{\theta,m} = \sum_{j=1}^{||m||} \alpha_{jm} \sum_{l=1}^{5} c_{jml} 1_{d_{ijm}=l}$$

Where θ_i comes from an OLS regression using the estimating equation:

$$d\ln f_\theta(J) - \overline{d\ln f_\theta(J)} = \pi_\theta + \sum_i \beta_{\theta,i}(S^\star_{\theta,i}(J) - \theta_i \overline{S^\star_{\theta,i}(J)}) + \nu_\theta(J)$$

2 Defining education groups

Our current results group education levels into three broad groups that I will often call Low, Mid, and High.

Table 2: Add caption

Label	**GCSE qualification level**
Low	Below GCSE A
Mid	GCSE A* / trade qualification
High	Bachelor +

3 Classifying jobs

We say that an occupation j is a core job of education group e if two conditions are met:

1. Education group e is overrepresented in the occupation relative to the overall population. That is:

$$s_e(j) \geq \overline{s}_e$$

   where $s_e(j)$ denotes the employment share of the education group e in job j, and $\overline{s}_e$ is its employment share in the population.

2. The employment share of group e in job j is at least 4 the employment share of any other education group that is overrepresented in the occupation.

$$s_e(j) \geq 4 s_{e'}(j)$$

for any other education group e' such that $s_{e'}(j) \geq \overline{s}_{e'}$.

4 Computing θs

4.1 Data I use

First I restrict data to only:

1. occupations that are core jobs in two consecutive SES-waves.
2. people with education levels matching the job-classification. For example, I restrict to observations of individuals with low-education in low-education core-jobs.

Using this restricted dataset, I occupational employment shares by education level:

$$s_e(j) = \frac{l_e(j)}{\sum_{j'} l_e(j')}$$

where l denotes employment and the summation is over jobs that stay in the core of education group e in two consecutive SES-waves.

5 Solution procedure

Out of equation 32 we have:

$$\frac{\partial \ln f_\theta(J)}{\partial A_i} - \frac{\partial \ln f_\theta(J')}{\partial A_i} = \frac{\varepsilon}{\varepsilon - 1}\left[\frac{\ln y_\theta^\star(J)}{\partial \ln A_i} - \frac{\ln y_\theta^\star(J')}{\partial \ln A_i}\right]$$

Moreover, out of question 44 we have

$$\frac{\partial \ln y_\theta^\star(J)}{\partial \ln A_i} = \theta_i S_{\theta,i}^\star(J)$$

Plugging into 32 we have:

$$\frac{\partial \ln f_\theta(J)}{\partial A_i} - \frac{\partial \ln f_\theta(J')}{\partial A_i} = \frac{\varepsilon}{\varepsilon - 1}\left[\theta_i S_{\theta,i}^\star(J) - \theta_i S_{\theta,i}^\star(J')\right] \tag{5}$$

Thus:

$$d \ln f_\theta(J) - d \ln f_\theta(J') = \frac{\varepsilon}{\varepsilon - 1} \sum_i \left[\theta_i S_{\theta,i}^\star(J) - \theta_i S_{\theta,i}^\star(J')\right] d \ln A_i + \theta_M S_{\theta,M}^\star(J) - \theta_M S_{\theta,M}^\star(J')$$

Summing over jobs and dividing by the number of jobs we have:

$$d\ln f_\theta(J) - \overline{d\ln f_\theta(J)} = \frac{\varepsilon}{\varepsilon-1}\sum_i \left[\theta_i S^\star_{\theta,i}(J) - \theta_i \overline{S^\star_{\theta,i}(J)}\right] d\ln A_i$$

This equation calls for the following regression specification:

$$d\ln f_\theta(J) - \overline{d\ln f_\theta(J)} \;=\; \alpha_\theta + \sum_i \beta_{\theta,i}(S^\star_{\theta,i}(J) - \theta_i \overline{S^\star_{\theta,i}(J)}) + \nu_\theta(J)$$

Then:

- Under the assumption that $\theta_i = 1$, $\frac{\varepsilon}{\varepsilon-1} d\ln A_i$ is identified out of the low education group.
- Rest of education groups identify θ_i.

5.1 Procedure

1. Guess $S_{\theta,i}(J)$.
2. Estimate θ_i out of core jobs.
3. Given θ_i estimate $S_{\theta,i}(J)$.
4. Return to 1.

5.2 What functions do I need to write

5.2.1 Estimation of θ_i

Let y_θ be the $J \times 1$ vector containing the vector of $d\ln f_\theta(J) - d\ln f_\theta(J')$. Let S_θ the $J \times I$ matrix of skill indexes $S^\star_{\theta,i}(J) - S^\star_{\theta,i}(J')$. Then:

$$\beta_\theta = \frac{\epsilon}{\epsilon - 1}[\theta_1 d\ln A_1 \ldots \theta_I d\ln A_I]'$$

I estimate β_θ by OLS:

$$\beta_\theta = (X'_\theta X_\theta)^{-1} X_\theta y_\theta$$

Using the appropriate block diagonal matrix I can estimate all the vectors at the same time. For this I need:

- The usual OLS function
- The function to block diagonalize the matrix that I already wrote.

Next, I need to back out the θ_i. For this I need to do:

$$\beta_1 = \frac{\epsilon}{\epsilon - 1}[d \ln A_1 \dots d \ln A_I]'$$

Then:

$$\theta = \beta_\theta \oslash \beta_1$$

For this I need:

- Function splitting the vector by education level.
- Function estimating β_θ: `estimate_beta_theta`
- Function estimating the θ `estimate_theta`.
- Function estimating averages of skill indexes by education level: `average_skill_use`.

5.2.2 How do I estimate the scales then?

There are a set of O skill questions in the SES survey that we have partitioned into M mutually exclusive groups that we index by m. Within each partition, we index the skill questions by j. Let d_{ijm} be individual's i answer for the skill question jm. $d_{ijm} \in \{1, 2, 3, 4, 5\}$. The problem is then:

$$\min_{\alpha_{jm}, c_{jml}} \frac{1}{N} \left[\sum_{m=1}^{I} \theta_j S_{\theta,m} \right]^2 \text{ s.t. } S_{\theta,m} = v$$

- I think this is mostly done. I just have to modify the loss function for this.

References