Selection bias and segregation indices: the international comparison of segregation levels

Ricardo Mora[1]

2023 Spanish Stata Users Group meeting Madrid
23 th October

[1]Universidad Carlos III, Madrid, Spain. Corresponding author. E-mail: ricmora@uc3m.es

Introduction

- Occupational segregation is the differing distribution of men and women across jobs.
- **Challenge**: Changes in female work participation influence occupational segregation. This makes interpretation of international or time differences in segregation measures difficult.
- **Traditional "Solutions":**
  - **Use information on the working population and** ignore issue.
  - **Use information on the working population and** measure segregation using a segregation index which is independent of these percentages (a property known as "Composition Invariance").

The significance of segregation indices

- Regular debates over the merits of various indices.
  - James and Taeuber 1985, Watts 1992, Reardon & Firebaugh 2002, Hutchens 2003, Frankel and Volij 2010.
- Composition Invariance (for example, Gini, Dissimilarity, and Hutchens indices):
  - Advantage: Given a sample, index computation changes cannot be influenced by female work rates.
  - Problem 1: Restricts the concept of segregation, potentially limiting research objectives.
  - Problem 2: Implicitly assumes equal occupational segregation patterns for working and non-working populations.

Other Segregation Indices

- Many indices lack the CI property. Examples: Theil's Entropy, Mutual Information index, Relative Diversity.
- Cohen (2004)'s proposal: Include 'Housework' in occupational categories. (Also Hook and Petit 2016.)
- Guinea-Martin, Mora and Ruiz-Castillo (2018): Economic vs Time vs Occupational segregation using a unit-decomposable index (Mutual Information) .
    - Both genders always equally represented, so no need for Composition Invariance.
    - Practical Problems:
        - Non-occupational categories limited and often vague.
        - Need for decomposable indices limits choice: Gini and Dissimilarity are excluded.
        - No measure of occupational segregation for the entire population.

Proposal

- Maximum Likelihood estimation of occupational segregation for the entire population dealing with non-ignorable non-response as in Ramahlo and Smith 2013.
- Can be applied to any segregation index.
- Requires:
  - gender frequencies per occupation in the working population
  - gender participation rates in socio-demographic groups
- Three scenarios:
  - Missing completely at random: non-parametric ML estimation leads to traditional approach.
  - Missing at random: non-parametric ML estimation requires individual characteristics, including participation rates, for the entire population.
  - Endogenous selection: ML estimation requires additional assumptions (in this talk, I center on parametric models).

Data and methodology

Data source and index measurement

- Data from 25 European labor force surveys from 2013 (most recent year with info on field of study).
- All individuals aged 25-29 up to 50-54.
- Labour market participation status (in the entire population).
- Occupational categories: three-digit International Standard Classification of Occupations (2008) (in the working population).
- Cells by country: Five year age intervals, three levels of education, nine fields of study, and other background information (number of children and previous job).
- Stata implementation with several indices of segregation: Gini, Dissimilarity (Duncan & Duncan), Simpson (Relative diversity), Hutchens, Theil's H, and Mutual Information.

Overview of the model

- Individuals can be women or men.
- They can choose to work or not.
- Those who work must select one of J occupations.
- Additional individual characteristics (e.g., education) available.
- Objective: Determine an index S_A that quantifies occupational segregation in population A.

The core problem

- $\{\pi_{jg}\}_A$ represents the joint distribution of occupations and gender in population A.

- The discussion centers on indices influenced only by this joint distribution: $S_A = S\left(\{\pi_{jg}\}_A\right)$
  - ML estimation of S_A, $\widehat{S}_A^{ML}$, is $S\left(\{\widehat{\pi}_{jg}^{ML}\}_A\right)$

- Individuals opt not to work if their best occupational choice isn't favorable relative to non-working.

- Missing information: Preferred occupation of non-workers.
  - Participation in the job market is a nonresponse missing data mechanism.

Case 1: Ignorable Non-Response (MCAR)

- Participation is independent of occupation, gender, and worker type.
- Sample job-gender frequencies within the working population, $\frac{\#(j,g,x,work=1)}{\#(x,work=1)}$, are Maximum Likelihood (ML) estimates for the entire population.
- These form the foundation for a consistent and efficient ML estimation of each segregation index: $S_A = S\left(\left\{\frac{\#(j,g,x,work=1)}{\#(x,work=1)}\right\}_A\right)$.
- Bootstrap techniques can calculate standard errors (Deutsch et al. 1994, Boisso et al. 1994, Ransom 2000, Allen et al. 2015).
- **Problem:** Best occupation preferences likely differ between the working and total population.

Case 2: Selection on Observables (MAR)

- Participation is conditionally independent of occupation, given gender **and type**.
  - We have info of individual characteristics that perfectly identify the type of each individual.
- Traditional approaches (using only working-population information) biases the segregation index.
  - Example: If female participation rises with education, the traditional method over-weighs highly educated women and the index might under-represent segregation if it's lower among educated groups. This negative bias should be larger in countries with relative low participation rates.
- ML solution under selection on observables:
  - Compute occupation-gender relative frequencies by type **in the working sample**.
  - Average these relative frequencies using as weights gender *cum* type of worker joint shares **in the entire sample**.

Case 3: Endogenous selection

- Missing at random is problematic if type info is incomplete. In that case, participation varies based on occupation, given gender and **observed individual type**.
  - This is a problem of endogenous sample selection and leads to inconsistent estimates of the index of segregation both in the traditional approach and also if we assume selection on observables.

- Unfortunately, the model assuming that participation is conditionally dependent on occupation, gender, and type, lacks identification without extra assumptions.
  - For each gender and type, the ML estimator only exploits the following condition: $\frac{\#(j,w=1,g,x)}{\#(w=0,g,x)} = \frac{\widehat{\Pr}^{ML}(w=1,j|g,x)}{\widehat{\Pr}^{ML}(w=0|g,x)}$.
    - These are less conditions than the number of parameters.

Parametric identification

- Option 1: Probability of participation depends on occupation, gender, and type of worker additive effects:
  $\Pr(w = 1|j, g, x) = G(\beta_j + \alpha_{fem} + \gamma_x)$
  - gender differences in participation rates are constant across occupations and types.

- Option 2: Probability of female participation depends on female-occupation and type of worker additive effects:
  $\Pr(w = 1|j, g = female, x) = G(\beta_0 + \alpha_{fem,j} + \gamma_x)$
  - male participation rates are missing at random.
  - endogenous selection only occurs in the female population.

- Option 3: Probability of female participation depends on how popular preferred occupation is in the male population:
  $\Pr(w = 1|j, g = female, x) = G\left(\beta_0 + \alpha_f \pi_{j|male} + \gamma_x\right)$
  - male participation rates are missing at random.
  - endogenous selection only occurs in the female population.

- $G(\cdot)$ is known (i.e., logit, probit,...)

Sample identification

- Options 1 and 2 are numerically unstable when the number of occupations is large (convergence is routinely not achieved)

- Option 3:
  - In the sample of male workers, estimate $\widehat{\gamma}_x$ and $\widehat{\pi}_{j|male}$
  - Plug these consistent estimates in the sample of women and estimate remaining parameters by ML estimation.
  - Algorithm usually converges (in parameters or log likelihood) in less than 10 iterations).
  - Likelihood is concave at maximum.
  - Variance-covariance estimator of $\widehat{\pi}_{j|female,x}$ is unstable (and with zero entries)

Stata implementation

Command `segsel`

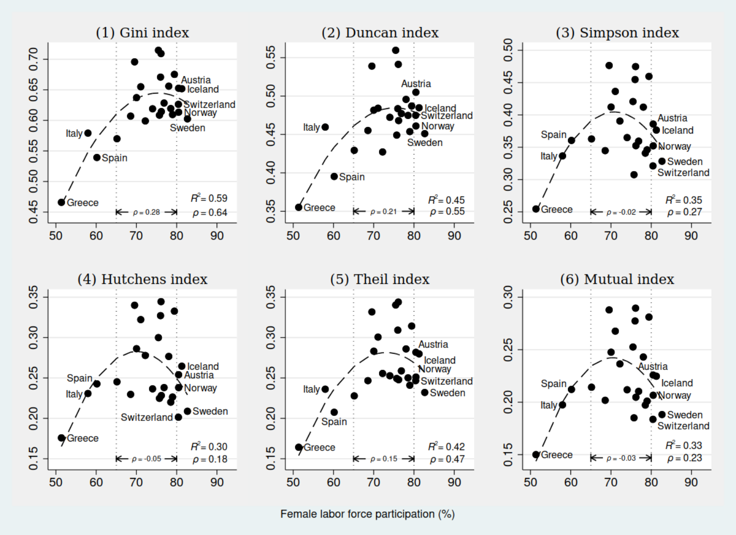
- Computes ML estimates of π_{jg} for the entire population. These estimates are stored in a ereturn matrix.
    - Hence, computation of the segregation index becomes a two-step procedure in Stata:
        - First step: estimate π_{jg} by Maximum Likelihood.
        - Second step: compute $S(\{\pi_{jg}\})$ using other Stata comands, such as **`seg`** (to compute Gini, Dissimilarity, Theil's H), **`hutchens`** (to compute Hutchens), or **`dseg`** (to compute the Mutual and Relative diversity).
- Current version includes:
    - The missing completely at random case: relative frequencies in the working population.
    - The missing at random case: weighted average of relative frequencies by type in the working population with weights equal to the relative gender and type frequencies in the entire population.
    - Three versions of the logit parametric case for endogenous selection: gf0 and gf1.
- Additional outcomes: test of ignorability

Illustration with command seg

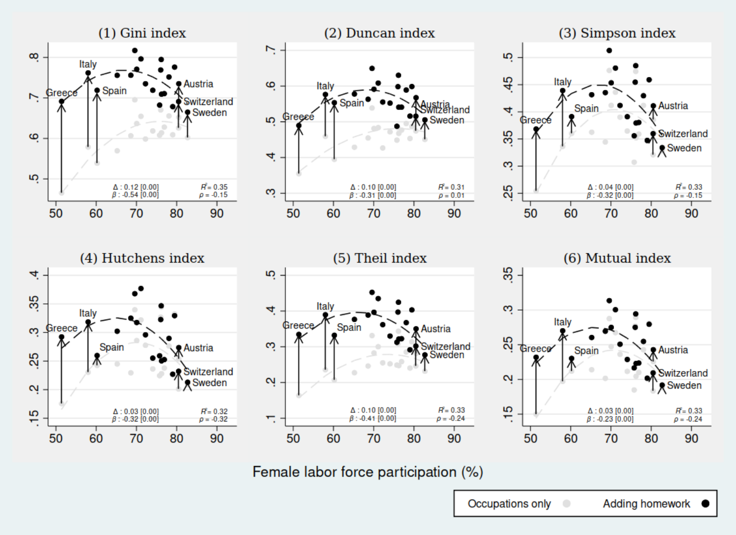
```
segsel occupation [fweight=nobs], groups(sex) model(pes,logit3) ///
        selection(work) evaltype(gf1) quietly
// Stata variables from ereturn matrices:
svmat e(Pr_jg), names("Pr_j") // vars:  Pr_j1 & Pr_j2, J obs.
svmat e(N), names( "N") // vars:  N, 1 obs.
// Keeping estimated probabilities and sample frequencies by gender and
occupation
keep Pr_j* N
keep if Pr_j1!=.
// Filling all J observations in N
replace N = N[_n-1] in 2/l
// Estimated frequencies by occupation and gender
gen nobs1 = int(Pr_j1 * N)
gen nobs2 = int(Pr_j2 * N)
// Indices computation
seg nobs1 nobs2, g d unit(_n) generate(g Gini d Duncan)
```

Key findings

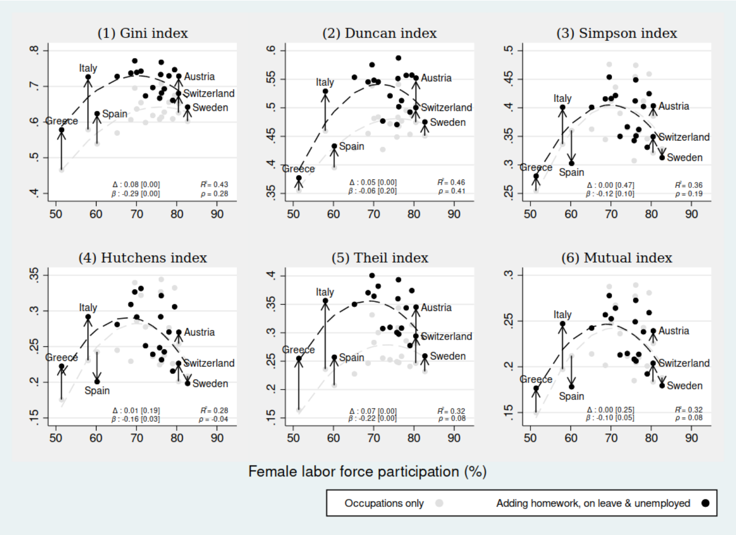
Traditional measures of occupational segregation



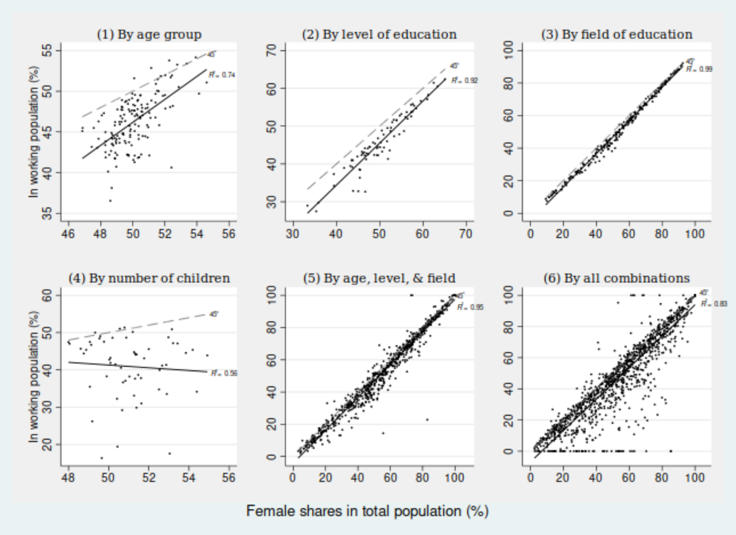
Broader approach: adding homework



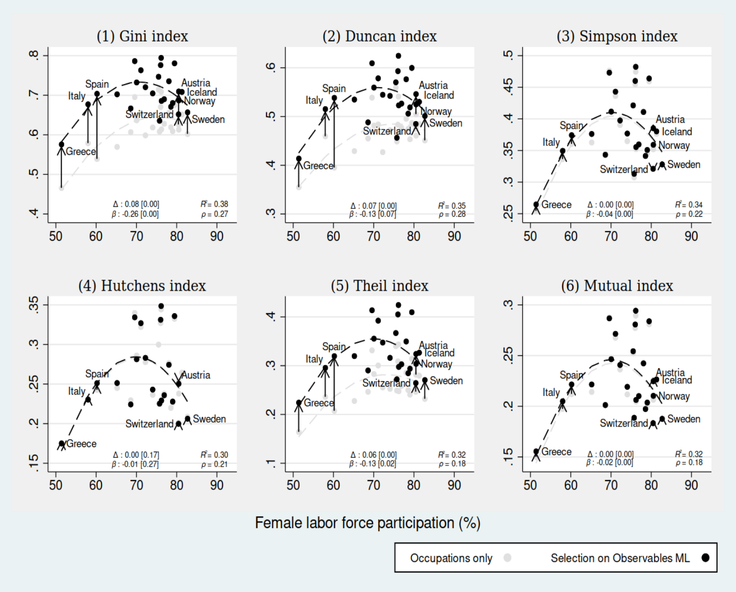
Broader approach: adding other categories



Female labor force participation and individual types



Selection on observables



Endogenous selection

- Preliminary results using the third option:

$$\Pr\left(w = 1|j, g = female, x\right) = G\left(\beta_0 + \alpha_f \pi_{j|male} + \gamma_x\right)$$

  - Parameter α_f captures how the probability of participation of a woman is associated to the popularity of her preferred occupational choice among men.

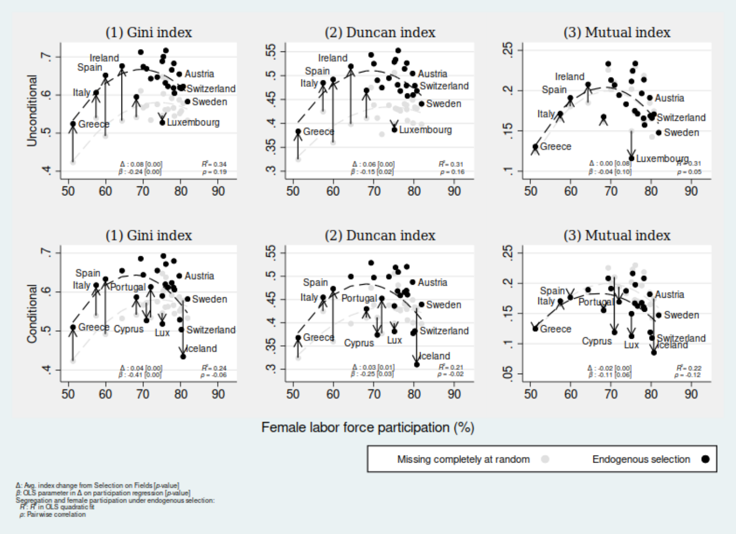
- Occupational categories: two-digit International Standard Classification of Occupations (2008) (in the working population).

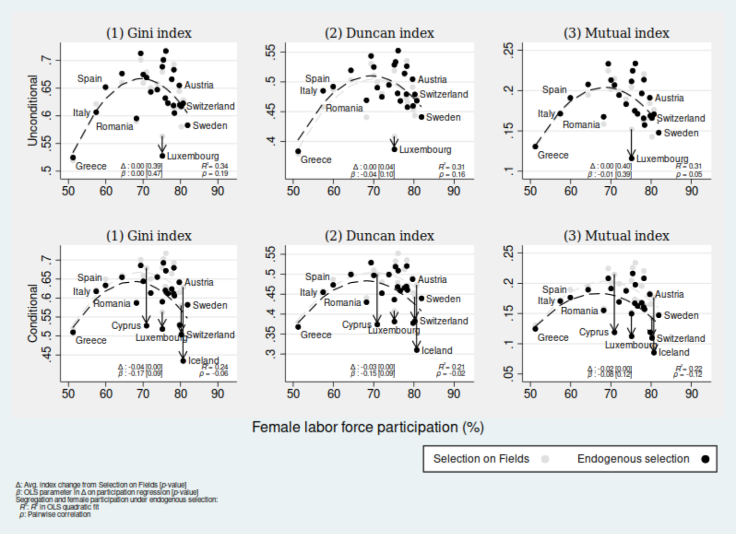
- Cells by country: five cells as the interaction of levels and fields of study

- Gini, Dissimilarity (Duncan & Duncan), and Mutual Information.

Table: Endogenous Bias in Female Labor Force:
The Role of Occupations Popular Among Men. MLE.

	Unconditional				Conditional on field			
	$\widehat{\alpha}^{ML}$	Std.Err.	z	p-value	$\widehat{\alpha}^{ML}$	Std.Err.	z	p-value
Italy	-29.870	0.0000	.	.	-0.000	0.0041	0.000	0.999
Czech Republic	-25.921	0.0146	1771.661	0.000	**-0.056**	0.0156	3.599	0.000
Estonia	-22.306	0.0293	760.307	0.000	-0.000	0.0375	0.000	0.999
Germany	-19.181	0.0033	5898.864	0.000	**-0.547**	0.0047	117.061	0.000
Hungary	-17.374	0.0217	801.668	0.000	-0.000	0.0123	0.000	0.999
Spain	-15.150	0.0071	2127.147	0.000	**-0.038**	0.0041	9.188	0.000
Norway	-14.148	0.0128	1108.084	0.000	**-0.044**	0.0179	2.440	0.015
Austria	-12.800	0.0119	1077.437	0.000	-0.000	0.0134	0.000	0.999
Romania	-11.959	0.0267	447.288	0.000	-0.000	0.0055	0.000	0.999
Portugal	-9.083	0.0104	872.173	0.000	-0.000	0.0105	0.005	0.996
Ireland	-7.464	0.0168	445.326	0.000	-0.000	0.0158	0.000	0.999
Latvia	-5.746	0.0230	250.092	0.000	-0.000	0.0285	0.000	0.999
Switzerland	-4.377	0.0177	247.673	0.000	-0.000	0.0160	0.000	0.999
Sweden	-0.250	0.0124	20.252	0.000	-0.000	0.0151	0.000	0.999
Belgium	-0.021	0.0115	1.807	0.071	**-0.060**	0.0124	4.790	0.000
Greece	-0.014	0.0071	2.013	0.044	-0.000	0.0069	0.000	0.999
France	-0.011	0.0050	2.205	0.027	-0.000	0.0050	0.000	0.999
Slovakia	-0.010	0.0148	0.708	0.479	-0.000	0.0157	0.000	0.999
Iceland	-0.006	0.0729	0.087	0.931	-0.011	0.0757	0.139	0.889
Finland	-0.006	0.0164	0.337	0.736	-0.003	0.0177	0.174	0.862
Denmark	-0.003	0.0141	0.238	0.812	-0.000	0.0163	0.000	0.999
Latvia	-0.000	0.0195	0.011	0.991	-0.000	0.0222	0.000	0.999
Netherlands	-0.000	0.0092	0.000	0.999	**-0.073**	0.0099	7.404	0.000
Cyprus	-0.000	0.0309	0.000	0.999	**-0.057**	0.0321	1.771	0.076
Luxembourg	-0.000	0.0450	0.000	0.999	-0.020	0.0468	0.430	0.667


(1) Gini index
(2) Duncan index
(3) Mutual index
Unconditional
Conditional
Ireland
Spain
Italy
Greece
Portugal
Cyprus
Lux
Austria
Switzerland
Sweden
Luxembourg
Iceland
Female labor force participation (%)
Missing completely at random
Endogenous selection
Δ: Avg. index change from Selection on Fields [p-value]
β: OLS parameter in Δ on participation regression [p-value]
Segregation and female participation under endogenous selection:
R²: R² in OLS quadratic fit
ρ: Pairwise correlation


(1) Gini index
(2) Duncan index
(3) Mutual index
Unconditional
Conditional
Spain
Italy
Romania
Greece
Austria
Switzerland
Sweden
Luxembourg
Cyprus
Iceland
Female labor force participation (%)
Selection on Fields
Endogenous selection
Δ: Avg. index change from Selection on Fields [p-value]
β: OLS parameter in Δ on participation regression [p-value]
Segregation and female participation under endogenous selection:
R²: R² in OLS quadratic fit
ρ: Pairwise correlation

Conclusions and discussion

Main takeaways/Conclusions

- This paper proposes an estimator of occupational segregation for the population as a whole which can be applied to any segregation index and does not require detailed individual information.
- Selection into participation in the labor market is viewed as a nonresponse missing data mechanism whereby the missing items are the occupational categories of non-participants.
- The fundamental methodological aspect of the proposal is to estimate for each individual that does not participate in the labor market the probability that he/she has to work in each occupation.
- Several scenarios regarding the missing mechanism are considered and ML estimation is implemented using a new Stata command.
- An illustration with European data shows that selection into participation is not ignorable in the absence of additional information.