Estimating Population Size from Register Data: a Capture-Recapture Approach

Lucy Brown[1]: lyb3@kent.ac.uk

Eleni Matechou[2] Bruno Santos[3] Eleonora Mussino[4]

1 University of Kent, UK
2 Queen Mary University of London, UK
3 Universidade de Lisboa, Portugal
4 Stockholm University, Sweden

Motivating Case Study

[1] Andrea Monti, Sven Drefahl, Eleonora Mussino, and Juho Härkönen. Over-coverage in population registers leads to bias in demographic estimates. Population Studies, 74(3):451–469, 2020

[2] Eleonora Mussino, Bruno Santos, Andrea Monti, Eleni Matechou, and Sven Drefahl. Multiple systems estimation for studying over-coverage and its heterogeneity in population registers. Quality & Quantity, 1–24, 2023

[3] Bruno Santos, Eleonora Mussino, Sven Drefahl, and Eleni Matechou. Using population register data and capture-recapture models to estimate over-coverage in Sweden. Scientific Reports, 14(1):1-10, 2024

Overcoverage:

- Due to imperfect emigration and/or death registration
- Leads to serious bias in population estimates
- Negatively influences policy-making and research

Overcoverage Estimation:

- Existing approaches[1][2] rely on multiple systems estimation (MSE) and only consider annual snapshots of the register data
- Instead, we have employed a longitudinal approach[3], following individuals, and hence registers, over different years

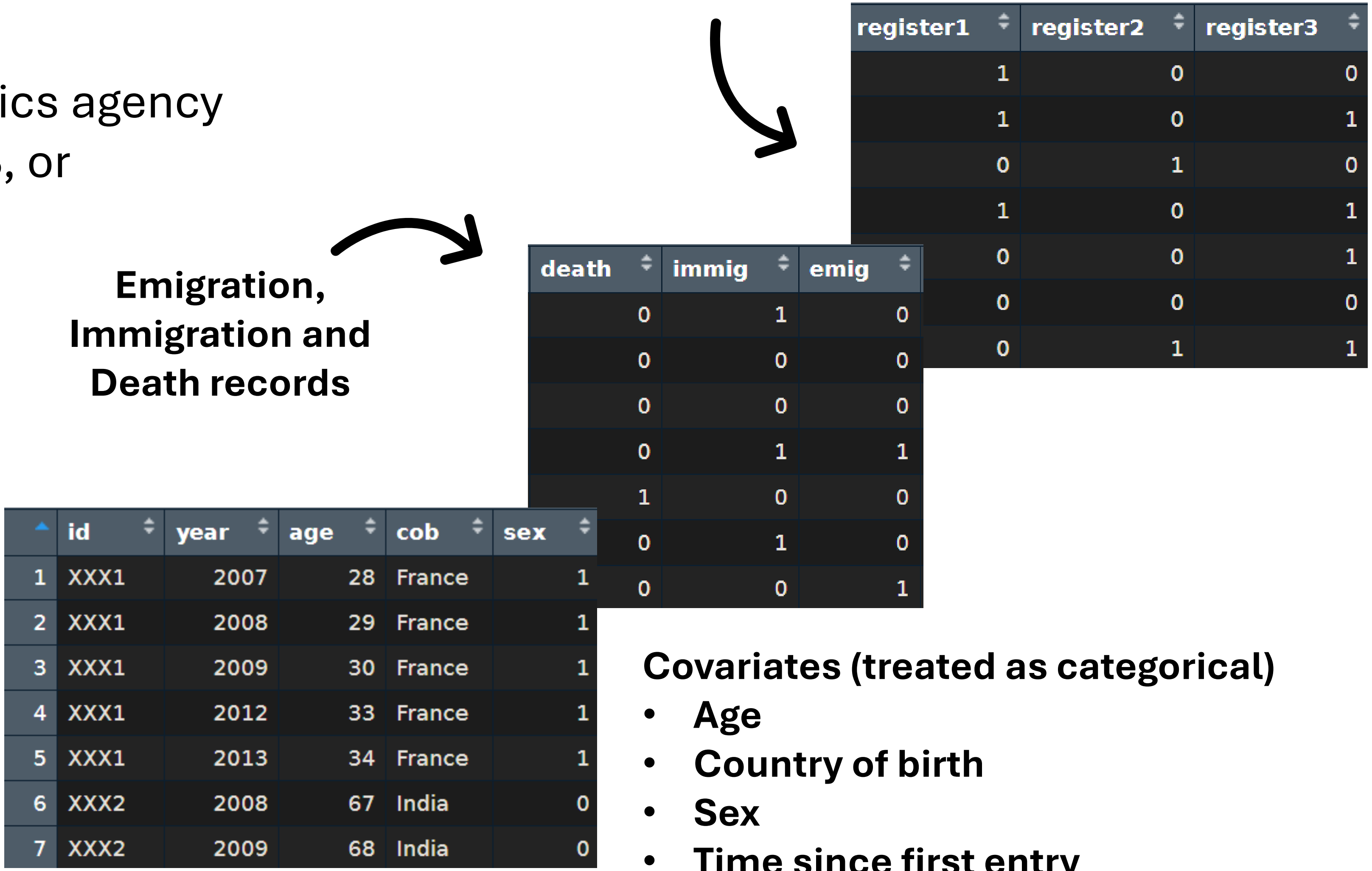
Administrative Register Data

- Provided by a national statistics agency (e.g. Statistics Sweden – SCB, or Statistics Norway - SSB)

- All foreign-born residents who first entered the country as adults during the study years (e.g. for Sweden 2003-2016, or for Norway 2007-2023)

An arbitrary number of registers

Emigration, Immigration and Death records



register1	register2	register3
1	0	0
1	0	1
0	1	0
1	0	1
0	0	1
0	0	0
0	1	1

death	immig	emig
0	1	0
0	0	0
0	0	0
0	1	1
1	0	0
0	1	0
0	0	1

	id	year	age	cob	sex
1	XXX1	2007	28	France	1
2	XXX1	2008	29	France	1
3	XXX1	2009	30	France	1
4	XXX1	2012	33	France	1
5	XXX1	2013	34	France	1
6	XXX2	2008	67	India	0
7	XXX2	2009	68	India	0

Covariates (treated as categorical)

- **Age**
- **Country of birth**
- **Sex**
- **Time since first entry**

Observations

An individual is observed on some **combination of R registers**, on which they can be observed (1) or not observed (0), i.e. 2 possible outcomes for each register

$\Longrightarrow$ 2^R register combinations

Example Person #1

$$(0,1) \longrightarrow (1,1) \longrightarrow \underbrace{\underset{\text{Death recorded}}{\boxed{(0,0)}} \longrightarrow (0,0) \longrightarrow (0,0)}_{\text{Dead}}$$

$$\text{alive} \longrightarrow \text{alive} \longrightarrow \text{dead} \longrightarrow \text{dead} \longrightarrow \text{dead}$$

Example Person #2

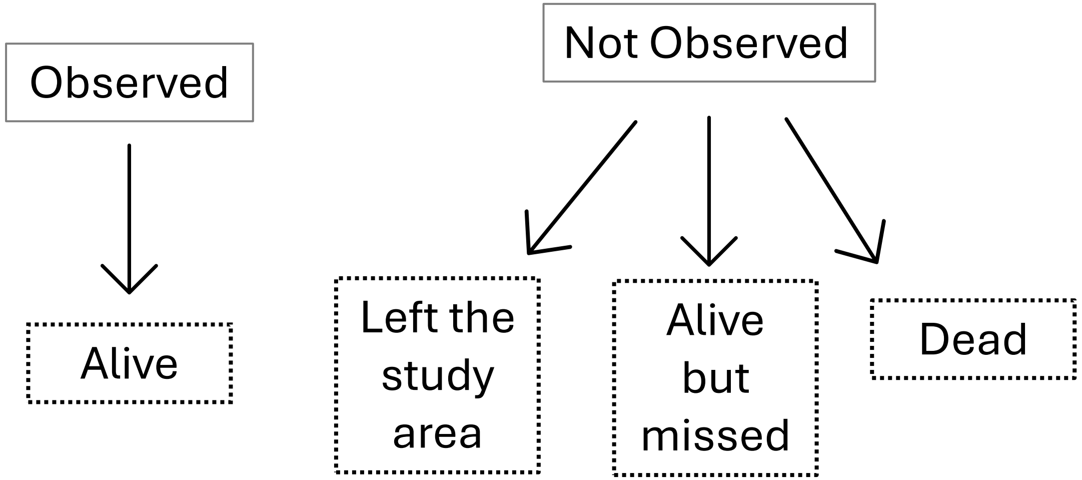
$$(1,1) \longrightarrow (0,1) \longrightarrow \underbrace{(0,0) \longrightarrow (0,0)}_{\text{Have they been missed or are they not there anymore?}} \longrightarrow (1,0)$$

$$\text{alive} \longrightarrow \text{alive} \longrightarrow ? \longrightarrow ? \longrightarrow \text{alive}$$

Capture-Recapture Models

Capture Recapture (CR) Models:

1. An initial capture is made
2. Unmarked individuals are marked in some unique way (tags, rings etc)
3. Individuals are released back into the population
4. Subsequent captures are made, allowing individuals to be tracked over time



Observed

Alive

Not Observed

Left the study area

Alive but missed

Dead

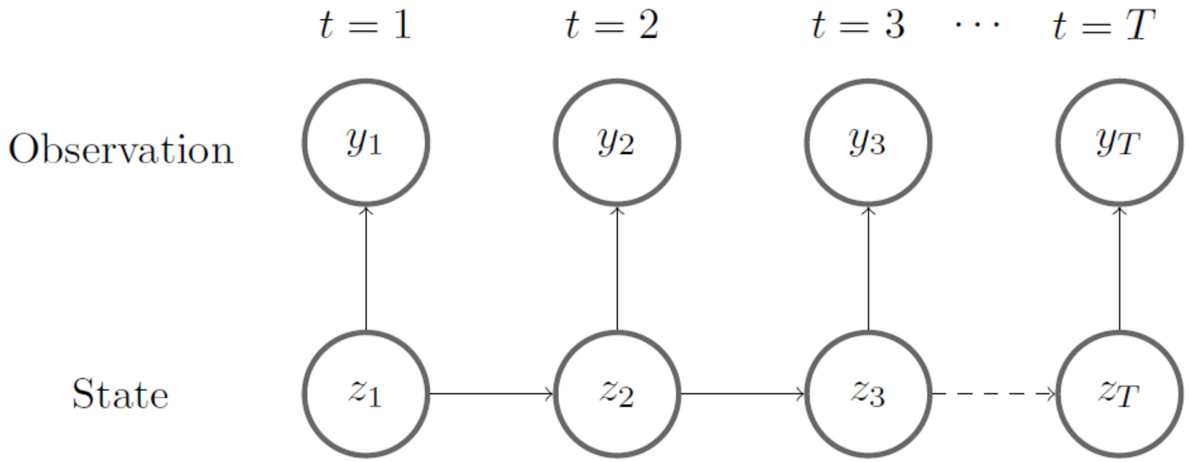
- CR models are commonly used in ecology to estimate population size for wild animals
- An individual's true state z_t is unobservable (latent)

Kenneth H. Pollock. Capture-recapture models. Journal of the American Statistical Association, 95(449):293–296, 2000. ISSN 01621459, 1537274X. URLhttp://www.jstor.org/stable/2669550

HMM Formulation

J. L. Laake. Capture-recapture analysis with hidden markov models. 2013.

- Hidden Markov model (HMM) formulation marginalises over the latent state to calculate the marginal likelihood
  - The marginal likelihood is the probability of the observed data and the latent states, given the model parameters
- Computationally expensive so use the Forward Algorithm
- This iterative method is very efficient and allows latent states to be inferred



Probability of history at the previous time point

Probability of transitioning to each new state given the state at $t-1$

$$\boldsymbol{\alpha_1 = \delta\, P(y_1)}$$

$$\boldsymbol{\alpha_t = \alpha_{t-1}\, \Gamma_{t-1}\, P(y_t)}$$

Observation probabilities dependent on the current state

$$\Rightarrow \quad \boldsymbol{L_T = \alpha_T 1'}$$

Simple CR Model

$$\Gamma = \begin{array}{c} \\ \end{array}\begin{bmatrix} \phi & 1-\phi \\ 0 & 1 \end{bmatrix} \begin{array}{l} z_t = \text{alive} \\ z_t = \text{dead} \end{array}$$

- States z_t: alive (1) or dead (0)
- Observations y_t: observed (1) or not observed (0)
- Probability of survival: ϕ
- Probability of observation: p
- All individuals are alive and observed when initially captured, i.e. initial state $\delta = [1,0]$

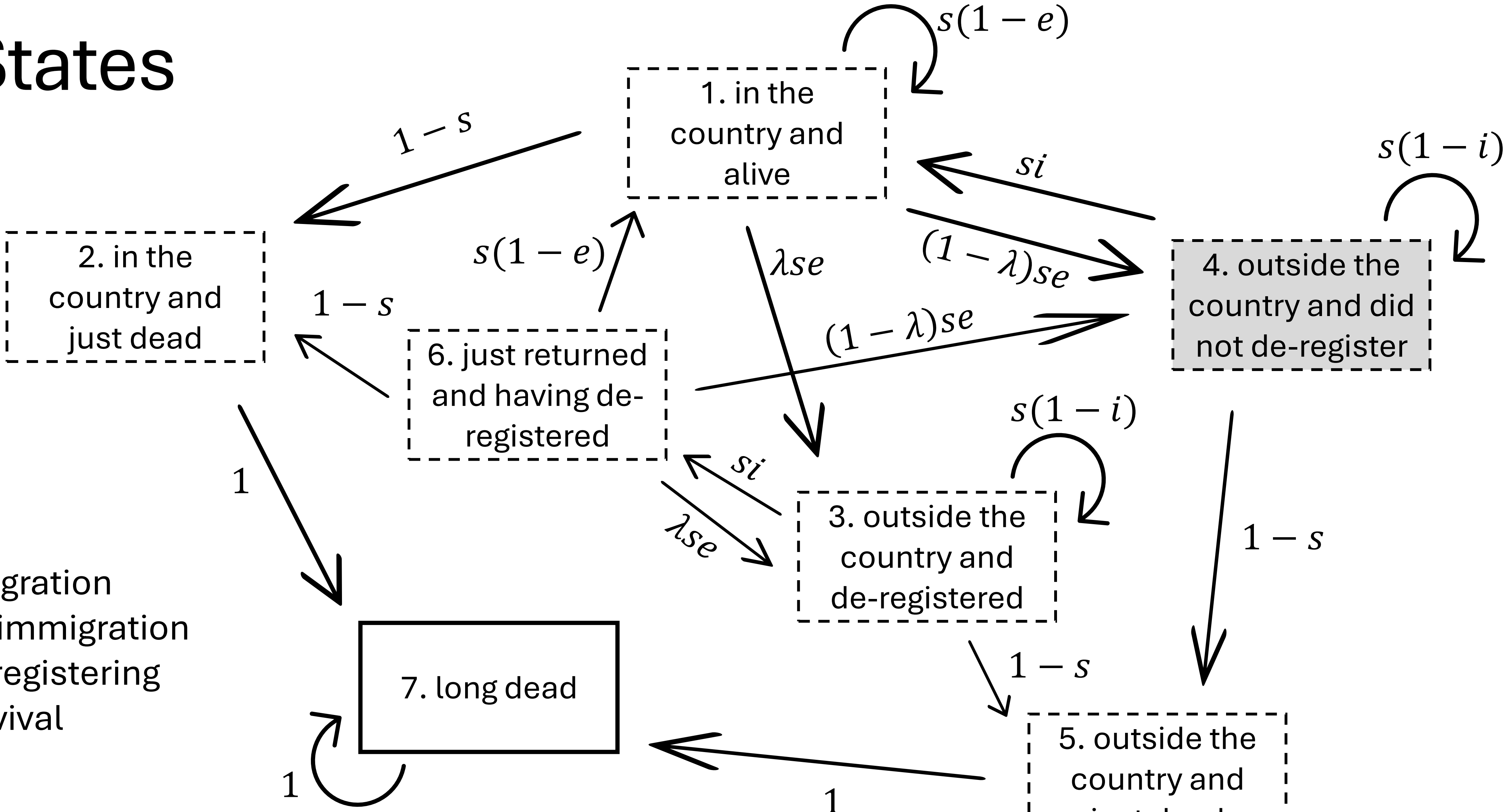
$$\mathrm{P}(0) = \begin{bmatrix} 1-p & 0 \\ 0 & 1 \end{bmatrix} \quad \mathrm{P}(1) = \begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix}$$

If an individual has capture history $y_t = [1,0,1]$, they have a likelihood:

$$L_T = \delta\, P(y_1)\, \Gamma\, P(y_2)\, \Gamma\, \mathrm{P}(\mathrm{y}_3) 1'$$

$$L_T = [1,0] \begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \phi & 1-\phi \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1-p & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \phi & 1-\phi \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

States



$s(1-e)$
1. in the country and alive
$1-s$
si
$s(1-i)$
2. in the country and just dead
$s(1-e)$
λse
$(1-\lambda)se$
4. outside the country and did not de-register
$1-s$
$(1-\lambda)se$
6. just returned and having de-registered
$s(1-i)$
1
si
λse
3. outside the country and de-registered
$1-s$
7. long dead
$1-s$
1
5. outside the country and just dead
1

e = emigration
i = (re-)immigration
λ = de-registering
s = survival

Transition Matrix

$$\begin{bmatrix} s(1-e) & 1-s & \lambda se & (1-\lambda)se & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & s(1-i) & 0 & 1-s & si & 0 \\ si & 0 & 0 & s(1-i) & 1-s & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ s(1-e) & 1-s & \lambda se & (1-\lambda)se & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

1. In the country and alive
2. In the country and just dead
3. Outside the country and de-registered
4. Outside the country but didn’t de-register

5. Outside the country and just dead
6. Just returned to the country having de-registered
7. Long dead

Model Coefficients Specification

- Parameters in the transition matrix are specified using logistic regression:

$$\mathrm{logit}(\theta) = \beta_0 + \underbrace{\beta_1 C_1 + \beta_2 C_2 +}_{\text{Individual- and time-dependent covariates C}} \; \dots + \underbrace{\epsilon}_{\text{Random effects}} \qquad \text{where } \epsilon \sim N(0, \sigma^2)$$

- Parameters in the observation matrix are specified using multinomial regression[4]:

$$X = \left[\underbrace{\begin{matrix} 1 & \cdots & 1 \\ 1 & \cdots & 1 \\ & \vdots & \\ 0 & \cdots & 0 \end{matrix}}_{\text{Registers}} \quad \underbrace{\begin{matrix} 1 & \cdots & 1 \\ 1 & \cdots & 0 \\ & \vdots & \\ 0 & \cdots & 0 \end{matrix}}_{\text{Covariates}} \quad \underbrace{\begin{matrix} 1 & \cdots & 1 \\ 1 & \cdots & 0 \\ & \vdots & \\ 0 & \cdots & 0 \end{matrix}}_{\text{Interactions}} \right]$$

$$p_j = \frac{\exp(\gamma X_{j*} + \epsilon_j)}{1 + \sum_h \exp(\gamma X_{h*} + \epsilon_h)} \quad \text{where } \epsilon \sim N(0, \sigma^2)$$

$$j = 1, \dots, J; \quad h = 1, \dots, J - 1$$

where p_j are the observation probabilities of each register combination

[4] A. Agresti. An introduction to categorical data analysis. John Wiley & Sons, 2nd edition, 2007.

Model Assumptions

- All individuals are initially alive and observed registering
- If an individual is observed, they must be in the country
- If an individual dies in the country, their death is recorded with probability 1
- If an individual de-registers when emigrating, then they must be observed re-registering the year they return
- Individuals have a maximum of one migration event each year
- Individuals must be present and observable for at least one year **after** initially entering

Application to Norway

(preliminary results are for 5% of the population)

Norwegian Register Data

- All foreign-born residents who first entered Norway between the years 2007 and 2021 as adults
- 548,092 individuals observed between the years 2008 and 2023

Registers:

- Change in marital status
- Birth of a child
- Obtaining Norwegian citizenship
- Internal moves recorded
- Main status in the labour market
- Individual income

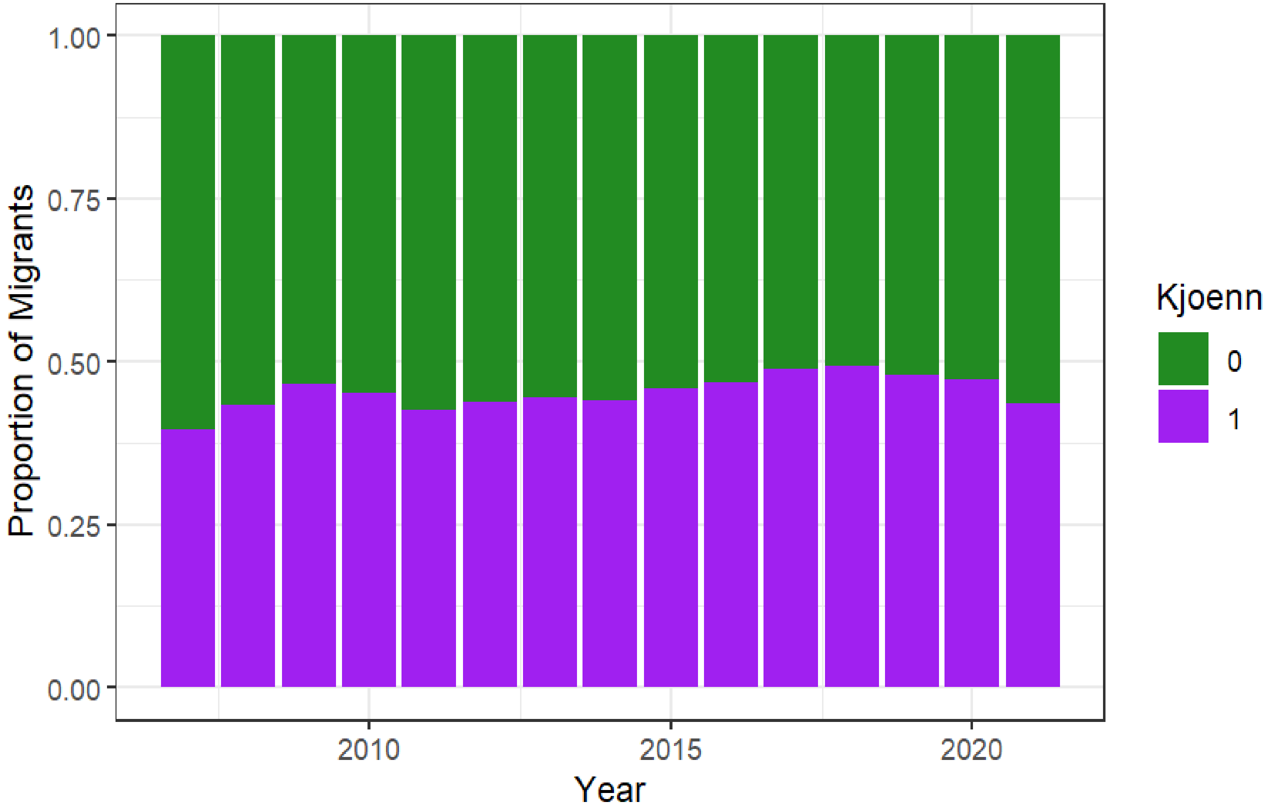
Covariates:

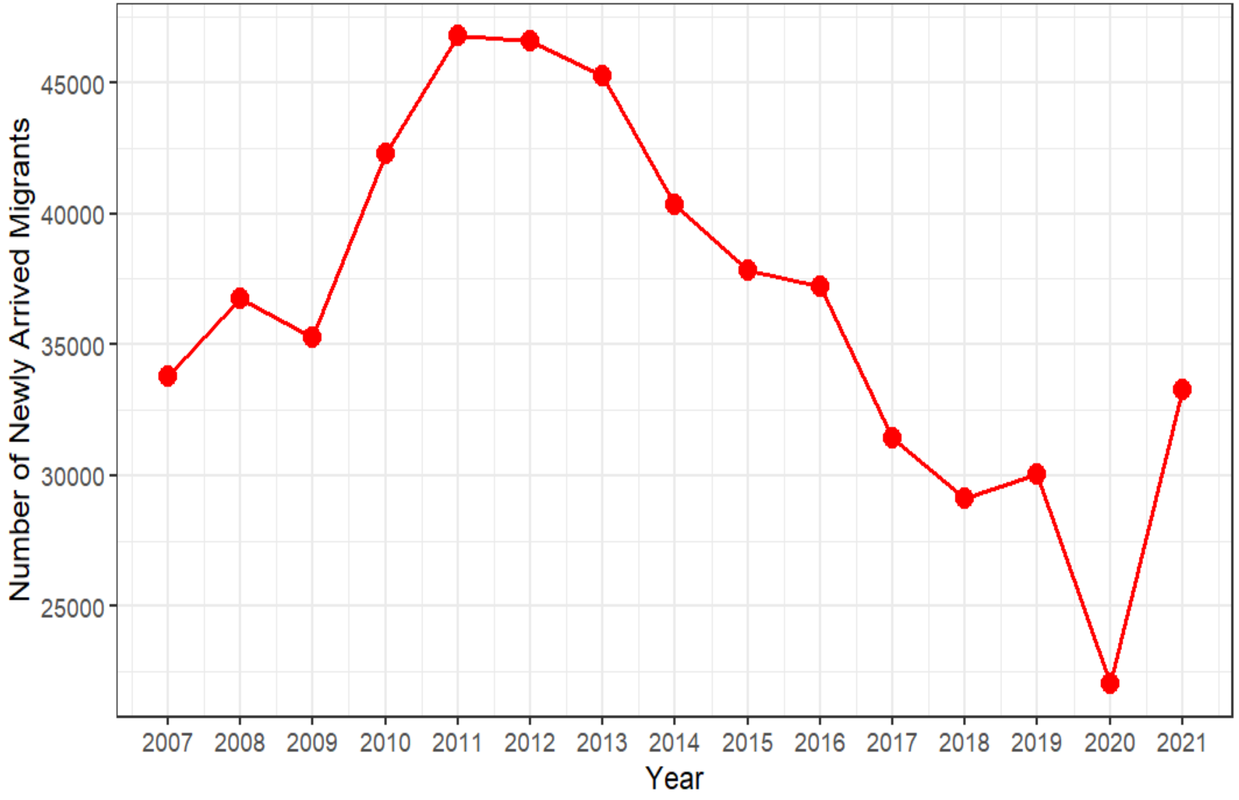
- Kjoenn = {0, 1}
- Age
  - 18 – 35 years old
  - 36 – 60 years old
  - 60+ years old
- Country of birth
  - CG1
  - CG2
  - CG3
  - Nordic

Possible additions include:

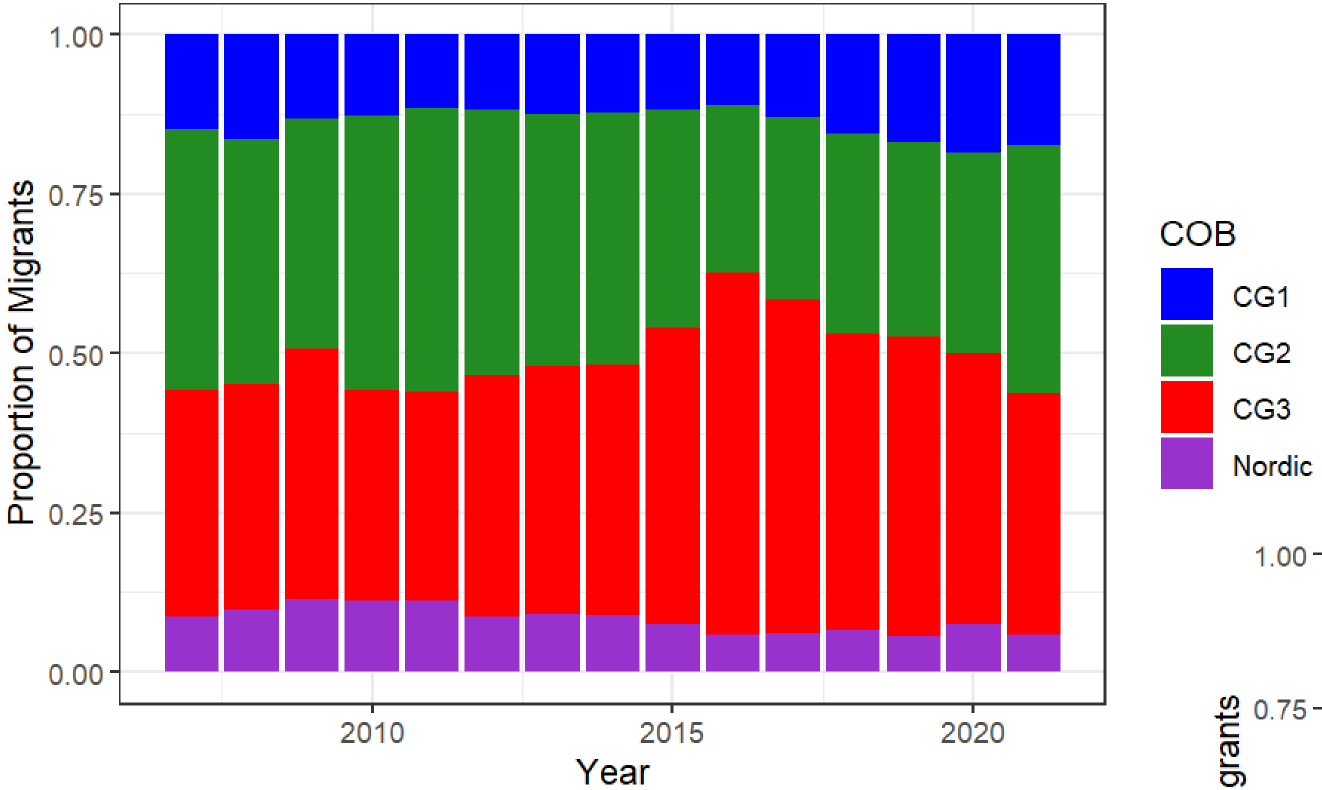
- Time since first entering the country (covariate)
- Type of migration (covariate)
- Adding “Secondary status in the labour market” (register)
- Splitting up “Labour market status” into employment, education, pension etc (register)

Distribution of Sex of Newly Arrived Migrants



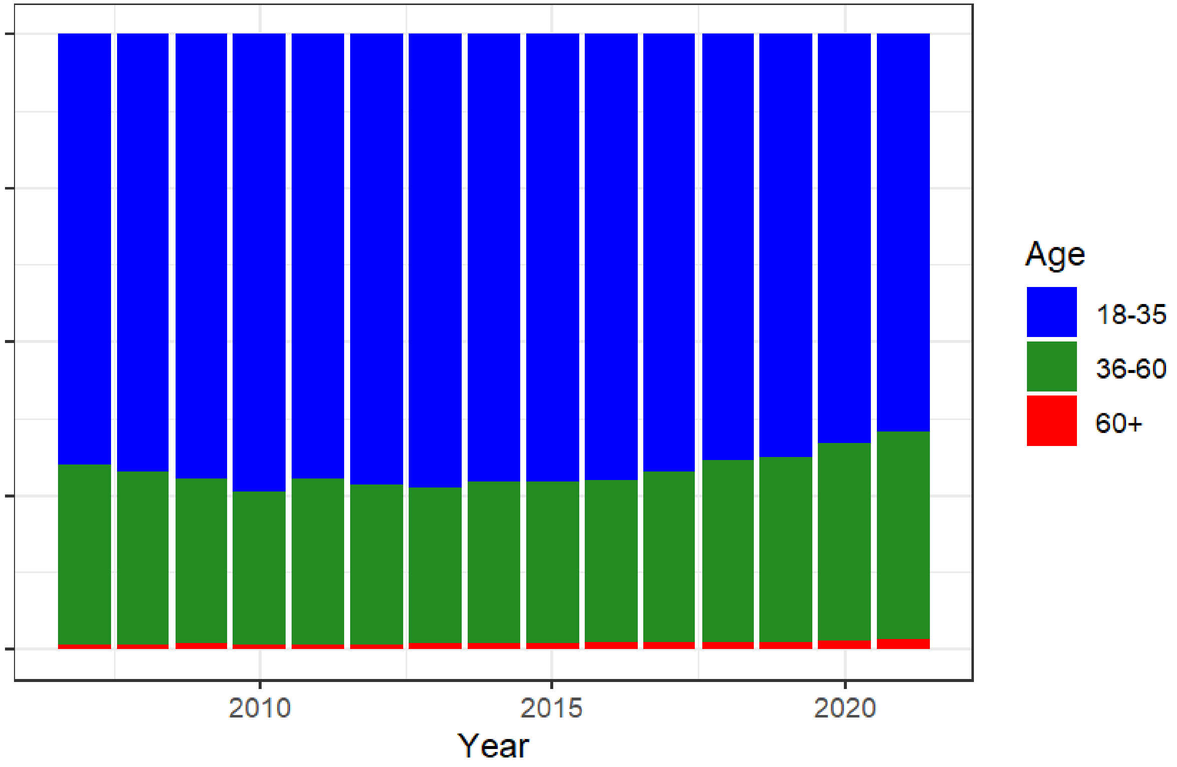


Number of Newly Arrived Migrants Each Year

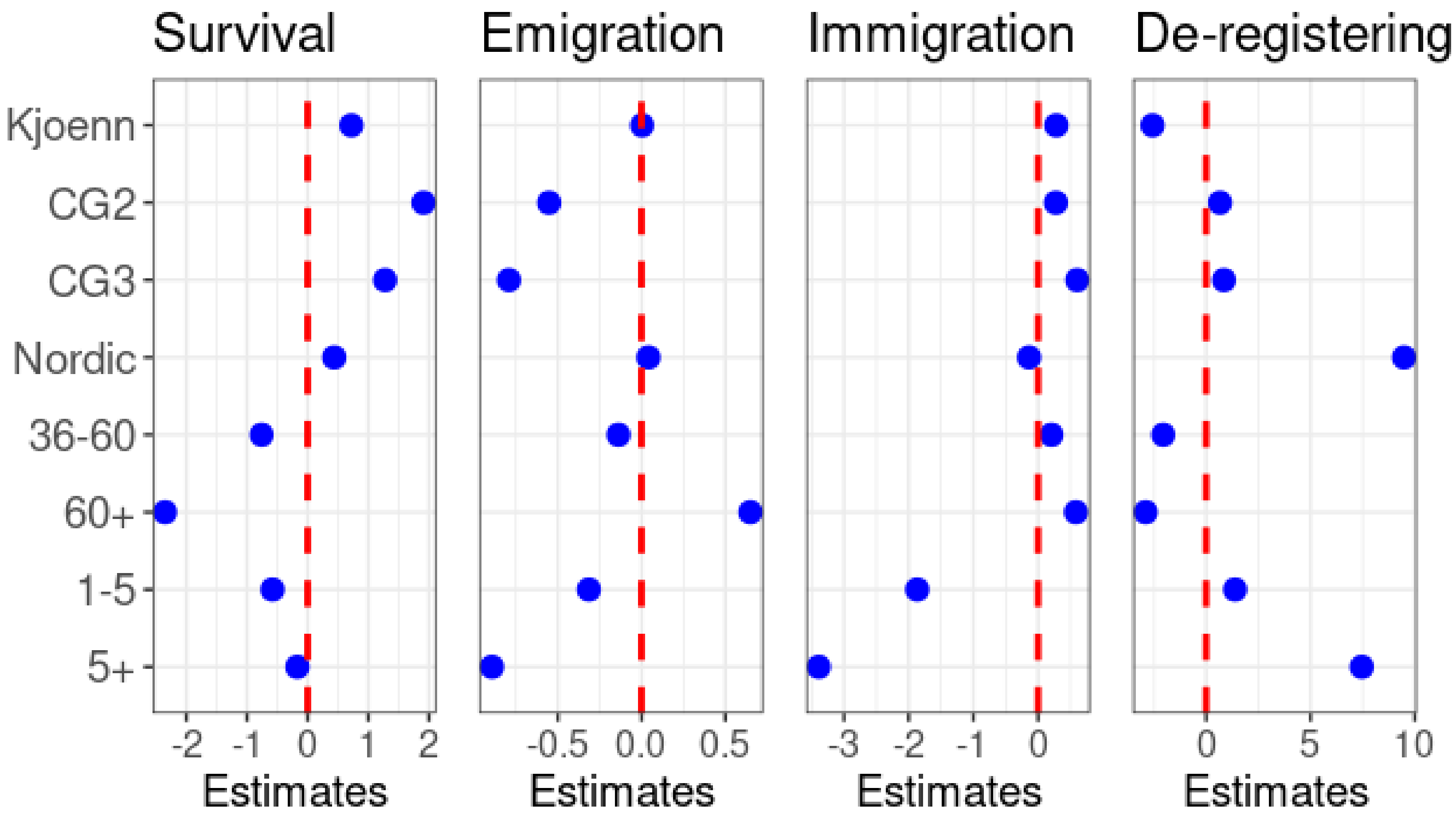


Distribution of Country of Birth of Newly Arrived Migrants

Distribution of Age of Newly Arrived Migrants



Life Event Probabilities



These estimates are on the logistic scale:

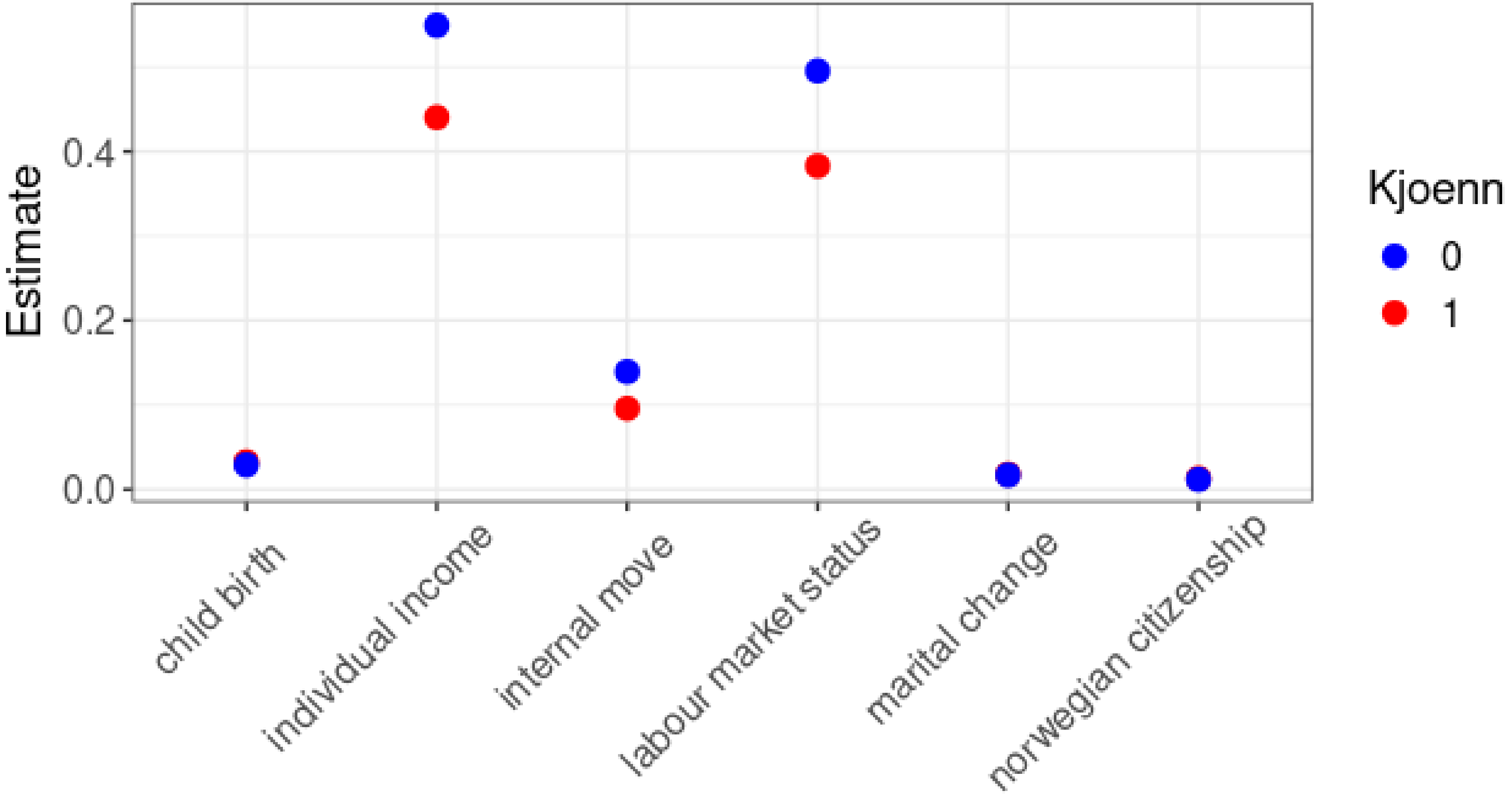
$$logit(\theta) = \beta_0 + \beta_1 * Kjoenn + \beta_2 * CG2 + \ ...$$

Baseline: individuals with kjoenn 0, aged between 18 and 35 who were born in country group CG1 and first entered the country that year

Baseline values:

s = 0.9986209
e = 0.1176378
i = 0.1953825
λ= 0.8000718

Register Observation Probabilities



Next Steps

- Finalise data and model
- Incorporate bootstrapping via Bag of Little Bootstraps [5] to obtain confidence intervals and improve computational time
- Viterbi Algorithm [6] to obtain the "optimal path" for each individual
  - Compute estimated population size for each year
- Estimate over-coverage

Additionally: Forward/Forward-Backward Algorithm [6] to obtain probability of presence in the country each year, for every individual

[5] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. Journal of the Royal Statistical Society Series B: Statistical Methodology, 76(4):795–816, 2014

[6] Daniel Jurafsky and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd edition, 2025. URL https://web.stanford.edu/ jurafsky/slp3/. Appendix A.

Thank you for your attention!

Email: lyb3@kent.ac.uk