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

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

Eleni Matechou[1] Bruno Santos[2] Eleonora Mussino[3]

1 University of Kent, UK
2 University of Lisbon, Portugal
3 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, pages 1–24, 2023

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, following individuals, and hence registers, over different years

Swedish Register Data

- Provided by the Swedish National Institute of Statistics (Statistics Sweden - SCB)
- All foreign-born residents who first entered Sweden between the years 2003 and 2016 as adults

(This is not real data)

Nine Registers

emp	stud	intmove	faminc	amf	child
0	0	0	0	1	0
1	0	0	1	0	0
1	0	0	1	0	0
1	0	0	1	0	0
1	0	0	1	0	0
0	1	0	0	0	0
0	1	0	0	0	0

Emigration, (Re-)Immigration and Death records

immig	emig	reimmig	marr	div	swecit
1	0	0	0	0	0
0	1	0	0	0	0
1	0	1	0	0	0
0	0	0	1	0	0
0	1	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

	id	year	age	cob	sex	death
1	XXX1	2003	27	France	1	0
2	XXX1	2004	28	France	1	0
3	XXX1	2007	31	France	1	0
4	XXX1	2008	32	France	1	0
5	XXX1	2009	33	France	1	0
6	XXX2	2014	68	India	0	0
7	XXX2	2015	69	India	0	1

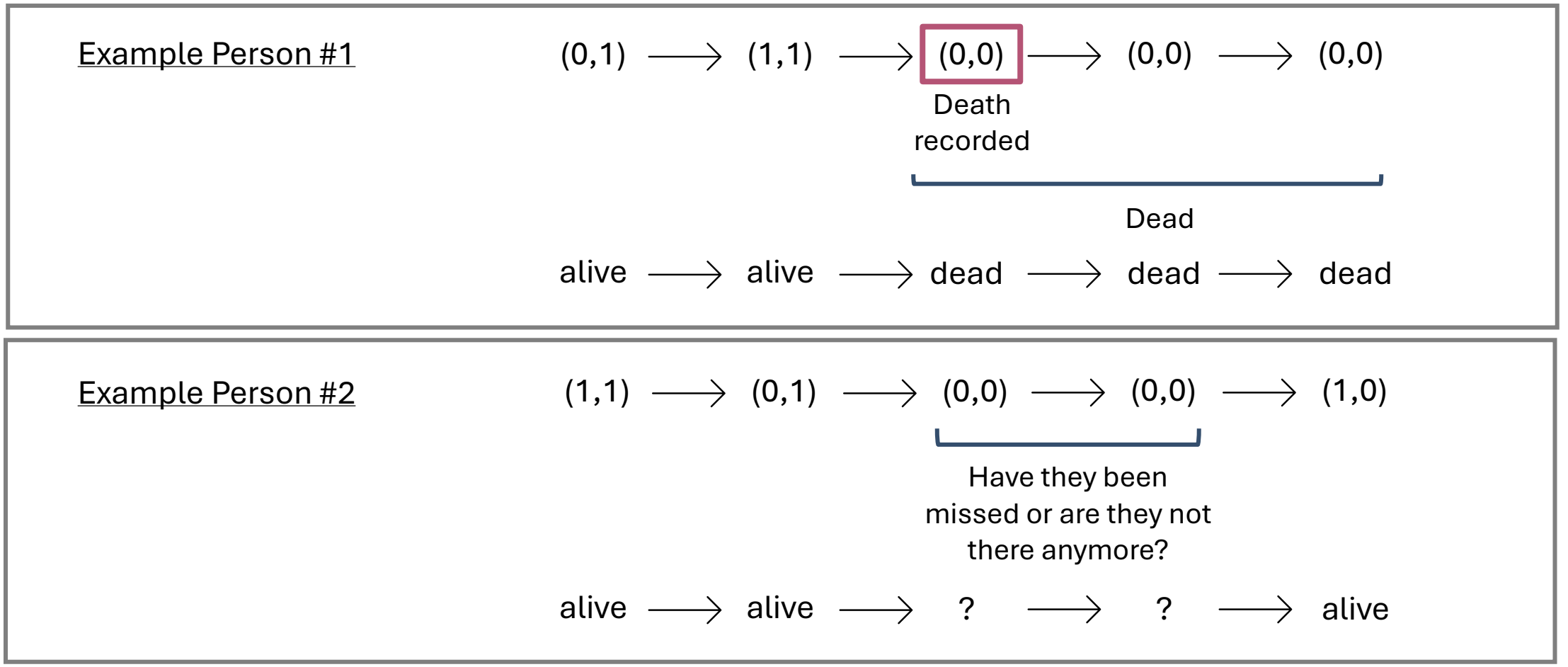
Covariates (treated as categorical):

- **Sex**
- **Country of birth**
- **Age**
- **Time Since first entering Sweden**

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

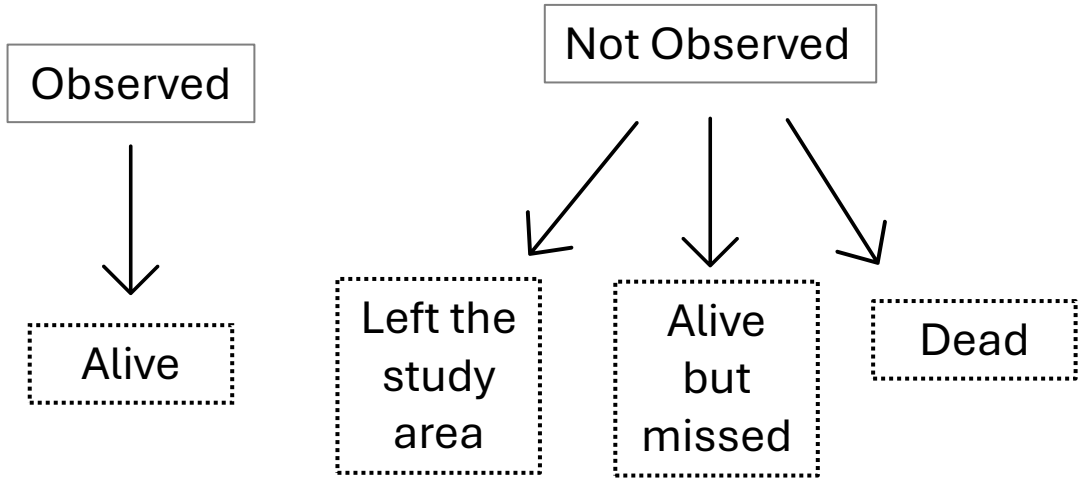
$\Longrightarrow$ 2^R register combinations



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



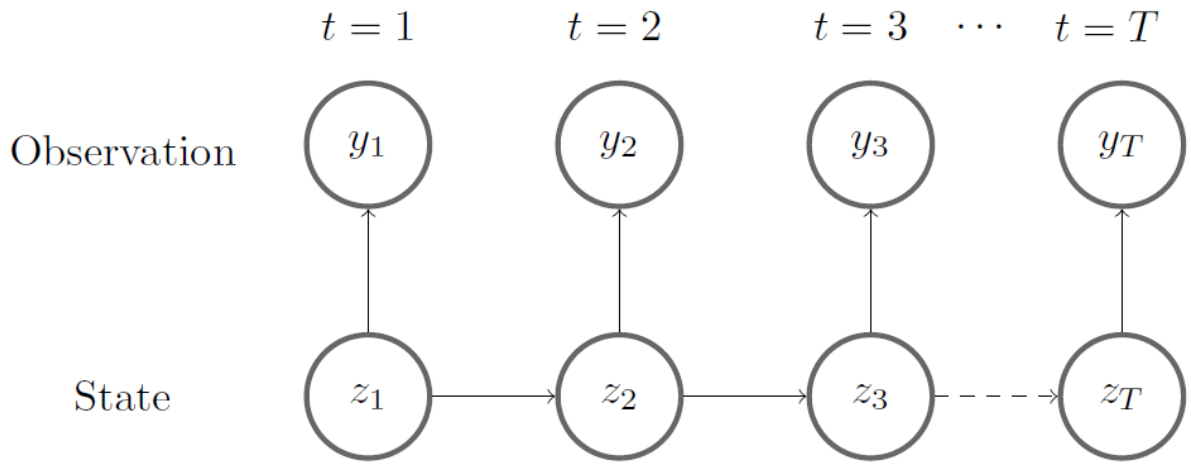
- CR models are commonly used in ecology to estimate population size for wild animals
- An individual's true state z_t is unobservable (latent)

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

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

Simple CR Model

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

- 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

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

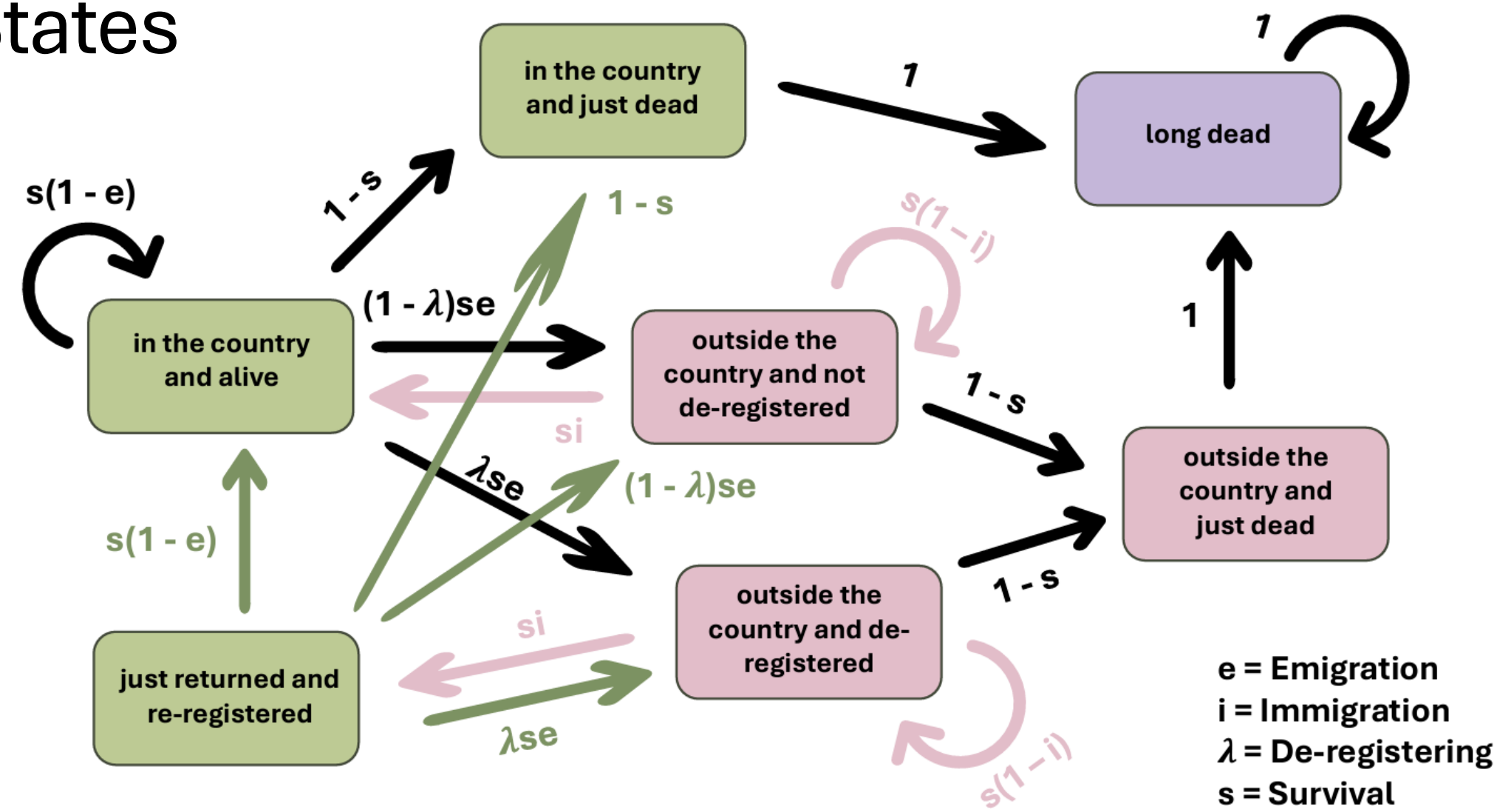
- All individuals are alive and observed when initially captured, i.e. initial state $\delta = [1,0]$

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


in the country and just dead
long dead
1
1
s(1 - e)
1 - s
1 - s
s(1 – i)
(1 - λ)se
in the country and alive
outside the country and not de-registered
1
si
1 - s
λse
(1 - λ)se
outside the country and just dead
s(1 - e)
1 - s
outside the country and de-registered
si
just returned and re-registered
λse
s(1 – i)
e = Emigration
i = 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:

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

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

$$X = \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \\ 1 & \cdots & 1 & 1 & \cdots & 0 & 1 & \cdots & 0 \\ & \vdots & & & \vdots & & & \vdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Registers | Covariates | Interactions

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

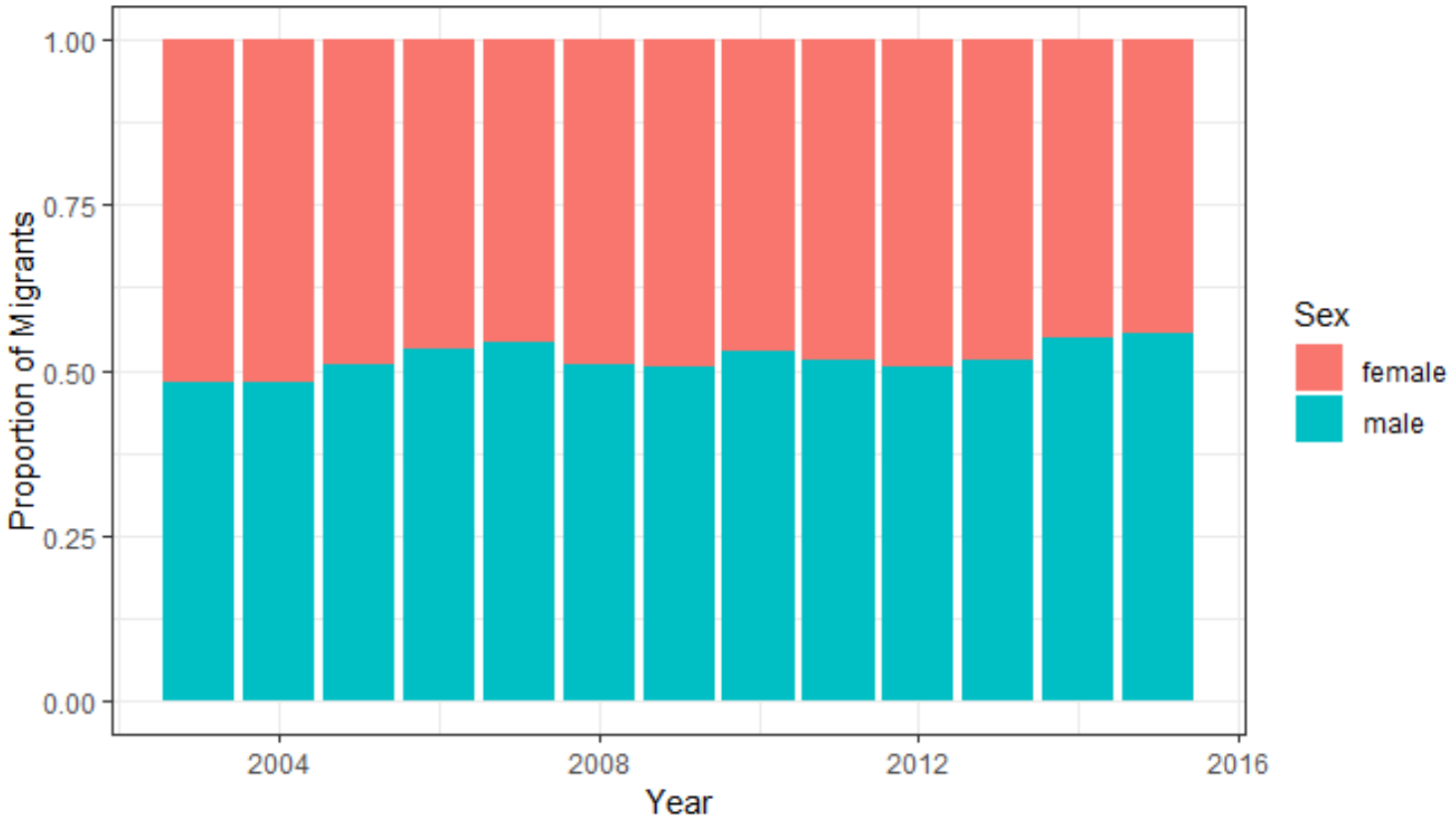
$$j = 1, \ldots, J; \quad h = 1, \ldots, J-1$$

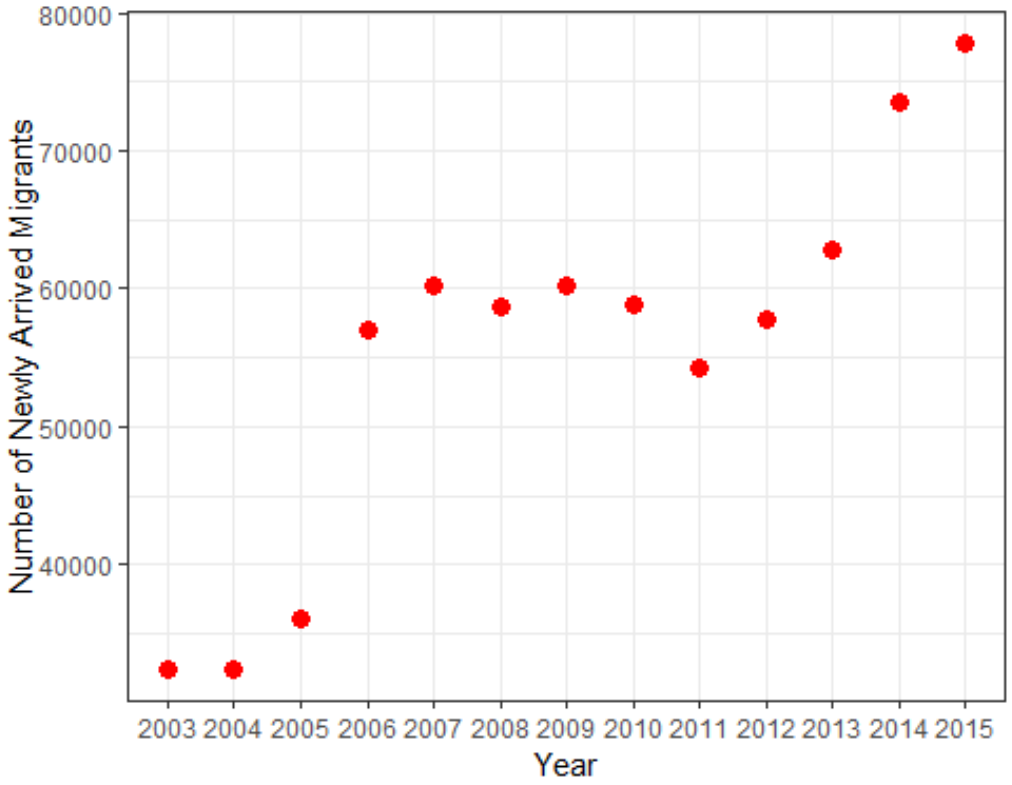
where p_j are the observation probabilities of each register combination

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

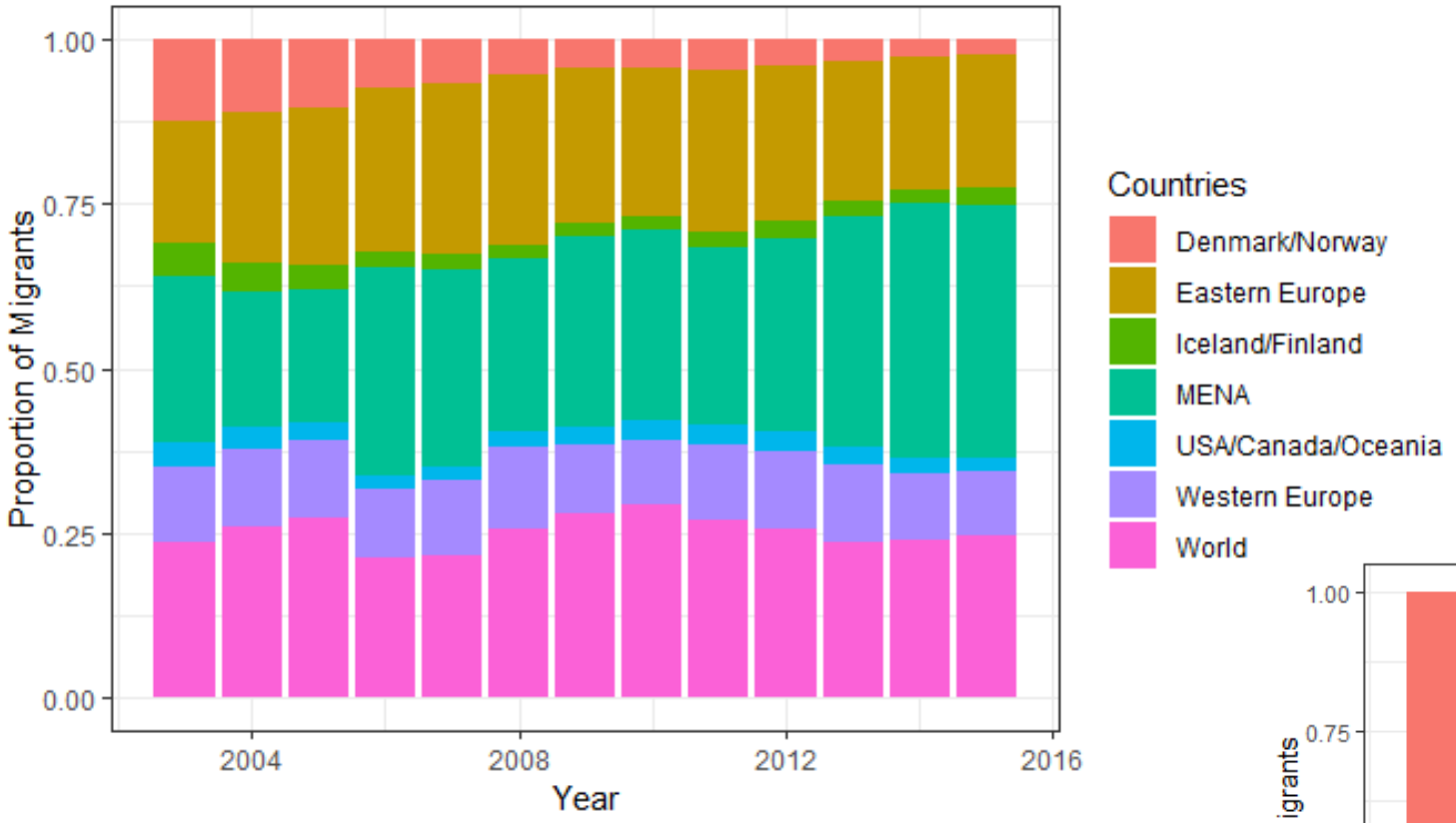
Results

Distribution of sex for newly arrived migrants



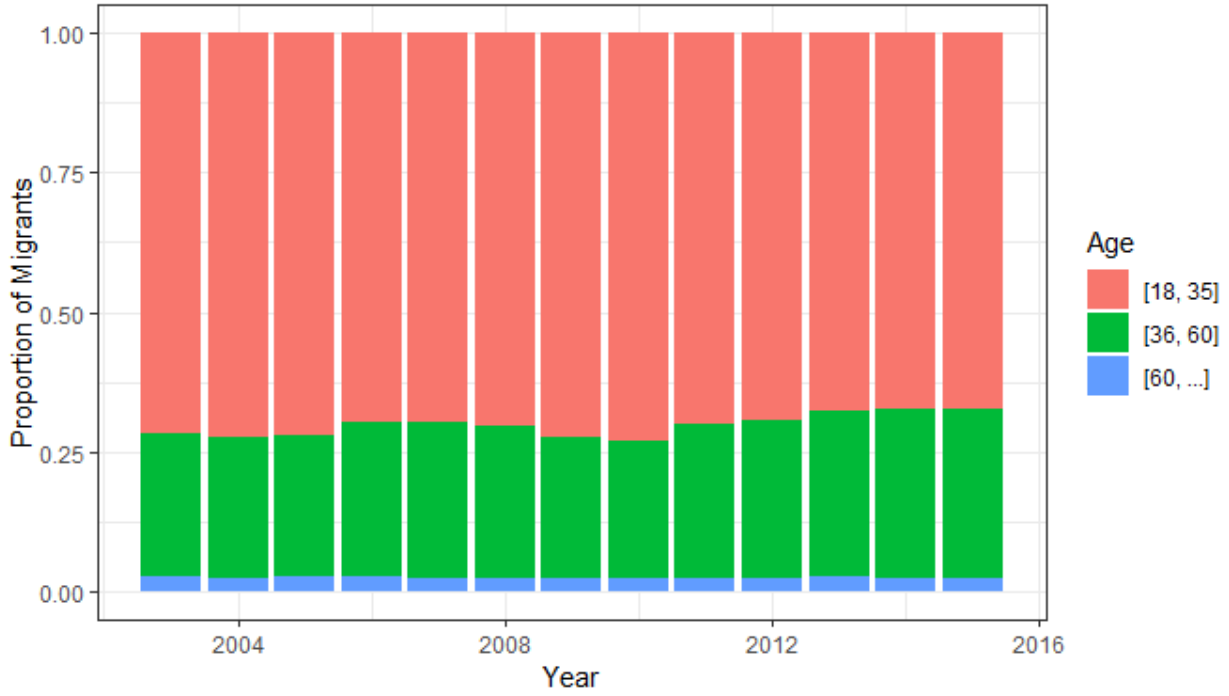


Number of newly arrived migrants

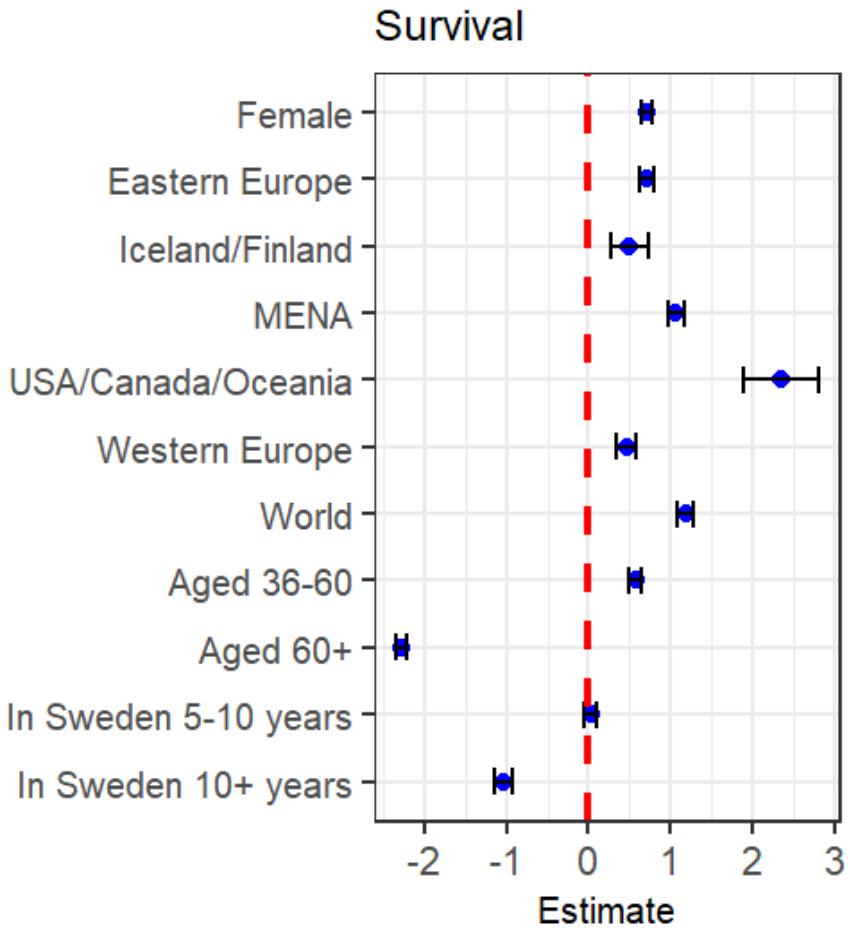


Distribution of country of birth for newly arrived migrants

Distribution of age for newly arrived migrants



Survival Probability



<u>Baseline:</u> men aged 18-35, born in Denmark/Norway who first entered Sweden less than 5 years ago

$$\text{logit}(\lambda) = 5.617 \quad \Rightarrow \quad \lambda = 0.996$$

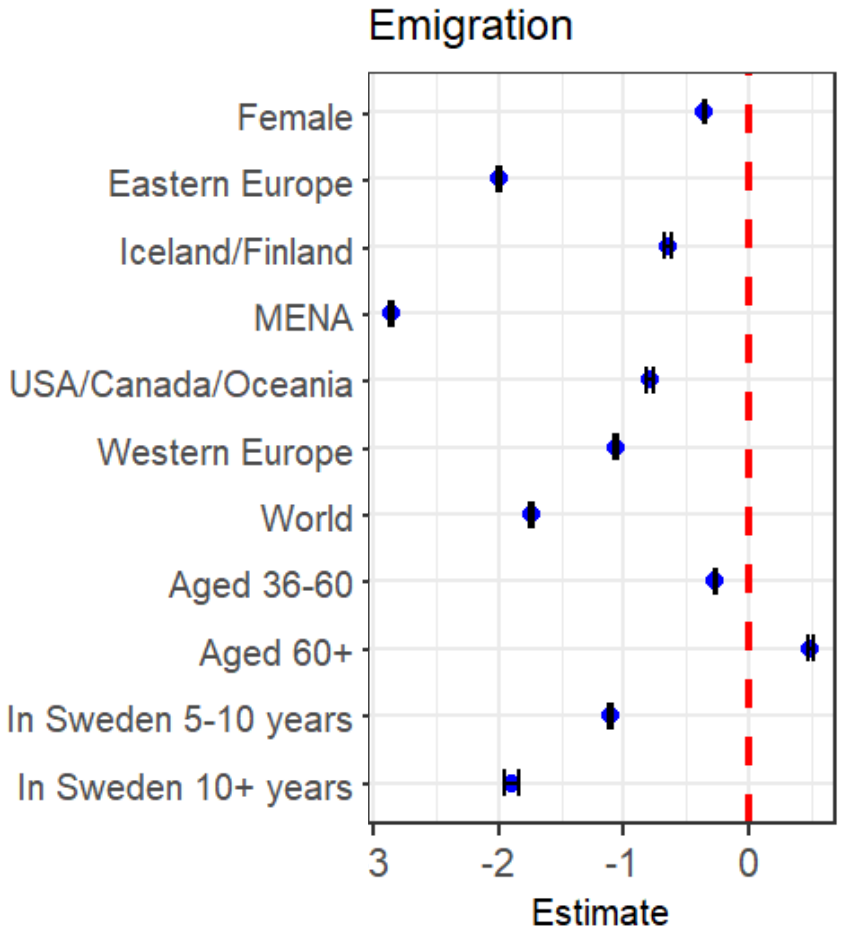
Women aged **36-60**, born in **USA/Canada/Oceania** who first entered Sweden **5-10 years ago**

$$\text{logit}(\lambda) = 5.617 + \mathbf{0.710} + \mathbf{2.340} + \mathbf{0.572} + \mathbf{0.023}$$
$$\Rightarrow \quad \lambda = 0.9999$$

Men **aged 60+**, born in Denmark/Norway who first entered Sweden **10+ years ago**

$$\text{logit}(\lambda) = 5.617 - \mathbf{2.272} - \mathbf{1.028}$$
$$\Rightarrow \quad \lambda = 0.910$$

Emigration Probability



<u>Baseline:</u> men aged 18-35, born in Denmark/Norway who first entered Sweden less than 5 years ago

$$\text{logit}(\lambda) = -0.427 \quad \Rightarrow \quad \lambda = 0.395$$

Men aged **60+**, born in Denmark/Norway who first entered Sweden less than 5 years ago

$$\text{logit}(\lambda) = -0.427 + \mathbf{0.485}$$

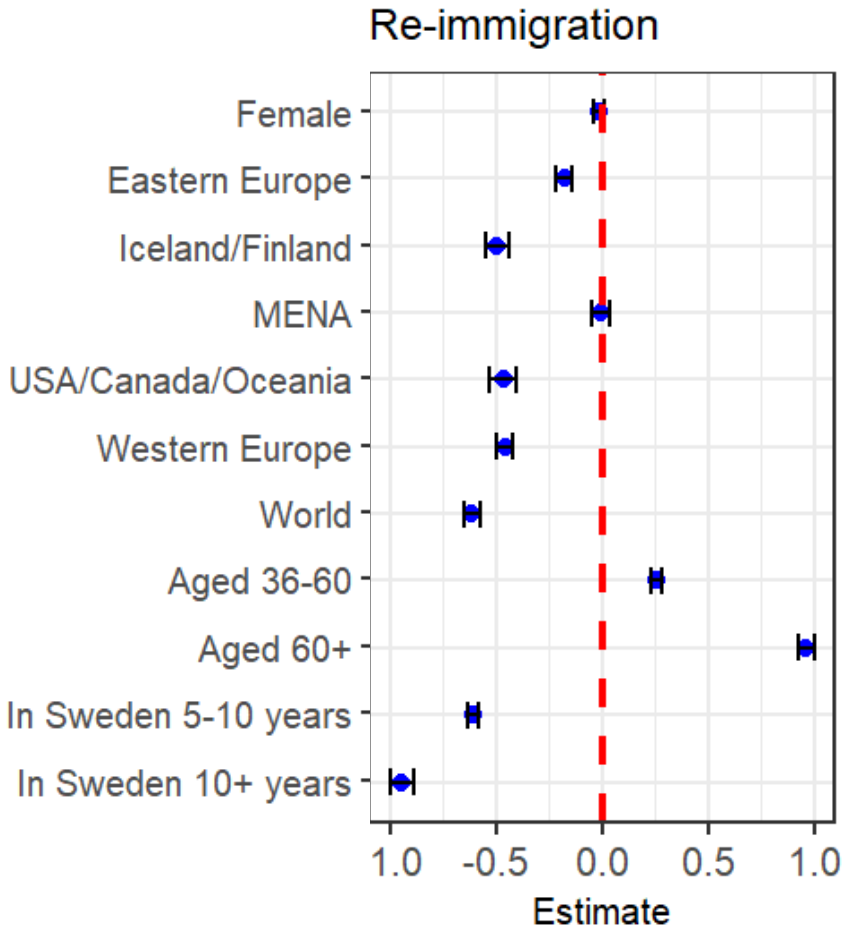
$$\Rightarrow \quad \lambda = 0.515$$

Women aged **36-60**, born in **MENA** who first entered Sweden **10+ years ago**

$$\text{logit}(\lambda) = -0.427 \; \mathbf{-0.356} \; \mathbf{-2.855} \; \mathbf{-0.273} \; \mathbf{-1.899}$$

$$\Rightarrow \quad \lambda = 0.00299$$

Re-Immigration Probability



<u>Baseline:</u> men aged 18-35, born in Denmark/Norway who first entered Sweden less than 5 years ago

$$\text{logit}(\lambda) = -2.978 \quad \Rightarrow \quad \lambda = 0.0532$$

Men aged **60+**, born in Denmark/Norway who first entered Sweden less than 5 years ago

$$\text{logit}(\lambda) = -2.978 + \mathbf{0.963}$$

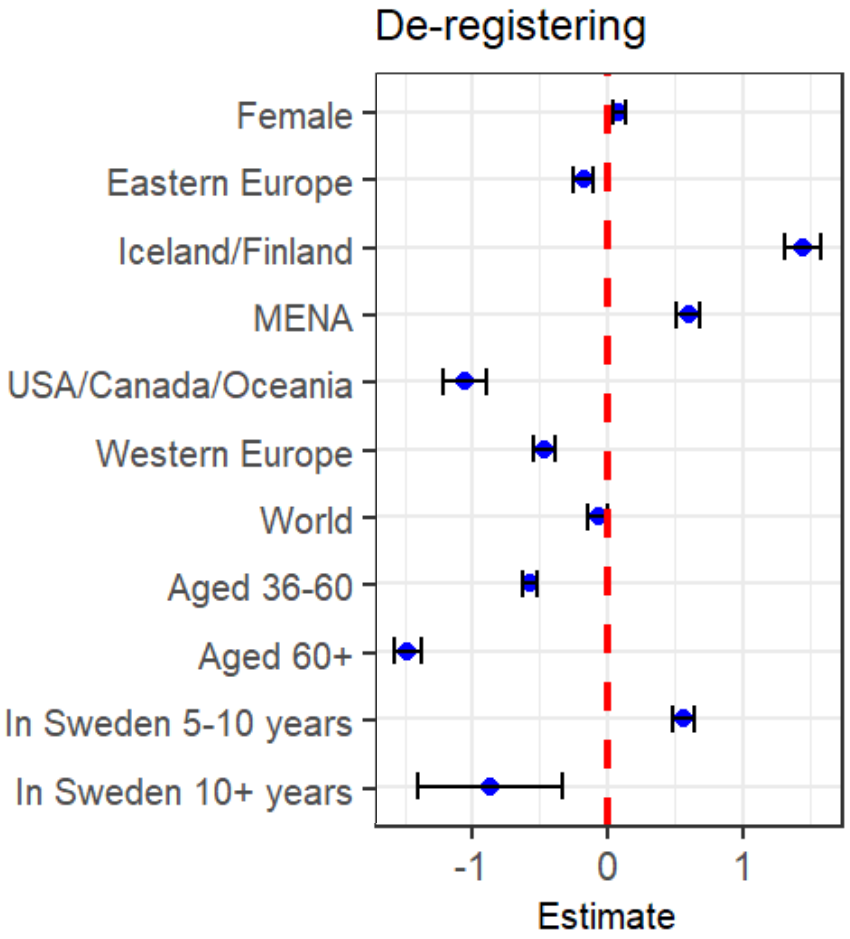
$$\Rightarrow \quad \lambda = 0.128$$

Women aged 18-35, born in **the rest of the World** who first entered Sweden **10+ years ago**

$$\text{logit}(\lambda) = -2.978 - 0.0159 - 0.613 - 0.946$$

$$\Rightarrow \quad \lambda = 0.0115$$

De-registering Probability



Baseline: men aged 18-35, born in Denmark/Norway who first entered Sweden less than 5 years ago

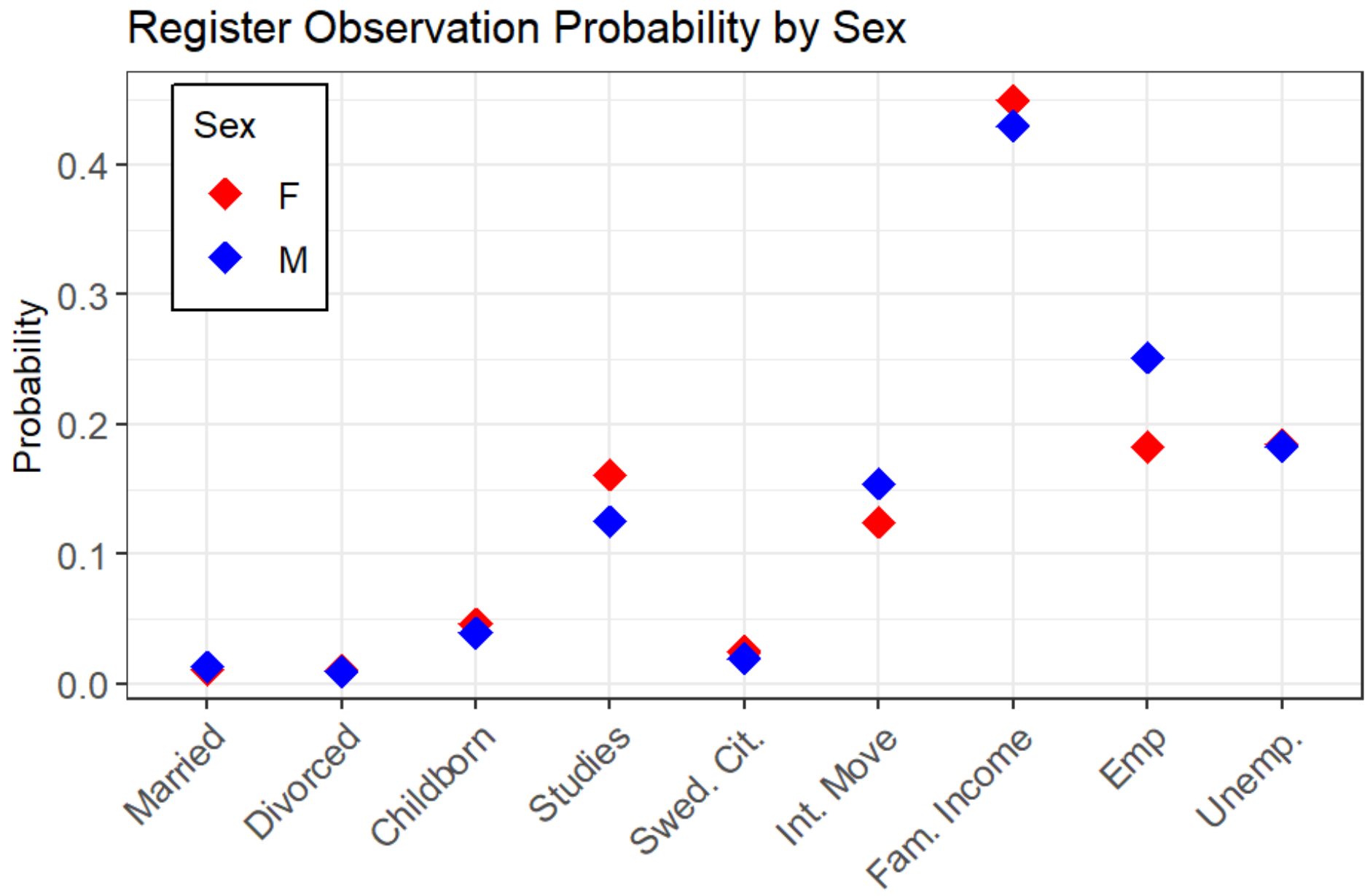
$$\text{logit}(\lambda) = -0.332 \quad \Rightarrow \quad \lambda = 0.418$$

Women aged 18-35, born in **Iceland/Finland** who first entered Sweden **5-10 years ago**

$$\text{logit}(\lambda) = -0.332 + \mathbf{0.088} + \mathbf{1.452} + \mathbf{0.568}$$
$$\Rightarrow \quad \lambda = 0.855$$

Men **aged 60+**, born in **USA/Canada/Oceania** who first entered Sweden **10+ years ago**

$$\text{logit}(\lambda) = -0.332 - \mathbf{1.053} - \mathbf{1.474} - \mathbf{0.865}$$
$$\Rightarrow \quad \lambda = 0.024$$


Register Observation Probability by Sex
Sex
F
M
Probability
0.0
0.1
0.2
0.3
0.4
Married
Divorced
Childborn
Studies
Swed. Cit.
Int. Move
Fam. Income
Emp
Unemp.

Register Combination Probabilities

Highest Probability Register Combinations	Male	Female
Family Income and Employment	0.0965	0.0762
Family Income	0.0368	0.0611
Family Income and Active Unemployment	0.0305	0.0388

Lowest Probability Register Combinations	Male	Female
All except Internal Move and Family Income	3.6107 x 10^-12	2.3061 x 10^-12
All except Family Income	6.2981 x 10^-12	2.3061 x 10^-12
All except Family Income and Active Unemployment	1.0539 x 10^-11	6.1599 x 10^-12

Probability of being unobserved

- Male: 0.0251
- Female: 0.0220

→

Probability of being observed in at least one register

- Male: 0.9749
- Female: 0.9780

Future Work

★ Calculate population size estimates using the Viterbi Algorithm

★ Considering further extensions such as dependence between family units

★ Application to equivalent data from Norway, provided by Statistics Norway

- Additional complexity that checks are regularly done and individuals manually removed if not present in the country

Thank you for your attention!

Email: lyb3@kent.ac.uk