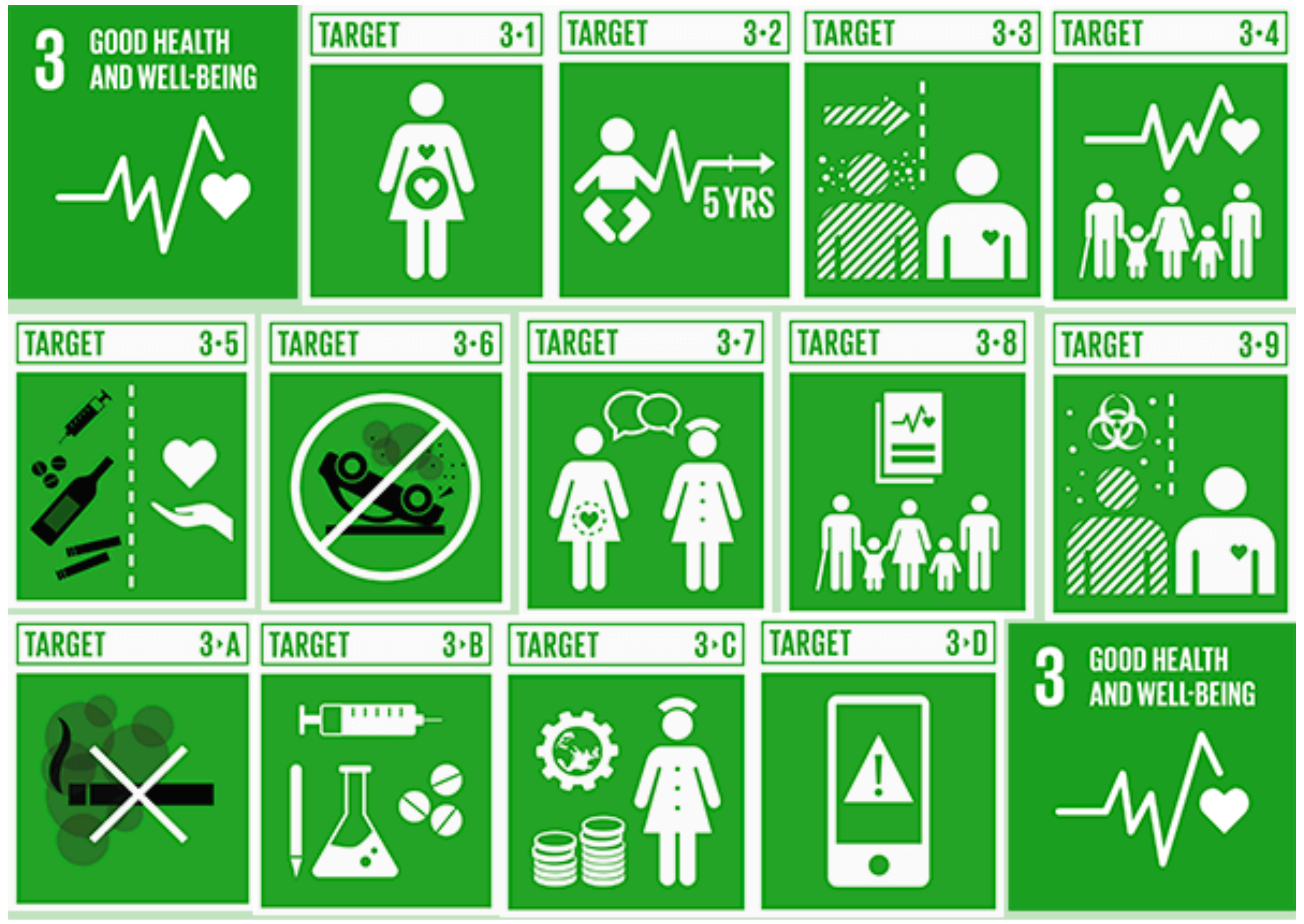
Balancing different information and modelling decisions in the Bayesian estimation of demographic quantities

Marija Pejchinovska (1) and Monica Alexander (1,2)
(1) Statistical Sciences
(2) Sociology
University of Toronto

‘Data science: Estimation and forecasting’, BSPS 2023

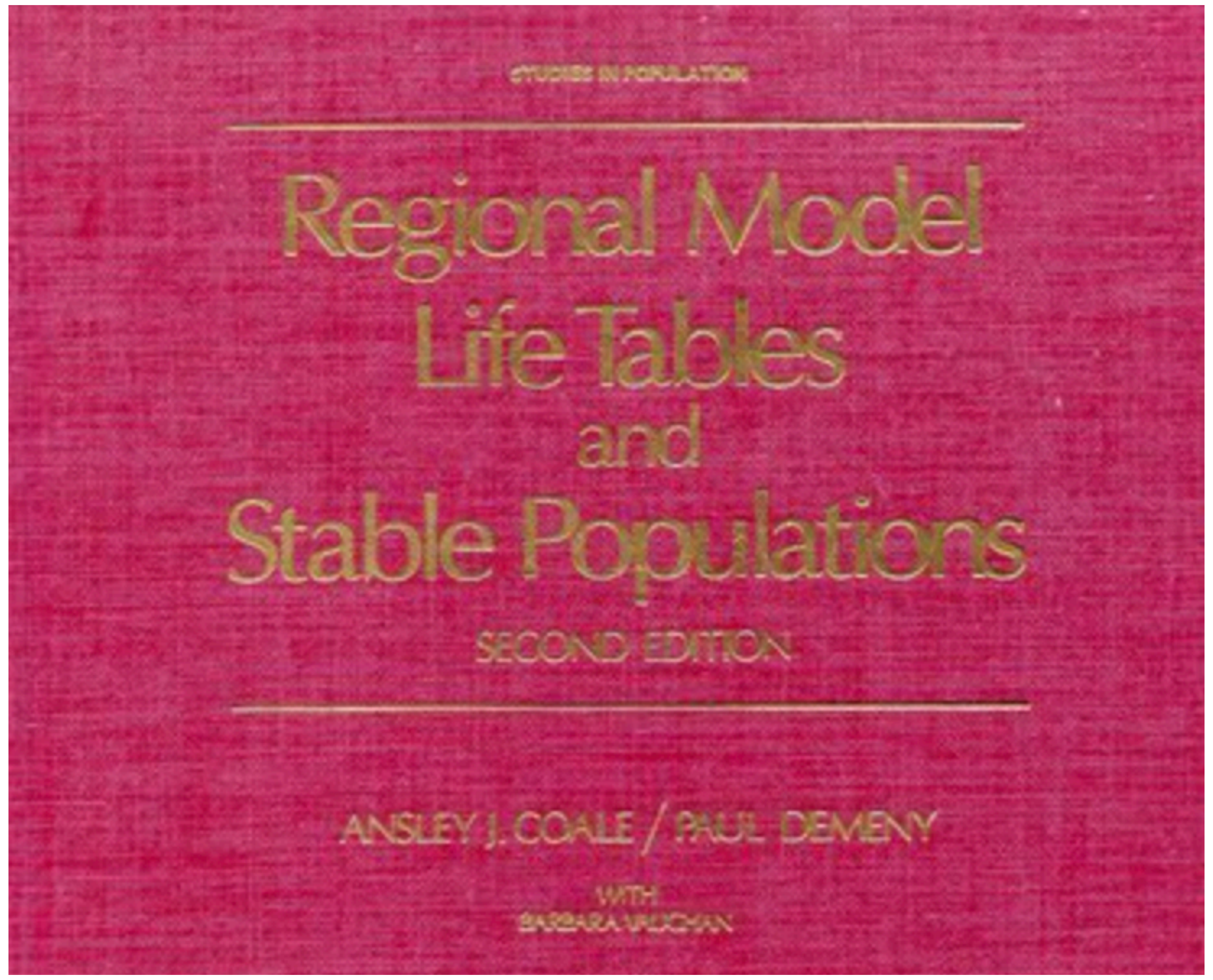
Background

- We are in the ‘Sustainable Development Goals’ era
- Detailed and specific health and development goals for all UN member countries, to be reached by 2030
- Specific targets require monitoring of key indicators on an increasingly granular spatial and temporal scale
- But this is challenging because of lack of available data, particularly in LMICs



Increased demand for timely estimates and projections

- We need methods of obtaining estimates of outcomes of interest in data sparse contexts
- Long history of this in demography!
- (Deterministic) Inferences about populations with limited data, based on systematic empirical patterns in populations with high-quality data
- But recent efforts have focused more on data-driven statistical approaches, utilizing what limited data we have
- A range of approaches, based on different research groups, philosophies



Motivation for this project

- How sensitive are estimates and projections to different modelling decisions and information sources, and when does it matter the most?
- How can we better communicate and quantify uncertainty in model choice?
- How can we systematize and validate model choice?

The rest of this presentation

- Explore these ideas with a motivating example: estimating the neonatal mortality rate in all countries worldwide

- Review existing UN model, suggest sensible alternatives, motivated by both data-driven patterns and previously observed empirical patterns

- Quantify sensitivity of estimates to model choice

- Discuss implications, next steps

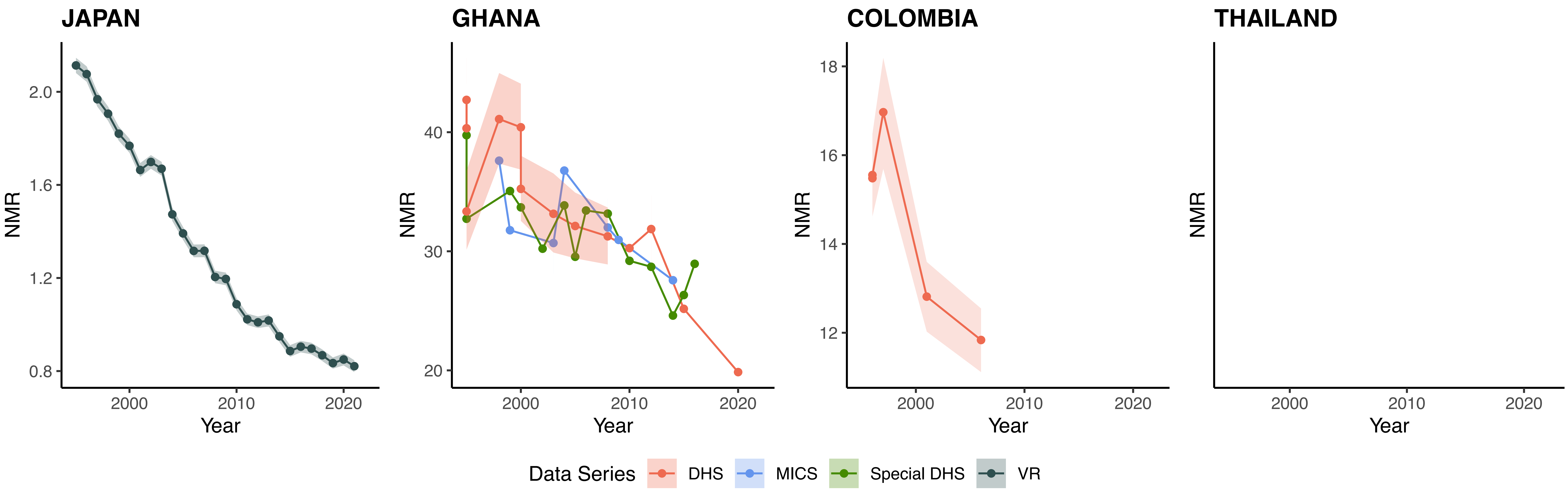
Work in progress! More questions than answers at this point!

Estimating the neonatal mortality rate in all countries worldwide

Neonatal mortality rate (NMR)

- Deaths in the first 28 days of life (per 1000 live births)
- Part of SDG goal 3.2 (<12 deaths/1000)
- UNICEF mandated to produce country-specific estimates and projections to 2030 annually
- Collate data based on civil registration and vital statistics systems, surveys (DHS, MICS)
- Current approach is a Bayesian penalized splines regression model, developed by Alexander and Alkema (2018)

Data contexts



Current model: key points

- We model the log ratio of neonatal mortality to other child mortality (1-59 months), $\log R_{c,t}$
- **Data model** relates the observed ratio r_i to the 'true' ratio $R_{c[i],t[i]}$, allowing for different observed standard errors based on data source
- **Process model:** $\log R_{c,t} = \log(f(U_{c,t})) + \log(P_{c,t})$

Current model: key points

- We model the log ratio of neonatal mortality to other child mortality (1-59 months), $\log R_{c,t}$
- **Data model** relates the observed ratio r_i to the 'true' ratio $R_{c[i],t[i]}$, allowing for different observed standard errors based on data source
- **Process model:** $\log R_{c,t} = \log(f(U_{c,t})) + \log(P_{c,t})$

'Expected' log ratio dictated by some function of U5MR

Current model: key points

- We model the log ratio of neonatal mortality to other child mortality (1-59 months), $\log R_{c,t}$
- **Data model** relates the observed ratio r_i to the 'true' ratio $R_{c[i],t[i]}$, allowing for different observed standard errors based on data source
- **Process model:** $\log R_{c,t} = \log(f(U_{c,t})) + \log(P_{c,t})$

Country-specific effect, modelled with first-order penalized B-Splines

Current model: key points

- We model the log ratio of neonatal mortality to other child mortality (1-59 months), $\log R_{c,t}$

- **Data model** relates the observed ratio r_i to the 'true' ratio $R_{c[i],t[i]}$, allowing for different observed standard errors based on data source

- **Process model:** $\log R_{c,t} = \boxed{\log(f(U_{c,t}))} + \log(P_{c,t})$

Country-specific effect, modelled with first-order penalized B-Splines

Relationship between ratio and U5MR (Model 0)

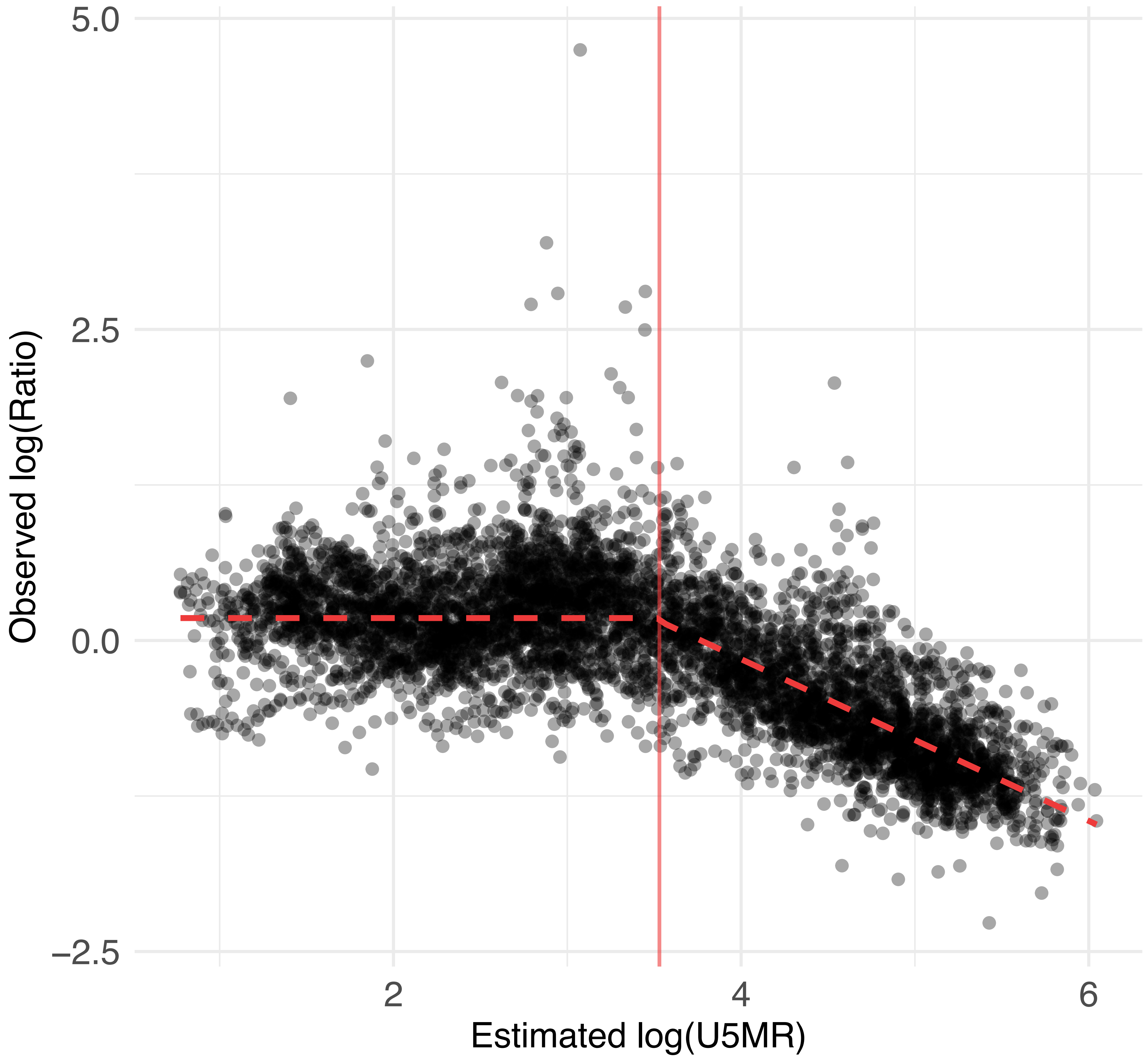
- Current model is

$\log(f(U_{c,t})) = \beta_0$ if $\log U_{c,t} \leq \theta$ and

$\log(f(U_{c,t})) = \beta_0 + \beta_1 \log U_{c,t}$ if $\log U_{c,t} > \theta$

- Where $U_{c,t}$ is the under-five mortality rate in that country and year

- Motivated by scatter plot of available data

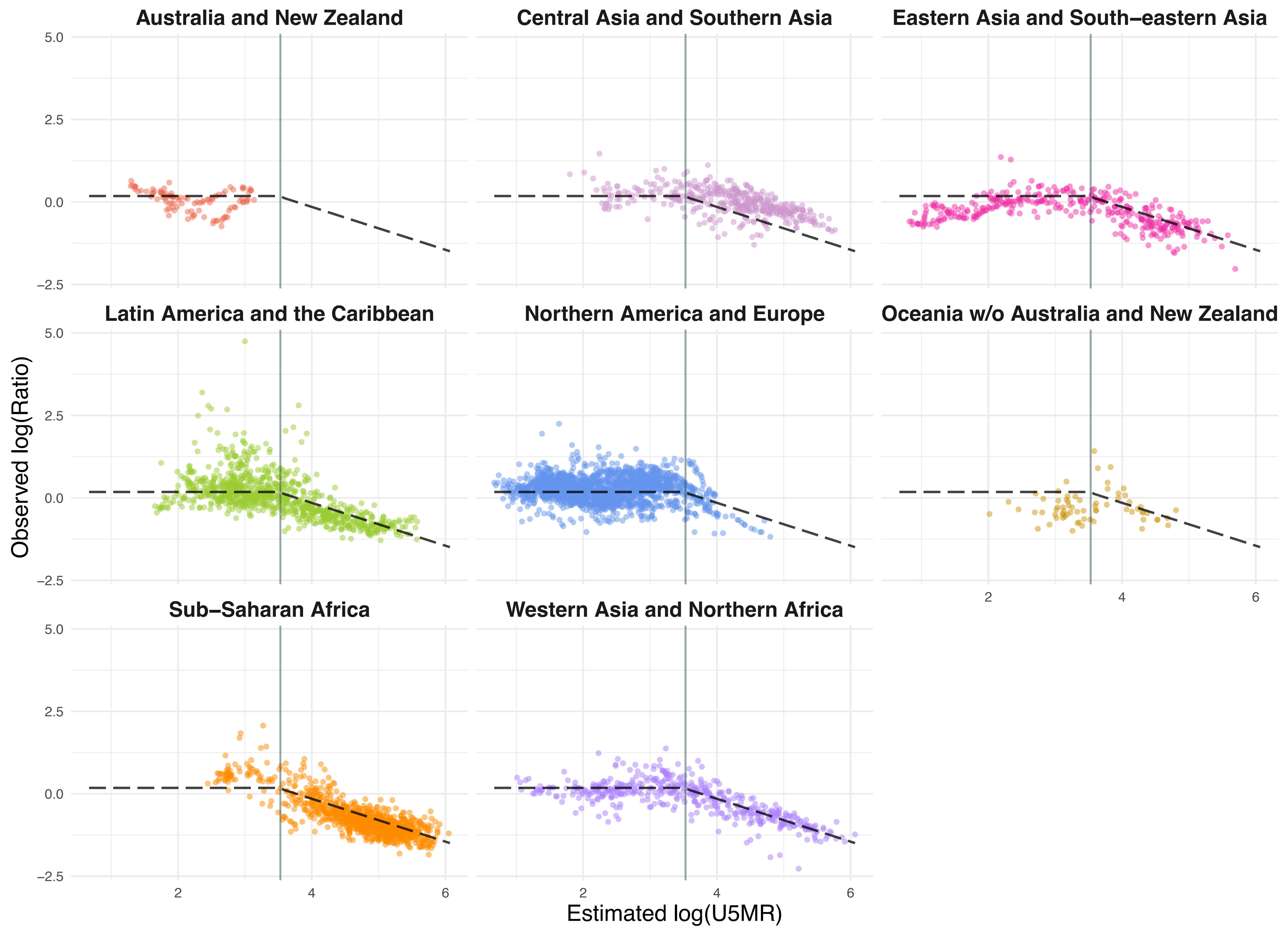


But wait

Australia and New Zealand
Central Asia and Southern Asia
Eastern Asia and South–eastern Asia
Latin America and the Caribbean
Northern America and Europe
Oceania w/o Australia and New Zealand
Sub–Saharan Africa
Western Asia and Northern Africa
Observed log(Ratio)
Estimated log(U5MR)
5.0
2.5
0.0
−2.5
2
4
6

A data-driven alternative (Model 1)

- Allow for systematic regional differences
- Expected level is a simple linear relationship that varies by region
- $\log(f(U_{c,t})) = \alpha_r + \beta_r \log U_{c,t}$



An alternative based on prior empirical observations

- Inspired by recent work by Guillot et al. (2022)
- Based on highly granular dataset on child mortality by age compiled from 25 countries over the years 1841–2016
- Developed ‘log-quad’ model which relates child mortality at various ages:

$$\log q_x = a_x + b_x \log q_5 + c_x \log q_5^2 + e_x$$

- Estimated coefficient values for mortality at different ages are provided in their paper

Table 2 Coefficients of the log-quadratic model estimated with the final U5MD, by sex and for both sexes combined

	Females				Males				Both Sexes Combined			
	a_x	b_x	c_x	v_x	a_x	b_x	c_x	v_x	a_x	b_x	c_x	v_x
7d	–3.6874	–0.3064	–0.1462	–0.4771	–3.2265	–0.1892	–0.1425	–0.4727	–3.4443	–0.2496	–0.1451	–0.4766
14d	–3.0879	–0.0624	–0.1165	–0.4244	–2.7078	0.0330	–0.1134	–0.4238	–2.8860	–0.0154	–0.1155	–0.4252
21d	–2.6890	0.1033	–0.0968	–0.3918	–2.3603	0.1846	–0.0944	–0.3931	–2.5139	0.1435	–0.0960	–0.3932
28d	–2.4645	0.1925	–0.0864	–0.3673	–2.1500	0.2725	–0.0835	–0.3699	–2.2961	0.2325	–0.0853	–0.3693
2m	–1.9445	0.3793	–0.0653	–0.2860	–1.6300	0.4729	–0.0594	–0.2907	–1.7720	0.4287	–0.0624	–0.2883
3m	–1.7128	0.4418	–0.0591	–0.2310	–1.4171	0.5317	–0.0532	–0.2338	–1.5505	0.4892	–0.0562	–0.2318
4m	–1.5420	0.4857	–0.0551	–0.1926	–1.2680	0.5695	–0.0494	–0.1940	–1.3920	0.5297	–0.0523	–0.1923
5m	–1.3830	0.5344	–0.0501	–0.1663	–1.1330	0.6115	–0.0448	–0.1650	–1.2457	0.5752	–0.0475	–0.1643
6m	–1.2361	0.5824	–0.0453	–0.1461	–1.0026	0.6566	–0.0398	–0.1449	–1.1068	0.6222	–0.0425	–0.1442
7m	–1.1008	0.6282	–0.0406	–0.1311	–0.8833	0.6995	–0.0352	–0.1291	–0.9801	0.6666	–0.0378	–0.1287
8m	–0.9867	0.6671	–0.0367	–0.1190	–0.7805	0.7374	–0.0310	–0.1167	–0.8718	0.7052	–0.0337	–0.1164
9m	–0.8881	0.7011	–0.0332	–0.1079	–0.6904	0.7711	–0.0272	–0.1068	–0.7770	0.7396	–0.0300	–0.1062
10m	–0.7996	0.7325	–0.0299	–0.0998	–0.6133	0.7998	–0.0241	–0.0980	–0.6948	0.7695	–0.0268	–0.0978
11m	–0.7223	0.7603	–0.0269	–0.0923	–0.5478	0.8246	–0.0213	–0.0911	–0.6237	0.7959	–0.0239	–0.0905
12m	–0.6532	0.7854	–0.0243	–0.0863	–0.4867	0.8482	–0.0187	–0.0854	–0.5591	0.8202	–0.0212	–0.0846
15m	–0.4909	0.8439	–0.0181	–0.0710	–0.3465	0.9020	–0.0126	–0.0709	–0.4086	0.8764	–0.0151	–0.0698
18m	–0.3833	0.8835	–0.0136	–0.0601	–0.2600	0.9347	–0.0087	–0.0598	–0.3126	0.9123	–0.0109	–0.0588
21m	–0.3063	0.9119	–0.0103	–0.0514	–0.2031	0.9557	–0.0060	–0.0509	–0.2468	0.9367	–0.0079	–0.0500
2y	–0.2444	0.9335	–0.0078	–0.0433	–0.1565	0.9719	–0.0040	–0.0431	–0.1933	0.9554	–0.0057	–0.0423
3y	–0.1202	0.9696	–0.0037	–0.0207	–0.0660	0.9954	–0.0010	–0.0225	–0.0885	0.9844	–0.0022	–0.0216
4y	–0.0432	0.9912	–0.0011	–0.0085	–0.0266	0.9992	–0.0003	–0.0087	–0.0333	0.9958	–0.0007	–0.0086
5y	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000

An alternative based on prior empirical observations (Model 2)

Use the log-quad model to inform the functional form of the model but also the priors on coefficients:

$$\log(f(U_{c,t})) = \alpha_r + \beta_r \log U_{c,t} + \gamma_r \log U_{c,t}^2$$

With

$$\alpha_r \sim N(a_{28} + v_{28}, 2^2)$$

$$\beta_r \sim N(b_{28}, 0.5^2)$$

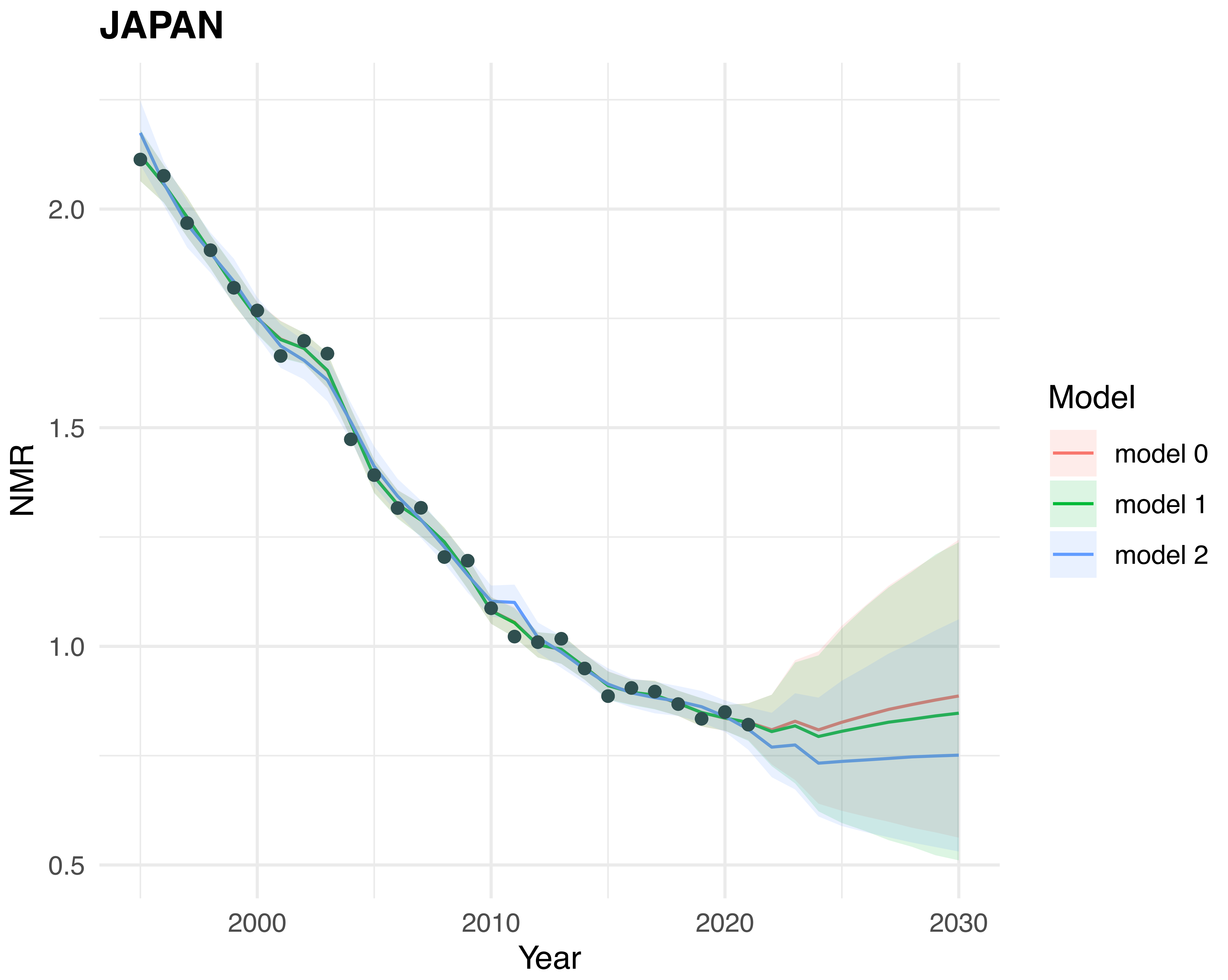
$$\gamma_r \sim N(c_{28}, 0.2^2)$$

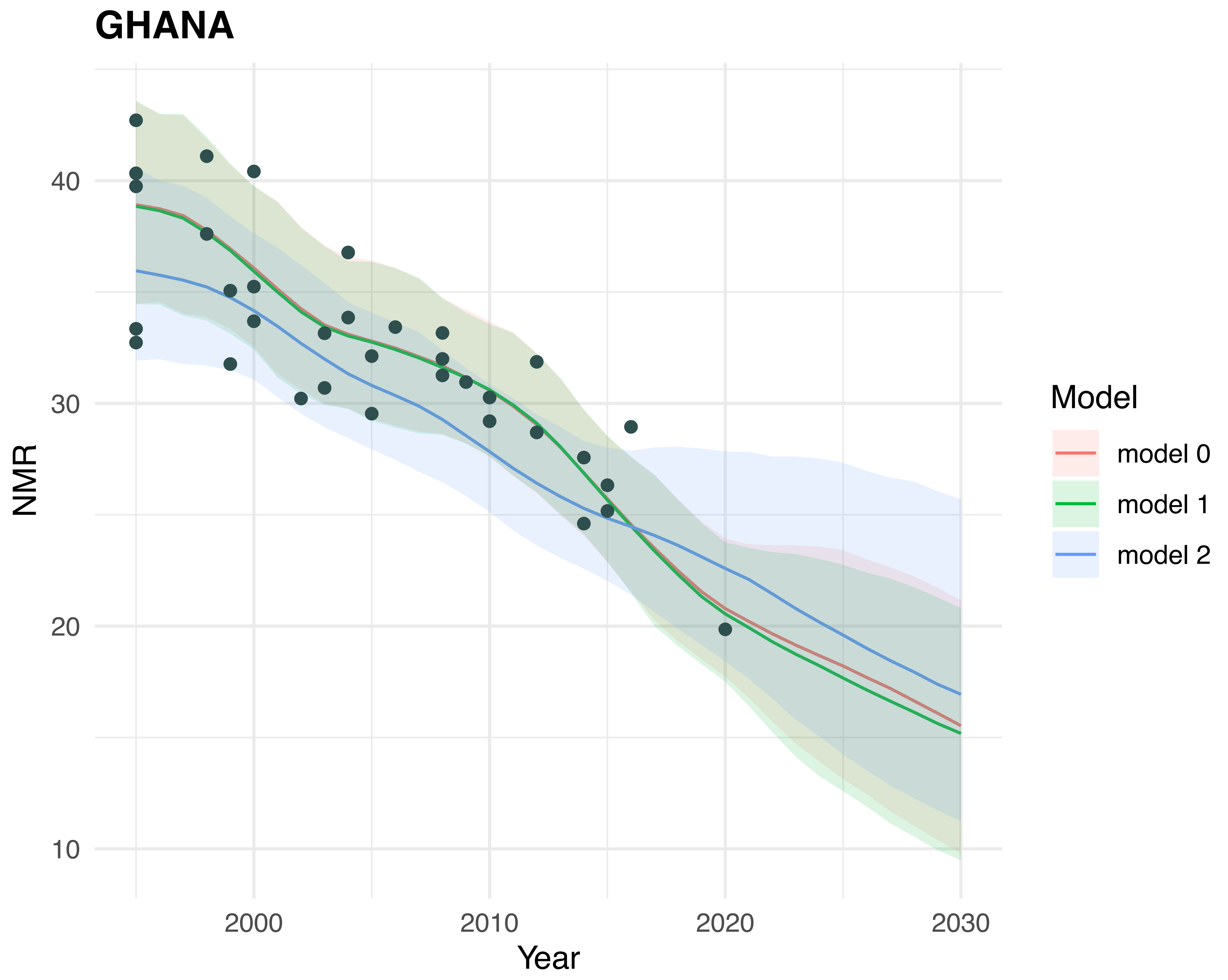
Summary of models

- Model 0: Global relationship; linear relationship with log U5MR with changing slope
- Model 1: Regional relationship; linear relationship with log U5MR
- Model 2: Regional relationship, informed by prior empirical observations, quadratic relationship with U5MR

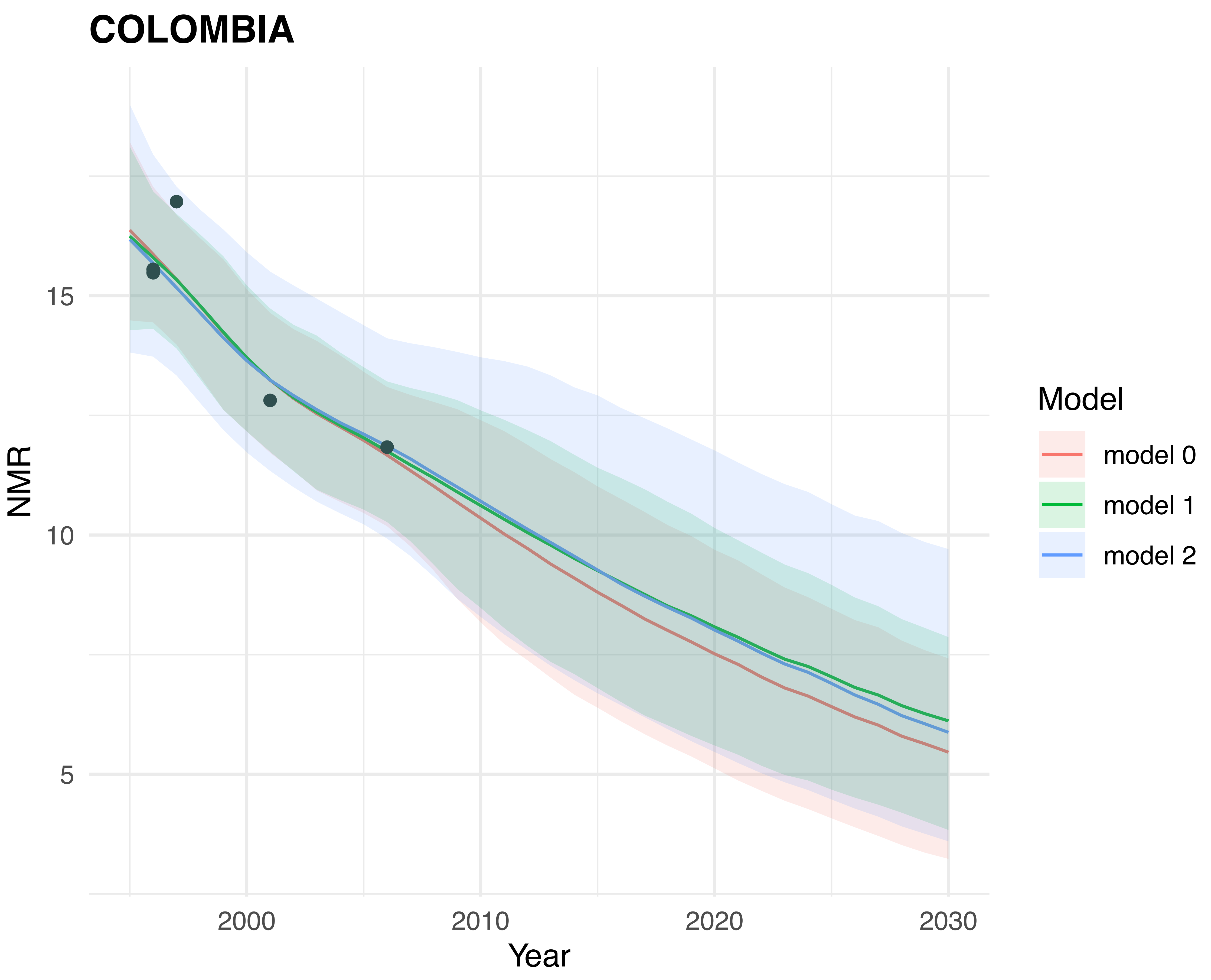
Results

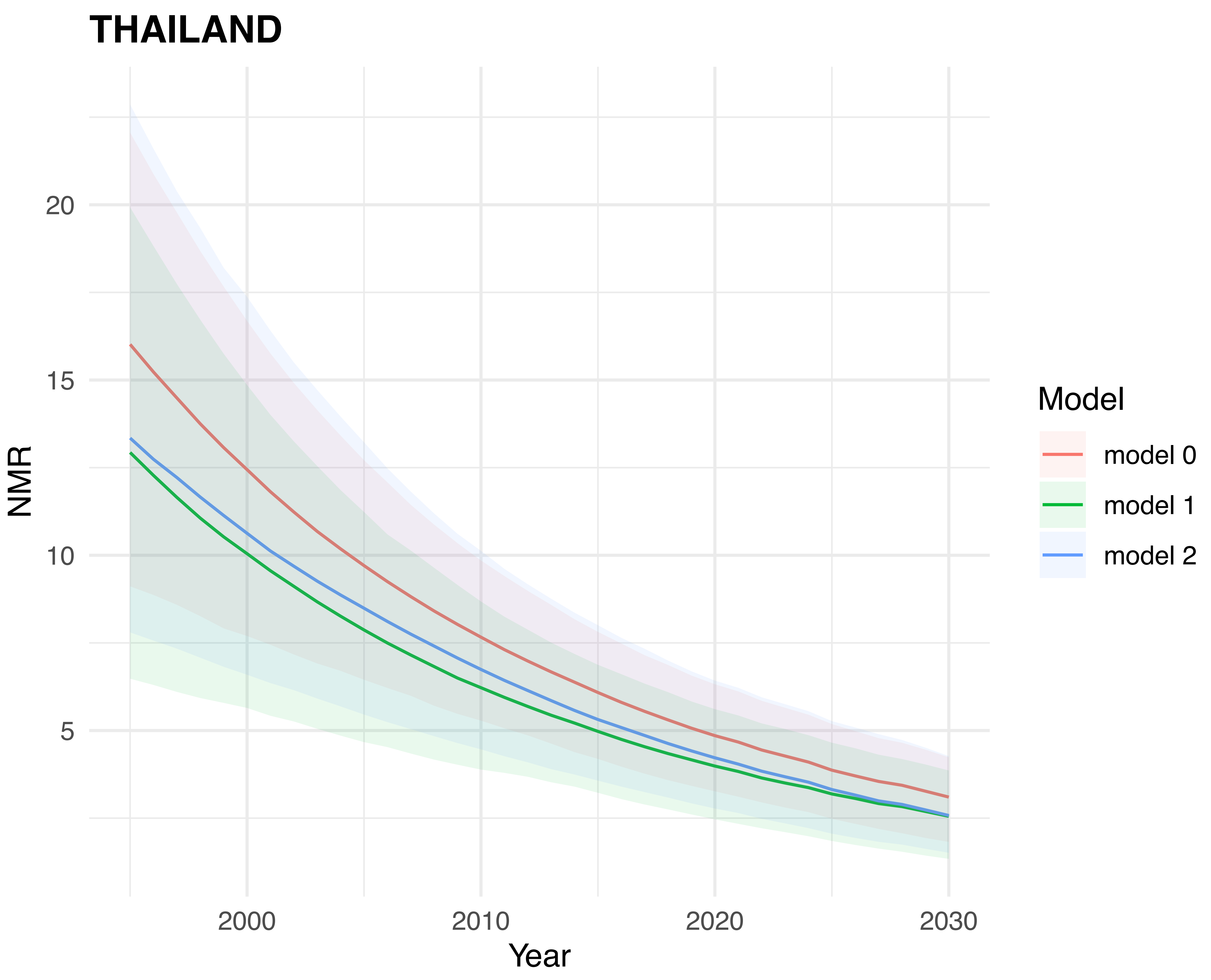
Case study countries





Case study countries





Comments

- The choice of ‘expected’ function matters most when there’s no data (and for projections)
- Projections can be quite different
  - Implications for target hitting, speed of decline
- Results suggest empirical-based priors in Model 2 may not be appropriate in high-mortality contexts
- Biggest differences overall between Model 0 and Model 2
- Uncertainty intervals overlap, but combined uncertainty is larger

Ways forward and future work

- Bayesian model averaging for better quantification of uncertainty
- Simulation

Bigger picture:

- Systematic framework for model comparison (Susmann, Alexander and Alkema)
- What to call estimates in situations where we have no data?
- Using results of uncertainty and estimation to advocate for better data collection

Thanks!

monica.alexander@utoronto.ca

monicaalexander.com

@monjalexander

MJAlexander

Summary metrics

- Comparing the models overall (against Model 0) on NMR scale
  - $MAE_{0|1} = 0.407, MAE_{0|2} = 0.76$
  - $MSE_{0|1} = 0.935, MSE_{0|2} = 14.7$
  - $RMSE_{0|1} = 0.967, RMSE_{0|2} = 3.84$