MULTI-LEVEL MODELLING OF GEOGRAPHICALLY AGGREGATED HEALTH DATA: A CASE STUDY ON MALIGNANT MELANOMA MORTALITY AND UV EXPOSURE IN THE EUROPEAN COMMUNITY

IAN H. LANGFORD[1]*, GRAHAM BENTHAM[1] AND ANNÉ-LISE McDONALD[1,2]

[1] *Centre for Social and Economic Research on the Global Environment, University of East Anglia and University College, London, U.K.*

[2] *Health Policy and Practice Unit, School of Health and Social Work, University of East Anglia, Norwich, U.K.*

Sara Serafino's project for Bayesian Statistics

Goals of the project

- Analyse the data using a Bayesian approach
- Build a model for the number of male deaths, taking into account the hierarchical data structure
- Check the model and comment the results
- Compare the results with those of Langford et al. (1998):

what can be said about the effect of the UVB dose on the malignant melanoma mortality?

Data

354 observations on 6 variables:

- **Nation:** Belgium, West Germany (WG), Denmark, France, United Kingdom (UK), Italy, Ireland, Luxembourg, Netherlands
- **Region:** Region ID - a factor from 1 to 79, skipping 26
- **County:** County ID - a factor from 1 to 354
- **Deaths:** Number of male deaths due to malignant melanoma during 1971–1980 (for some nations it is from 1975-1976 onwards)
- **Expected:** Number of expected deaths
- **UVB:** Centered measure of the UVB dose reaching the earth's surface in each county.

Data

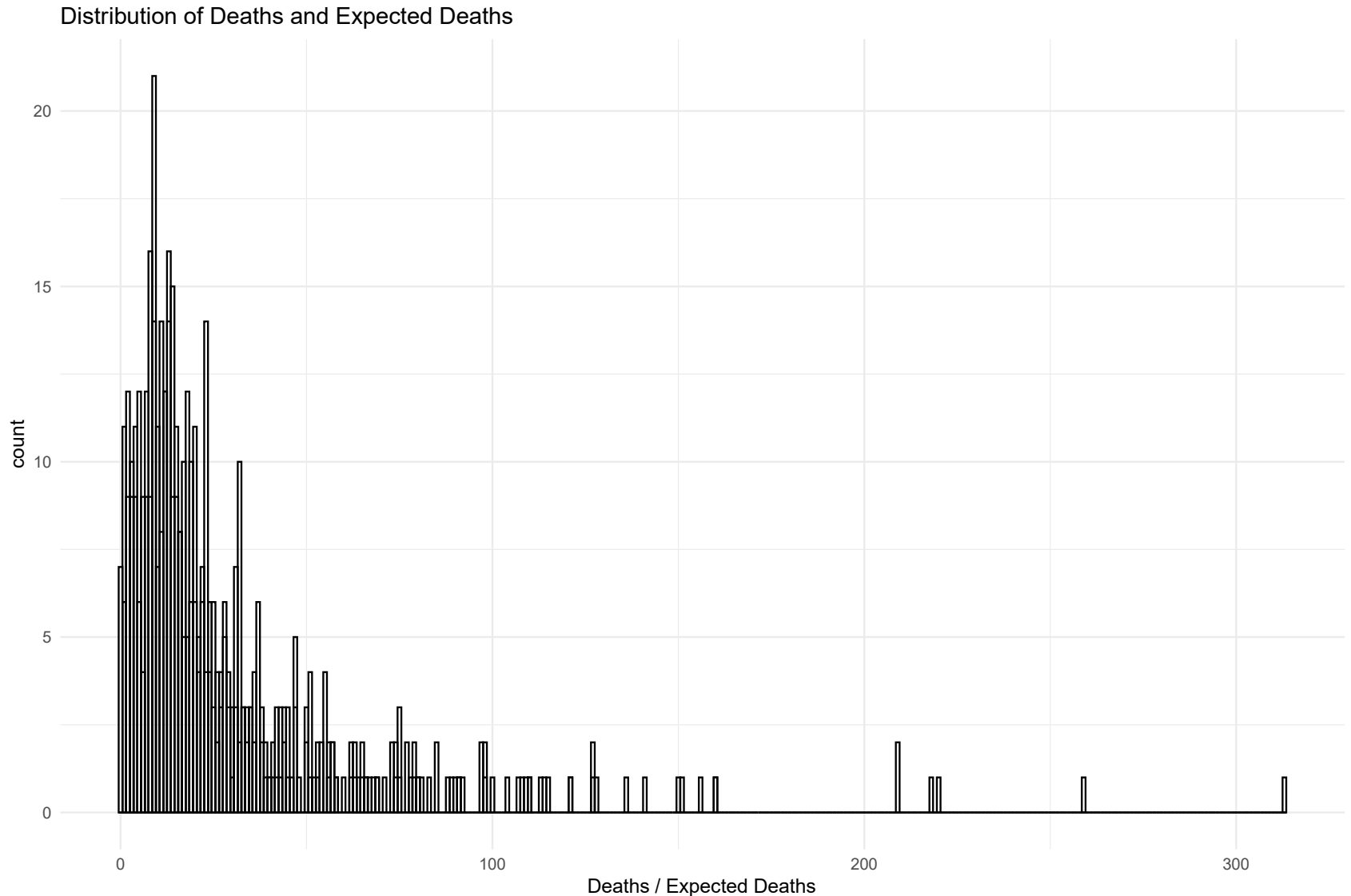
Table I. The geographical hierarchy of the IARC mortality data

Nation	Regions	Counties
1. Belgium	3	11
2. West Germany	11	30
3. Denmark	3	14
4. France	21	94
5. United Kingdom	11	70
6. Italy	20	95
7. Ireland	4	26
8. Luxembourg	1	3
9. Netherlands	4	11

Geographical hierarchy: counties within regions within nations

multi-level model hierarchy

Explorative analysis



If they were too different, there could be some unconsidered factors influencing the observed deaths.

Explorative analysis

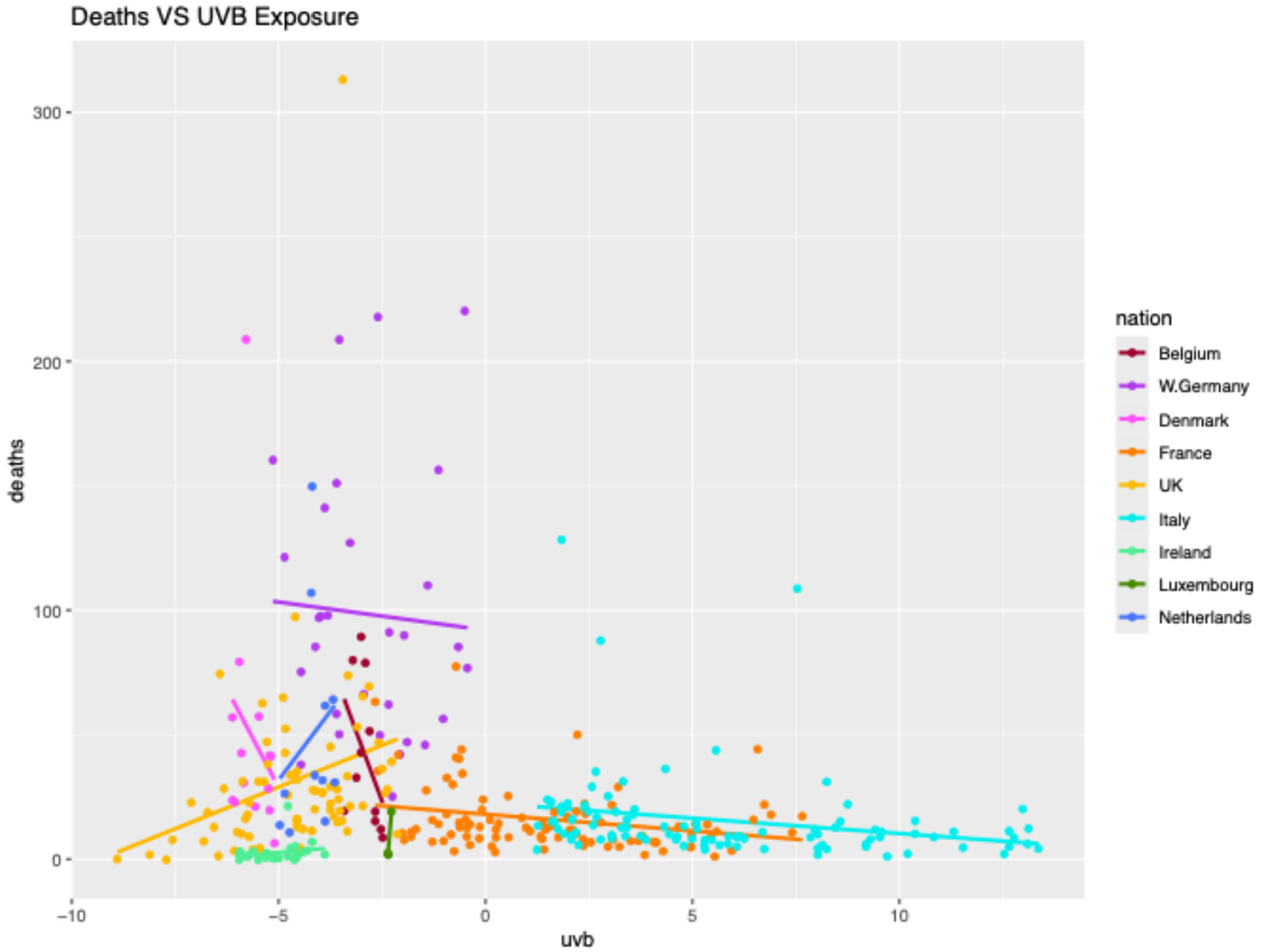
Negative correlation with complete pooling

On the contrary, accounting for hierarchy at nation level…

Explorative analysis

Nation	deaths	people /km² *
Belgium	449	384
WG	2949	234
Denmark	681	139
France	1495	101
UK	2179	281
Italy	1462	195
Ireland	67	65
Luxembourg	23	255
Netherlands	546	520

* From most recent data, just for the big picture



Explorative analysis

UK, Ireland, Netherlands, Luxembourg:

- Low exposure to UVB
- Positive relationship

Belgium, WG, Denmark:

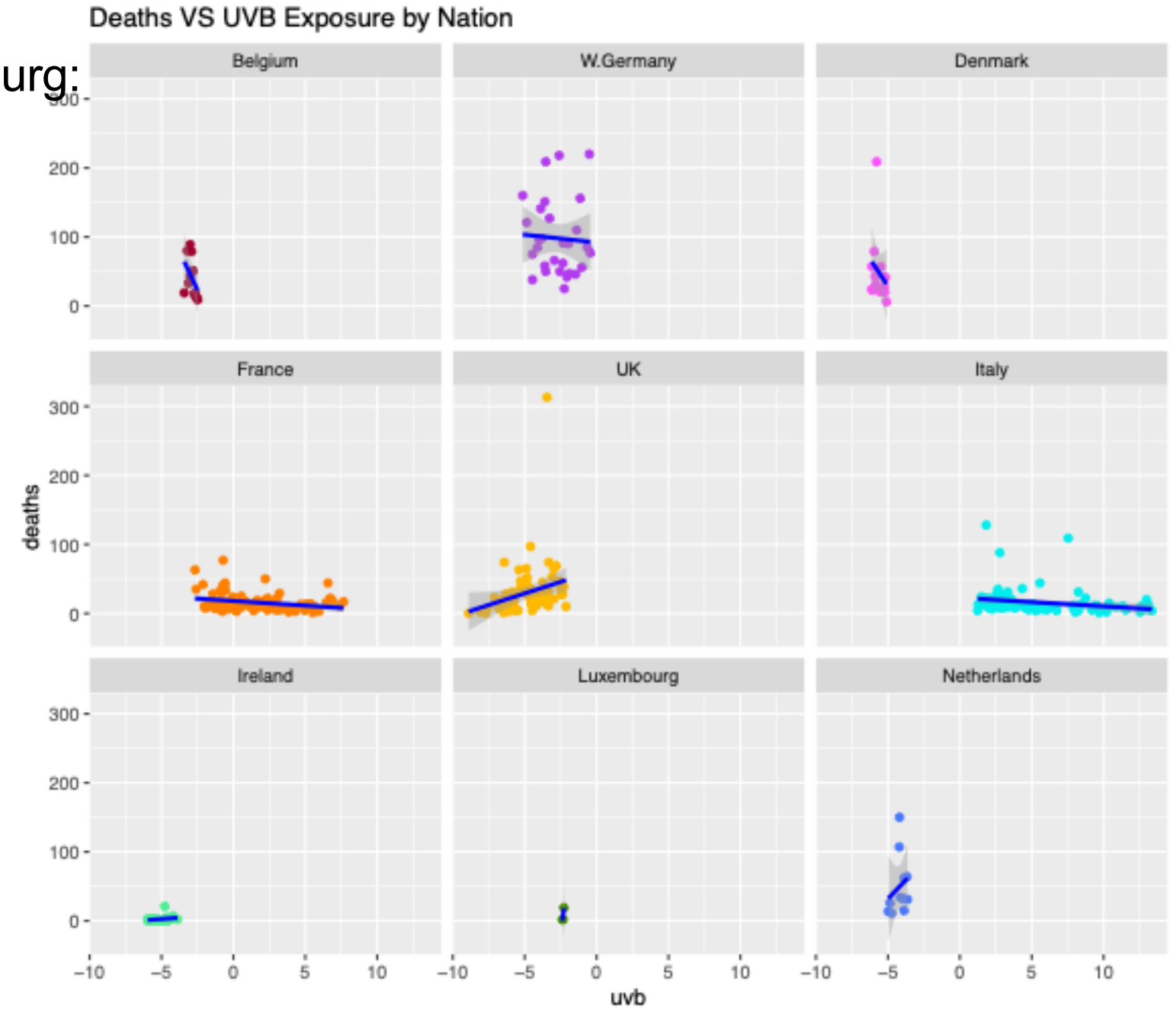
- Low exposure to UVB
- Negative relationship

Italy:

- High(est) exposure to UVB
- Negative relationship

France:

- Mid exposure to UVB
- (slightly) Negative relationship



Explorative analysis

UK, Ireland, Netherlands, Luxembourg:

 Low exposure to UVB

 Positive relationship

Belgium, WG, Denmark:

 Low exposure to UVB

 Negative relationship

Italy:

 High(est) exposure to UVB

 Negative relationship

France:

 Mid exposure to UVB

 (slightly) Negative relationship

Northern Europeans more at risk of intermittent high exposures.

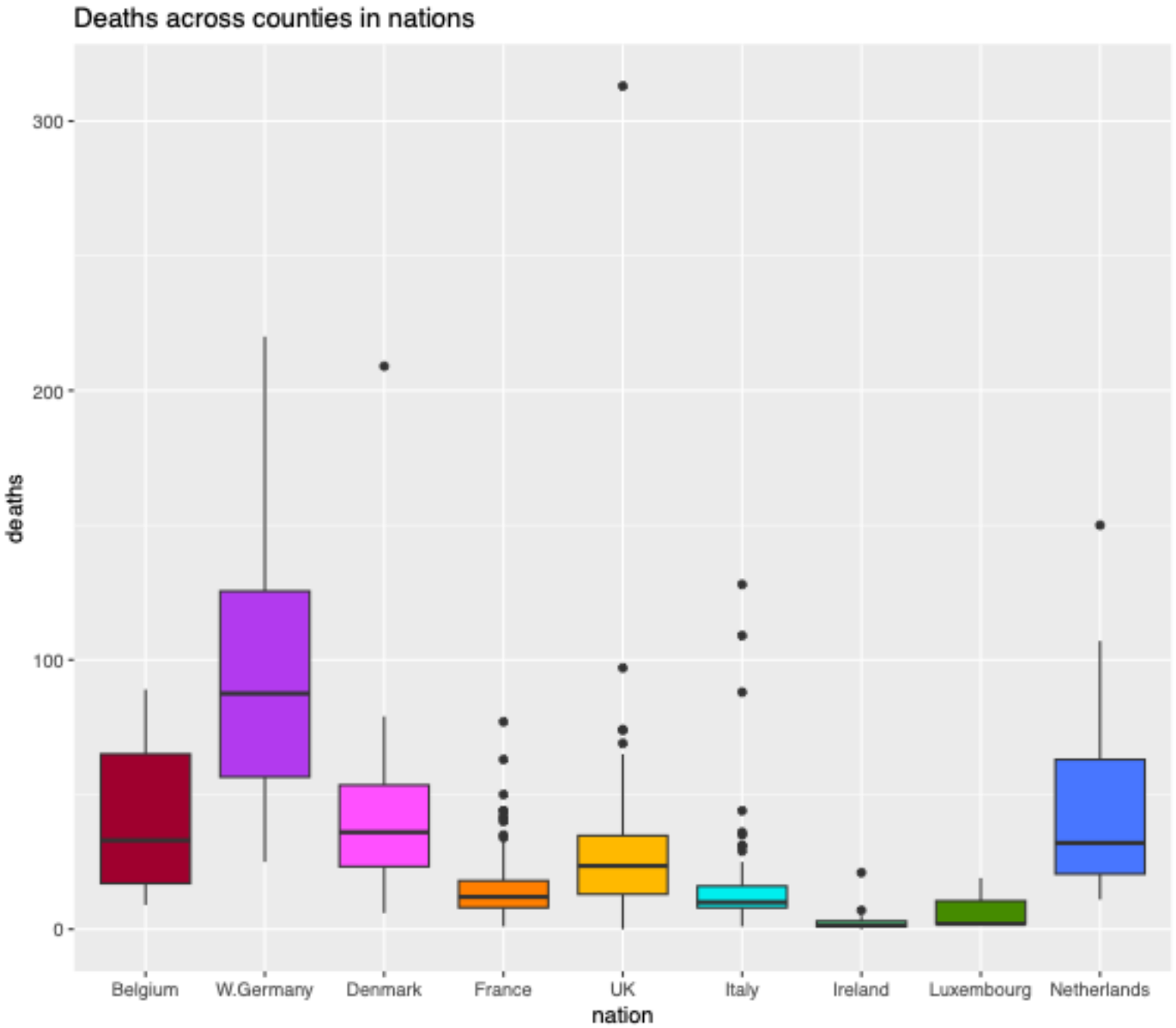
Less sun exposure at home & more recreational travel to warmer climates?

Other factors influencing the behaviour:

- Use of sunscreen
- Effective UVB exposure
- History of deceased
- Origin different from nation

Differences within nations – especially Italy and France – depending on north-south and genetic factors

Explorative analysis



- Boxes extend from 25% to 75% of data.
- Whiskers extend from minimum to maximum within 1.5 times its box.
- Outliers are regions with deaths exceptionally higher than the average.

- West Germany has the highest number of deaths and variation
- Ireland and Luxembourg have lower mortality rates with less variation

Methods

Multi-level model based on generalized least squares estimation:

$$Y = X\beta + Z\theta$$

- X design matrix associated with a vector of fixed parameters β
- Z design matrix associated with a vector of random parameters θ
- Y vector of responses

Methods

At level 1, Y is a Poisson distributed response vector of observed cases (O), hence we need to include an offset of expected numbers of cases (E) with a logarithmic link function:

$$O \sim Poisson(\mu)$$
$$\ln(\mu) = \ln(E) + X\beta + Z\theta$$

Since counties have very heterogeneous populations (hence variance), it is theoretically more appropriate to consider level 1 random effects as following a negative binomial distribution. However, the paper found no improvement of fit with this.

At level 2 and 3, the random parameters follow a normal distribution.

Frequentist approach

Generalized linear mixed model with Poisson regression, an intercept for each region and nation and an offset term to take into account the expected deaths:

```
M1 <- glmer(deaths ~ uvb + (1 | region) +
                                    + (1 | nation),
            Mmmec,
            poisson,
            offset = log(expected))
```

Frequentist approach

```
     AIC      BIC   logLik deviance df.resid
  2198.7   2214.2  -1095.3   2190.7      350

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9440 -0.7788 -0.0071  0.6277  3.9102

Random effects:
 Groups Name        Variance Std.Dev.
 region (Intercept) 0.04829  0.2198
 nation (Intercept) 0.13708  0.3702
Number of obs: 354, groups:  region, 78; nation, 9

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.06398    0.13351  -0.479   0.6318
uvb         -0.02822    0.01139  -2.478   0.0132
```

- Must compare with others
- AIC and BIC are not the best for hierarchical models

Symmetrical residuals but with too extreme values: overdispersion

Variance between nations and regions within nations

 Negative relationship

Bayesian approach (stan_glmer)

Like before, but the Bayes approach relies on Hamiltonian Monte Carlo sampling from the posterior distribution:

```
M1.rstanarm <- stan_glmer (deaths ~ uvb +
                                    + (1 | region) + (1 | nation),
                                    Mmmec,
                                    poisson,
                                    offset = log(expected))
```

Bayesian approach (stan_glmer)

```
Estimates:
                mean   sd   10%
(Intercept)    -0.1    0.2 -0.3
uvb             0.0    0.0  0.0
```

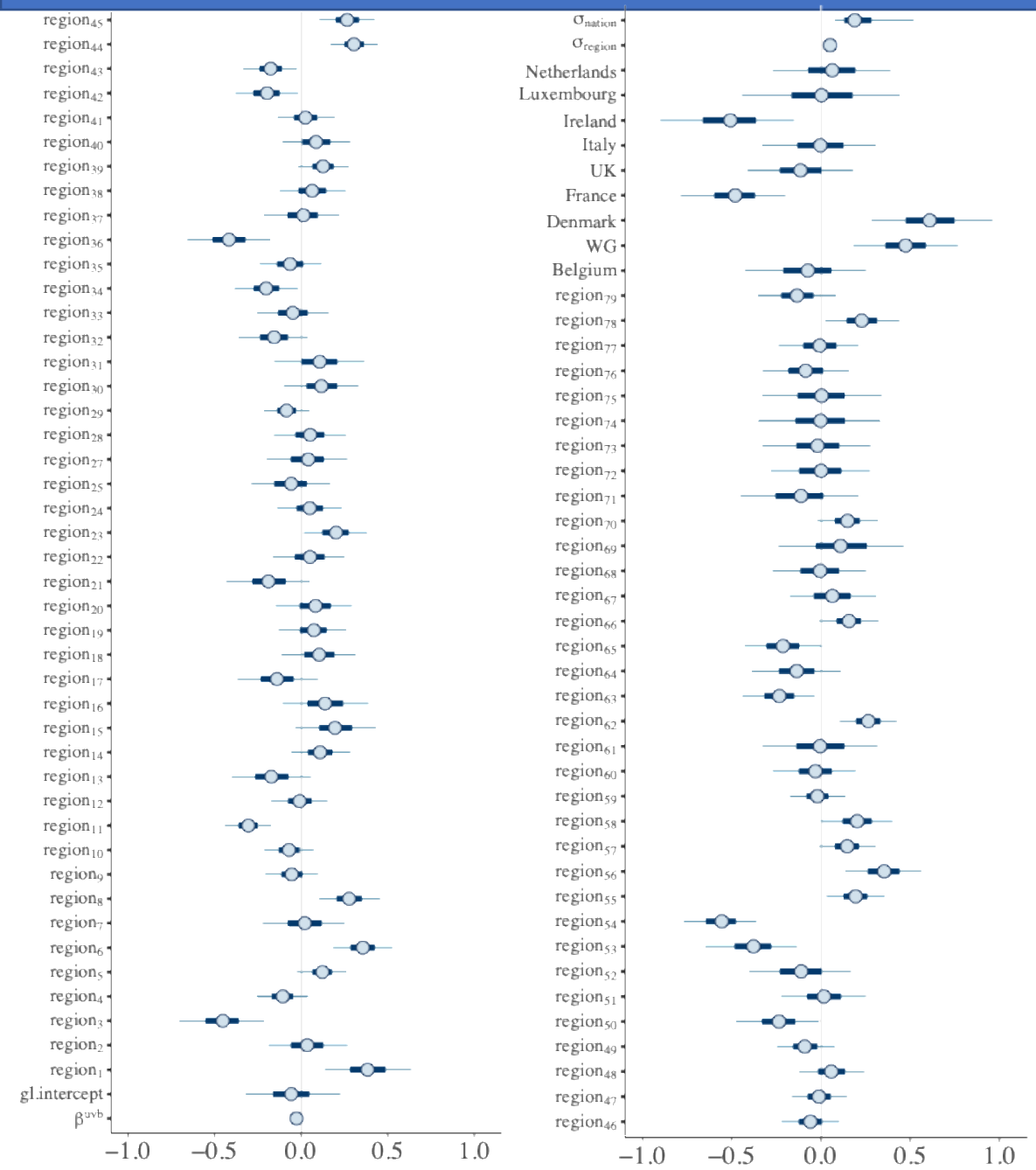
UVB does not have a meaningful effect (contrary to M1)

```
Fit Diagnostics:
          mean   sd   10%   50%   90%
mean_PPD 27.8    0.4 27.3  27.8  28.3
```

```
MCMC diagnostics
                                mcse Rhat n_eff
(Intercept)                     0.0  1.0  1304
uvb                             0.0  1.0  1596
mean_PPD                        0.0  1.0  3909
log-posterior                   0.3  1.0   892
```

Converges

Bayesian approach (stan_glmer)



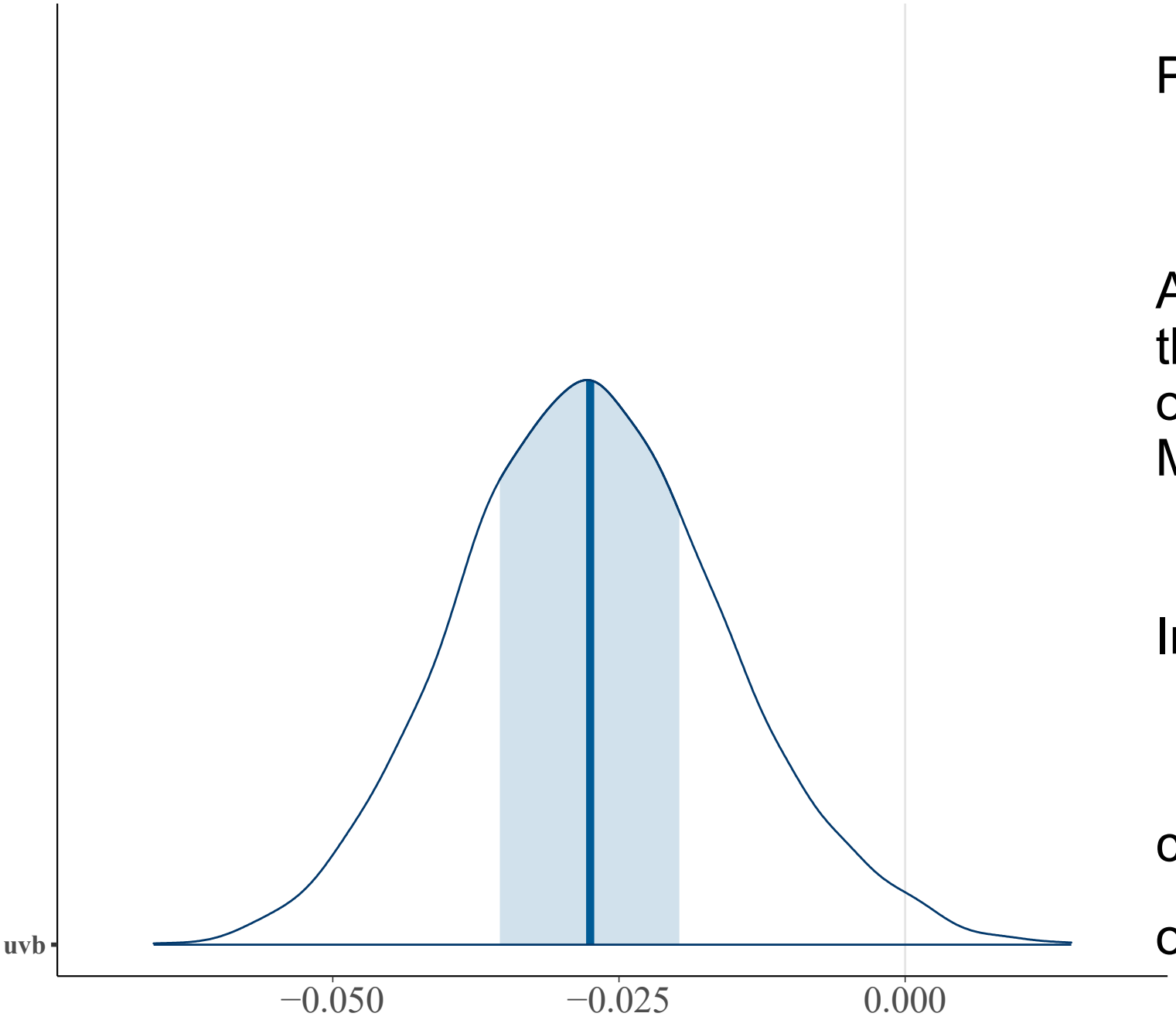
Credibility intervals:

Solid blue line: 50% credible interval

Tinner light blue line: 95% credible interval

Captures well the uncertainty of estimated parameters

Bayesian approach (stan_glmer)



Posterior estimate of UVB for M1:

$$-0.03443$$

A positive difference of 1 unit in this predictor has a linear effect of -0.03443 on the probability of MM death.

In terms of odds ratios:

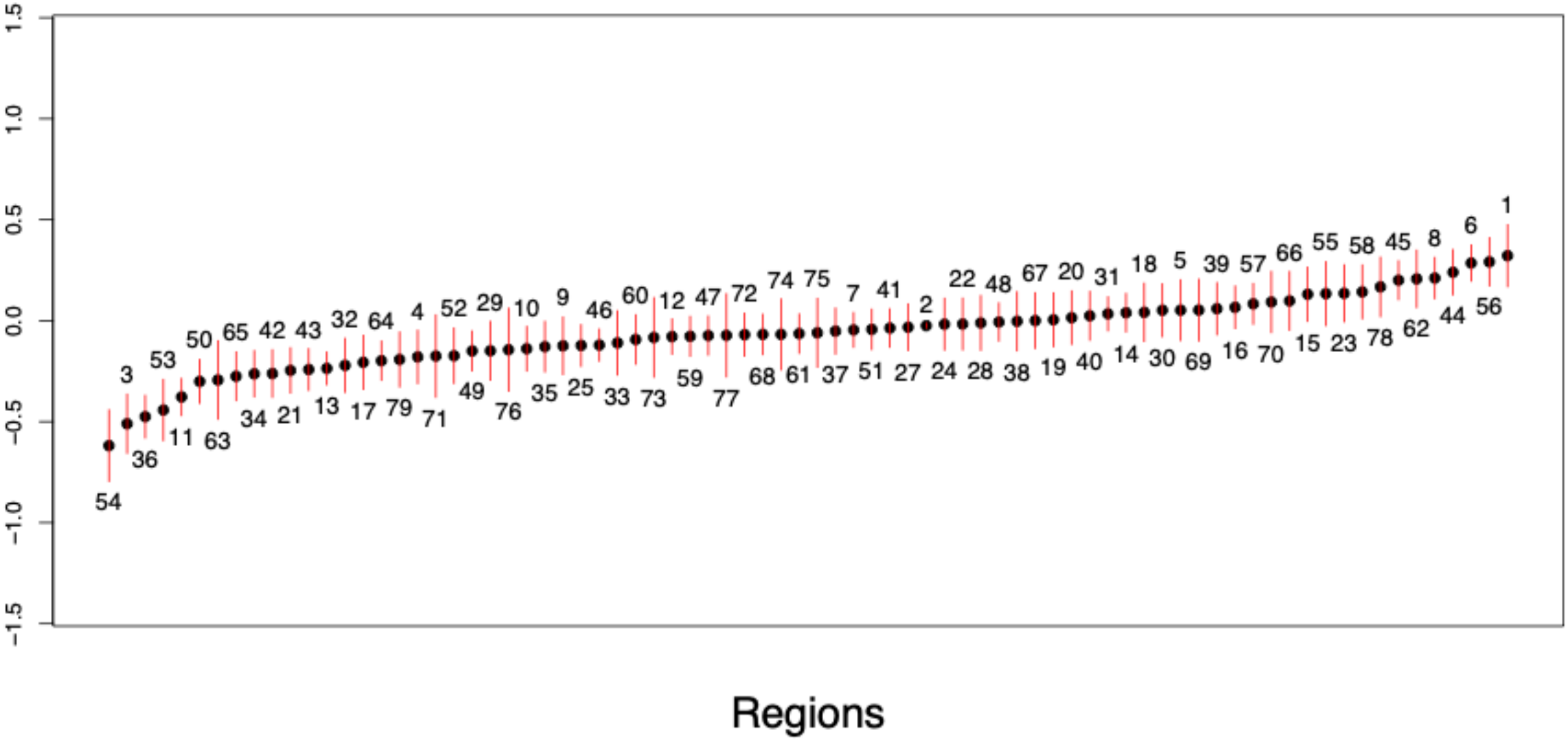
$$e^{-0.03443} \approx 0.96616$$

causing a decrease in the odds of 3.4%

Bayesian approach (stan_glmer)

Random effects for the regions:

Posterior means of the region intercepts ± standard error of random effect



Bayesian approach (model A)

Variance components model with UVBI in the fixed part of the model so that, for the *i*th county in the *j*th region in the *k*th nation:

$$\ln(O_{ijk}) = \ln(E_{ijk}) + \beta_0 + \beta_1 UVBI_{ijk} + s_k + u_{jk} + e_{ijk}$$
$$s_k \sim N(0, \sigma_s^2) \qquad u_{jk} \sim N(0, \sigma_u^2)$$

- β_0 intercept term
- β_1 mean (fixed) effect of UVBI
- s_k, u_{jk}, e_{ijk} random terms associated with the intercept at levels 3, 2, 1 respectively

Bayesian approach (model A)

Direct measurements of UVB at earth's surface are too sparse for reliable estimates ⟹ $UVBI_{ijk}$ calculated for each county

Table III. UVB index: descriptive statistics for each nation

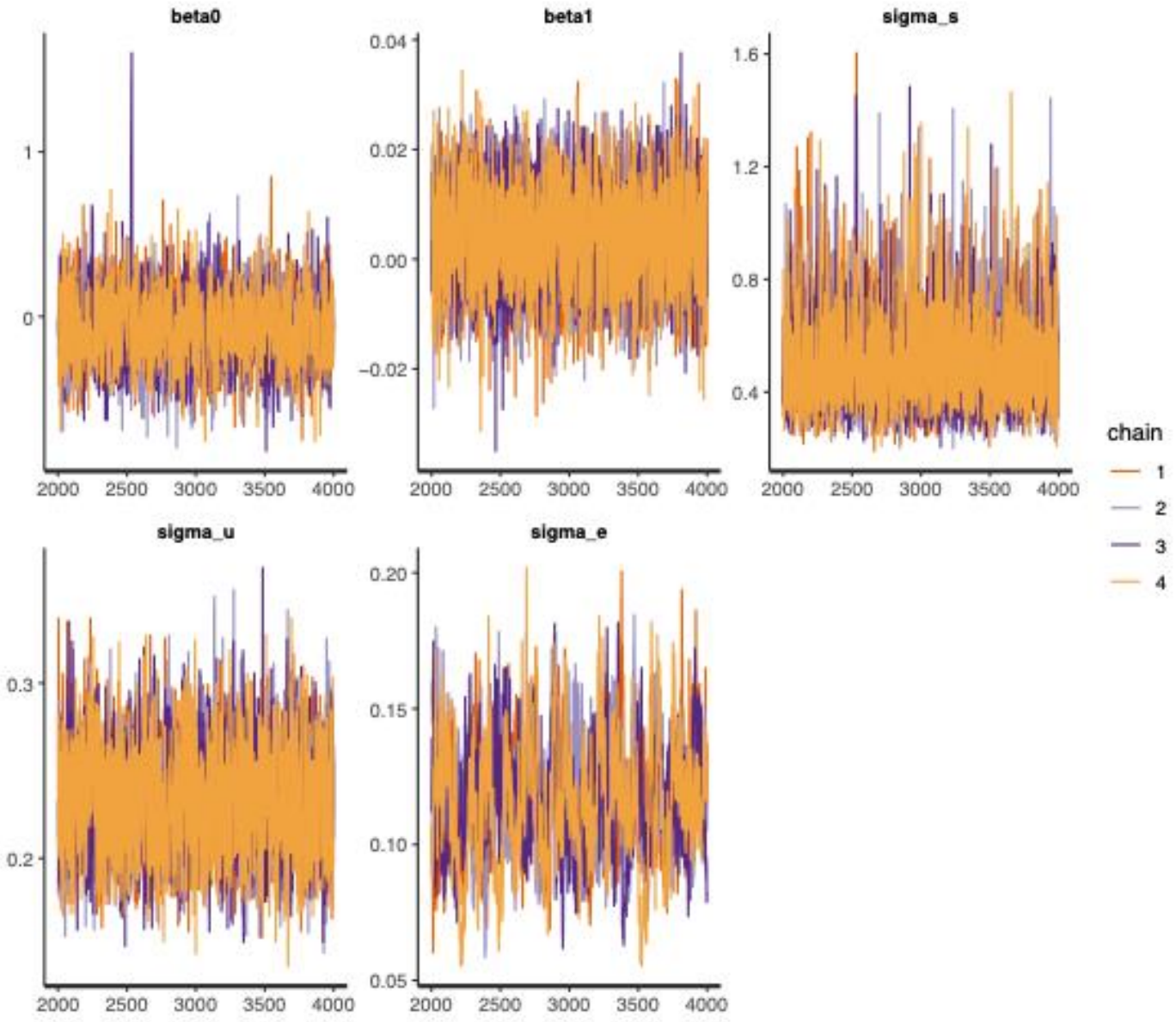
Nation	*N*	Mean	Standard deviation	Min	Max
Belgium	11	12·70	0·29	12·17	13·10
West Germany	30	12·79	1·35	10·45	15·15
Denmark	14	9·96	0·38	9·47	10·49
France	94	17·18	2·59	12·92	23·24
United Kingdom	70	10·91	1·50	6·69	13·46
Italy	95	21·45	3·51	16·83	28·95
Ireland	26	10·54	0·60	9·64	11·70
Luxembourg	3	13·26	0·05	13·22	13·31
Netherlands	11	11·40	0·47	10·62	11·94

Use a truncated norm to generate random deviates for each county starting from nation statistics

Bayesian approach (model A)

$$\beta_0, \beta_1, \sigma_s, \sigma_u, \sigma_e \sim N(0,1)$$

	Model A Estimate	(SE)
Fixed part		
β_0	0·0103	(0·134)
β_1 (UVBI)	− 0·0360	(0·0107)
β_2 (RDENS)		
β_3 (RGDP)		
Random part		
Level 3: nations		
σ_s^2	0·140	(0·0733)
σ_{st}		
σ_t^2		
Level 2: regions		
σ_u^2	0·0424	(0·00956)
σ_{uv}		
σ_v^2		
Level 1: counties		
σ_{e1}^2	1.11	(0·0937)
σ_{e2}^2		

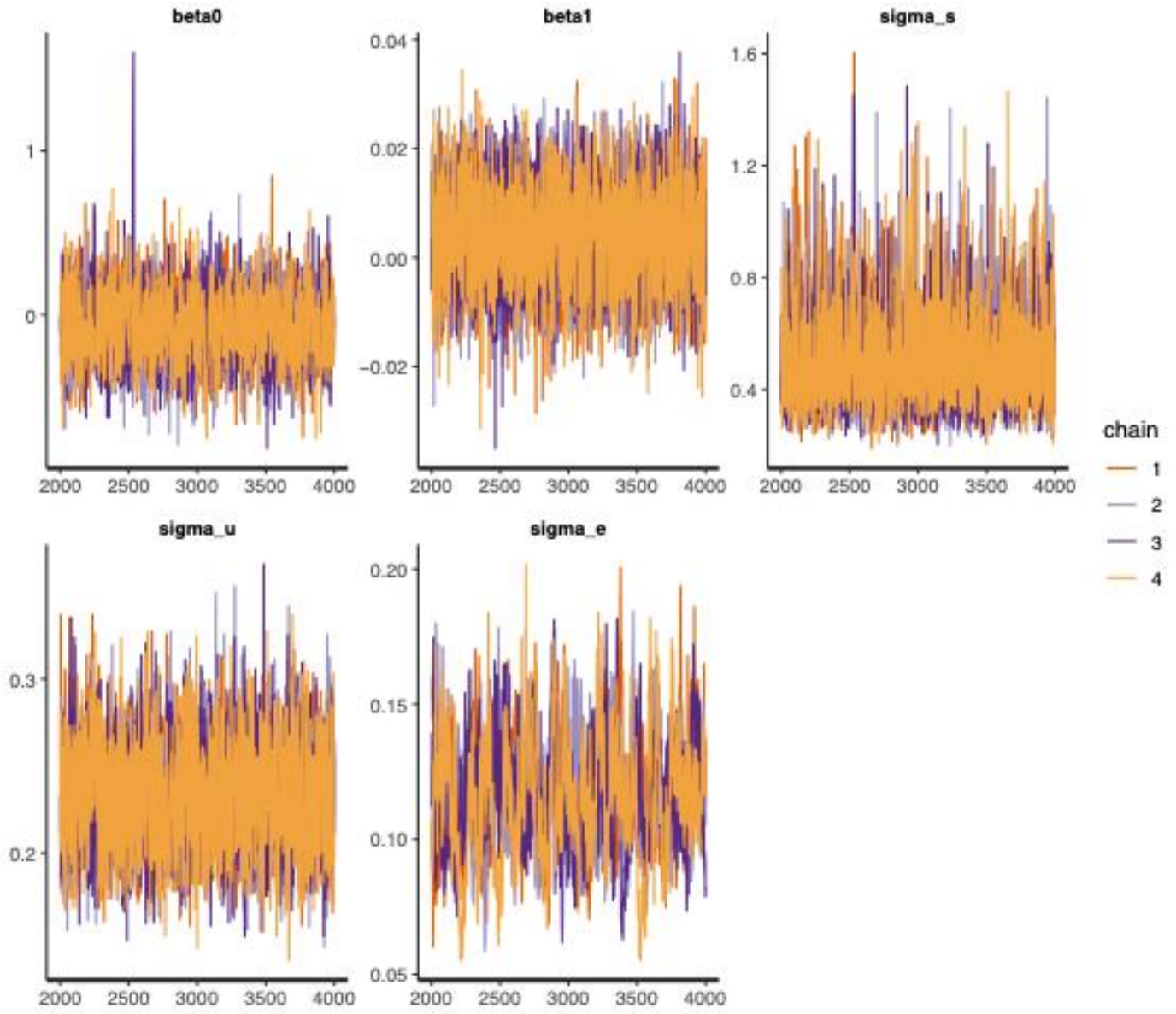


Bayesian approach (model A)

$$\beta_0, \beta_1, \sigma_s, \sigma_u, \sigma_e \sim N(0,1)$$

variable	rhat	ess_bulk	ess_tail
<chr>	<dbl>	<dbl>	<dbl>
beta0	1.00	1992.	2630.
beta1	1.00	6337.	5855.
sigma_s	1.00	5310.	4764.
sigma_u	1.00	2671.	5089.
sigma_e	1.01	287.	420.

With 8000 iterations instead of 4000, the warning about bulk and tail ESS too low disappears and $\hat{R}_{\sigma_e} = 1.00$



stan_glmer VS model A

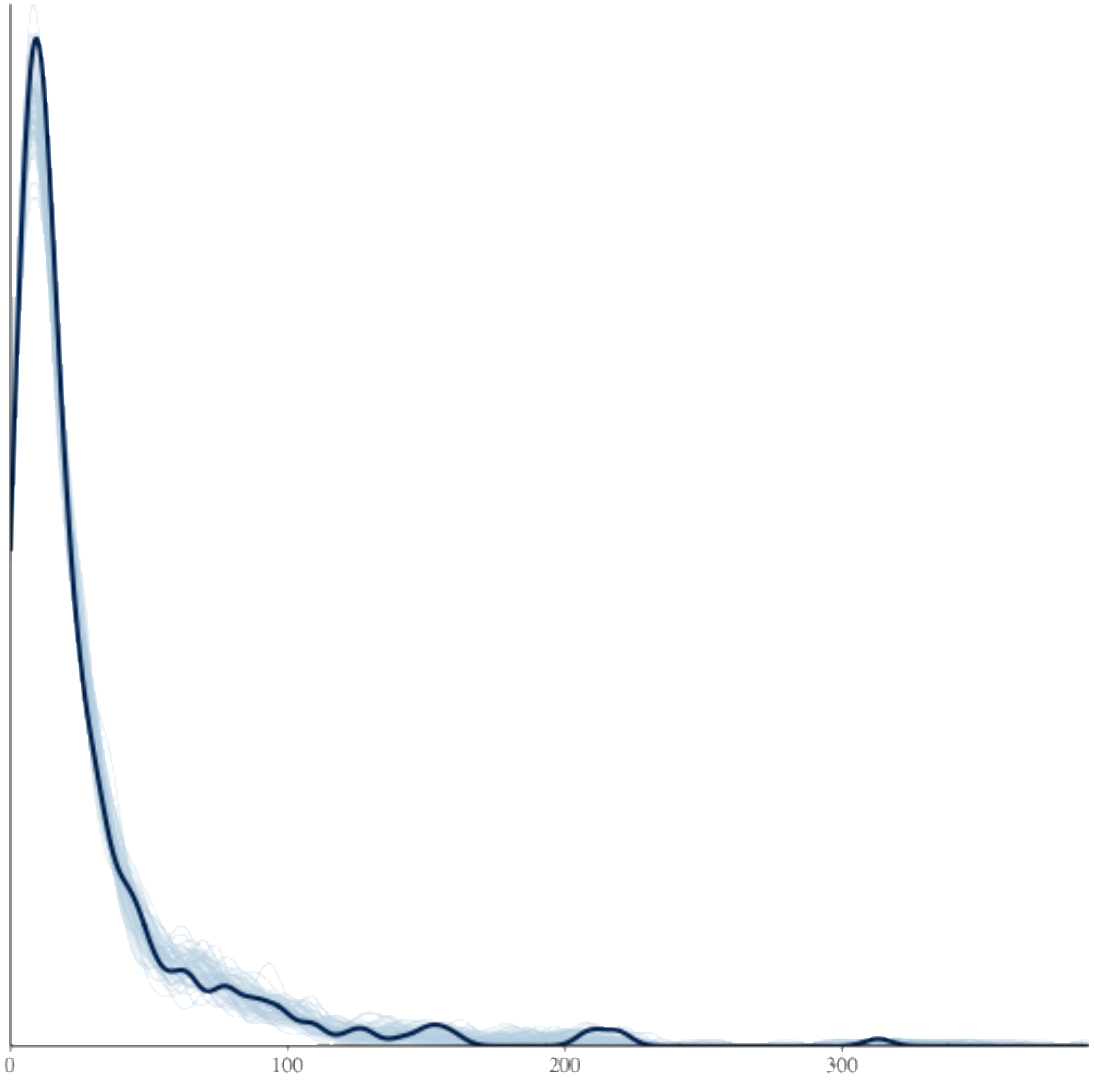
```
M1.rstanarm <- stan_glmer (deaths ~
uvb +
            + (1 | region)
            + (1 | nation),

            Mmmec,

            poisson,

            offset = log(expected))
```

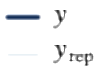
```
comp_model_A <- stan_model('poisson_regression.stan’)

fit_model_A <- sampling(comp_model_A, data = stan_data,
seed = 123, iter=4000)
```

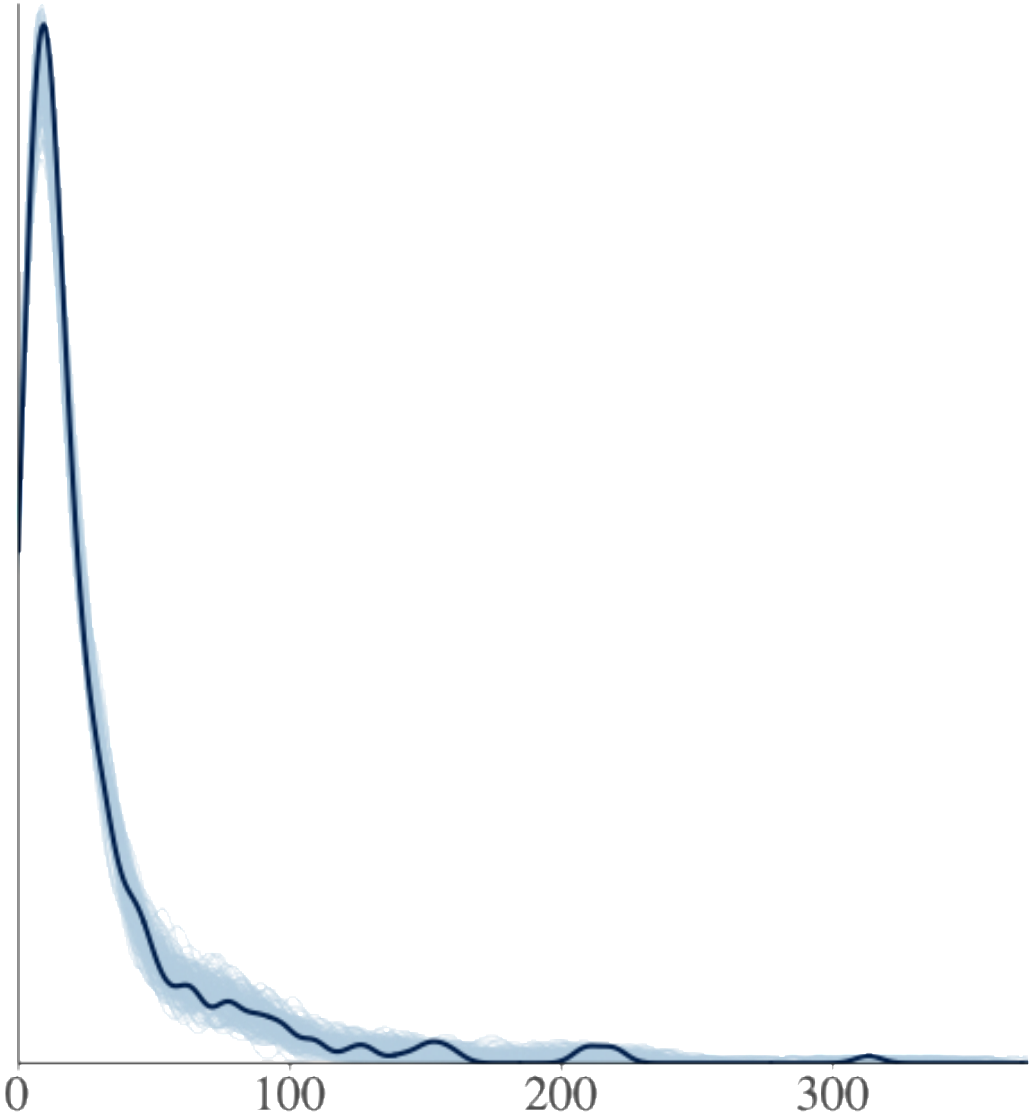
poisson_regression.stan:

```
. . .
eta[n] = log(expected[n]) + beta0 + beta1 * UVBI[n] + s[k[n]]
+ u[j[n]] + e[n];

deaths ~ poisson_log(eta)
```

stan_glmer VS model A: densities

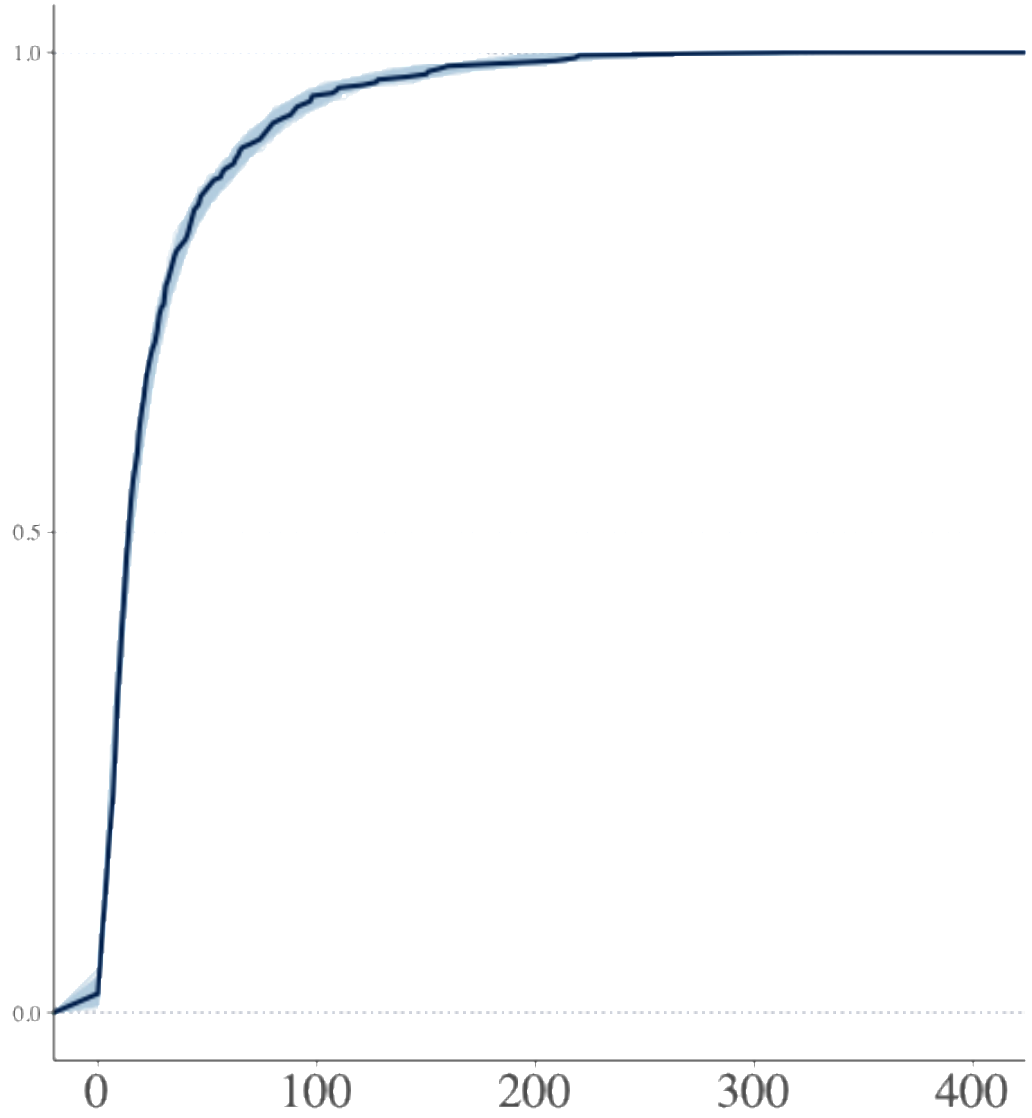




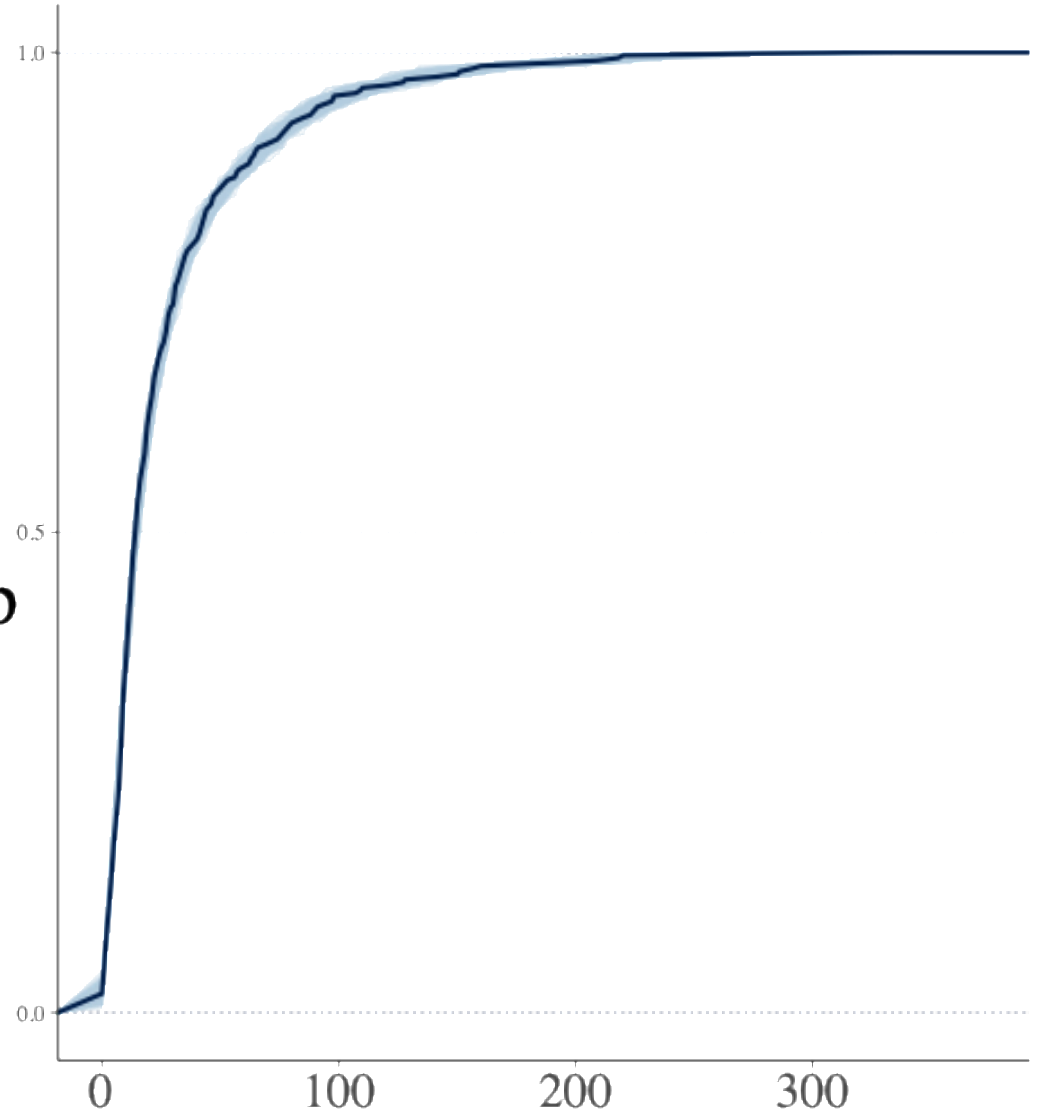


y
y_{rep}

stan_glmer VS model A: Empirical Cumulative Distribution Function

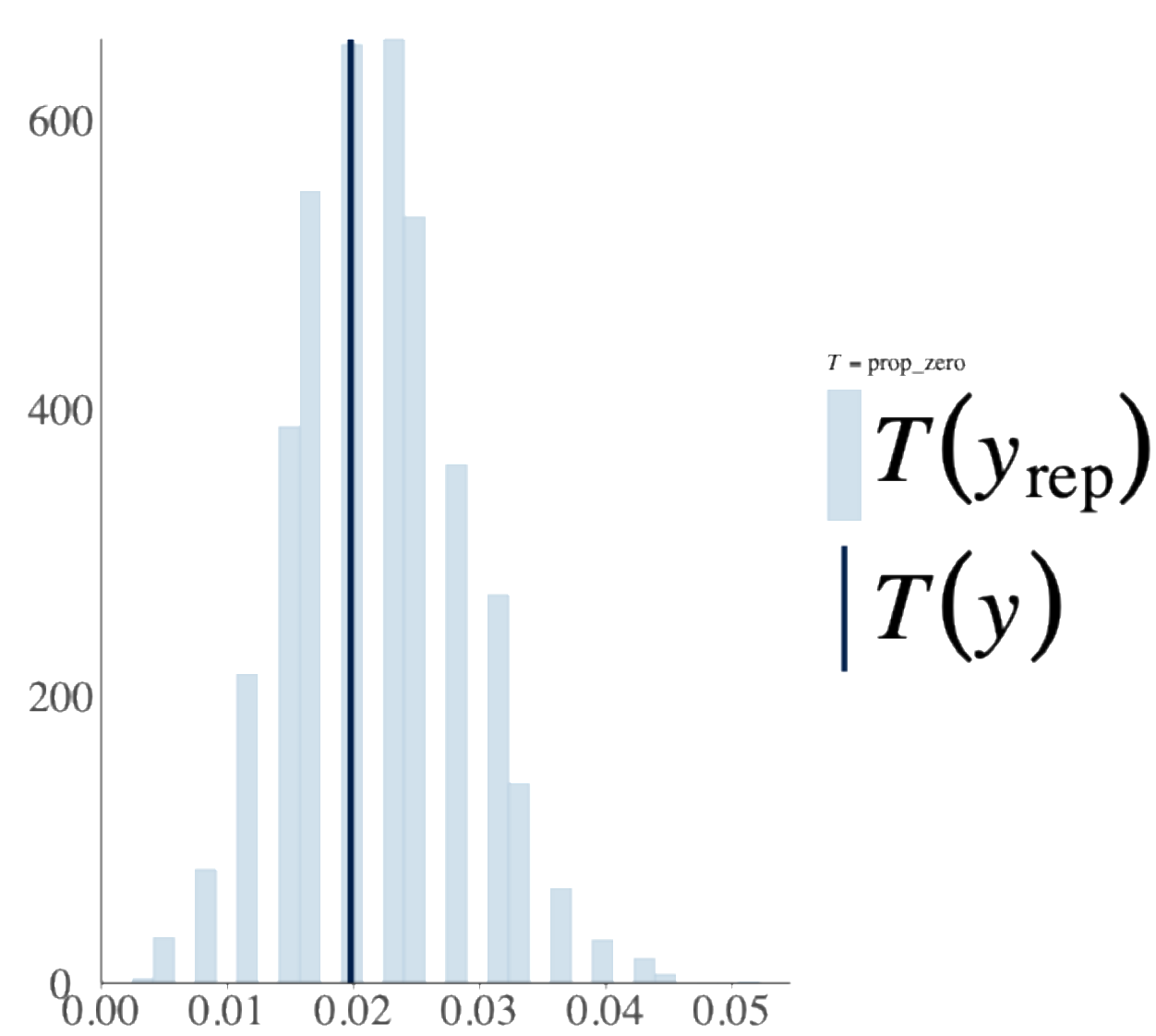


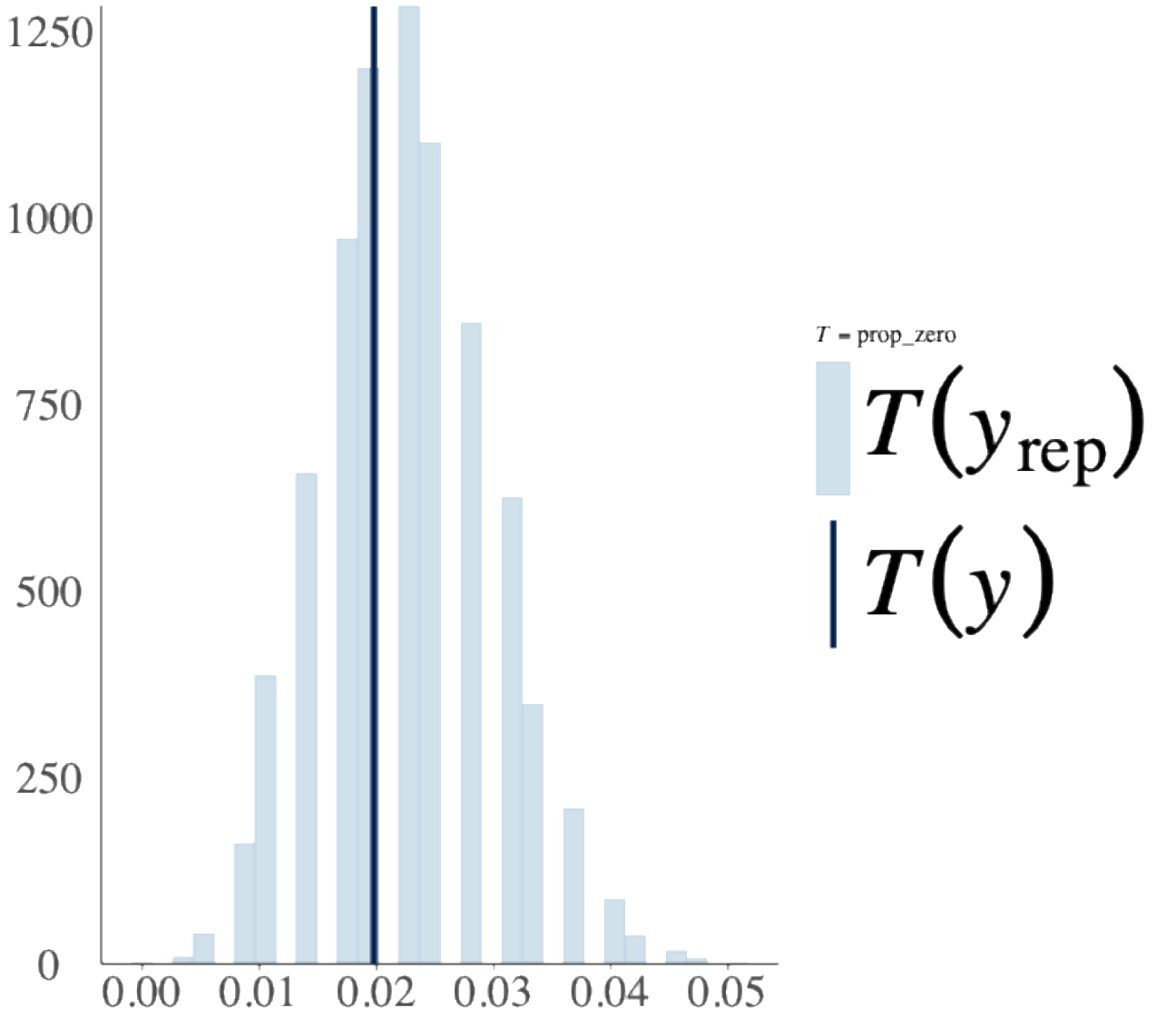
y
y_{rep}



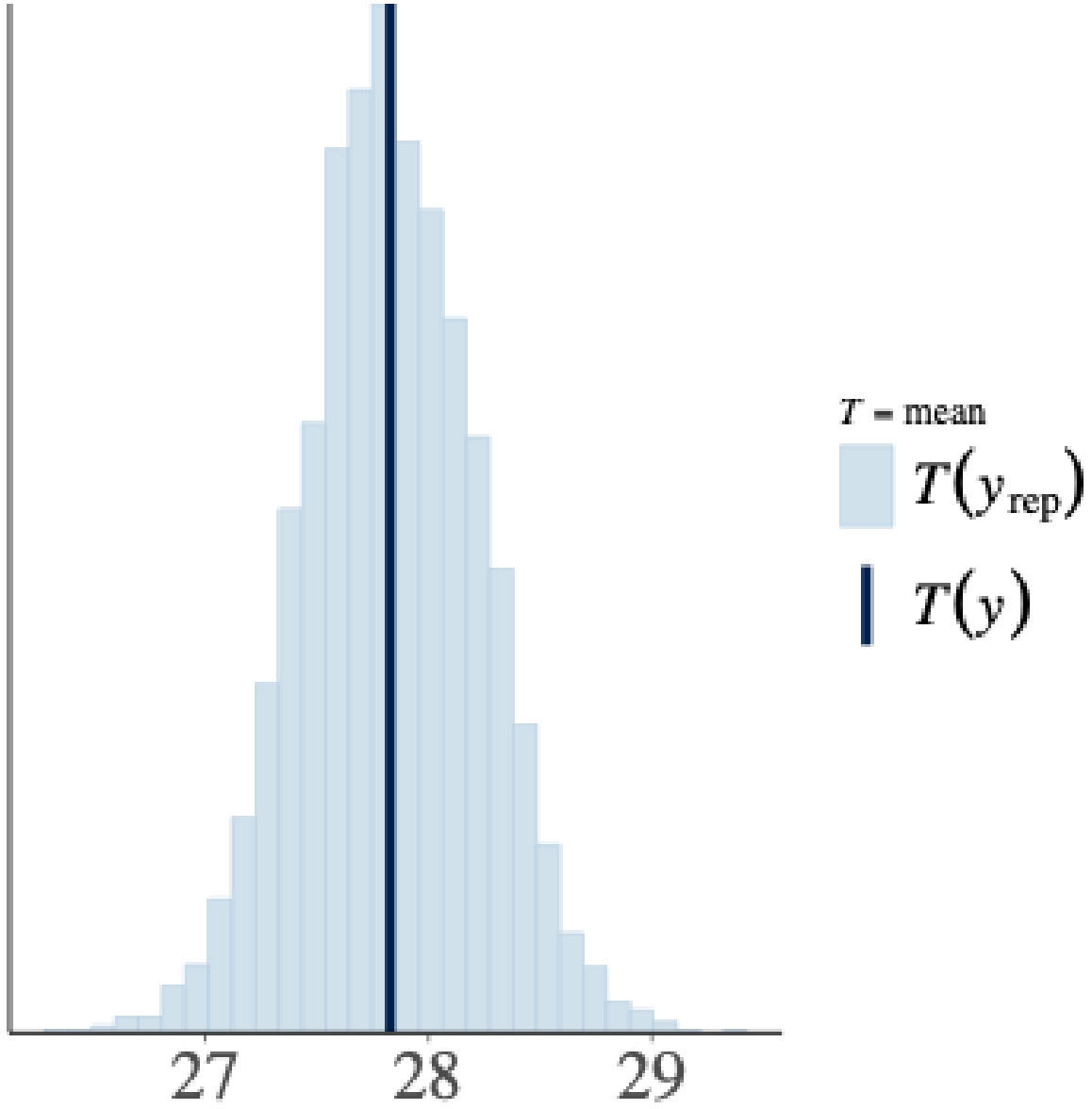
y
y_{rep}

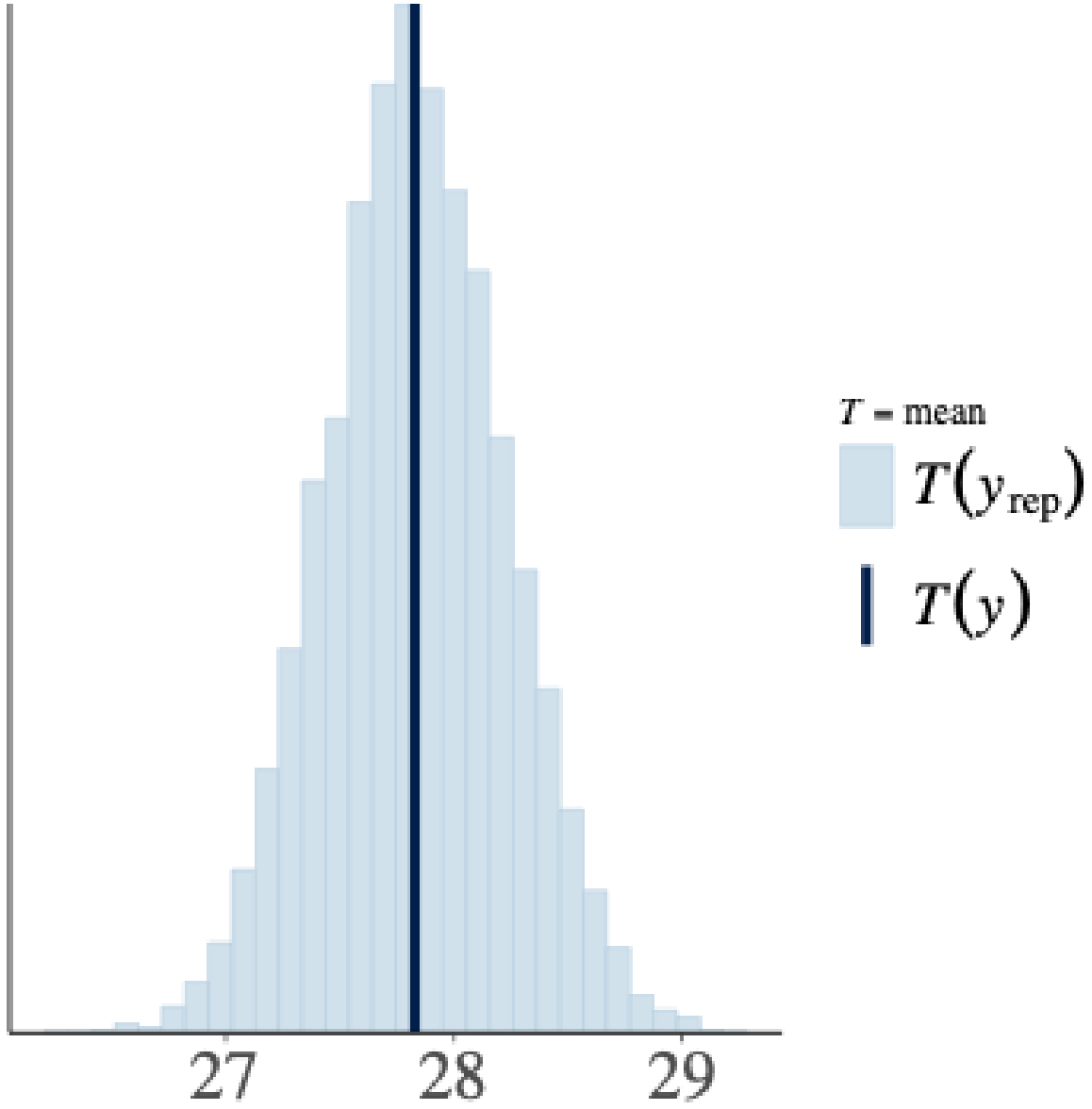
stan_glmer VS model A: proportion of zeros



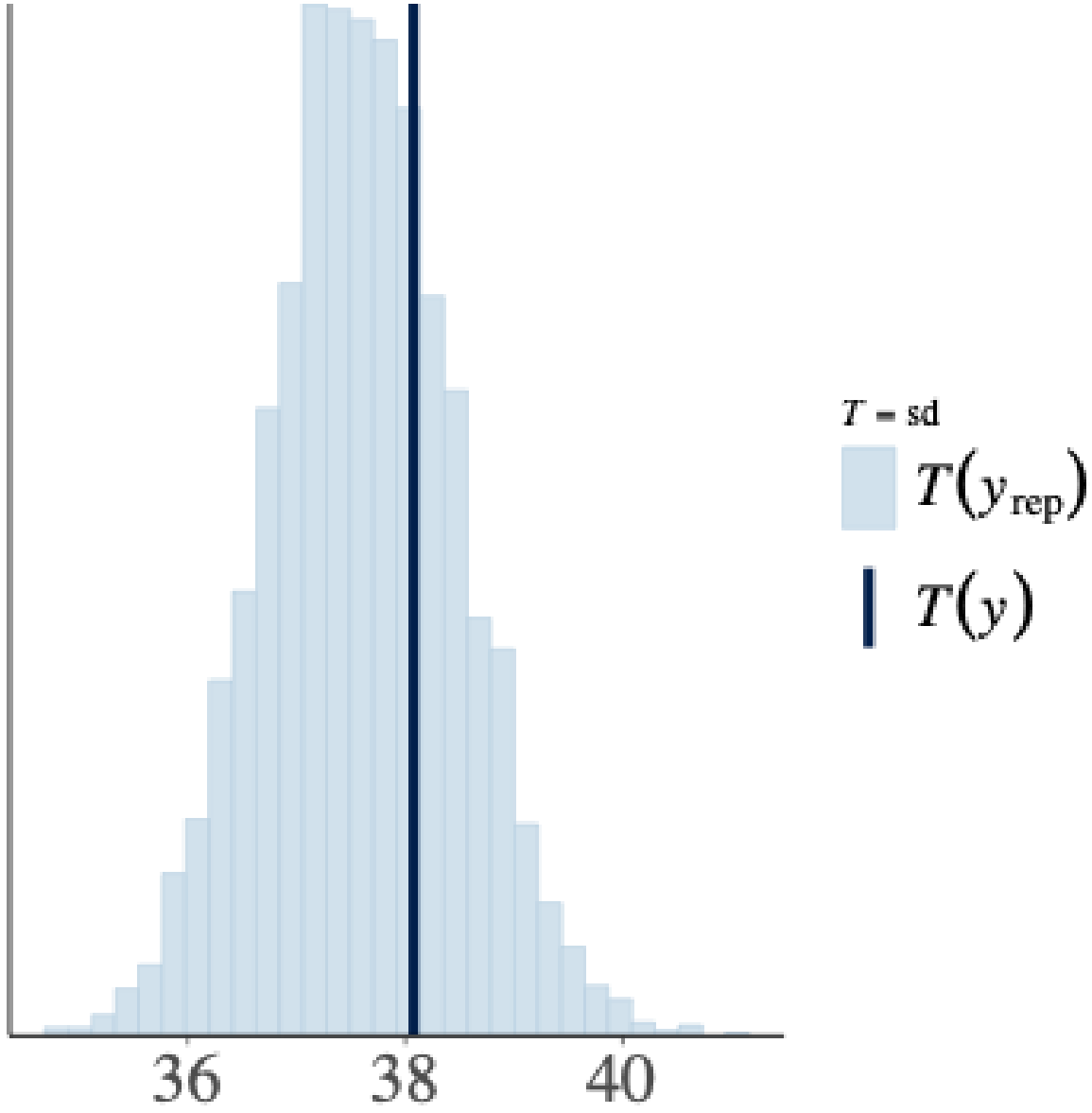


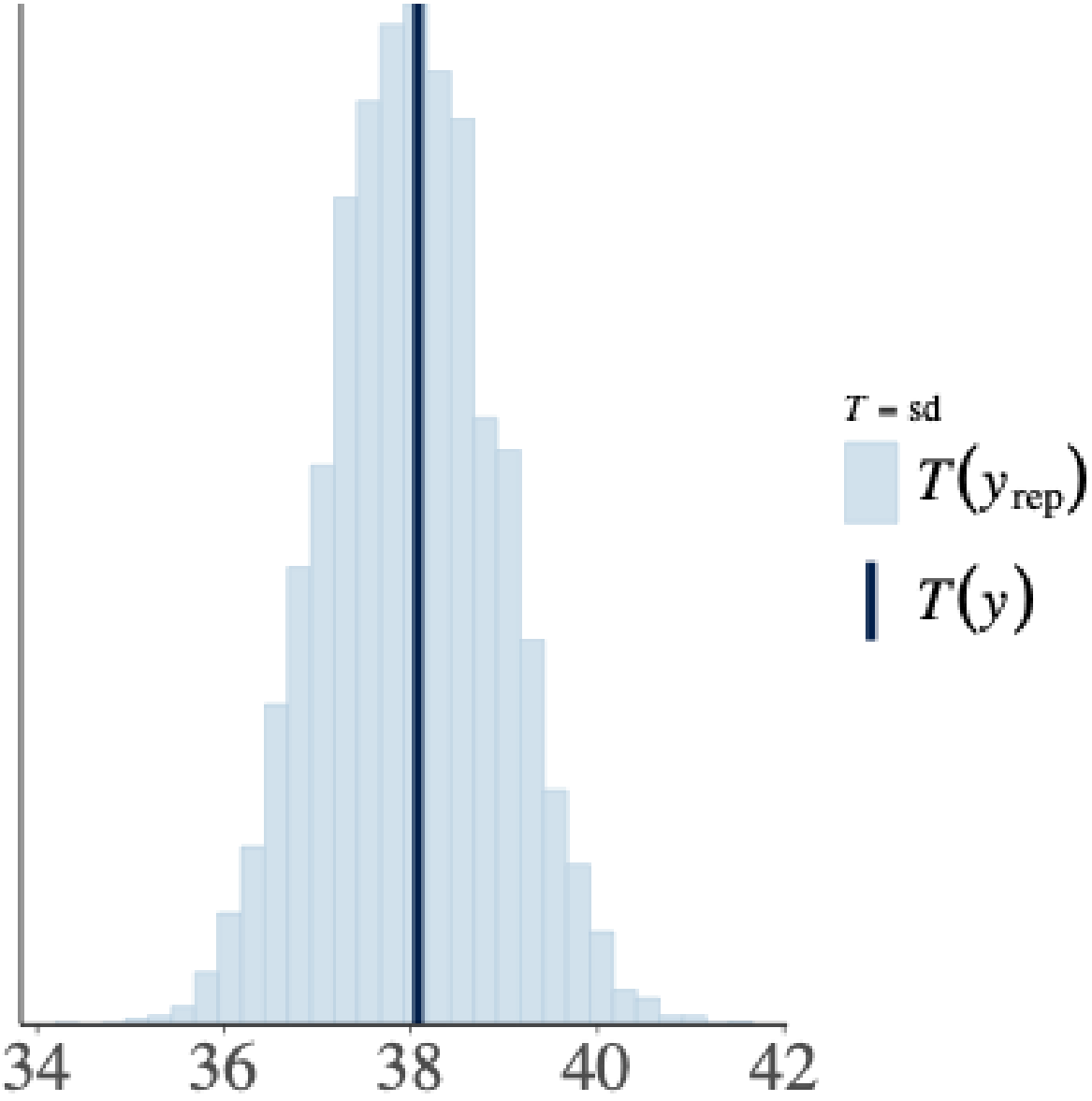
stan_glmer VS model A: mean



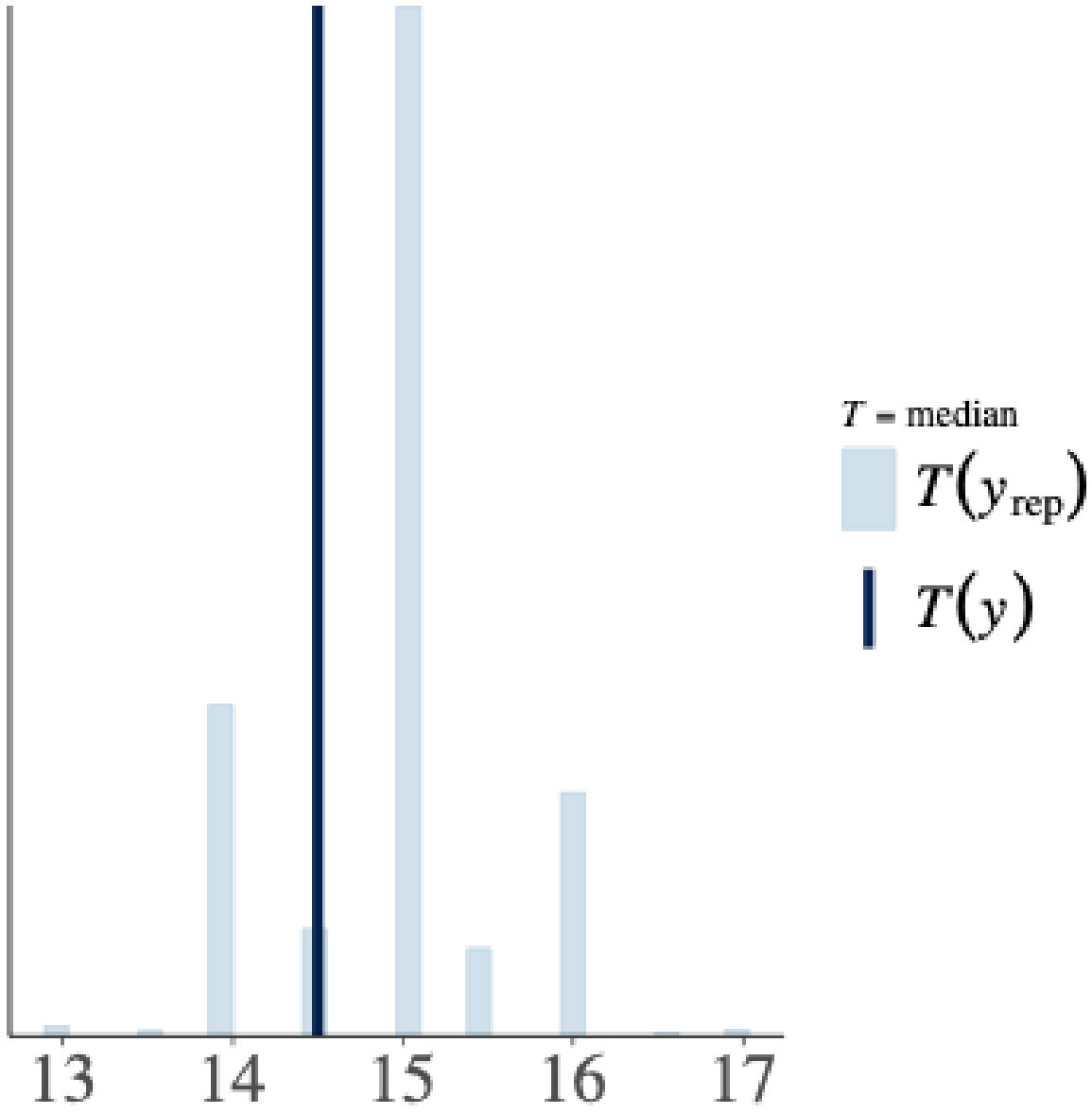


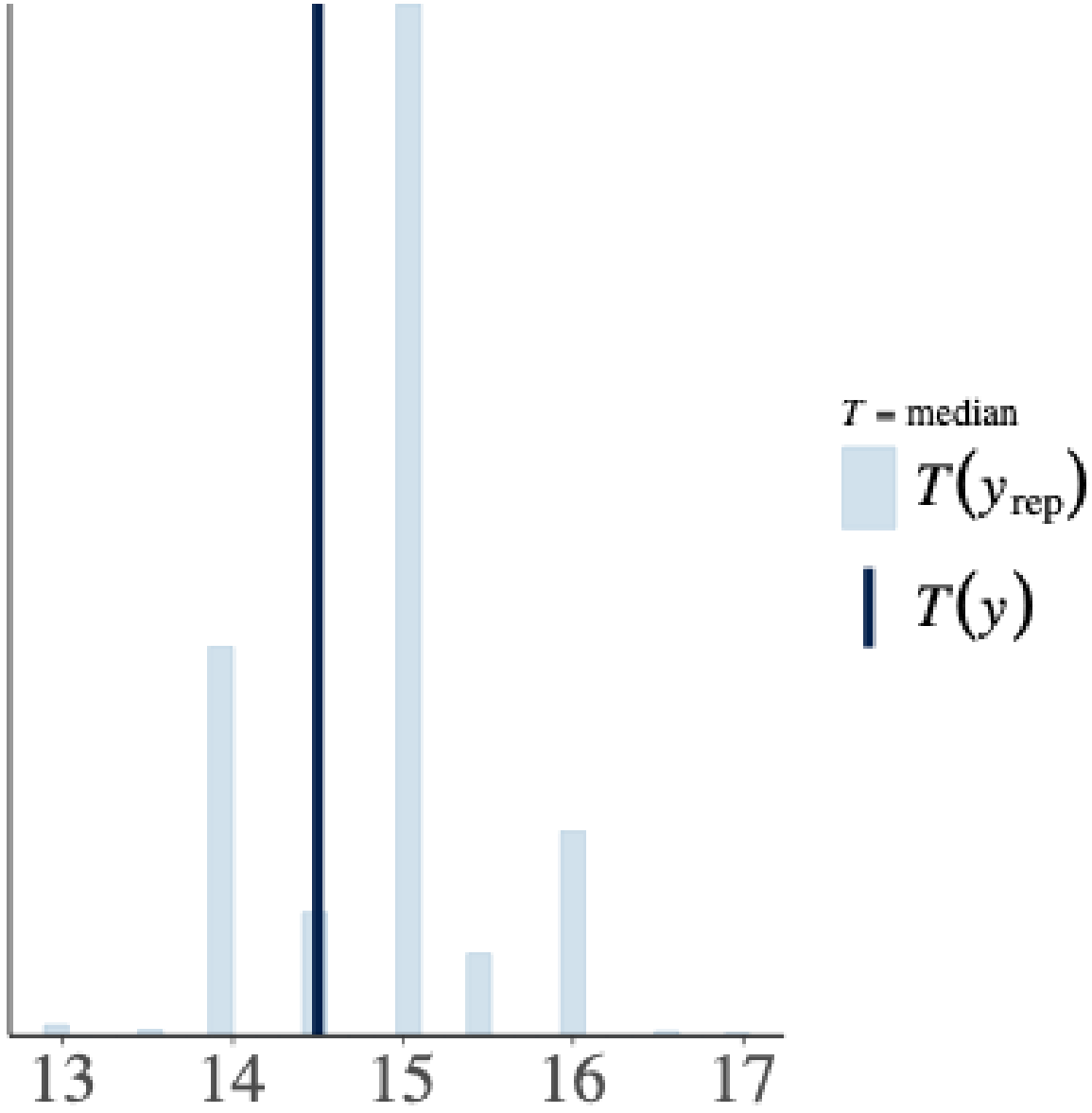
stan_glmer VS model A: standard deviation



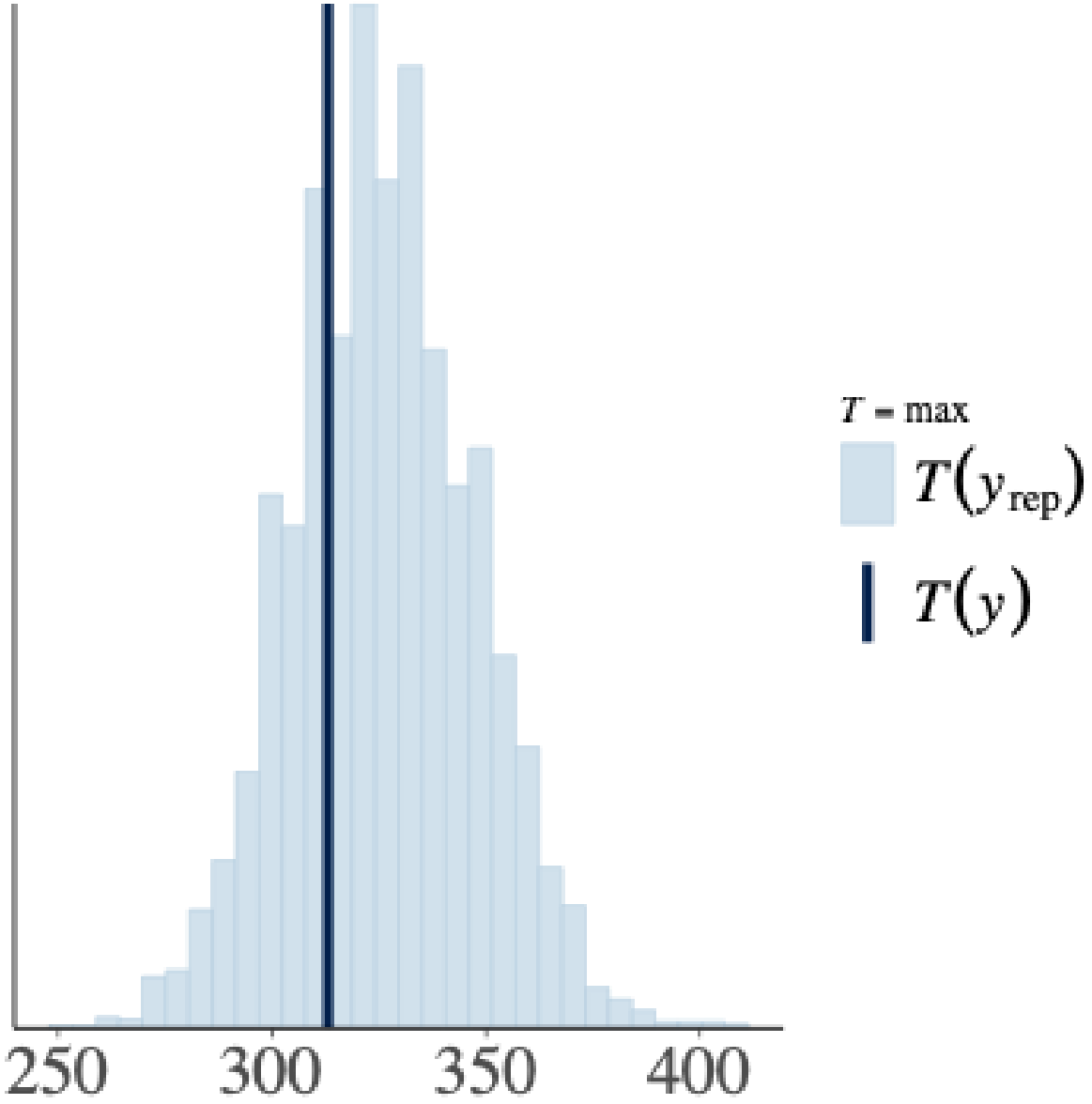


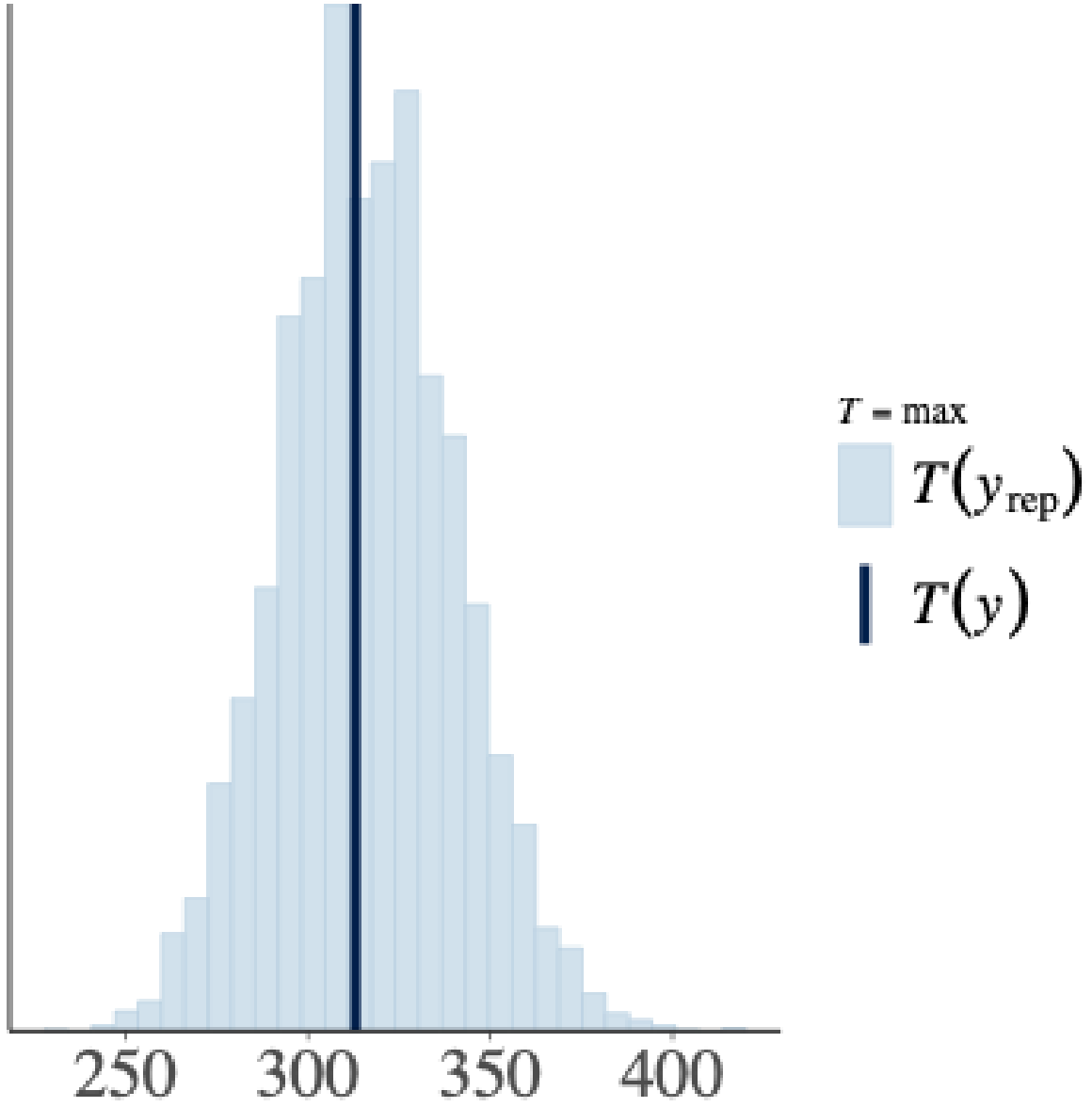
stan_glmer VS model A: median



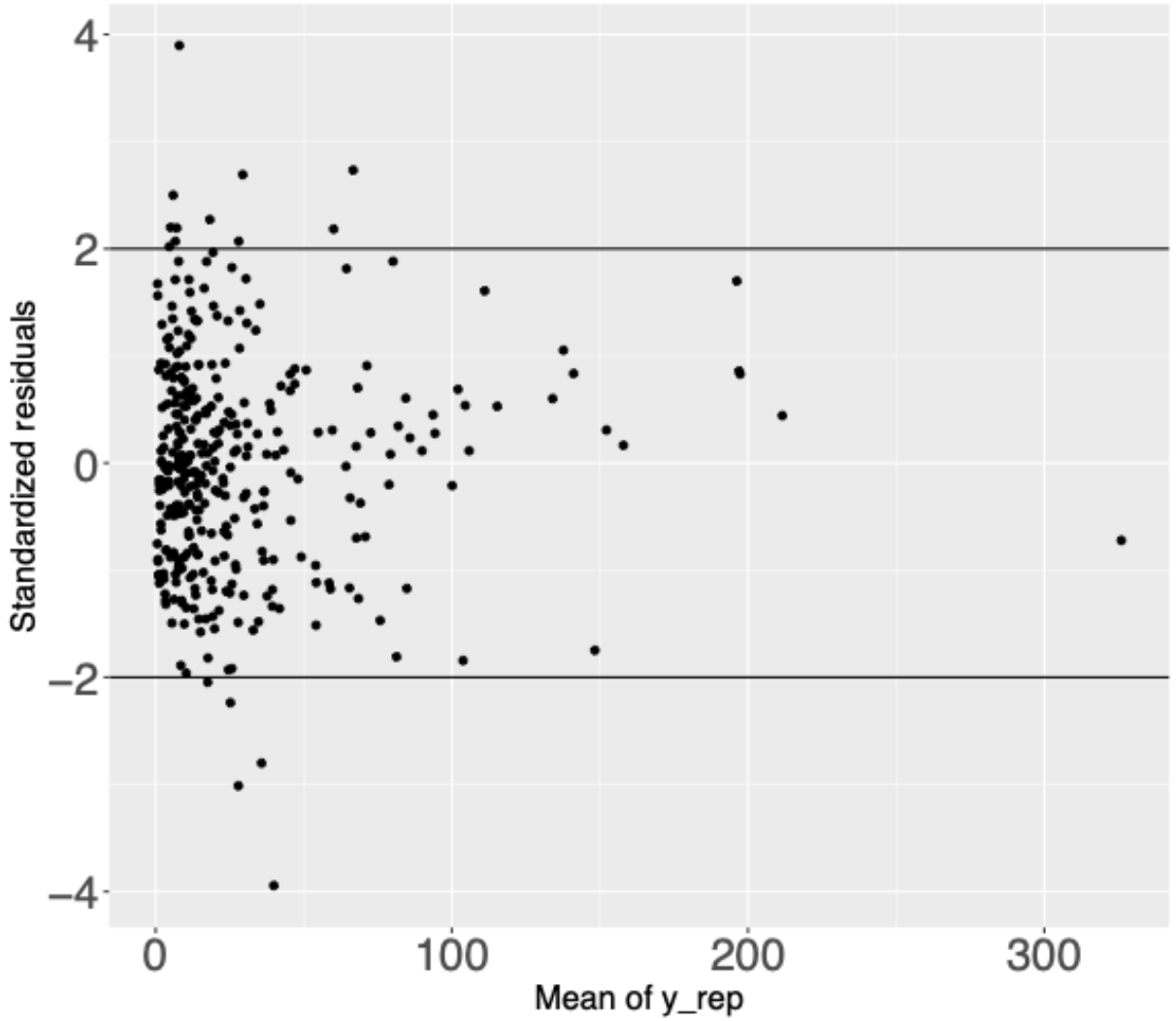


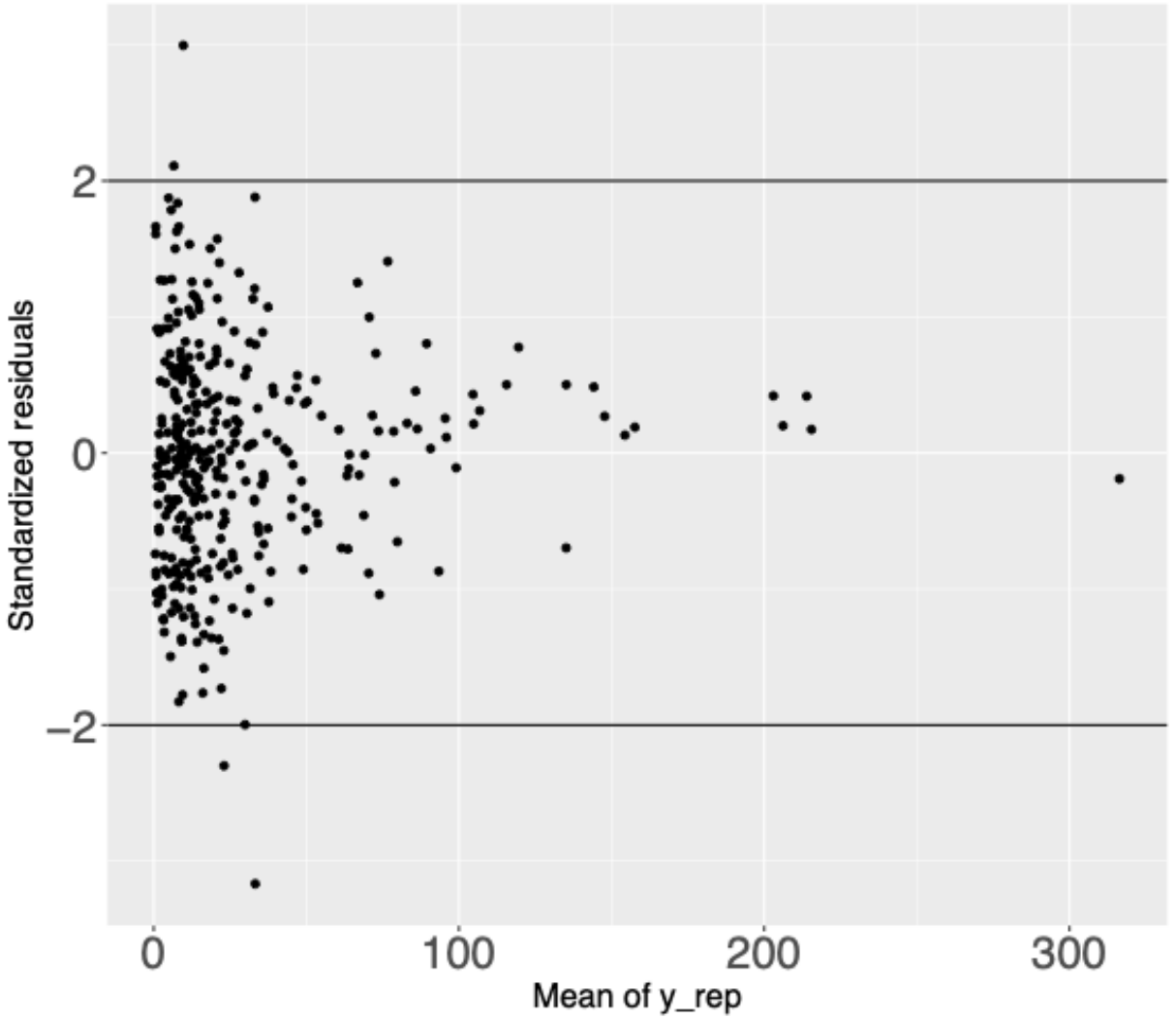
stan_glmer VS model A: maximum



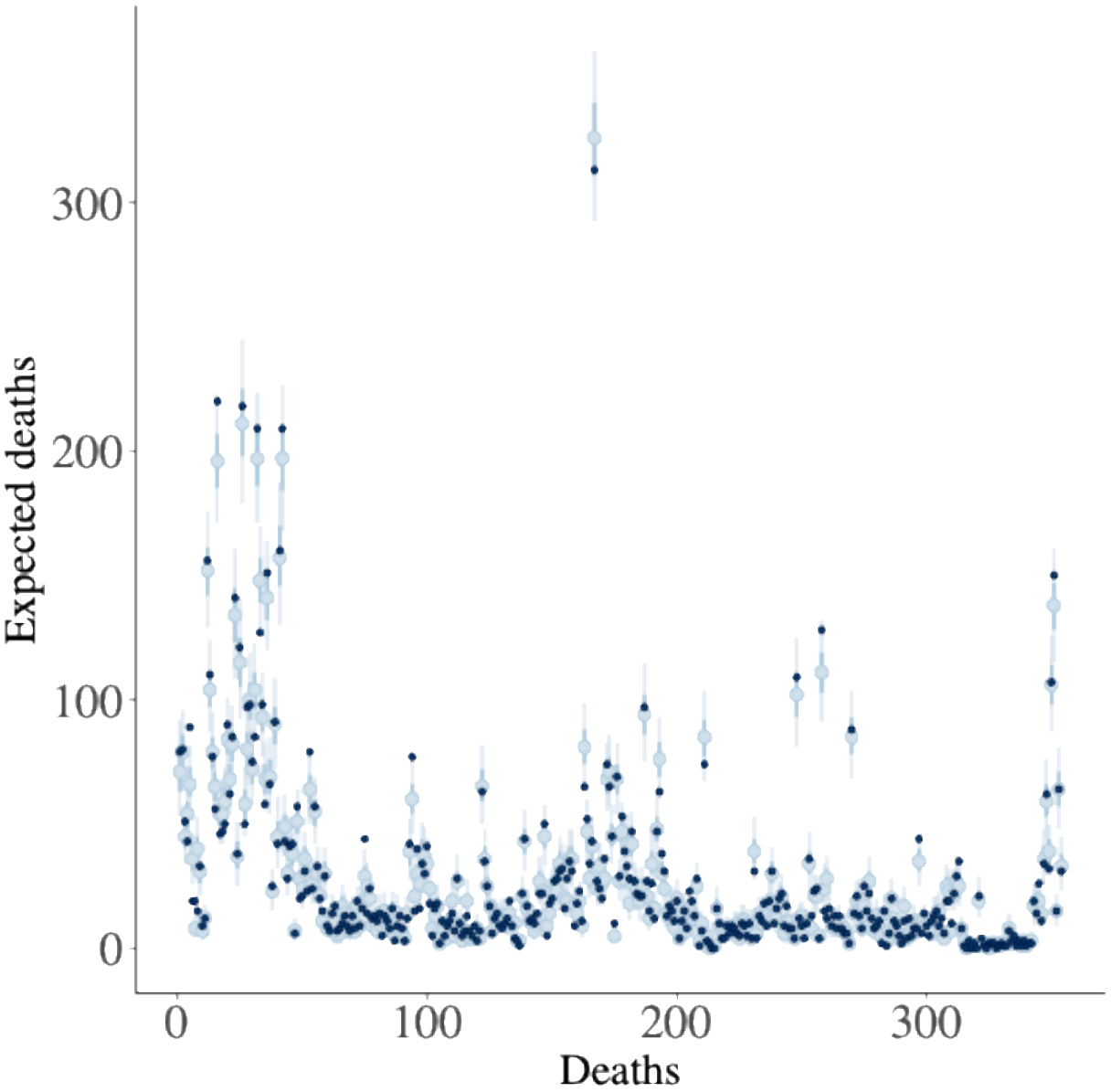


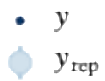
stan_glmer VS model A: standardized residuals

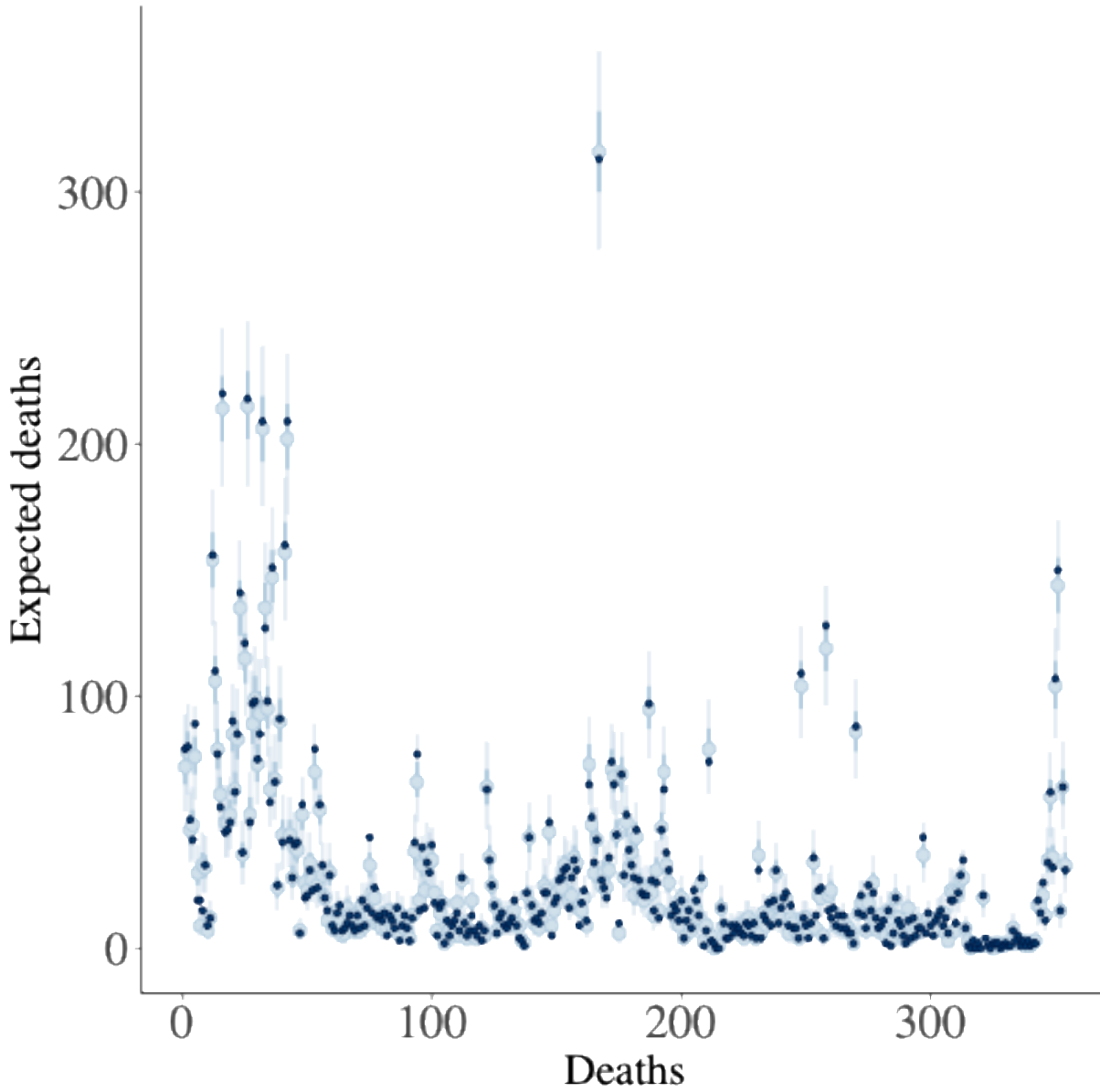




stan_glmer VS model A: predictive intervals







stan_glmer VS model A: Leave One Out Cross Validation

stan_glmer computed from 4000 by 354 log-likelihood matrix.

```
         Estimate   SE
elpd_loo  -1069.3 25.7
p_loo        77.4  8.5
looic      2138.6 51.3

Pareto k diagnostic values:
                         Count Pct.
(-Inf, 0.7]   (good)     340   96.0%
   (0.7, 1]   (bad)       10    2.8%
   (1, Inf)   (very bad)   4    1.1%
```

My model A computed from 8000 (or 4000) by 354 log-likelihood matrix.

```
         Estimate   SE
elpd_loo  -1054.2 21.3
p_loo       112.6  9.5
looic      2108.3 42.6

Pareto k diagnostic values:
                         Count Pct.
(-Inf, 0.7]   (good)     317   89.5%
   (0.7, 1]   (bad)       35    9.9%
   (1, Inf)   (very bad)   2    0.6%
```

```
         Estimate   SE
elpd_loo  -1053.6 20.9
p_loo       112.6  9.0
looic      2107.2 41.8
```

→ Overall they seem interchangeable

VS Langford et al. (1998)

	mean	squared	sd
beta0	−0.07		0.21
beta1	0.00		0.01
sigma_s	0.50	0.25	0.16
sigma_u	0.23	0.053	0.03
sigma_e	0.12	0.014	0.02

Parameters of my model A
≠
Parameters of Langford et al.'s model A

Differences may be due to:

- UVBI for counties were generated from nation stats instead of direct measurements
- β priors were weakly informative $\sim N(0,1)$ instead of not being used
- Bayesian approach instead of frequentist

	Model A	
	Estimate	(SE)
Fixed part		
β_0	0·0103	(0·134)
β_1 (UVBI)	− 0·0360	(0·0107)
β_2 (RDENS)		
β_3 (RGDP)		
Random part		
Level 3: nations		
σ_s^2	0·140	(0·0733)
σ_{st}		
σ_t^2		
Level 2: regions		
σ_u^2	0·0424	(0·00956)
σ_{uv}		
σ_v^2		
Level 1: counties		
σ_{e1}^2	1.11	(0·0937)

VS Langford et al. (1998)

UK, Ireland, Belgium, Netherlands:

Positive relationship

WG, Denmark:

Higher mortality rates

Negative relationship

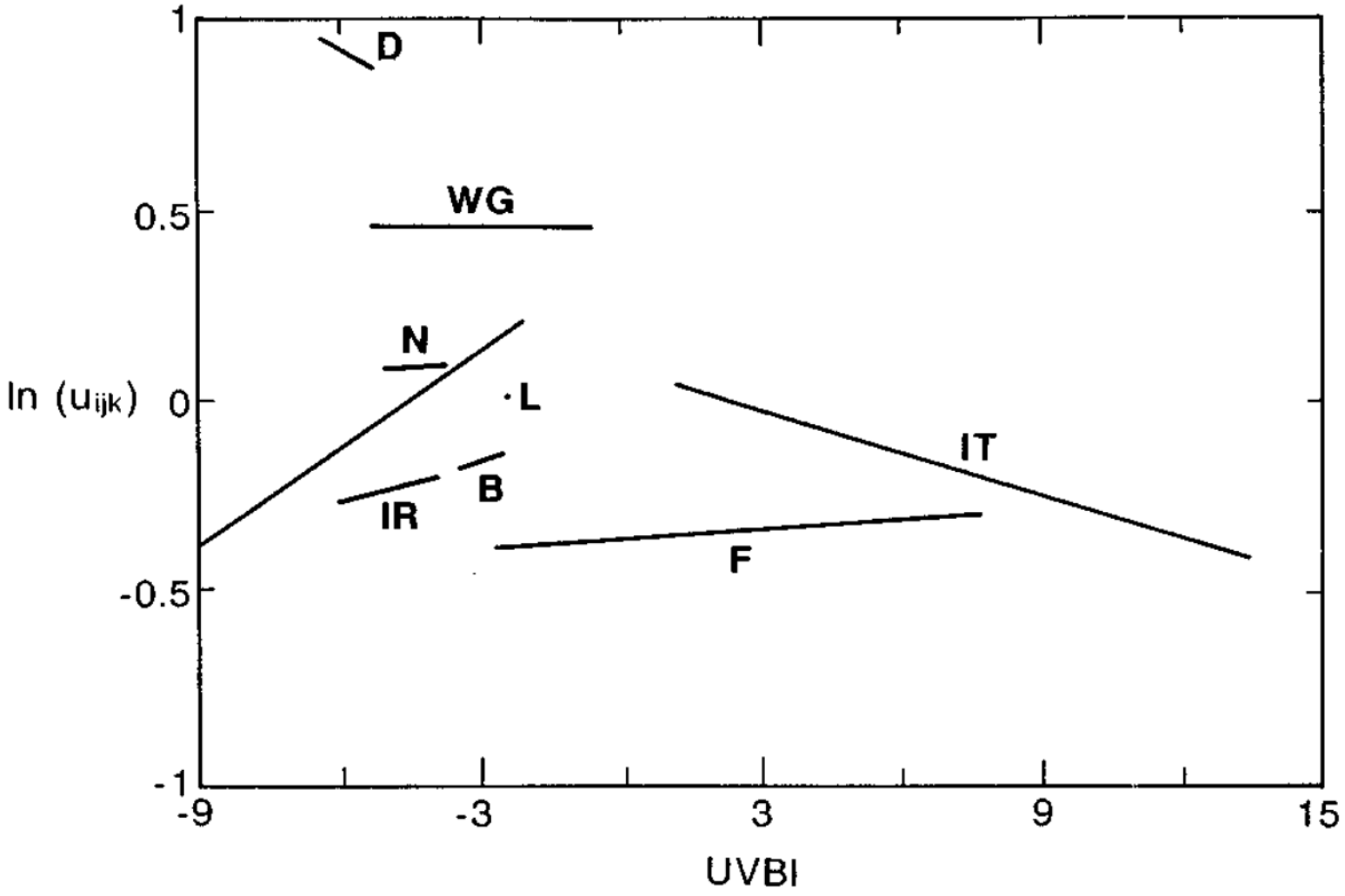
Italy:

Highest exposure to UVB

Negative relationship

France:

Little relationship



VS Langford et al. (1998)

UK, Ireland, Belgium, Netherlands:

 Positive relationship

WG, Denmark:

 Higher mortality rates

 Negative relationship

Italy:

 Highest exposure to UVB

 Negative relationship

France:

 Little relationship

From explorative analysis:

UK, Ireland, Netherlands, Luxembourg:

 Positive relationship

Belgium, WG, Denmark:

 Negative relationship

Italy:

 High(est) exposure to UVB

 Negative relationship

France:

 (slightly) Negative relationship