

Model Building

Jonathan Ma

2025-12-01

Table of Contents

1	Loading the Data In.....	2
2	Model Initialization with Prior	3
2.1	Model Specification.....	3
2.2	Model Fine Tuning.....	4

1 Loading the Data In

We double check that data structure is correct, pull a specific cancer to analyse (breast in this case), and reformat anything so our model runs smoothly.

```
## 'data.frame': 37530 obs. of 13 variables:  
## $ id      : int 1028 1034 1152 2172 4489 5518 9181 9314 9415 10092 ...  
## $ region  : Factor w/ 3 levels "Large","Nonmetro",...: 1 1 1 1 1 1 1 1 1 1 ...  
1 ...  
## $ stage   : Factor w/ 3 levels "Distant","Localized",...: 2 2 2 2 2 2 2 2 2 3 ...  
2 3 ...  
## $ age     : Factor w/ 20 levels "00 years","01-04 years",...: 14 16 17 19  
15 13 17 18 18 17 ...  
## $ sex     : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 1 1 1 1 ...  
## $ raceth  : Factor w/ 3 levels "B","O","W": 3 1 3 2 3 1 2 3 2 3 ...  
## $ grade   : Factor w/ 2 levels "End","Start": 2 2 2 1 2 2 2 2 1 2 ...  
## $ size    : int 30 30 38 27 22 49 22 50 27 27 ...  
## $ year    : int 2017 2016 2016 2017 2017 2016 2017 2017 2017 2016 ...  
## $ marry   : Factor w/ 2 levels "Married","Unmarried": 1 2 1 2 2 2 1 1 2  
2 ...  
## $ site    : int 20 25 35 20 10 10 20 40 35 30 ...  
## $ regionid: int 1 1 1 1 1 1 1 1 1 ...  
## $ patientid: int 1 2 3 4 5 6 7 8 9 10 ...  
  
## 'data.frame': 3031 obs. of 13 variables:  
## $ id      : int 27728 47284 369220 389055 513533 533849 686099 744683 7  
52091 788601 ...  
## $ region  : Factor w/ 3 levels "Large","Nonmetro",...: 1 1 1 1 1 1 1 1 1 1 ...  
1 ...  
## $ stage   : Factor w/ 3 levels "Distant","Localized",...: 3 3 3 2 3 3 3 3 2  
3 2 ...  
## $ age     : Factor w/ 20 levels "00 years","01-04 years",...: 15 19 18 9  
14 18 16 18 18 18 ...  
## $ sex     : Factor w/ 2 levels "Female","Male": 2 1 2 1 1 2 1 1 1 1 ...  
## $ raceth  : Factor w/ 3 levels "B","O","W": 3 3 2 2 2 3 2 3 3 3 ...  
## $ grade   : Factor w/ 2 levels "End","Start": 2 2 2 2 2 2 2 2 2 1 ...  
## $ size    : int 30 40 42 21 30 30 51 50 40 50 ...  
## $ year    : int 2017 2017 2017 2017 2017 2017 2016 2017 2016 2016 ...  
## $ marry   : Factor w/ 2 levels "Married","Unmarried": 1 2 1 2 2 1 1 2 2  
1 ...  
## $ site    : int 50 50 50 50 50 50 50 50 50 50 ...  
## $ regionid: int 1 1 1 1 1 1 1 1 1 ...  
## $ patientid: int 14 17 61 65 103 108 164 197 209 238 ...  
  
## 'data.frame': 3031 obs. of 10 variables:  
## $ age     : Factor w/ 20 levels "00 years","01-04 years",...: 15 19 18 9  
14 18 16 18 18 18 ...  
## $ sex     : Factor w/ 2 levels "Female","Male": 2 1 2 1 1 2 1 1 1 1 ...  
## $ raceth  : Factor w/ 3 levels "B","O","W": 3 3 2 2 2 3 2 3 3 3 ...  
## $ grade   : Factor w/ 2 levels "End","Start": 2 2 2 2 2 2 2 2 2 1 ...
```

```

## $ size      : int  30 40 42 21 30 30 51 50 40 50 ...
## $ year      : int  2017 2017 2017 2017 2017 2017 2016 2017 2016 2016 ...
## $ marry     : Factor w/ 2 levels "Married","Unmarried": 1 2 1 2 2 1 1 2 2
1 ...
## $ regionid  : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
## $ patientid: Factor w/ 3026 levels "14","17","61",..: 1 2 3 4 5 6 7 8 9 1
0 ...
## $ latestage: int  1 1 1 0 1 1 1 0 1 0 ...

```

2 Model Initialization with Prior

We will sample with No U-Turn Sampling (NUTS). We assigned weakly informative priors to all model parameters. Fixed effects received independent $\text{Normal}(0, 2^2)$ priors, encouraging shrinkage without overly restricting plausible values. The intercept was assigned a wider $\text{Normal}(0, 5^2)$ prior to reflect uncertainty in baseline log-odds. Group-level standard deviations (for patientid and regionid) used Student-t(3, 0, 2.5) priors, which provide regularization while allowing for potential group-level variability.

2.1 Model Specification

Logistic regression with:

- Outcome: latestage (1 = regional/distant, 0 = localized)
- Fixed effects: age, sex, raceth, grade, size_z, year_z, marry
- Random intercepts:
 - Patient-level: $(1 \mid \text{patientid})$
 - Region-level: $(1 \mid \text{regionid})$
- Estimation: NUTS via cmdstanr backend
- Priors:
 - $\beta_j \sim \mathcal{N}(0, 2^2)$
 - Intercept $\sim \mathcal{N}(0, 5^2)$
 - Group SDs $\sim t_3(0, 2.5)$ (approx IG)

```

## Running MCMC with 4 parallel chains...
##
## Chain 1 Iteration:    1 / 2000 [  0%] (Warmup)
## Chain 2 Iteration:    1 / 2000 [  0%] (Warmup)
## Chain 3 Iteration:    1 / 2000 [  0%] (Warmup)
## Chain 4 Iteration:    1 / 2000 [  0%] (Warmup)
## Chain 3 Iteration: 1001 / 2000 [ 50%] (Sampling)
## Chain 4 Iteration: 1001 / 2000 [ 50%] (Sampling)
## Chain 1 Iteration: 1001 / 2000 [ 50%] (Sampling)
## Chain 2 Iteration: 1001 / 2000 [ 50%] (Sampling)
## Chain 4 Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 4 finished in 132.2 seconds.
## Chain 1 Iteration: 2000 / 2000 [100%] (Sampling)

```

```

## Chain 1 finished in 136.5 seconds.
## Chain 3 Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 3 finished in 141.7 seconds.
## Chain 2 Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 2 finished in 158.9 seconds.
##
## All 4 chains finished successfully.
## Mean chain execution time: 142.3 seconds.
## Total execution time: 159.0 seconds.

```

2.2 Model Fine Tuning

Increasing Warmpup to 1500 and 3000 iterations, we got divergence on 1/6000 transitions (0.017%), and 2 chains haveing E-BFMI (Energy Bayesian Fraction of Missing Information) less than 0.3 which is a diagnostic of poor posterior exploration which usually happens due to poor scaling. To counter this, we will standardize size and year variables, and increase stepsize to 0.9995 to avoid divergences.

```

## Running MCMC with 4 parallel chains...
##
## Chain 1 Iteration: 1 / 3000 [ 0%] (Warmup)
## Chain 2 Iteration: 1 / 3000 [ 0%] (Warmup)
## Chain 3 Iteration: 1 / 3000 [ 0%] (Warmup)
## Chain 4 Iteration: 1 / 3000 [ 0%] (Warmup)
## Chain 4 Iteration: 1501 / 3000 [ 50%] (Sampling)
## Chain 3 Iteration: 1501 / 3000 [ 50%] (Sampling)
## Chain 2 Iteration: 1501 / 3000 [ 50%] (Sampling)
## Chain 1 Iteration: 1501 / 3000 [ 50%] (Sampling)
## Chain 4 Iteration: 3000 / 3000 [100%] (Sampling)
## Chain 4 finished in 165.6 seconds.
## Chain 3 Iteration: 3000 / 3000 [100%] (Sampling)
## Chain 3 finished in 173.5 seconds.
## Chain 2 Iteration: 3000 / 3000 [100%] (Sampling)
## Chain 2 finished in 244.7 seconds.
## Chain 1 Iteration: 3000 / 3000 [100%] (Sampling)
## Chain 1 finished in 245.1 seconds.
##
## All 4 chains finished successfully.
## Mean chain execution time: 207.2 seconds.
## Total execution time: 245.2 seconds.

```