Some prior predictive checks

```
library(tidyverse)
source("../../init.R")
```

Navigating to the US directory to load use the US data loading wrappers.

```
find_and_set_directory("US/exploration")
source("load_and_preprocess.R")
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths

##
## Attaching package: 'jsonlite'

## The following object is masked from 'package:purrr':
##
##     flatten

## Rows: 19311 Columns: 139
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (14): TRACT, NAME, ST, STATE_FIPS, CNTY_FIPS, STCOFIPS, STATE_ABBR, STA...
## dbl (125): total_ppl_acs20E, median_annual_incomeE, house_price_medianE, ren...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 20044 Columns: 124
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): STATE_ABBR, zip
## dbl (122): total_ppl_acs20E, median_annual_incomeE, house_price_medianE, bui...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
find_and_set_directory("US/exploration")
source("compute_descriptives.R")
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

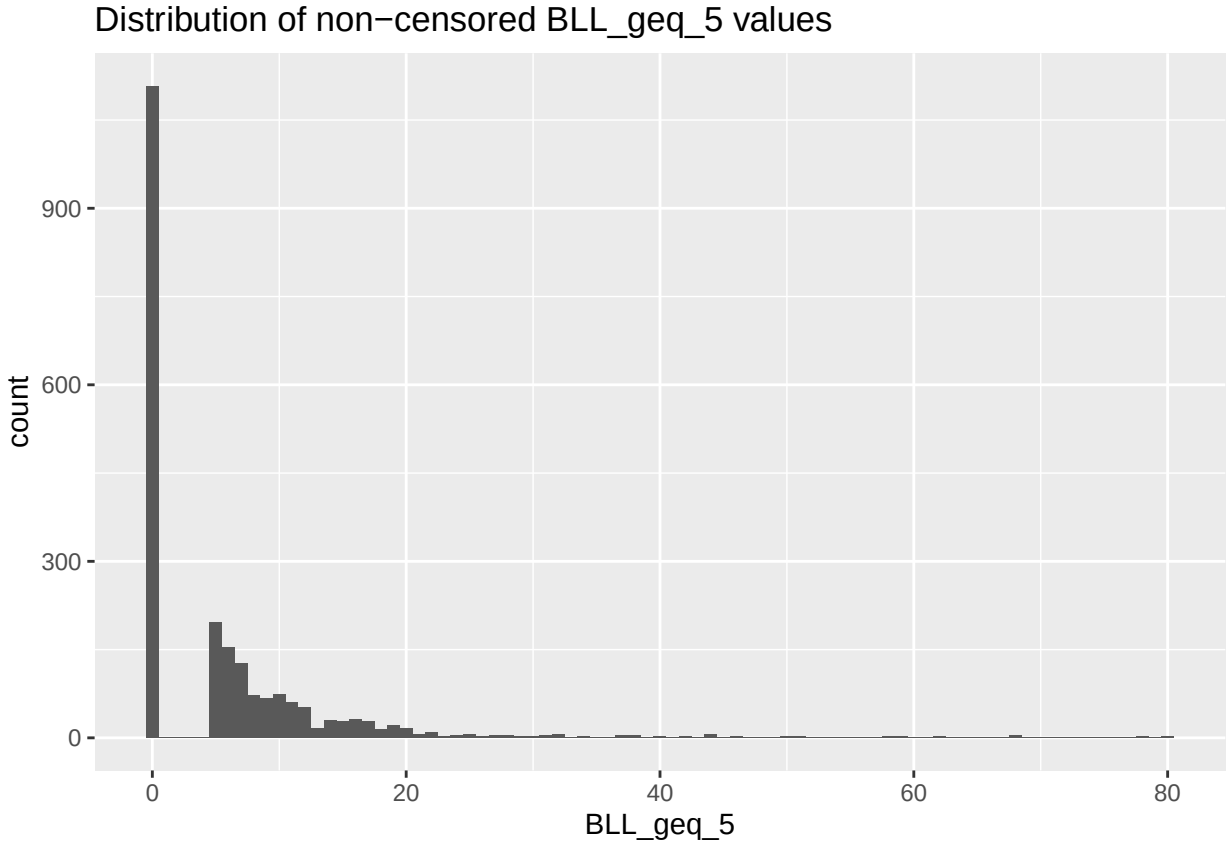
In particular, I want to focus on two tract states, Ohio and Massachusetts, as well as two ZIP states, New Jersey and Illinois.

Ohio

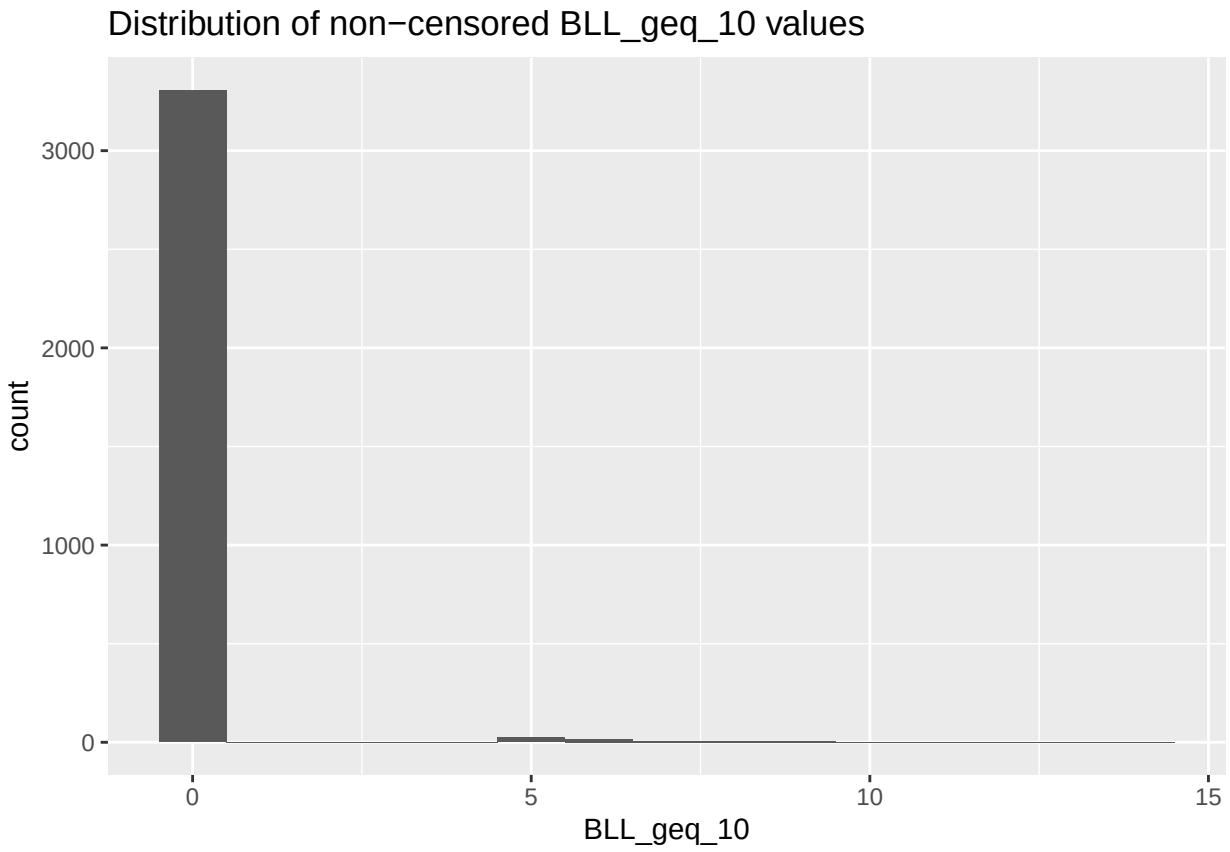
```
data_summary("OH", year = 2010)
```

```
## [1] "Loading OH from processed_data"

## Rows: 94292 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): state, BLL_geq_5, BLL_geq_10, tested
## dbl (2): tract, year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

## Warning: Using `across()` in `filter()` was deprecated in dplyr 1.0.8.
## i Please use `if_any()` or `if_all()` instead.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## [1] "Additional features added: "

## Warning in datasummary_skim_numeric(data, output = output, fmt = fmt, histogram
## = histogram, : The histogram argument is only supported for (a) output types
## "default", "html", or "kableExtra"; (b) writing to file paths with extensions
## ".html", ".jpg", or ".png"; and (c) Rmarkdown or knitr documents compiled to
## PDF or HTML. Use `histogram=FALSE` to silence this warning.

## [[1]]
## # A tibble: 1 x 18
##    year n_obs lead_cens_5 lead_nocens_5 lead_censR_5 lead_cens_10 lead_nocens_10
##   <dbl> <int>       <int>         <int>        <dbl>        <int>          <int>
## 1  2010  4722        2534          2188        0.537         1350           3372
## # i 11 more variables: lead_censR_10 <dbl>, test_cens <int>, test_censR <dbl>,
## #   sup_threshold_5 <dbl>, sup_threshold_10 <dbl>, median_lead_5 <dbl>,
## #   q75_lead_5 <dbl>, max_lead_5 <dbl>, median_lead_10 <dbl>,
## #   q75_lead_10 <dbl>, max_lead_10 <dbl>
##
## [[2]]
##
```

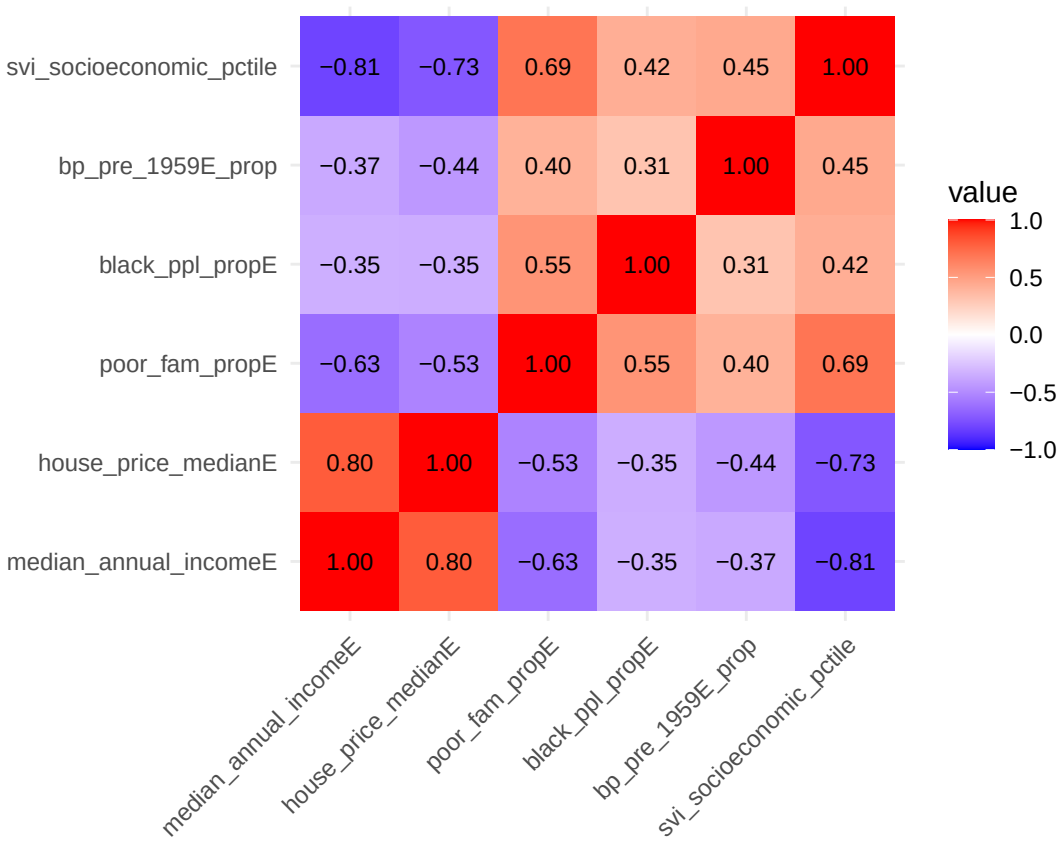
```
##
## |                          | Unique (#) | Missing (%) |  Mean  |   SD   |   Min   | Median |   Max   |
## |:-------------------------|-----------:|------------:|-------:|-------:|--------:|-------:|--------:|
## | BLL_geq_5                |         44 |           0 |    5.3 |    6.1 |     0.0 |    5.0 |    80.0 |
## | BLL_geq_10               |          8 |           0 |    1.5 |    2.3 |     0.0 |    0.0 |    14.0 |
## | tested                   |        163 |           0 |   42.3 |   32.4 |     5.0 |   35.0 |   372.0 |
## | year                     |          1 |           0 | 2010.0 |    0.0 |  2010.0 | 2010.0 |  2010.0 |
## | tested_ell               |          1 |           0 |    5.0 |    0.0 |     5.0 |    5.0 |     5.0 |
## | ell_5                    |          1 |           0 |    4.0 |    0.0 |     4.0 |    4.0 |     4.0 |
## | ell_10                   |          1 |           0 |    4.0 |    0.0 |     4.0 |    4.0 |     4.0 |
## | median_annual_incomeE    |       2271 |           0 |    0.0 |    1.0 |    -2.5 |   -0.0 |     5.5 |
## | house_price_medianE      |       1464 |           0 |    0.0 |    1.0 |    -1.6 |   -0.1 |     6.2 |
## | poor_fam_propE           |       2308 |           0 |   -0.0 |    1.0 |    -1.1 |   -0.4 |     5.6 |
## | black_ppl_propE          |       2157 |           0 |   -0.1 |    0.9 |    -0.6 |   -0.5 |     3.5 |
## | bp_pre_1959E_prop        |       2358 |           0 |   -0.0 |    1.0 |    -1.8 |   -0.1 |     2.0 |
## | svi_socioeconomic_pctile |       2104 |           0 |   -0.0 |    1.0 |    -2.2 |   -0.0 |     1.6 |
## | under_yo5_pplE           |        529 |           0 |  229.5 |  137.7 |     8.0 |  204.0 |  1263.0 |
## | ped_per_100k             |         69 |           0 |   87.0 |   67.0 |     0.0 |   68.1 |   243.7 |
##
## [[3]]
## [[3]][[1]]
```




```
##
## [[3]][[2]]
```


Distribution of non−censored BLL_geq_10 values
3000
2000
1000
0
count
0
5
10
15
BLL_geq_10

```
##
##
## [[4]]
```



I take one additional preprocessing steps on the predictors passed to the logit side (income, building period, and SVI), I also standardise the pediatrician rate.

```
# standardise pediatrician rate
oh_merged <- oh_merged |>
  mutate(ped_per_100k = (ped_per_100k - mean(ped_per_100k)) / sd(ped_per_100k))
```

```
# load STAN model
library(cmdstanr)
```

```
## This is cmdstanr version 0.6.1

## - CmdStanR documentation and vignettes: mc-stan.org/cmdstanr

## - CmdStan path: /Users/lasse/.cmdstan/cmdstan-2.33.0

## - CmdStan version: 2.33.0

##
## A newer version of CmdStan is available. See ?install_cmdstan() to install it.
## To disable this check set option or environment variable CMDSTANR_NO_VER_CHECK=TRUE.
```

```
find_and_set_directory("models/prior_predictive_checks")
ppc_stan <- cmdstan_model("poisson_thinned_exclusion_ppc.stan")
```

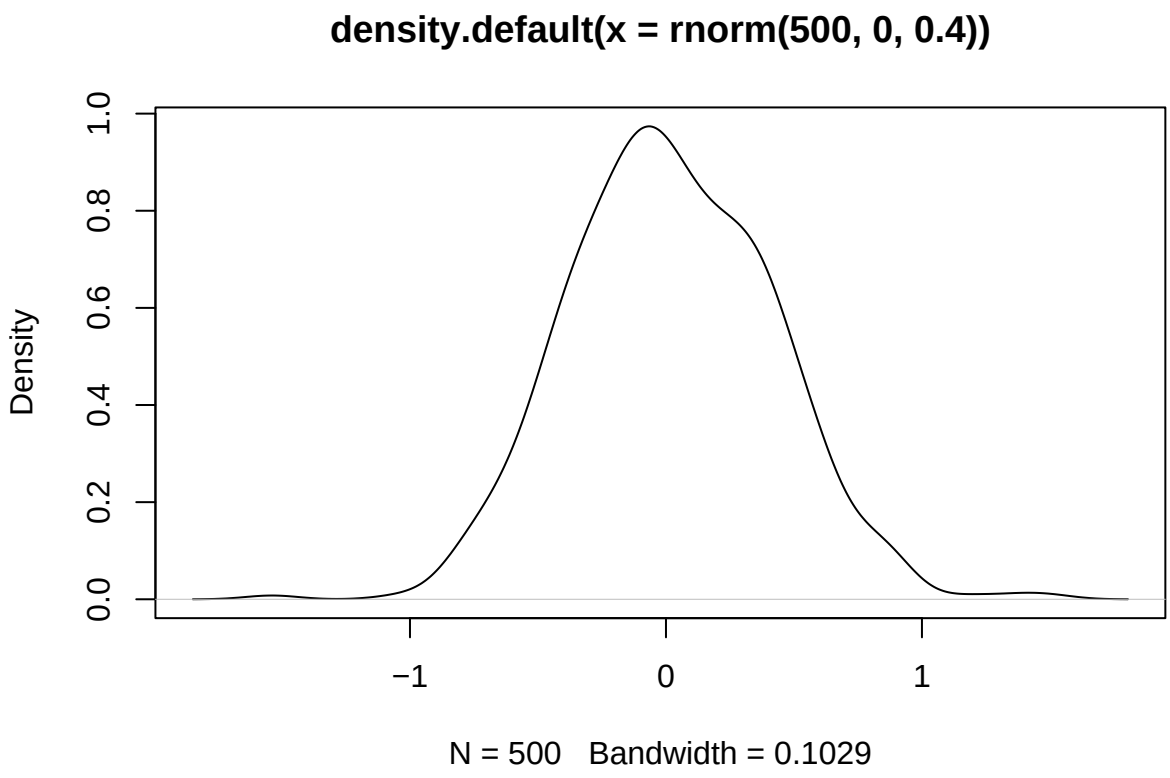
```
## Warning in readLines(stan_file): incomplete final line found on
## 'poisson_thinned_exclusion_ppc.stan'
```

```
# create function to prep data for given STAN model

stan_data <- function(state_final){
  list(
    N_obs = state_final |> filter(!BLL_geq_5_suppressed) |> count() |> pull(n),
    N_cens = state_final |> filter(BLL_geq_5_suppressed) |> count() |> pull(n),
    y_obs = state_final |> filter(!BLL_geq_5_suppressed) |> pull(BLL_geq_5),
    # obs
    median_income_obs = state_final |> filter(!BLL_geq_5_suppressed) |> pull(median_annual_incomeE),
    house_price_obs = state_final |> filter(!BLL_geq_5_suppressed) |> pull(house_price_medianE),
    poverty_obs = state_final |> filter(!BLL_geq_5_suppressed) |> pull(poor_fam_propE),
    black_prop_obs = state_final |> filter(!BLL_geq_5_suppressed) |> pull(black_ppl_propE),
    building_period_obs = state_final |> filter(!BLL_geq_5_suppressed) |> pull(bp_pre_1959E_prop),
    svi_obs = state_final |> filter(!BLL_geq_5_suppressed) |> pull(svi_socioeconomic_pctile),
    # censored
    median_income_cens = state_final |> filter(BLL_geq_5_suppressed) |> pull(median_annual_incomeE),
    house_price_cens = state_final |> filter(BLL_geq_5_suppressed) |> pull(house_price_medianE),
    poverty_cens = state_final |> filter(BLL_geq_5_suppressed) |> pull(poor_fam_propE),
    black_prop_cens = state_final |> filter(BLL_geq_5_suppressed) |> pull(black_ppl_propE),
    building_period_cens = state_final |> filter(BLL_geq_5_suppressed) |> pull(bp_pre_1959E_prop),
    svi_cens = state_final |> filter(BLL_geq_5_suppressed) |> pull(svi_socioeconomic_pctile),
    # pediatricians & kids
    z_obs = state_final |> filter(!BLL_geq_5_suppressed) |> pull(ped_per_100k),
    z_cens = state_final |> filter(BLL_geq_5_suppressed) |> pull(ped_per_100k),
    kids_obs = state_final |> filter(!BLL_geq_5_suppressed) |> pull(under_yo5_pplE),
    kids_cens = state_final |> filter(BLL_geq_5_suppressed) |> pull(under_yo5_pplE)
  )
}
```

Setting the priors as follows:

In the logit: - intercept of N(0, 1.2) for "uniform" log-odds - slopes of N(0,0.5) for all 4 predictors (income, building period, SVI, and pediatrician rate) In the poisson: - intercept of N(-1.58,2) based on the NHANES national average of 2.6% - slopes of N(0,0.5) for all intercepts - for higher variances some observations go beyond the rate tolerance - except income and the building period, which I center -0.1 and 0.1 respectively.

```
# plot N(0, 0.6)
plot(density(rnorm(500,0,0.4)))
```




```
# sample from prior predictive distribution
oh_stan <- stan_data(oh_merged)

ppc_samples <- ppc_stan$sample(
  data = oh_stan,
  iter_sampling = 500,
  refresh = NULL,
  fixed_param = TRUE) # required for RNG?
```

```
## Running MCMC with 4 sequential chains...
##
## Chain 1 Iteration:   1 / 500 [  0%]  (Sampling)
## Chain 1 Iteration: 100 / 500 [ 20%]  (Sampling)
## Chain 1 Iteration: 200 / 500 [ 40%]  (Sampling)
## Chain 1 Iteration: 300 / 500 [ 60%]  (Sampling)
## Chain 1 Iteration: 400 / 500 [ 80%]  (Sampling)
## Chain 1 Iteration: 500 / 500 [100%]  (Sampling)
## Chain 1 finished in 1.8 seconds.
## Chain 2 Iteration:   1 / 500 [  0%]  (Sampling)
## Chain 2 Iteration: 100 / 500 [ 20%]  (Sampling)
## Chain 2 Iteration: 200 / 500 [ 40%]  (Sampling)
## Chain 2 Iteration: 300 / 500 [ 60%]  (Sampling)
## Chain 2 Iteration: 400 / 500 [ 80%]  (Sampling)
## Chain 2 Iteration: 500 / 500 [100%]  (Sampling)
## Chain 2 finished in 1.8 seconds.
## Chain 3 Iteration:   1 / 500 [  0%]  (Sampling)
```

```
## Chain 3 Iteration: 100 / 500 [ 20%]  (Sampling)
## Chain 3 Iteration: 200 / 500 [ 40%]  (Sampling)
## Chain 3 Iteration: 300 / 500 [ 60%]  (Sampling)
## Chain 3 Iteration: 400 / 500 [ 80%]  (Sampling)
## Chain 3 Iteration: 500 / 500 [100%]  (Sampling)
## Chain 3 finished in 1.8 seconds.
## Chain 4 Iteration:   1 / 500 [  0%]  (Sampling)
## Chain 4 Iteration: 100 / 500 [ 20%]  (Sampling)
## Chain 4 Iteration: 200 / 500 [ 40%]  (Sampling)
## Chain 4 Iteration: 300 / 500 [ 60%]  (Sampling)
## Chain 4 Iteration: 400 / 500 [ 80%]  (Sampling)
## Chain 4 Iteration: 500 / 500 [100%]  (Sampling)
## Chain 4 finished in 1.8 seconds.
##
## All 4 chains finished successfully.
## Mean chain execution time: 1.8 seconds.
## Total execution time: 7.7 seconds.
```

```
ppc_samples$summary(variables = c("y_thinned"))
```

```
## # A tibble: 4,722 x 10
##    variable       mean median    sd   mad    q5   q95  rhat ess_bulk ess_tail
##    <chr>         <num>  <num> <num> <num> <num> <num> <num>    <num>    <num>
##  1 y_thinned[1]   1.80      0 25.0      0     0     4  1.00    1907.    1933.
##  2 y_thinned[2]   1.23      0 12.3      0     0     4  1.00    1827.    1816.
##  3 y_thinned[3]   1.64      0 17.4      0     0     4  1.00    2009.    1763.
##  4 y_thinned[4]   1.01      0  8.13     0     0     4  1.00    1953.    1898.
##  5 y_thinned[5]   1.00      0  8.30     0     0     4  1.00    2156.    2018.
##  6 y_thinned[6]   1.99      0 20.7      0     0     6  1.00    1994.    1821.
##  7 y_thinned[7]   1.04      0  8.88     0     0     4  1.00    2044.    1989.
##  8 y_thinned[8]   1.11      0 10.9      0     0     4  1.00    1700.    1758.
##  9 y_thinned[9]   1.02      0  8.95     0     0     4  1.00    2113.    1972.
## 10 y_thinned[10]  1.29      0  7.16     0     0     5  1.00    2032.    2000.
## # i 4,712 more rows
```

```
# add y_thinned from $summary to oh_merged
oh_merged_ppc <- oh_merged |>
  bind_cols(ppc_samples$summary(variables = c("y_thinned")) |>
              filter(str_detect(variable, "y_thinned")) |>
              select(c(mean, median, q5, q95)))
```

Plot the thinning rates

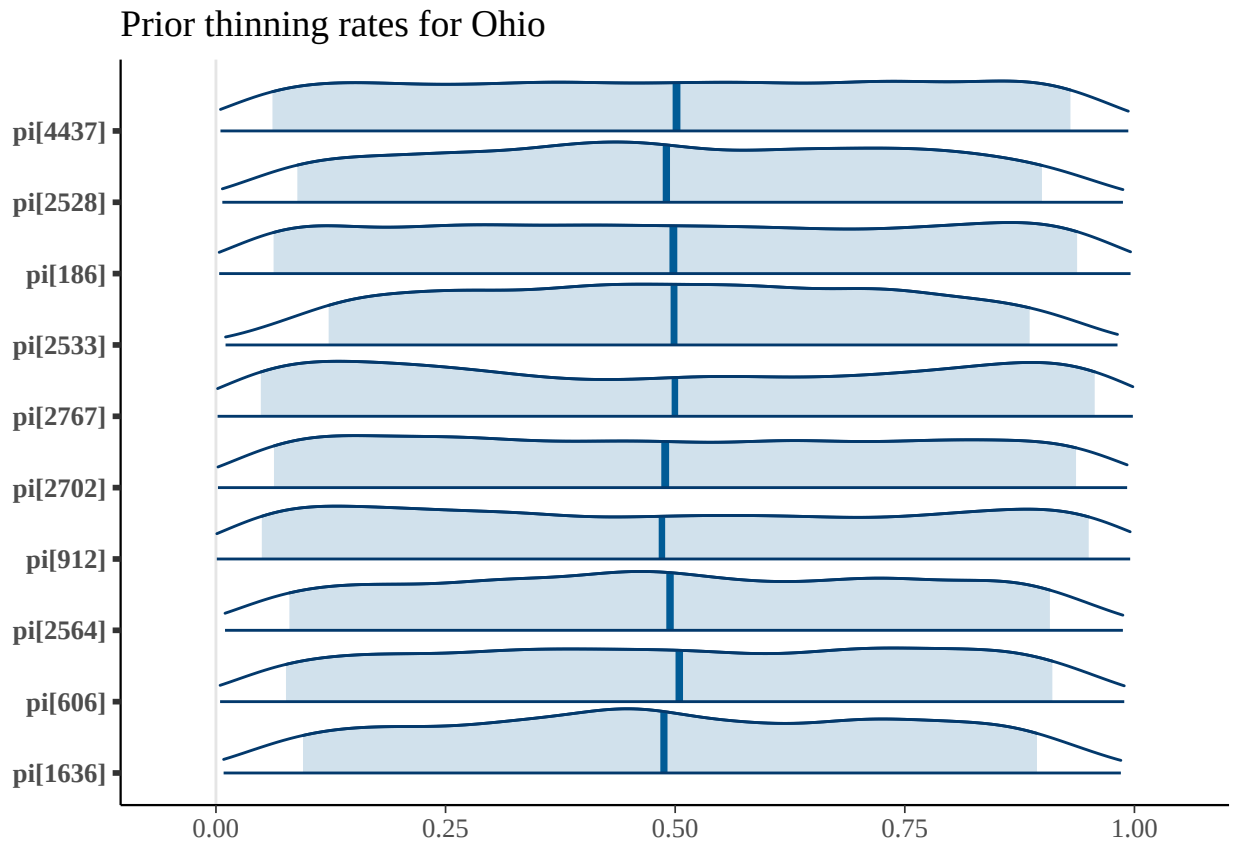
```
library(bayesplot)
```

```
## This is bayesplot version 1.10.0

## - Online documentation and vignettes at mc-stan.org/bayesplot

## - bayesplot theme set to bayesplot::theme_default()

##    * Does _not_ affect other ggplot2 plots
```

```
##     * See ?bayesplot_theme_set for details on theme setting
```

```
ppc_samples$draws(format = "draws_df") |>
  # select columns that contain "pi"
  select(starts_with("pi")) |>
  # randomly select 100 columns of those (for speed)
  select(sample(1:nrow(oh_merged_ppc), 10)) |>
  mcmc_areas(prob = 0.9) +
  ggtitle("Prior thinning rates for Ohio")
```

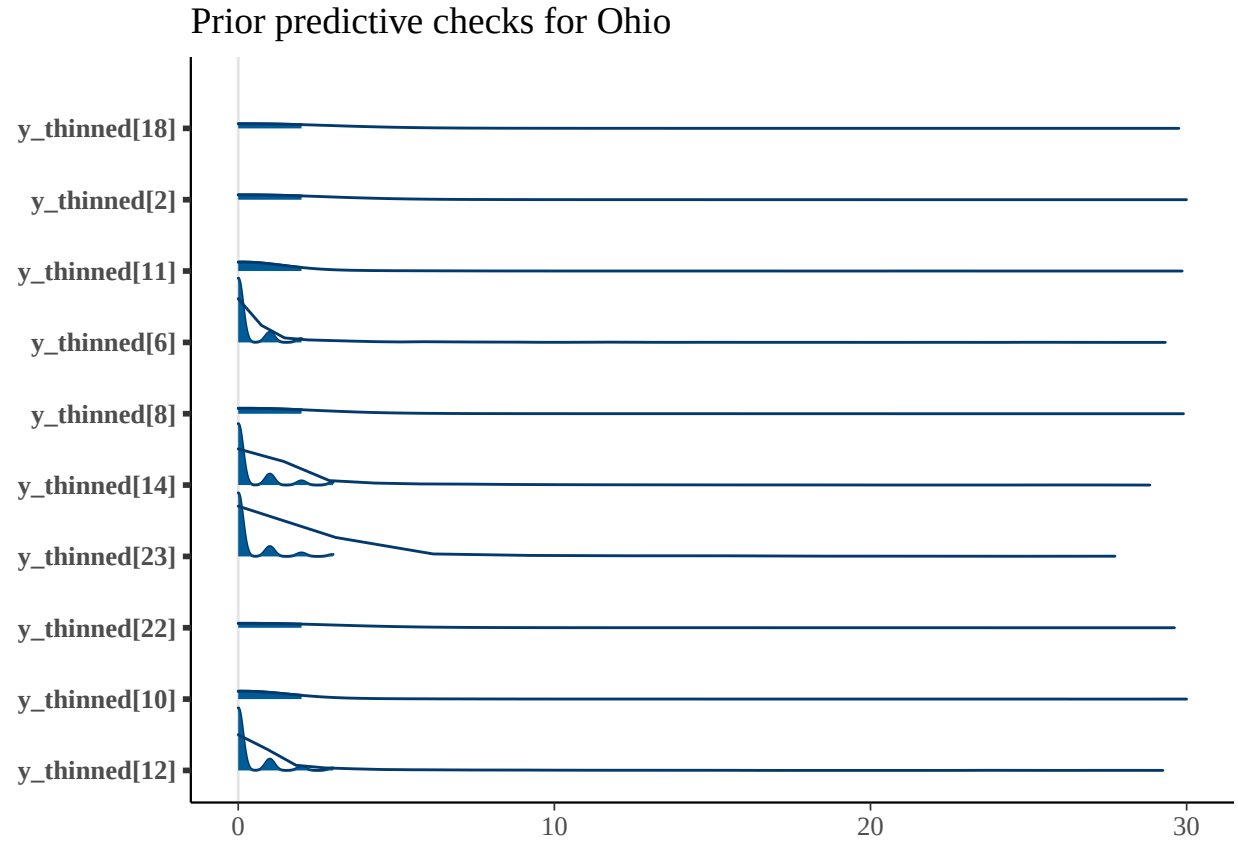
```
## Warning: Dropping 'draws_df' class as required metadata was removed.
```




```
library(bayesplot)

# extract and plot draws
ppc_samples$draws(format = "draws_df", variables = c("y_thinned")) |>
  # select columns that contain "y_thinned"
  select(starts_with("y_thinned")) |>
  # randomly select 100 columns of those (for speed)
  select(sample(1:length(oh_merged), 10)) |>
  mcmc_areas(prob = 0.8, alpha = 0.1) +
  ggtitle("Prior predictive checks for Ohio") +
  xlim(0,30)
```

```
## Warning: The following arguments were unrecognized and ignored: alpha
```

```
## Warning: Dropping 'draws_df' class as required metadata was removed.

## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.

## Warning: Removed 10 rows containing missing values (`geom_segment()`).
```



Massachusetts

```
data_summary("MA", year = 2010)
```

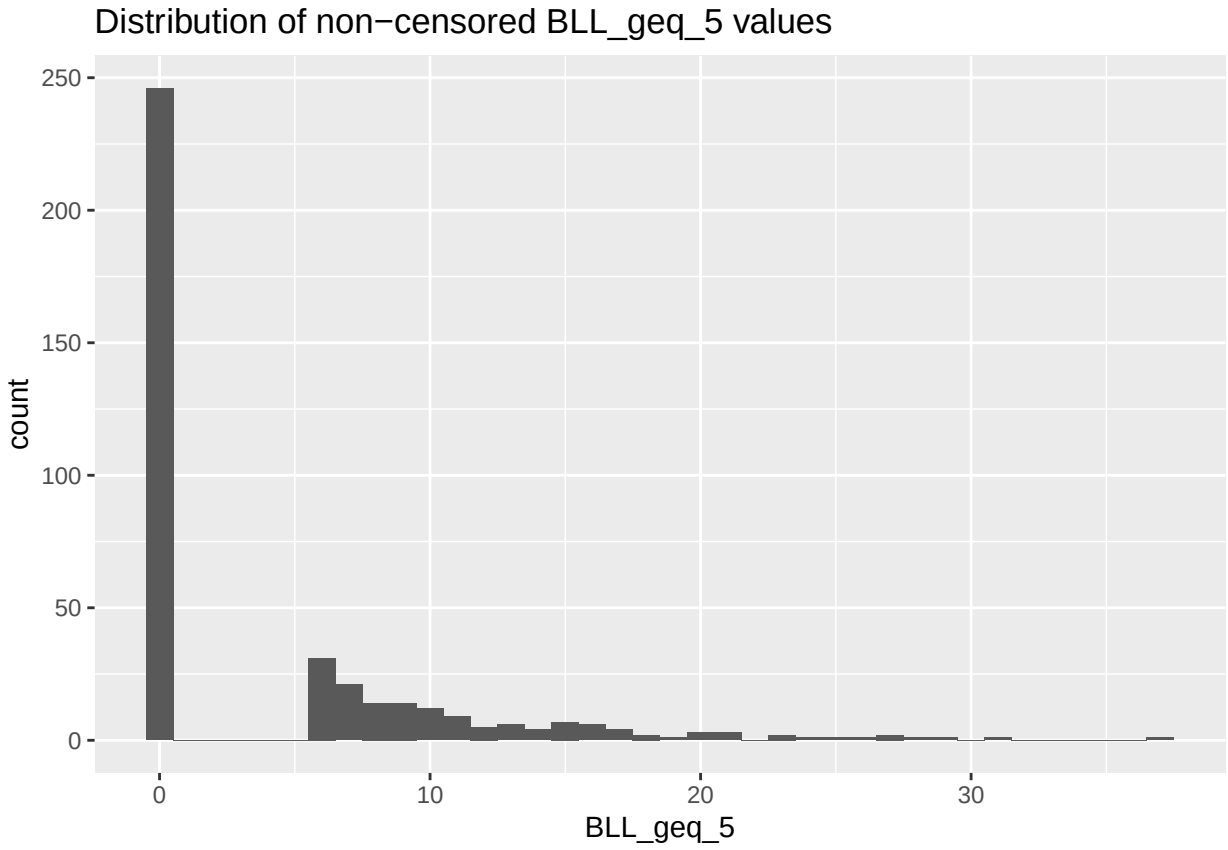
```
## [1] "Loading MA from processed_data"

## Rows: 16192 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): state, town, BLL_geq_5
## dbl (3): year, tested, tract
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## Warning: Using ‘across()‘ in ‘filter()‘ was deprecated in dplyr 1.0.8.
## i Please use ‘if_any()‘ or ‘if_all()‘ instead.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

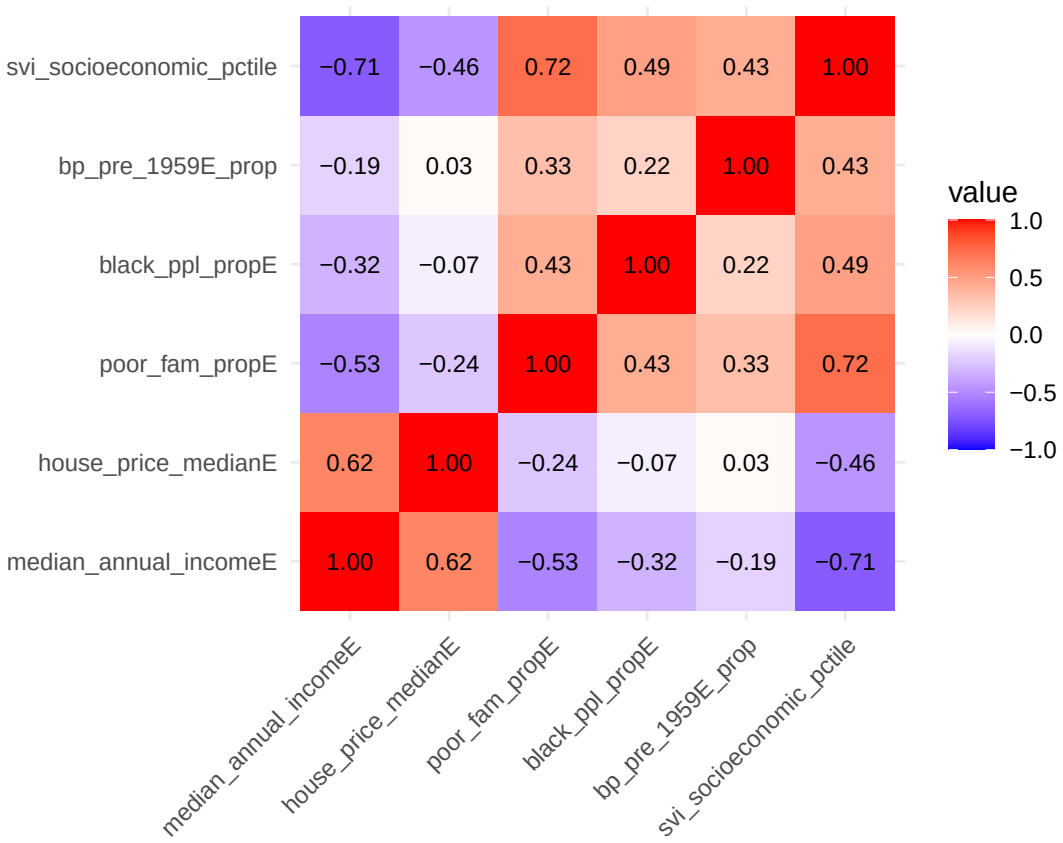
```
## [1] "Additional features added: "
```

```
## Warning in datasummary_skim_numeric(data, output = output, fmt = fmt, histogram
## = histogram, : The histogram argument is only supported for (a) output types
## "default", "html", or "kableExtra"; (b) writing to file paths with extensions
## ".html", ".jpg", or ".png"; and (c) Rmarkdown or knitr documents compiled to
## PDF or HTML. Use ‘histogram=FALSE‘ to silence this warning.
```

```
## [[1]]
## # A tibble: 1 x 11
##    year n_obs lead_cens_5 lead_nocens_5 lead_censR_5 test_cens test_censR
##   <dbl> <int>       <int>         <int>        <dbl>     <int>      <dbl>
## 1  2010   966         567           399        0.587         0          0
## # i 4 more variables: sup_threshold_5 <dbl>, median_lead_5 <dbl>,
## #   q75_lead_5 <dbl>, max_lead_5 <dbl>
##
## [[2]]
##
##
## |                          | Unique (#) | Missing (%) |  Mean  |   SD   |   Min  | Median |   Max  |
## |:-------------------------|-----------:|------------:|-------:|-------:|-------:|-------:|-------:|
## | year                     |          1 |           0 | 2010.0 |   0.0  | 2010.0 | 2010.0 | 2010.0 |
## | BLL_geq_5                |         27 |           0 |    4.7 |   4.3  |    0.0 |    5.0 |   37.0 |
## | tested                   |        276 |           0 |  141.4 |  70.4  |    3.0 |  133.0 |  411.0 |
## | tested_ell               |          1 |           0 |    1.0 |   0.0  |    1.0 |    1.0 |    1.0 |
## | ell_5                    |          1 |           0 |    4.0 |   0.0  |    4.0 |    4.0 |    4.0 |
## | median_annual_incomeE    |        953 |           0 |    0.1 |   1.0  |   -2.3 |   -0.0 |    5.0 |
## | house_price_medianE      |        889 |           0 |    0.1 |   1.0  |   -1.5 |   -0.2 |    6.6 |
## | poor_fam_propE           |        958 |           0 |   -0.1 |   0.9  |   -0.9 |   -0.4 |    5.7 |
## | black_ppl_propE          |        908 |           0 |   -0.0 |   1.0  |   -0.6 |   -0.4 |    5.8 |
## | bp_pre_1959E_prop        |        964 |           0 |   -0.1 |   1.0  |   -2.3 |   -0.1 |    2.0 |
## | svi_socioeconomic_pctile |        913 |           0 |   -0.1 |   0.9  |   -1.3 |   -0.4 |    2.1 |
## | under_yo5_pplE           |        419 |           0 |  263.0 | 134.5  |   14.0 |  243.5 | 1057.0 |
## | ped_per_100k             |         14 |           0 |  139.7 |  93.0  |    0.0 |  101.1 |  386.2 |
##
## [[3]]
## [[3]][[1]]
```


Distribution of non−censored BLL_geq_5 values
count
250
200
150
100
50
0
0
10
20
30
BLL_geq_5

```
##
##
## [[4]]
```




```
# standardise pediatrician rate
ma_merged <- ma_merged |>
  mutate(ped_per_100k = (ped_per_100k - mean(ped_per_100k)) / sd(ped_per_100k))
```

```
# sample from prior predictive distribution
ma_stan <- stan_data(ma_merged)

ppc_samples <- ppc_stan$sample(
  data = ma_stan,
  iter_sampling = 500,
  refresh = NULL,
  fixed_param = TRUE)
```

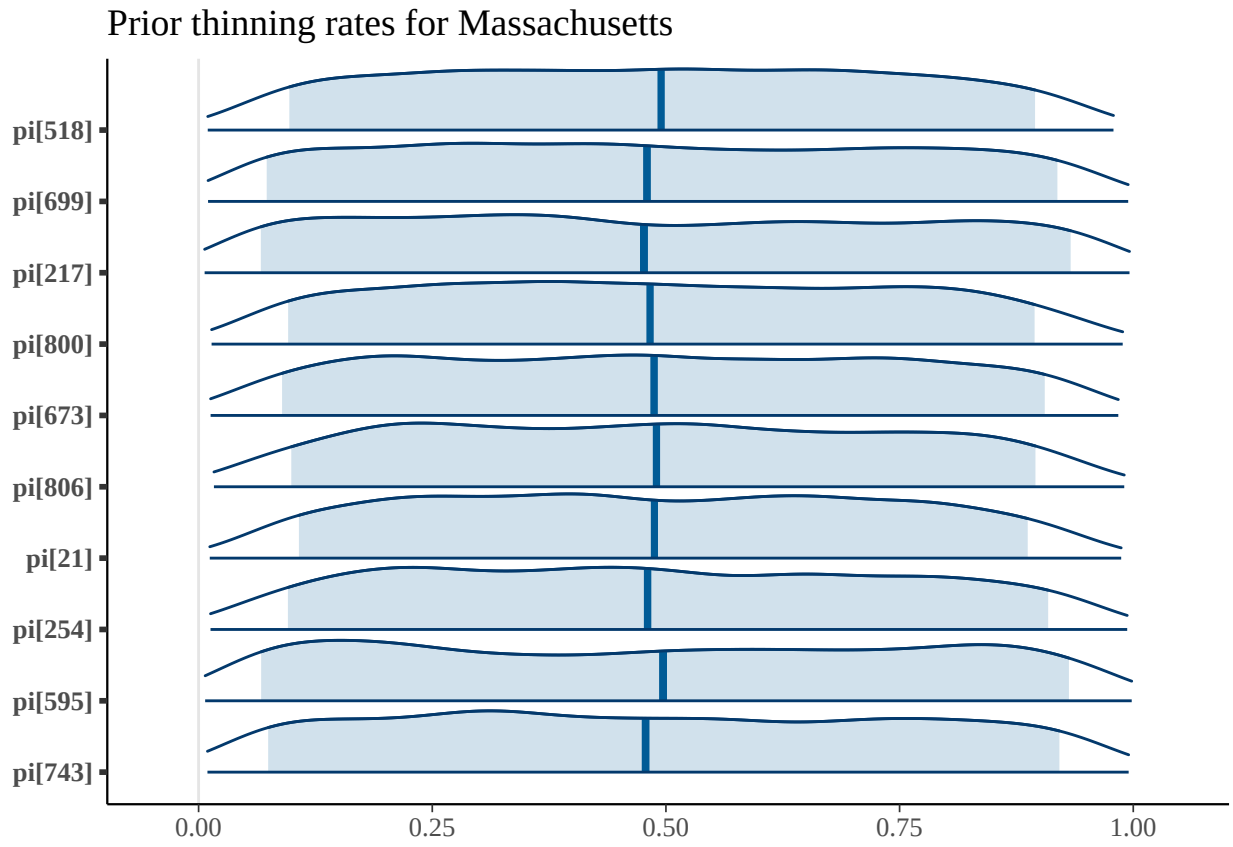
```
## Running MCMC with 4 sequential chains...
##
## Chain 1 Iteration:   1 / 500 [  0%]  (Sampling)
## Chain 1 Iteration: 100 / 500 [ 20%]  (Sampling)
## Chain 1 Iteration: 200 / 500 [ 40%]  (Sampling)
## Chain 1 Iteration: 300 / 500 [ 60%]  (Sampling)
## Chain 1 Iteration: 400 / 500 [ 80%]  (Sampling)
## Chain 1 Iteration: 500 / 500 [100%]  (Sampling)
## Chain 1 finished in 0.4 seconds.
## Chain 2 Iteration:   1 / 500 [  0%]  (Sampling)
## Chain 2 Iteration: 100 / 500 [ 20%]  (Sampling)
## Chain 2 Iteration: 200 / 500 [ 40%]  (Sampling)
```

```
## Chain 2 Iteration: 300 / 500 [ 60%]  (Sampling)
## Chain 2 Iteration: 400 / 500 [ 80%]  (Sampling)
## Chain 2 Iteration: 500 / 500 [100%]  (Sampling)
## Chain 2 finished in 0.4 seconds.
## Chain 3 Iteration:   1 / 500 [  0%]  (Sampling)
## Chain 3 Iteration: 100 / 500 [ 20%]  (Sampling)
## Chain 3 Iteration: 200 / 500 [ 40%]  (Sampling)
## Chain 3 Iteration: 300 / 500 [ 60%]  (Sampling)
## Chain 3 Iteration: 400 / 500 [ 80%]  (Sampling)
## Chain 3 Iteration: 500 / 500 [100%]  (Sampling)
## Chain 3 finished in 0.4 seconds.
## Chain 4 Iteration:   1 / 500 [  0%]  (Sampling)
## Chain 4 Iteration: 100 / 500 [ 20%]  (Sampling)
## Chain 4 Iteration: 200 / 500 [ 40%]  (Sampling)
## Chain 4 Iteration: 300 / 500 [ 60%]  (Sampling)
## Chain 4 Iteration: 400 / 500 [ 80%]  (Sampling)
## Chain 4 Iteration: 500 / 500 [100%]  (Sampling)
## Chain 4 finished in 0.4 seconds.
##
## All 4 chains finished successfully.
## Mean chain execution time: 0.4 seconds.
## Total execution time: 1.9 seconds.
```

```r
# add y_thinned from $summary to ma_merged
ma_merged_ppc <- ma_merged |>
  bind_cols(ppc_samples$summary(variables = c("y_thinned")) |>
              filter(str_detect(variable, "y_thinned")) |>
              select(c(mean, median, q5, q95)))
```

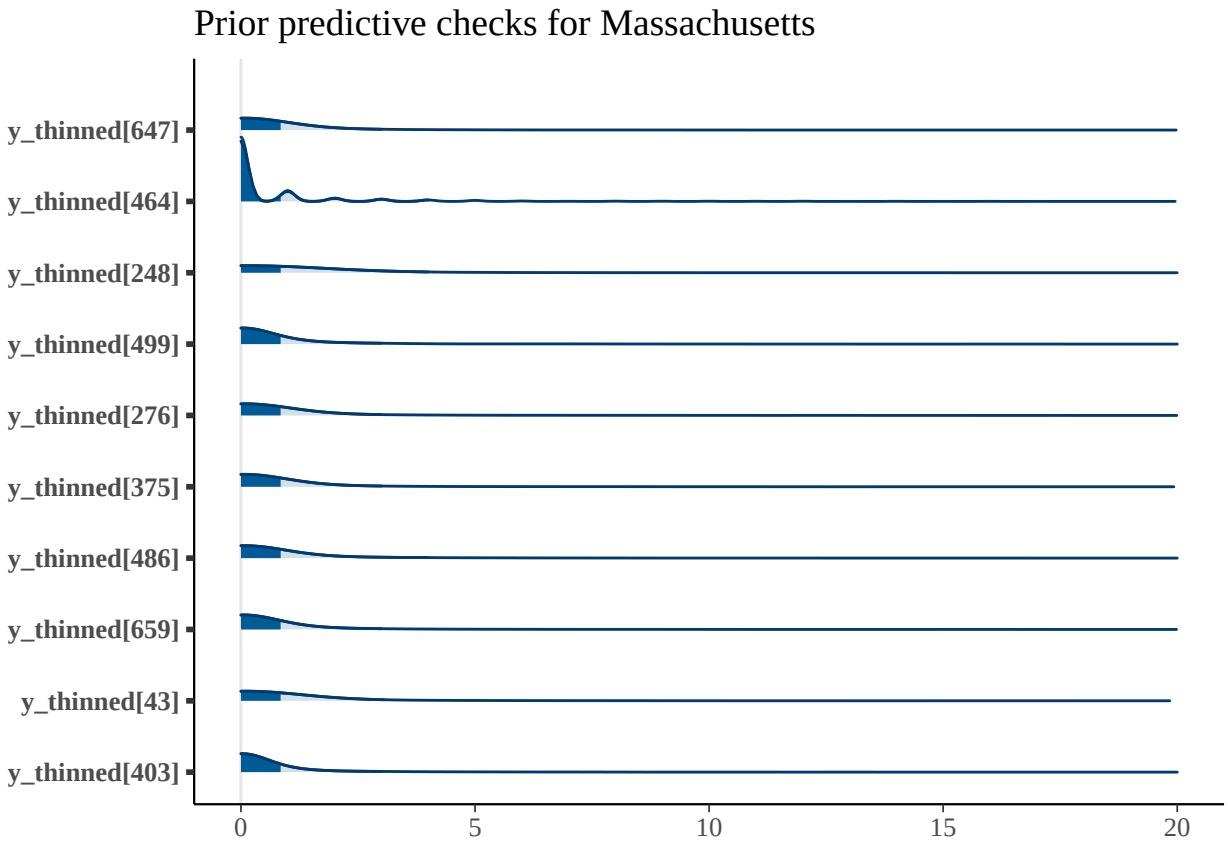
```r
# extract prior thinning rates
ppc_samples$draws(format = "draws_df") |>
  # select columns that contain "pi"
  select(starts_with("pi")) |>
  # randomly select 100 columns of those (for speed)
  select(sample(1:nrow(ma_merged), 10)) |>
  mcmc_areas(prob = 0.9) +
  ggtitle("Prior thinning rates for Massachusetts")
```

```
## Warning: Dropping 'draws_df' class as required metadata was removed.
```




```
# extract and plot draws
ppc_samples$draws(format = "draws_df", variables = c("y_thinned")) |>
  # select columns that contain "y_thinned"
  select(starts_with("y_thinned")) |>
  # randomly select 100 columns of those (for speed)
  select(sample(1:nrow(ma_merged_ppc), 10)) |>
  mcmc_areas(prob = 0.9, alpha = 0.1) +
  ggtitle("Prior predictive checks for Massachusetts") +
  xlim(0,20)
```

```
## Warning: The following arguments were unrecognized and ignored: alpha

## Warning: Dropping 'draws_df' class as required metadata was removed.

## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.

## Warning: Removed 10 rows containing missing values (`geom_segment()`).
```


Prior predictive checks for Massachusetts
y_thinned[647]
y_thinned[464]
y_thinned[248]
y_thinned[499]
y_thinned[276]
y_thinned[375]
y_thinned[486]
y_thinned[659]
y_thinned[43]
y_thinned[403]
0
5
10
15
20