Prepare data for BRMS fits focusing on x0

Michael Gilchrist

Created: 2023-02-22; Code run on 2023-03-22

Goal

- Process data as in `2023-02-07_brms_two.piece_focus.on.x0.fittings/` for other routines

Recap

Set up

Install libraries

```
# install packages user might not have by replacing FALSE with TRUE

## load libraries
library(stats)
library(MASS) # provides negative binomial fitting:  glm.nb
library(ggplot2)
library(tidyverse)
library(viridisLite)
```

Load Data

```
resave_data <- FALSE # flag for saving data_stan and possibly others in future

input_dir <- file.path("input")
output_dir <- file.path("output")

load(file.path(input_dir, "data.processing_2022-12-15.Rda"),
     verbose = TRUE)
```

```
## Loading objects:
##   motif_data
##   motif_data_40C
##   motif_stats
##   motif_stats_40C
##   bird_bill_data
```

```
motif_data
```

```
## # A tibble: 146 x 28
##    male  round trial_round motif~1 motif~2 temp_~3 humid~4 chamber date  counter
##    <fct> <dbl>       <dbl>   <int>   <dbl>   <dbl>   <dbl> <fct>   <chr> <chr>
##  1 T229      1           1       0 0          45.8      NA 6       02/1~ KIM
##  2 T229      1           2      24 0.0131     42.3      NA 6       02/1~ KIM
##  3 T229      1           3     114 0.0622     40.7      NA 6       02/1~ KIM
##  4 T229      1           4     198 0.108      26.2      NA 6       02/1~ KIM
##  5 T229      1           5     315 0.172      34.9      NA 6       02/2~ KIM
##  6 T231      1           1      57 0.0431     42.8      NA 2       02/1~ RAS
##  7 T231      1           2       7 0.00529    45.0      NA 2       02/1~ RAS
##  8 T231      1           3      86 0.0650     41.1      NA 2       02/1~ KIM
##  9 T231      1           4      24 0.0181     27.2      NA 2       02/1~ RAS
## 10 T231      1           5     215 0.162      36.5      NA 2       02/2~ RAS
## # ... with 136 more rows, 18 more variables: test_order <int>,
## #   temp_target <dbl>, temp_median <dbl>, humdity_mean <dbl>, motif_rate <dbl>,
## #   mass <dbl>, n_obs_completed <lgl>, motif_count_plus_1 <int>,
## #   log_motif_count_plus_1 <dbl>, temp <dbl>, n_obs_round <int>, n_obs <int>,
## #   trial <int>, motif_prop_round <dbl>, weights <dbl>, svp <dbl>, vpd <dbl>,
## #   vpd_offset <dbl>, and abbreviated variable names 1: motif_count,
## #   2: motif_prop, 3: temp_mean, 4: humidity_mean
```

Process Data

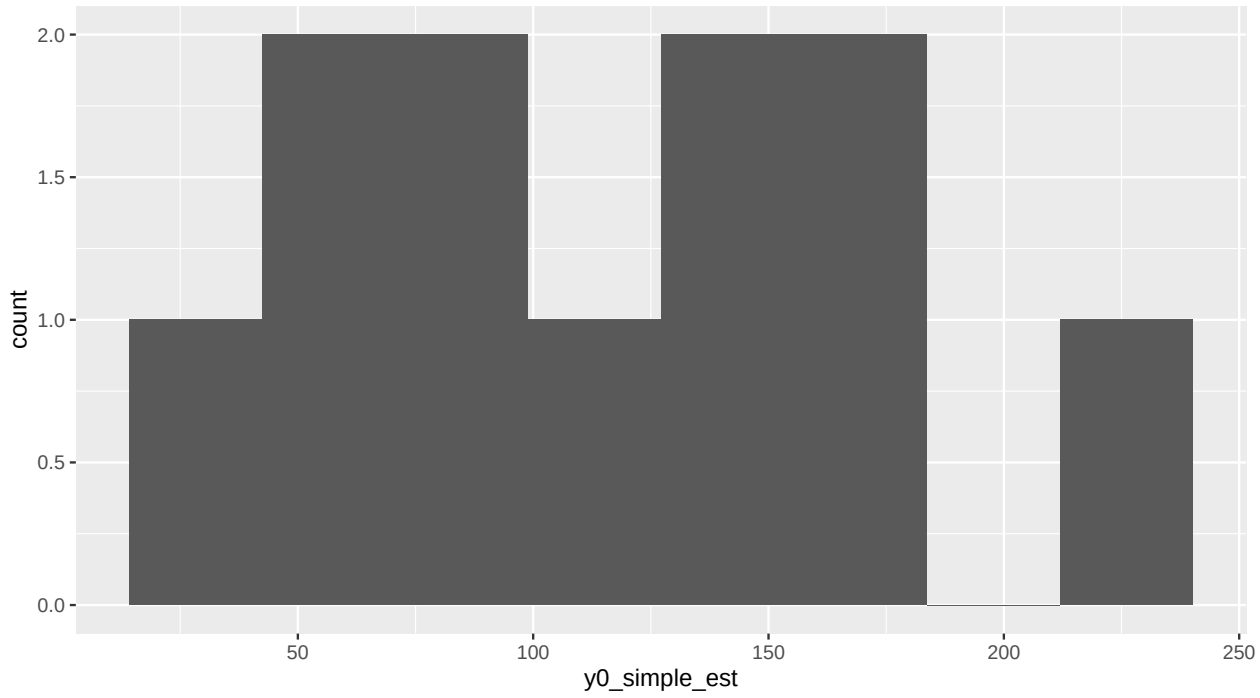
Create Working Dataset

```
filter_data <- TRUE

if(filter_data) {
    males_filtered_disp <- motif_stats_40C %>%
        filter(dispersion < 50) %>%
        pull(male)

    males_filtered_mean <- motif_stats %>%
        filter(mean > 10) %>% # changing from 10 to 40 removes previous male 7 (T258)
        pull(male)

    male_vector <- intersect(males_filtered_mean, males_filtered_disp)
} else {
    male_vector <- motif_data %>% select(male) %>% distinct()
}

data_ind <- motif_data %>%
  filter(male %in% male_vector) %>%
  mutate(male = droplevels(male)) %>%
  mutate(index = as.integer(male)) %>%
  mutate(male = as.character(male)) %>%
  arrange(index) %>%
  select(male, index, motif_count, temp_target, temp, round, trial_round, date, counter) %>%
```

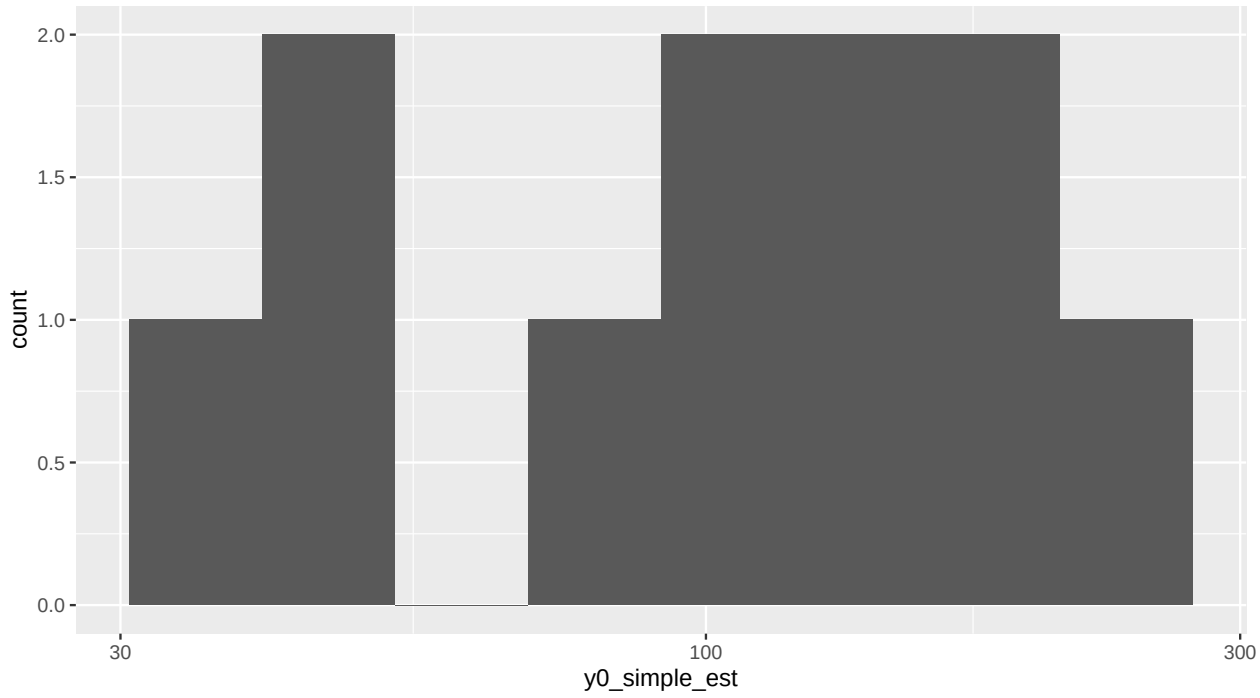
```
  ##    left_join(index_shape, by = "index") %>%
  group_by(male) %>% mutate(y0_simple_est = mean(motif_count), phi_ind = var(motif_count)/y0_simple_est)
  ungroup() %>%
  mutate()

save(data_ind, file = file.path(output_dir, "data_ind.Rda"))
write_csv(data_ind, file = file.path(output_dir, "data_ind.csv"))

stats_ind <- motif_stats %>%
    filter(male %in% male_vector)

save(stats_ind, file = file.path(output_dir, "stats_ind.Rda"))
write_csv(stats_ind, file = file.path(output_dir, "stats_ind.csv"))

summary(data_ind)
```

```
##      male                index         motif_count   temp_target      temp
##  Length:107         Min.   : 1.00   Min.   :  0   Min.   :27   Min.   :25.7
##  Class :character   1st Qu.: 3.00   1st Qu.: 56   1st Qu.:35   1st Qu.:34.3
##  Mode  :character   Median : 5.00   Median : 90   Median :40   Median :38.8
##                     Mean   : 5.14   Mean   :116   Mean   :38   Mean   :37.6
##                     3rd Qu.: 8.00   3rd Qu.:175   3rd Qu.:42   3rd Qu.:41.1
##                     Max.   :11.00   Max.   :425   Max.   :44   Max.   :45.6
##      round       trial_round       date              counter
##  Min.   :1.00   Min.   :1.00   Length:107         Length:107
##  1st Qu.:1.00   1st Qu.:2.00   Class :character   Class :character
##  Median :2.00   Median :3.00   Mode  :character   Mode  :character
##  Mean   :2.15   Mean   :2.83
##  3rd Qu.:3.00   3rd Qu.:4.00
##  Max.   :3.00   Max.   :6.00
##  y0_simple_est      phi_ind
##  Min.   : 34.2   Min.   :  9.7
##  1st Qu.: 76.4   1st Qu.: 12.3
##  Median :106.2   Median : 22.5
##  Mean   :116.5   Mean   : 35.5
##  3rd Qu.:161.2   3rd Qu.: 58.9
##  Max.   :232.0   Max.   :103.7
```

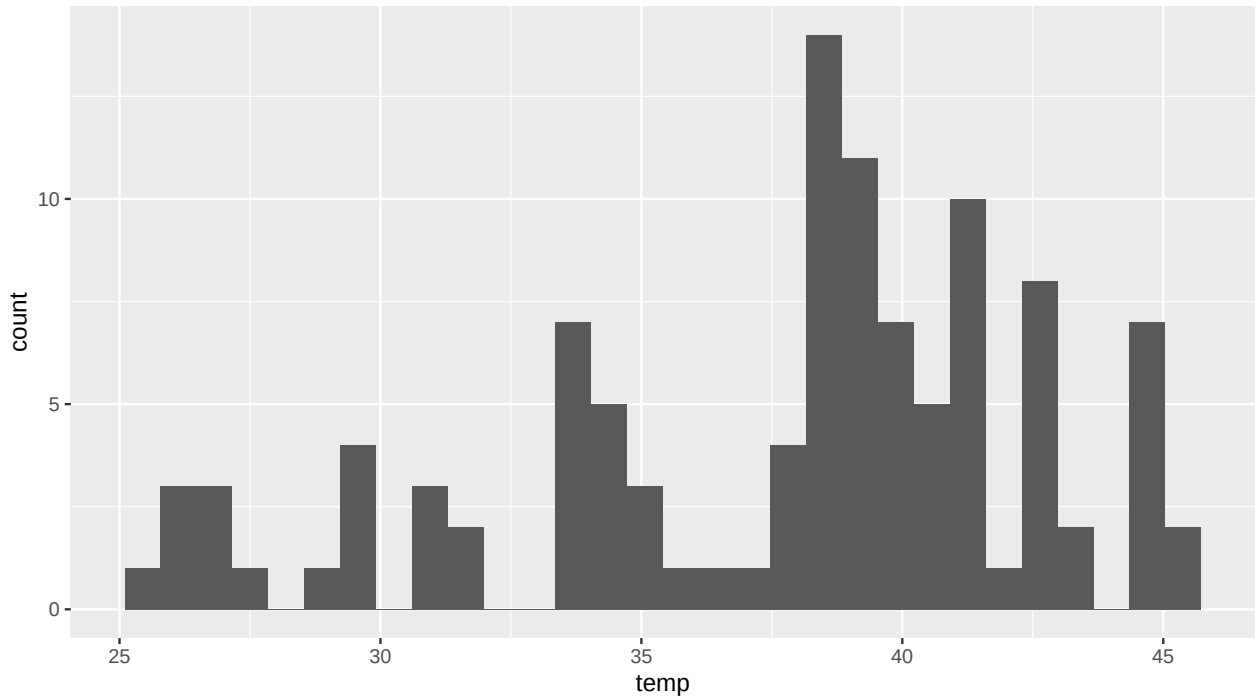
```
y0_hist <- data_ind %>%
  select(y0_simple_est) %>%
  unique() %>%
  ggplot(aes(y0_simple_est)) + geom_histogram(bins = 8)
print(y0_hist)
```




```
log_y0_hist <- y0_hist +
  scale_x_log10()
print(log_y0_hist)
```




```
temp_hist <- ggplot(data_ind, aes(temp)) + geom_histogram()
print(temp_hist)
```




```
n_male <- length(unique(data_ind$male))

## for calculation of 'theta_bar' (size in `rnbinom`) see note in `Set Up Simulate Data`
summary_stats <- data_ind %>%
    ungroup() %>%
    summarize(y0_bar = mean(y0_simple_est),
              y0_sd = sd(y0_simple_est),
              log_y0_bar = mean(log(y0_simple_est)),
              log_y0_sd = sd(log(y0_simple_est)),
              phi_bar = median(phi_ind), # qpoisson over dispersion
              theta_bar = y0_bar^2/(y0_sd^2 - y0_bar)) # nb shape parameter
comment(summary_stats) <- "summary stats for observed bird motifs"

save(summary_stats, file = file.path(output_dir, "obs_summary_stats.Rda"))
```

Fit Models

See other folders created on this day.